I T H E A

International Journal

INFORMATION CONTENT & PROCESSING

2014 Volume 1 Number 3

International Journal
INFORMATION CONTENT & PROCESSING
Volume 1 / 2014, Number 3

IJ ICP is official publisher of the scientific papers of the members of the ITHEA® International Scientific Society

IJ ICP rules for preparing the manuscripts are compulsory.
The **rules for the papers** for ITHEA International Journals as well as the **subscription fees** are given on www.ithea.org .
The papers should be submitted by ITHEA® Submission system http://ij.ithea.org .
Responsibility for papers published in IJ IMA belongs to authors.

International Journal "INFORMATION CONTENT AND PROCESSING" Volume 1, Number 3, 2014

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politechnika de Madrid, Spain,
Hasselt University, Belgium
Institute of Informatics Problems of the RAS, Russia,
St. Petersburg Institute of Informatics, RAS, Russia
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia.

Publisher: **ITHEA®**
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Technical editor: Ina Markova

Printed in Bulgaria

ISSN 2367-5128 (printed) **ISSN 2367-5152 (online)**

УЛУЧШЕННЫЕ CART ТЕХНОЛОГИИ ГЕНЕРАЦИИ ЧАСТИЧНО СИНТЕТИЧЕСКИХ ДАННЫХ

Левон Асланян, Вардан Топчян

Аннотация: *Работа посвящена исследованию вопросов анализа персональных данных обеспечивающих конфиденциальность данных. Предполагается что даны частично критические социологические данные и перед представлением этих данных общественности требуется их модифицировать так, чтобы конфиденциальные данные не раскрывались, и чтобы анализ этих данных не отличался от анализа исходных данных. Работа строит улучшенные алгоритмы класс деревьев классификации и регрессии, которые предоставляют решение задачи генерации так называемых синтетических данных. Новое решение учитывает структуры областей конфиденциальности и проводит оптимизацию дерева замены данных на синтетические.*

Ключевые слова: *классификация, регрессия, раскрытие данных, синтетические данные.*

ACM Classification Keywords*: H.1 Information Systems – Models and principles, I.2.0 Artificial intelligence.*

Введение

Предоставление экономических, социальных данных общественным структурам является неотъемлемой частью деятельности государственных статистических организаций. Открытый доступ к данным имеет большие преимущества. Прежде всего такие данные могут явится источником осуществления разного рода исследований, в том числе и в учебных целях. Тем не менее ограничение риска раскрытия (disclosure limitation) конфиденциальной информации продолжает оставаться одной из главных задач статистических организаций потому что даже при удалении очевидных идентификаторов таких как имя или адрес не исключают возможности доступа к персональным данным. Ведь как показано многими авторами благодаря сопоставлениям значений общих ключевых атрибутов в нескольких таблицах данных можно выявить определенные персональные данные. Для решения данной задачи/проблемы часто прибегают к модификации (perturbation) исходных данных или к их замене другими, новыми данными. Эти данные генерируются на основе разных моделей и алгоритмов [4]. Часто модифицируются значения отдельных атрибутов/дескрипторов. Таким образом защищая отдельные поля информации, они могут привести к побочному эффекту, к искажению связей между разными сегментами множества данных, что в свою очередь может привести к ошибочным выводам на этапе статистического анализа данных.

Альтернативным подходом решения поставленной задачи, который одновременно пытается сохранить функциональные связи между сегментами множества данных, является подход генерации так называемых полных синтетических данных (synthetic data generation, SDG) [5]. В этом случае, статистическая организация должна, во-первых, произвольно и независимо отмечать общий формат и критическое содержание единиц информации и включать их в соответствующее предполагаемое множество синтетических данных, во-вторых: по выбранной стратегии/алгоритму устанавливать новые,

синтетические, значения в единицах информации, и в-третьих: предоставить общественности некоторое количество множеств, сгенерированных синтетических данных. Известны различные методы [6] генерации полных синтетических данных, обеспечивающих получение значимых результатов с использованием стандартных статистических методов.

Несмотря на отмеченные преимущества института полных синтетических данных, процесс их генерации довольно трудоемкий. Не понятен также подход когда изменяются неконфиденциальная составляющая исходной информации. В связи с этим часто прибегают к использованию схемы генерации частично синтетических данных [6], представляющих из себя сочетание оригинальных и синтетических данных. Потребность в генерации частично синтетических данных возникает в тех случаях, когда статистическое агентство стремится защитить конфиденциальность для определенных записей. С этой целью, генерируются синтетические значения лишь для определенных атрибутов, а значения остальных не изменяются.

Как и в случае полных синтетических данных, частично синтетические данные так же обеспечивают ограничение риска раскрытия информации, позволяя получать значимые результаты с использованием стандартных статистических методов. Отметим, что, в силу своей природы, применение частично синтетических данных, обеспечивает более точные результаты статистических вычислений. По той же причине, риск раскрытия информации выше по сравнению с полными синтетическими данными. Однако, известны различные алгоритмы [5, 6] для их генерации, используемые многими статистическими организациями (U.S. Federal Reserve Board, U.S. Bureau of the Census, Statistical agencies of Germany and New Zealand, etc.), что говорит о перспективах данного метода.

Анализ существующих алгоритмов генерации синтетических данных показывает их эвристическую структуру. Таким образом обоснованием является эксперимент и нет теоретической обоснованности использования того или другого подхода. Вместе с тем область конфиденциальности данных задачи хорошо интерпретируема и она подлежит формальному описанию. Данная работа, впервые, сформулирует формальную модель критических данных и попытается построить улучшенные алгоритмы генерации синтетических данных следя за сохранением как отдельных значений параметров задачи так и за совместными значениями групп параметров и атрибутов. В теоретическом плане, как это замечено отдельными авторами, сформулированные задачи схожи с вероятностными задачами восстановления отсутствующих значений (missing value). В таких схемах возможно получение оценок ошибки однако практические задачи не обладают достаточной информацией для восстановления вероятностных распределений и наша цель не в получении таких оценок а в формализации и использовании дополнительных свойств задачи для сформирования более адекватного прикладного результата.

Выше представленное послужило основой для нашего изучения методов генерации частично синтетических данных [1, 2]. Для осуществления идеи этой работы нам необходимо выбрать и остановиться на одном из подходящих методов генерации синтетических данных. Как показывает анализ литературных данных [5, 6], приемлемых является [6], работа которого основана на использовании деревьев CART (*Classification and Regression Trees*). Прежде чем перейти к более подробному рассмотрению этого алгоритма, дадим краткое описание формата наших данных и предполагаемых синтетических данных задачи.

Формат частично синтетических данных

Процесс генерации частично синтетических данных состоит из двух этапов: (1) предварительная обработка (preprocessing) входных данных, (2) замещение отмеченных, критических значений на синтетические. Формально данный процесс можно описать следующим образом.

Пусть, $\mathcal{U}$ множество отдельных элементов, из которых составлены входные данные, $\mathcal{U} = \{U_1, U_2, \dots, U_N\}$, где каждый элемент характеризуется множеством атрибутов $\mathcal{A} = \{A_1, A_2, \dots, A_p\}$:

$$U_i = \left(a_{i1}, a_{i2}, \dots, a_{ij}, \dots, a_{ip}\right), (1 \le i \le N, 1 \le j \le p).$$

На этапе препроцессинга данных произвольным образом отмечается определенное количество элементов множества $\mathcal{U}$ для текущего наблюдения (observation) и отмечаются конфиденциальные атрибуты (строки и столбцы матрицы $\mathcal{U}$ соответственно), и устанавливаются пороговые условия для атрибутов.

Пусть, $n\ (n \le N)$ есть количество произвольно выбранных элементов множества $\mathcal{U}$, а $d\ (d \le p)$ – количество конфиденциальных атрибутов. Обозначим выбранные элементы и конфиденциальные атрибуты через $\{U_{i_1}, U_{i_2} \dots, U_{i_n}\}$ и $\{A_{j_1}, A_{j_2}, \dots, A_{j_d}\}$ соответственно. С целью определения этих элементов и атрибутов введем вспомогательные наборы индикаторов $I = (I_1, I_2, \dots, I_N)$ и $J = (J_1, J_2, \dots, J_p)$:

$$I_r = \begin{cases} 1, U_i \in \{U_{i_1}, U_{i_2} \dots, U_{i_n}\} \\ 0, U_i \notin \{U_{i_1}, U_{i_2} \dots, U_{i_n}\} \end{cases} 1 \le i \le N,$$

$$J_k = \begin{cases} 1, A_j \in \{A_{j_1}, A_{j_2} \dots, A_{i_d}\} \\ 0, A_j \notin \{A_{j_1}, A_{j_2} \dots, A_{i_d}\} \end{cases} 1 \le j \le p.$$

Схематически данный процесс представлен на Рис. 1.

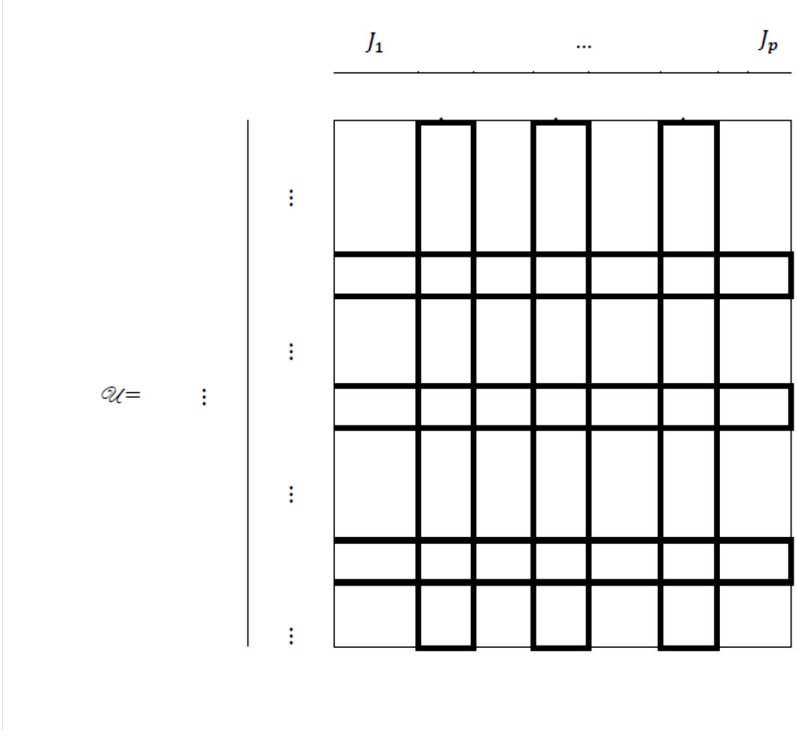


Рис. 1 Исходные данные задачи с выделением критических данных

По существу, в результате определяется расщепление $U_{obs} = (U_{rep}, U_{nrep})$ исходной матрицы наблюдений размерами $[n \times p]$ составленного из рассматриваемых *(observed)* единиц информации. Здесь U_{rep} – представляет собой матрицу размера $[n \times d]$ значений конфеденциальных атрибутов $A_{j_1}, A_{j_2}, \dots, A_{j_d}$ (replaced vs. not replaced). А U_{nrep} – $[n \times (p-d)]$ матрица значений остальных атрибутов (атрибутов значения которых не подвергаются замещению). Для упрощения представления матрицы U_{obs} ее столбцы соответствующим образом можно переставить (Рис. 2).

$U_{obs} =$

	U_{rep}			U_{nrep}		
	A_{j_1}	...	A_{j_d}	$A_{j(d+1)}$	...	A_{j_p}
U_{i_1}	$a_{i_1 j_1}$	...	$a_{i_1 j_d}$	$a_{i_1 j(d+1)}$	...	$a_{i_1 j_p}$
U_{i_2}	$a_{i_2 j_1}$	...	$a_{i_2 j_d}$	$a_{i_2 j(d+1)}$	...	$a_{i_2 j_p}$
⋮						
U_{i_n}	$a_{i_n j_1}$	...	$a_{i_n j_1}$	$a_{i_n j(d+1)}$	...	$a_{i_n j_p}$

Рис. 2 Переставленная исходная матрица

Далее, для определения критических областей значений атрибутов $A_{j_1}, A_{j_2}, \dots, A_{j_d}$ устанавливаются соответствующие пороговые условия (threshold conditions) посредством некоторого множества $\mathscr{C} = \{C_1, \dots, C_d\}$ (Рис. 3).

$U_{rep} =$

	C_1	...	C_d
	A_{j_1}	...	A_{j_d}
U_{i_1}	$y_{i_1 j_1}$	...	$y_{i_1 j_d}$
U_{i_2}	$y_{i_2 j_1}$	...	$y_{i_2 j_d}$
⋮			
U_{i_n}	$y_{i_n j_1}$	...	$y_{i_n j_d}$

$\mathscr{C} = \{C_1, \dots, C_d\}$

Рис. 3 Критическая матрица и пороговые значения атрибутов

На основании этих условий определяется индикаторная матрица $Z\,[n \times d]$, характеризующая необходимость изменения значений конфиденциальных атрибутов (Рис. 4). А именно, значение элемента z_{rk} берется равным единице, если значение соответствующего атрибута A_{j_k} , $a_{i_r j_k}$, удовлетворяет пороговому условию C_k, $z_{rk} = 1\ (1 \le r \le n)$, $(1 \le k \le p)$. В обратном случае $z_{rk} = 0$.

Индикаторная (0,1) матрица в принципе может иметь произвольную форму. Понятно что обычно условия C_k представляются простыми логико-арифметическими выражениями и что этим характеризуется структура самой матрицы Z. Однако в общем случае матрица произвольная и вопрос ее эффективного применения связан с задачами ее оптимального представления. Здесь можно рассмотреть схему представления матрицы в виде суммы малого числа матриц простой структуры, представление при помощи иерархического дерева или другие виды представления. Индикаторная (0,1) матрица таким образом имеет вид

$$Z = \begin{array}{|c|c|c|} \hline z_{1j_1} & \cdots & z_{1j_d} \\ \hline z_{2j_1} & \cdots & z_{2j_d} \\ \hline & & \\ \vdots & & \\ \hline z_{rj_1} & \cdots & z_{zj_d} \\ \hline & & \\ \vdots & & \\ \hline z_{nj_1} & \cdots & z_{nj_d} \\ \hline \end{array}$$

Рис. 4 Индикаторная матрица

и согласно замеченному, с этим связано некоторое адекватное структурное представления данной матрицы.

Суммарным результатом предварительной обработки множества входных данных $\mathcal{U}$и вспомогательных индикаторных наборов I, J являются матрица U_{obs} , соответственно матрицы U_{rep} и U_{nrep} , и индикаторная матрица Z. Полученный набор данных обозначим через $D = (U_{rep}, U_{nrep}, Z)$.

Вторая, в принципе основная часть процесса генерации частично синтетических данных представляет из себя процесс изменения/замещения критических значений. А именно, на основании полученного набора данных $D = (U_{rep}, U_{nrep}, Z)$ и выбранного алгоритма, производятся замещения соответствующих значений матрицы U_{rep} на новые, синтетические значения. Процесс замещения производится независимо некоторое количество m раз, в результате чего генерируются m различных множеств частично синтетических данных:

$$SD_t = \left(U_{syn}^t, U_{nrep}\right),\ 1 \le t \le m\ ,$$

где $U_{syn}^t -$ матрица с установленными синтетическими значениями $t-$го наблюдаемого множества. Отметим, что матрица U_{nrep} одинакова для всех множеств $SD_t, 1 \le t \le m$. Это характерное но не обязательное условие задачи. При желании можно синтезировать и не критические значения и к тому есть своя причина. Первая причина сложностная - она упрощает работу алгоритмов замещения данных. Вторая связана с специфичным раскрытием, когда добывая нужное количество некритических значений атакующий пытается восстановить ее связи к критическим данным.

Таким образом, полученные множества частично синтетических данных $SD_1, SD_2, \dots, SD_m$ являются теми данными, которые предоставляются соответствующим организациям и общественным структурам. На этапе оценки предлагаемой модели и алгоритма замещения данных необходимо подтвердить (validation) что по совокупности предполагаемых методов анализа результаты обработки синтетических данных не будет отличаться от оригинальных результатов по исходным данным.

Описание алгоритма CART в виде удобном для SDG

Теоретико-графовое понятие дерева служит основой ряда моделей поиска, классификации, предоставления привилегий, и др. [8-12]. Примеры известных древовидных моделей принятия решений включают ID3, C4.5, CART, CHAID. В статистике, наряду с иерархическим кластерным анализом используются процедуры типа Bagging, Random Forest, Boosted Trees, которые строят и оптимизируют классы деревьев. Деревья являются существенной компонентой алгоритмов сжатия данных (gif, lzw), систем вычислений, распознавания, семантических анализов данных и др. Наша ближайшая задача в этих терминах характеризуется как задача построения моделей иерархических ресурсов данных и задача иерархической классификации с минимизацией ошибки. Генерирующая идея работы заключается в ограничении решающих правил используемых в вершинах деревьев классом условий и ограничений, характеризующих критические данные задачи. Это возможно и эффективно потому что в основной модели дерево строится по критерию минимизации ошибки и только после этого снижается ее сложность учитывая критические и не критические значения переменных. Построение модели требует некоторую детализацию описания древовидных структур и процедур описания критических данных, к которомы и сейчас мы переходим.

Алгоритм CART является одним из представителей древовидных моделей обработки данных. Для нашего случая алгоритм предназначен для генерации частично синтетических множеств данных, которые

возможно использовать для вычисления простых статистических величин переменных (математического ожидания, дисперсии и т.д.) и построения классификационных и линейных регрессионных моделей. Работа алгоритма основана на бинарных деревьях с условиями на вершинах. Они используются с целью управления условного распределения критических значений конфиденциальных атрибутов.

Деревья CART используются для прогнозирования значений зависимой переменной на основании набора предикторов. Принцип построения CART заключается в рекурсивном разбиении множества рассматриваемых элементов данных на подмножества, однородные относительно зависимой переменной. А именно, на каждом шаге определяется наилучшее условие по некоторому предиктору и производится разбиение текущего множества (growing). В результате, в листьях полученного дерева будут содержатся элементы данных с одинаковым значением зависимой переменной. Поскольку, полученное дерево может состоять из неоправданно большого количества узлов и ветвей, то для достижения приемлемого размера этих деревьев производится их отсечение (pruning) на основании некоторого критерия оптимальности. По существу, листья дерева CART представляют условное распределение зависимой переменной для рассматриваемого набора предикторов.

Построение дерева. При описании работы алгоритма мы будем придерживаться обозначений, введенных в предыдущей части. Не нарушая общности, допустим, что матрица U_{obs}, полученная в результате предварительной обработки данных, состоит из первых n элементов множества $\mathscr{U}$. Учитывая то, что порядок атрибутов не фиксирован по смыслу нашей задачи, в этапе обработки данных их переставлением мы можем добиться того, чтобы конфиденциальные данные оказались только в первых d столбцах данных, т.е. они определены атрибутами

$$A_{conf} = \{A_1, A_2, \dots, A_d\},\ A_{conf} \subseteq \mathscr{A}$$

В алгоритме, генерация синтетических данных осуществляется последовательно, путем наращивания, по каждому конфиденциальному атрибуту. В связи с этим, на первом этапе работы производится упорядочивание атрибутов $A_1, A_2, \dots, A_d$ по мере убывания количества критических значений по входным данным. С этой целью, для каждого атрибута A_k $(1 \le k \le d)$, на основании индикаторной матрицы Z, вычисляется следующее значение:

$$\sum_{i=1}^{n} z_{ik}\ .$$

Однако, не исключено, что для некоторой группы атрибутов данное значение может быть одинаковым. В этом случае, порядок для этих атрибутов устанавливается по мере важности каждого из них при построении соответствующих деревьев CART для остальных членов этой группы. Для наглядности рассмотрим частный случай с двумя атрибутами. Допустим, что A_b и A_c – атрибуты с одинаковым количеством критических значений,

$$\sum_{i=1}^{n} z_{ib} = \sum_{i=1}^{n} z_{ic}\ .$$

Для этих атрибутов строятся соответствующие им деревья *CART*, T_b и T_c (Рис. 5). Далее, для атрибута A_b определяется величина P_b, равная глубине в дереве T_b, где впервые встречается разбиение по атрибуту A_c (если такое разбиение отсутствует, то P_b условно берется равным $P_b = \infty$). Величина P_b характеризует то насколько сильным предиктором является A_c для атрибута A_b. А именно, чем меньше

значение P_b, тем больше зависимость атрибута A_b от A_c. Аналогичным образом определяется величина P_c для атрибута A_c. На основании полученных данных, порядок для этих атрибутов устанавливается по мере убывания величин P_a и P_b.

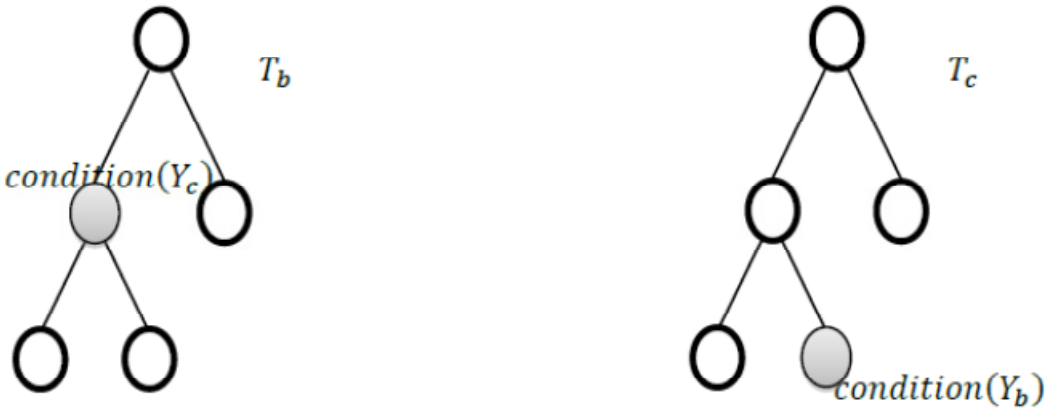


Рис. 5. Определение уровня разбиения по данному атрибуту

В результате, упорядоченные атрибуты обозначаются следующим образом: $A_{(1)}, A_{(2)}, \dots, A_{(d)}$.

Пусть, $A_{(k)}$ – текущий атрибут. С целью генерации синтетических значений для атрибута $A_{(k)}$, в первую очередь, строится дерево CART, $T_{(k)}$. Поскольку, $T_{(k)}$ используется с целью определения условного распределения атрибута $A_{(k)}$ в пространстве критических значений, то в качестве объектов для ее построения рассматриваются лишь те элементы U_{obs}, для которых $z_{i(k)} = 1$. Однако, если количество этих элементов не достаточно для построения корректной модели, тогда используются все элементы U_{obs}. А в качестве предикторов берутся остальные $(p-1)$ атрибутов, что обеспечивает максимальную информативность во время построения $T_{(k)}$. В отличие от традиционного метода построения деревьев CART, в данном алгоритме вместо механизма отсечения используется методика ранней остановки с применением проверки на нетривиальность разбиения, где в качестве критерия рассматриваются минимальное количество элементов и различных значений атрибута $A_{(k)}$.

Замещение значений. Далее, на основании полученного дерева $T_{(k)}$, осуществляются замещения значений атрибута $A_{(k)}$. В связи с тем, что в листьях $T_{(k)}$ содержатся элементы U_{obs}, однородные относительно значений $A_{(k)}$, то процесс замещения реализуется последовательно по листьям данного дерева. В данном алгоритме, замещения критических значений осуществляется благодаря методам перестановки (relocation) и переоценки (reevaluation) значений. Пусть, L есть текущий лист в дереве $T_{(k)}$, а $A_{(k)}^L = \{a_{(k)1}^L, a_{(k)2}^L, \dots, a_{(k)n_L}^L\}$ – множество значений атрибута $A_{(k)}$ в данном листе. Сначала осуществляется перестановка значений множества $A_{(k)}^L$. С этой целью применяется метод Байесовского бутстрапинга. Данный метод генерирует значения на основании некоторого множества возможных значений (donor pool). Для листа L в качестве данного множества рассматривается $A_{(k)}^L$. В согласии с процедурой Байесовского бутстрапинга, во-первых, генерируются $(n_L - 1)$ равномерно распределенные, произвольные числа в интервале $(0,1)$ и они упорядочеваются в порядке возрастания: $a_0 = 0, a_1, a_2, \dots, a_{(n_L-1)}, a_{n_L}$. Во-вторых, генерируются n_L таких же числе в интервале $(0,1]$, $u_1, u_2, \dots, u_i, \dots, u_{n_L}$, (Рис. 7), и наконец, для каждого u_i $(1 \le i \le n_L)$ определяется

интервал $(a_{j-1}, a_j]$, в котором оно содержится, $u_i \in (a_{j-1}, a_j]$, и соответствующее значение $a^L_{(k)i}$ заменяется на $a^L_{(k)j}$.

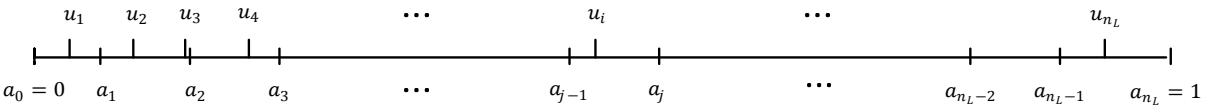


Рис. 6 Замещение значений по процедуре Байесовского бутстрапинга

Не исключено, что после перестановки значений атрибута $A_{(k)}$ некоторые из них могут остаться неизменными. В таких случаях, производится переоценка новых/переставленных значений $A_{(k)}$. С этой целью, в листе L определяется вероятностная плотность этих значений с помощью вычислителя плотности Гауссовского ядра (Gaussian kernel density estimator):

$$\hat{f}(x) = \frac{1}{hn_L}\sum_{i=1}^{n_L}\frac{1}{\sqrt{2\pi}}e^{\frac{-\left(x-a^L_{rep,(k)i}\right)^2}{2}},$$

$$h = \left(\frac{4\hat{\sigma}^5}{3n_L}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}{n_L}^{-\frac{1}{5}},$$

где $a^L_{rep,(k)1}, a^L_{rep,(k)2}, \dots, a^L_{rep,(k)n_L}$ – новые значения атрибута $A_{(k)}$ в листе L, а $\hat{\sigma}$ – среднеквадратическое отклонение этих значений. Затем, для каждого элемента данных произвольным образом выбирается значение вероятностной плотности из множества вычисленных. На основании этого значения вычисляются функция распределения $F(x)$ и ее обратная функция $F^{-1}(y)$:

$$F(x) = \int_{-\infty}^{x}\hat{f}(t)d(t),$$

$$F^{-1}(y) = \inf\{F(x) \geq y\}.$$

В результате, полученное значение функции $F^{-1}(y)$ устанавливается в качестве нового значения атрибута $A_{(k)}$ для рассматриваемого элемента данных.

Не исключено, что некоторые элементы данных, участвующие в построении дерева $T_{(k)}$ $(k > 1)$, могут содержать синтетические значения атрибутов $A_{(1)}, A_{(2)}, \dots, A_{(k-1)}$ построенные на предыдущих шагах алгоритма. Тогда, для обеспечения согласованности в замещениях, элементы данных с одинаковой комбинацией критических значений $A_{(k)}$ и новых значений $A_{(1)}, \dots, A_{(k-1)}$ рассматриваются отдельно от остальных. В тех случаях, когда эти элементы не содержатся в одном из листьев $T_{(k)}$, то алгоритм осуществляет обходит дерева снизу-вверх для определения внутренней вершины, в которой они содержатся. Замещения значений атрибута $A_{(k)}$ для этих элементов осуществляются в полученной вершине.

Таким образом, в результате последовательных замещений критических значений конфиденциальных атрибутов генерируется множество частично синтетических данных. Весь процесс повторяется

независимо m раз и полученные множества $SD_t, 1 \leq t \leq m$ предоставляются соответствующим организациям и общественным структурам.

Анализ алгоритма

Очевидно, что преимуществами приведенного алгоритма являются, во-первых, ее сложностные характеристики. Это связано с тем, что, в алгоритме осуществляется последовательная обработка конфиденциальных атрибутов, и это дает приемлемое приближение, и возможность получения качественных результатов даже при большом количестве конфиденциальных атрибутов и других данных. Во-вторых, использование деревьев CART дает возможность работать как с “качественными”, так и с количественными атрибутами, т.е. работать со смешанными данными (mixed data). Более того, благодаря этим деревьям возможно обнаружение сложных, не линейных функциональных связей между атрибутами данных. Однако, алгоритм не лишен недостатков. Так, ограничением алгоритма является его возможное применение при наличии относительно большого количества элементов данных. Поскольку, при существенно большом количестве последних возможно построение не точных а приближенных (усеченных) моделей CART, то это в свою очередь может явится причиной потери функциональных связей между различными атрибутами, что явно негативный фактор. Кроме того, согласно исследованиям [6] отсутствуют данные подтверждающие, что используемый порядок последовательного рассмотрения критических атрибутов является оптимальным. Не исключено, что при некотором другом порядке могли бы быть получены лучшие результаты. И наконец, алгоритм характеризуется некоторой нерациональностью, поскольку с целью отдельного рассмотрения элементов данных, содержащих некоторую комбинацию критических значений конфиденциальных атрибутов, алгоритм обходит соответствующее дерево CART, даже при их отсутствии. А это отрицательным образом сказывается на его производительности.

Основные причины - недостатки CART побудившие данное исследование тем не менее адресуют другие – нестандартные характеристики. Во первых, речь идет о трудоемкой процедуре отсечения (pruning) деревьев при помощи которой отсекаются часть ветвей при минимальной потере точности классификации или регрессионного приближения. Понятно что более логично выработать процедуры остановки построения дерева чем ее построить и потом отсечь. Это где то напоминает критерии кластеризации которые в иерархическом кластерном анализе управляют процесс завершения роста кластеров. По нашему предположению в задаче SDL по процедуре CART это может быть достигнуто путем применения на вершинах дерева не произвольных а критически характеризованных условий разделения текущей смеси объектов. Во вторых, порой очень сомнительны результаты замещений данных и непонятно что об этом не напоминают существующие публикации по теме. Имеется в виду следующее. Пусть определенное критическое значение оно только одно, и принимается на каком то количестве объектов. Существующие процедуры замещения не имеют пространства для замены значений ь значение одно и оно может быть замещено самым собой. Эли нужно расширить пространство замен, или нужно ввести меру соответствия задачи идее SDL. Мера соответствия для отдельного атрибута зависит от интервала применяемых значений и от количества различных значений, а общая мера для задачи интегрирует отдельные меры атрибутов. Ниже содержится продолжение описания улучшения CART учитывающее указанные недостатки внедрения стандартного метода.

Модификация алгоритма

Рассмотрим основную модель задачи. Пусть, $\mathcal{U} = \{U_1, U_2, \dots, U_N\}$ множество отдельных элементов, из которых составлены рассматриваемые входные данные задачи (*information units set*). Пусть отдельный элемент данных характеризуется множеством значений атрибутов множества $\mathcal{A} = \{A_1, A_2, \dots, A_p\}$.

Не нарушая общности, мы будем предполагать, что в данном наблюдении матрица U_{obs} состоит из первых n элементов множества $\mathcal{U}$. А конфиденциальные данные содержатся в первых d атрибутах:

$$A_{conf} = \{A_1, A_2, \dots, A_d\},\ A_{conf} \subseteq \mathcal{A}.$$

В принципе, в качестве конфиденциальных атрибутов могут быть рассмотрены как так называемые "качественные", так и количественные атрибуты. Однако, мы ограничимся рассмотрением только количественных атрибутов.

Для этих атрибутов вводится множество пороговых условий (*threshold conditions set)*, которые и определяют степень их конфиденциальности,

$$\mathcal{C} = \{C_1, C_2, \dots, C_d\}.$$

Условие C_j $(1 \le j \le d)$ определяет критическую область *(critical area)* значений атрибута A_j. Для простоты рассмотрений будем предполагать, что условия C_j задают числовые интервалы в области определения соответствующего атрибута A_j, хотя рассмотрение других структур ограничений может оказаться вполне естественным и полезным. Пусть, C_j определяет интервал $(\bar{c}_j, \bar{\bar{c}}_j)$ критических значений в области $(\bar{a}_j, \bar{\bar{a}}_j)$ определения атрибута A_j, $\bar{a}_j \le \bar{c}_j \le \bar{\bar{c}}_j \le \bar{\bar{a}}_j$.

Далее, мы предполагаем, что в качестве дополнительной информации нам дано множество $\mathcal{R}$, элементы которого констатируют наличие коррелированности между определенными группами атрибутов множества $\mathcal{A}$,

$$\mathcal{R} = \{R_1, R_2, \dots, R_t\}.$$

R_k $(1 \le k \le t)$ является подмножеством $\mathcal{A}$, $R_k \subseteq \mathcal{A}$, которое указывает на существование коррелированности (или выдвигает требование сохранения формы и степени коррелированности) между элементами этого множества атрибутов. Дальнейший анализ будет основан на предположении, что все атрибуты множества A_{conf} представлены в системе $\mathcal{R}$ и каждый ее элемент содержит хотя бы один конфиденциальный атрибут. Действительно, в обратном случае, если некоторый элемент R_k $(1 \le k \le t$ не содержит ни одного конфиденциального атрибута, то его рассмотрение не имеет смысла, ибо изменения критических значений атрибутов A_{conf} никак не отразятся на связи, представленного этим элементом. Кроме того, если некоторый атрибут A_j $(1 \le j \le d)$ не представлен в системе $\mathcal{R}$, то это означает, что A_j не коррелирован ни с одним из других атрибутов множества A_{conf}, что в свою очередь свидетельствует о том, что нет необходимости в рассмотрении элементов данных с комбинацией A_j и других конфиденциальных атрибутов.

Очевидно, что рассмотрение всех комбинаций конфиденциальных атрибутов не рационально, что выдвигает необходимость анализа системы $\mathcal{R}$ для выявления наиболее характерных комбинаций, которыми можно было ограничится. Пусть $d > 2$. Предположим также, что атрибуты, участвующие в определении коррелированности по $R_1, R_2, \dots, R_t$ только парные и содержат пересечения. Рассмотрим

подмножества $R_{k_1}, R_{k_2} \in \mathscr{R}$, характеризующие коррелированность между атрибутами A_{j_1}, A_{j_2} и A_{j_2}, A_{j_3}, $(1 < j_1 \neq j_2 \neq j_3 \leq m)$ соответственно, $R_{k_1} = \{A_{j_1}, A_{j_2}\}, R_{k_2} = \{A_{j_2}, A_{j_3}\}$. Допустим так же, что дополнительно не задана коррелированность между атрибутами A_{j_1} и A_{j_3}. В целях сохранения связи по R_{k_1} необходимо, что бы изменения значений A_{j_1} и A_{j_2} были согласованы. А именно, значения A_{j_1} должны быть изменены с учетом соответствующих значений атрибута A_{j_2} и наоборот. Аналогичные суждения имеют место для R_{k_2} и атрибутов A_{i_2}, A_{i_3}. Очевидно, что атрибут A_{j_2} зависит как от A_{j_1}, так и от A_{j_3}. Поэтому, для сохранения коррелированности по R_{k_1}, R_{k_2} значения атрибутов A_{j_1} и A_{j_3} так же должны изменены в согласии друг с другом. В результате, между атрибутами A_{j_1} и A_{j_3} возникает взаимосвязь, при условии рассмотрения атрибута A_{j_2}. Выше приведенные данные позволяют ввести следующее естественное определение.

Определение 1.

Скажем, что атрибуты A_{j_1} и A_{j_v} условно коррелированы, при условии рассмотрения атрибутов $A_{j_2}, A_{j_3}, \dots, A_{j_{v-1}}$, если существует набор парных коррелированностей $R_{k_1}, \dots, R_{k_{v-1}}$ так, что $R_{k_1} = \{A_{j_1}, A_{j_2}\}, R_{k_2} = \{A_{j_2}, A_{j_3}\}, \dots, R_{k_{v-1}} = \{A_{j_{v-1}}, A_{j_v}\}$.

Условную коррелированность атрибутов A_{j_1}, A_{j_v} обозначим через $R_{A_{j_1}, A_{j_2}, \dots, A_{j_v}} = \{A_{j_1}, A_{j_v}\}$.

Дальнейшим анализом системы $\mathscr{R}$ явилось изучение бинарного отношения между атрибутами, представленными в этой системе. Рассмотрим множество этих атрибутов обозначенное через A_{corr} (*correlated*).

Определение 2.

Скажем, что атрибут A_{j_1} входит в бинарное отношение α коррелированности с атрибутом A_{j_2}, $A_{j_1} \alpha A_{j_2}$, если A_{j_1} и A_{j_2} удовлетворяют одному из следующих условий:

- Атрибуты A_{j_1} и A_{j_2} совпадают: $A_{j_1} = A_{j_2} \Rightarrow A_{j_1} \alpha A_{j_2}$,
- Атрибуты A_{j_1}, A_{j_2} объявлены коррелированными множеством $\mathscr{R}$: $\exists R_k \in \mathscr{R}$, $R_k = \{A_{j_1}, A_{j_2}\}$ или $R_k = \{A_{j_2}, A_{j_1}\}$,
- Атрибуты A_{j_1}, A_{j_2} условно коррелированы: $\exists A_{j_3}, \dots, A_{v_v} \in A_{corr}$, такие что $R_{A_{j_1}, A_{j_3}, \dots, A_{j_v}, A_{j_2}} = \{A_{j_1}, A_{j_2}\} \Rightarrow A_{j_1} \alpha A_{j_2}$.

Очевидно, что α удовлетворяет свойствам рефлексивности и симметричности:

$$\forall A_{j_k} \in A_{corr} \Rightarrow A_{j_k} \alpha A_{j_k},$$

$$\forall A_{j_k}, A_{j_r} \in A_{corr}, A_{j_k} \alpha A_{j_r} \Rightarrow A_{j_r} \alpha A_{j_k}.$$

Покажем, что это отношение удовлетворяет также и свойству транзитивности, а именно:

$$\forall A_{j_k}, A_{j_r}, A_{j_s} \in A_{corr}, A_{j_k} \alpha A_{j_r}, A_{j_r} \alpha A_{j_s} \Rightarrow A_{j_k} \alpha A_{j_s}.$$

Так как $A_{j_k} \alpha A_{j_r}$, $A_{j_r} \alpha A_{j_s}$, то из определения отношения α следует, что между атрибутами A_{j_k}, A_{j_r} и A_{j_r}, A_{j_s} существует либо прямая, либо условная коррелированность. Тогда в силу **определения 1** атрибуты A_{j_k} и A_{j_s} будут условно коррелированными. А это в свою очередь означает, что A_{j_k} входит в отношение α с атрибутом A_{j_s}: $A_{j_k} \alpha A_{j_s}$.

Итак, отношение α удовлетворяет свойствам рефлексивности, симметричности и транзитивности, следовательно, оно является отношением эквивалентности. В этом случае, α разбивает множество A_{corr} на непересекающиеся классы эквивалентности:

$$A_{corr} = A_{corr}^1 \cup A_{corr}^2 \cup \ldots \cup A_{corr}^s,$$

$$A_{corr}^i \cap A_{corr}^j = \emptyset,\ 1 \le i \ne j \le s.$$

Причем, любые два атрибута одного и того же класса взаимосвязаны друг с другом, а между атрибутами различных классов коррелированность отсутствует.

Данный анализ позволяет заключит, что подобное разбиение множества A_{corr} на классы эквивалентности дает возможность ограничится рассмотрением возможных определенных комбинаций конфиденциальных атрибутов в пределах одного класса. Кроме того, дальнейшее рассмотрение конфиденциальных атрибутов целесообразней производить последовательно в каждом классе в отдельности.

Пусть, A_{corr}^i – текущий класс эквивалентности. Для удобства интерпретации, рассмотрим частный случай, когда класс A_{corr}^i состоит только из трех конфиденциальных атрибутов: $A_{corr}^i = \{A_{j_1}, A_{j_2}, A_{j_3}\}$; $A_{j_k} \in A_{conf}$ $(k = 1,2,3)$. Не нарушая общности, допустим, что порядок рассмотрения этих атрибутов в построении дерева расщеплений следующий: $A_{j_1} - A_{j_2} - A_{j_3}$. Дерево T_{j_1}, соответствующее атрибуту A_{j_1}, строится на основании множества элементов данных с критическими значениями этого атрибута, $U^{A_{j_1}} = \{U_k,\ \overline{c_{j_1}} \le a_{kj_1} \le \overline{\overline{c_{j_1}}}\}, U^{A_{j_1}} \subseteq U_{obs}$. Поскольку, элементы данных, содержащие комбинации критических значений атрибутов A_{j_1} и A_{j_2}, A_{j_3}, должны быть рассмотрены в отдельности от остальных элементов множества $U^{A_{j_1}}$, тогда целесообразней осуществить процедуру их отделения на начальной стадии/этапе построения дерева T_{j_1}. С этой целью, в первую очередь произвести разбиения множества $U^{A_{j_1}}$ по атрибутам A_{j_2}, A_{j_3} и в качестве условий разбиений рассматривать наличие критических значений этих атрибутов в элементах множества $U^{A_{j_1}}$ (Рис.7).

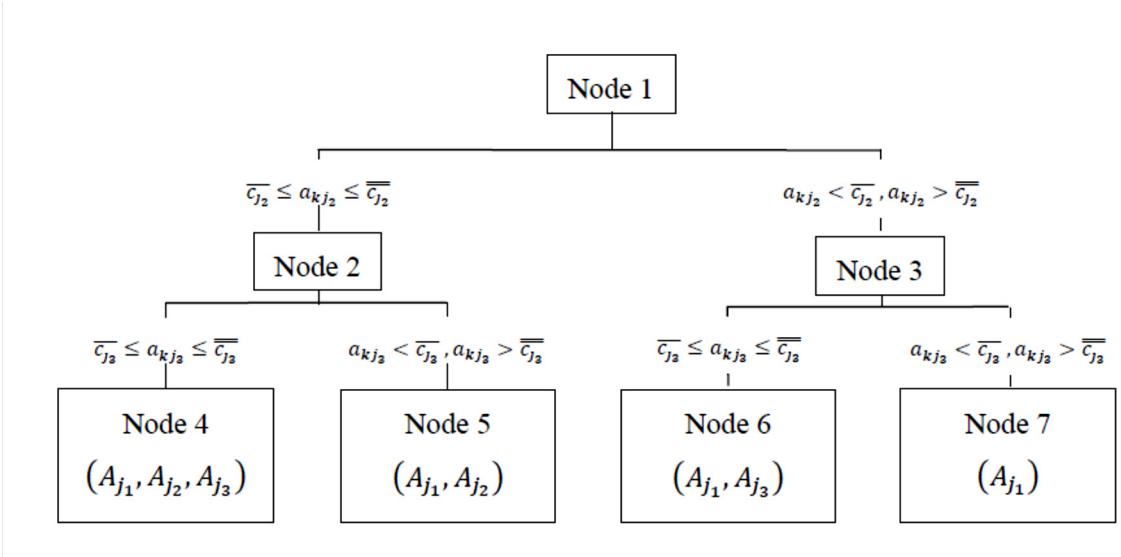


Рис. 7 Начало дерева по группе критических атрибутов

Далее, как видно из рисунка 7 в узлах Node 4, Node 5, Node 6 содержатся критические значения определенных комбинаций атрибутов, поэтому дальнейшие разбиения каждого из этих узлов необходимо осуществить таким образом, чтобы с одной стороны сохранить корреляции между элементами класса A^i_{corr}, а с другой стороны обеспечить однородность данных по соответствующей комбинации в узлах-потомках. Поскольку, в наших исследованиях мы ограничиваемся рассмотрением только количественных атрибутов в качестве конфиденциальных, то для дальнейших разбиений этих узлов мы применяем методику разбивающего иерархического кластерного анализа [1]. В этом случае, элементы данных рассматриваются как 3-мерные векторы, состоящие из значений атрибутов класса A^i_{corr}, что позволяет осуществлять разбиения с учетом корреляций между атрибутами этого класса. В качестве меры расстояния между элементами данных рассматривается Эвклидово расстояние, а в качестве меры однородности полученных подмножеств - мера RMSSTD (Root-Mean-Square Standard Deviation) [14], равная среднеквадратичному отклонению критических значений атрибутов соответствующей комбинации. Что касается узла Node 7, то, поскольку в нем содержатся критические значения только атрибута A_{j_1}, его разбиения будут осуществляться тем же способом, что и в деревьях CART.

По существу, листья дерева T_{j_1} будут содержат элементы данных однородные либо по атрибуту A_{j_1}, либо по некоторым комбинациям A_{j_1} и остальных атрибутов класса A^i_{corr}. В тех листьях, где содержатся критические значения только атрибута A_{j_1}, замещения могут быть осуществлены так же, как и в ранее представленном алгоритме (с использованием Байесовского бутстрапринга и вычислителя плотности Гауссовского ядра), а в остальных листьях - по наборам значений соответствующих атрибутов комбинации вместо последовательных замещений по каждому из них. Благодаря этому, по возможности сохраняется корреляция между атрибутами отдельных комбинаций.

Для атрибутов A_{j_2}, A_{j_3} процесс построения соответствующих деревьев T_{j_2}, T_{j_3} и дальнейшие замещения их критических значений осуществляются аналогичным образом. Очевидно, что с целью построения дерева T_{j_2} будут применены элементы данных множества $U^{A_{j_2}} \setminus U^{A_{j_1}}$. Это предполагает, что в этом дереве выборка данных будет основана на отделении элементов с комбинацией критических значений атрибутов $\left(A_{j_2}, A_{j_3}\right)$ от остальных элементов. Что касается дерева T_{j_3}, то анализ всех возможных комбинаций атрибутов класса A^i_{corr} в деревьях T_{j_1} и T_{j_2}, позволяет рассматривать T_{j_3} как дерево CART со множеством данных $U^{A_{j_3}} \setminus \left(U^{A_{j_1}} \cup U^{A_{j_2}}\right)$.

Приведенный простой пример позволяет заключить, что если класс эквивалентности A^i_{corr} состоит из m_i атрибутов, $A^i_{corr} = \left\{A_{j_1}, A_{j_2}, \dots, A_{j_{m_i}}\right\}$, из которых первые q являются конфиденциальными, тогда для атрибута A_{j_k} $(1 \le k \le q)$ бинарное дерево решений T_{j_k} строится на основании множества элементов $U^{A_{j_k}} \setminus \cup_{r=1}^{k-1} U^{A_{j_r}}$. Причем, при построении этого дерева, в первую очередь, рассматриваются разбиения по атрибутам $A_{j_{k+1}}, A_{j_{k+2}}, \dots, A_{j_q}$ с целью определения и отдельного рассмотрения элементов данных, содержащих комбинации критических значений атрибутов A_{j_k} и $A_{j_{k+1}}, A_{j_{k+2}}, \dots, A_{j_{m_i}}$. В результате, в листьях T_{j_k} будут содержатся элементы данных однородные либо по критическим значениям некоторой

комбинации атрибутов A_{j_k} и $A_{j_{k+1}}, A_{j_{k+2}}, \dots, A_{j_{m_i}}$, либо по критическим значениям только атрибута A_{j_k}. А замещения осуществляются либо по наборам значений соответствующих комбинаций, либо по атрибуту A_{j_k}.

Таким образом, представленные данные с очевидностью свидетельствуют, что при наличии дополнительной информации в виде системы $\mathscr{R}$ изначально могут быть детерминированы элементы данных, содержащие наиболее важные/первостепенные комбинации конфиденциальных атрибутов. Кроме того, основываясь на рассмотренных количественных атрибутах в качестве конфиденциальных, становится ясным возможность разбиения множества элементов, содержащую некоторую комбинацию классов на более однородные, по значениям атрибутов этой комбинации, и подмножества и дальнейшие замещения по наборам значений атрибутов позволят сохранить первостепенные корреляции между этими атрибутами по системе $\mathscr{R}$. Данные модельные структуры и анализ подтверждаются вычислительными экспериментами.

Заключение

Задачи сохранения конфиденциальности данных при распределенных вычислениях связаны с новыми теоретическими и прикладными исследованиями. С одной стороны, криптография пытается синтезировать схемы кодирования, в которых результат анализа этих данных совпадает с анализом исходных, не кодированных данных, однако известно что такие прикладные системы будут созданы не так скоро. Альтернативные модели анализа данных существуют и имеют эвристический характер. В данной работе исследовались схемы вычислений с сохранением конфиденциальности данных которые основаны на использовании моделей CART и введены дополнительные компоненты модели повышающие скорость вычислений и близость результатов анализа данных к исходным. Результат достигается путем исключения этапа урезания деревьев из процесса анализа данных, что основано на том, что в качестве условий разветвления деревьев используются условия определяющие конфиденциальность данных задачи.

Литература

1. L. Aslanyan, V. Topchyan, Hierarchical Cluster Analysis For Partially Synthetic Data Generation, Transactions of IIAP of NAS of RA, Mathematical Problems of Computer Science, submitted, 2013.
2. Vardan Topchyan, Statistical Disclosure Limitation of Public Use Data by Syntheses with Clustering, ITA 2013 – ITHEA ISS Joint International Events on Informatics, Winter Session, December 18 – 19, 2013, Sofia, Bulgaria, pp. 17.
3. Willenborg, L. and de Waal, T. (2001). Elements of Statistical Disclosure Control. New York: Springer-Verlag.
4. Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 462–468.
5. Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19, 1–16.
6. Reiter, J.P. (2005). Significance Tests for Multi-component Estimands from Multiply-imputed, Synthetic Microdata. Journal of Statistical Planning and Inference, 131, 365 - 377.

7. Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.
8. Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.
9. Barros, Rodrigo C., Basgalupp, M. P., Carvalho, A. C. P. L. F., Freitas, Alex A. (2011). A Survey of Evolutionary Algorithms for Decision-Tree Induction. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol. 42, n. 3, p. 291-312, May 2012.
10. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, Foundations of Machine Learning, MIT Press, 2012.
11. Donald E. Knuth, The Art of Computer Programming, Volume 3, Sorting and Searching, Second Edition (Reading, Massachusetts: Addison-Wesley, 1998), 780pp, ISBN 0-201-89685-0
12. A. V. Aho, J. E. Hopcroft, J. D. Ullman, Data Structures and Algorithms. Addison-Wesley, 1983. ISBN 0-201-00023-7.
13. M. Yu. Moshkov, Conditional tests, Problemi Kibernetiki (in Russian), issue 40, pp. 131-170, Moscow, Nauka, 1983.
14. Subhash Sharma: Applied multivariate techniques, John Wiley & Sons, Inc., 1996.
15. Little, R.J.A. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 407–426.
16. Kennickell, A.B. (1997). Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, (eds), Record Linkage Techniques, 1997, 248–267. Washington, D.C.: National Academy Press.
17. Abowd, J. M. and Woodcock, S. D. (2001). Disclosure Limitation in Longitudinal Linked Data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds). Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, 215–277. Amsterdam: North-Holland.
18. Liu, F. and Little, R.J.A. (2002). Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata. In Proceedings of the Joint Statistical Meetings of the American Statistical Association, 2133–2138.
19. Jorg Drechsler (2011). Synthetic Datasets for Statistical Disclosure Control. Theory and Implementation.
20. Reiter, J.P. (2005). Using CART to Generate Partially Synthetic, Public Use Microdata. Journal of Official Statistics, Vol. 21
21. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and Regression Trees. Belmont, CA: Wadsworh, Inc.
22. Reiter, J.P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology, 181–189.
23. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze, (2008) Introduction to Information Retrieval,378-382.
24. Rubin, D.B. (1981). The Bayesian Bootstrap. The Annals of Statistics, 9, 130–134.
25. Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis, (2002) Clustering Valifity Checking Methods: Part II, (19-27).
26. L. Aslanyan, H. Danoyan, P. Zelenko, Applied Best-Match Type Algorithms, Proceedings of the 7th annual scientific conference of RAU, Yerevan, pp. 56-62, 2013.
27. L. H. Aslanyan, H. E. Danoyan, Complexity of Elias algorithm based on codes with covering radius three, Proceedings of the Yerevan state university, 2013 №1, pp. 44-50.
28. L. H. Aslanyan, H. E. Danoyan, Complexity of Elias algorithm based on Hamming and extended Hamming codes, Reports of NAS RA, vol. 113, no. 2, pp. 151-158, 2013.
29. L. H. Aslanyan, H. E. Danoyan, “On the optimality of a hash-coding type search algorithm”, Proceedings of the 9th conference CSIT, Yerevan, Armenia, pp. 55-57, 2013.

Информация об авторах

Levon Aslanyan *– ITHEA ISS, Sofia, Bulgaria; EC Horizon2020 ICT NCP Armenia; Institute for informatics and automation problems of NAS RA, 1, P. Sevak street, Yerevan 0014, Armenia, e-mail: lasl@sci.am*

Major Fields of Scientific Research: Discrete optimization, Artificial intelligence, NLP, WSN, Privacy preserved computation

Vardan Topchyan *– Institute for informatics and automation problems of NAS RA, 1, P. Sevak street, Yerevan 0014, Armenia, e-mail: vardan.topchyan@gmail.com*

Major Fields of Scientific Research: Decision models, Homomorphic encryption, Privacy preserved computation

Enhanced Cart Technologies in Partial Synthetic Data Generation

Levon Aslanyan, Vardan Topchyan

Abstract: *This work aims at studying personal data analysis area, when confidentiality property of data is ensured. It is supposed that we are given partially critical social science data and prior to the submission of data to the public it is required to modify them so that confidential information is not disclosed, and that the analysis of these data did not differ from the analysis of raw data. Our work builds improved algorithms of class of classification and regression trees, which provide solution to the problem of generation of the so-called synthetic data. The new solution of generation takes into account the structure of the areas of privacy and is providing optimized tree replacement for synthetic data sets.*

Keywords: *classification, regression, data disclosure, synthetic data.*

СИСТЕМНЫЙ АНАЛИЗ НАПРАВЛЕНИЙ ВЫЧИСЛИТЕЛЬНОГО ИНТЕЛЛЕКТА

Юрий Зайченко, Михаил Згуровский

Аннотация: *Исследование и анализ новой области искуственного интеллекта - вычислительный интеллект (ВИ) представлены в статье. Основные компоненты ВИ-технологий, методов и приложений, как и их взаимосвязи определяются и анализируются в этой работе. Обсуждены взаимоотношения ВИ и мягких вычислений.*

Ключевые слова: *вычислительный интеллект, мягкие вычисления, нейронные сети, нечеткая логика, эволюционное моделирование.*

ACM Classification Keywords*: I.2 Artificial Intelligence, I.2.11 Distributed Artificial Intelligence*

Введение

Современный этап развития систем поддержки принятия решений характеризуется все большей их интеллектуализацией, причем процесс автоматизации творческих процессов принятия решений идет как вглубь, так и вширь, охватывая все новые и новые сферы, которые считались ранее целиком прерогативой человека. Стремление человечества переложить на компьютеры часть интеллектуальных задач, выполняемых ЛПР (лицом принимающим решения) появилось уже на заре компьютеризации. С появлением компьютеров стали возникать вопросы, м ожет ли машина «мыслить»? Возможно ли поручить ЭВМ, часть функций мыслительных функций человека, появились работы по моделированию отдельных функций человеческого мозга, в частности по распознаванию образов, работы по моделированию мышления и психики.

В 1950 году А. Тьюринг сформулировал свой знаменитый тест проверки на интеллектуальность компьютера. Если поместить человека и компьютер в разные комнаты, а оператор с использованием клавиатуры будет задавать им вопросы и если ответы компьютера и человека совпадут, то оператор не сможет отличить человека от компьютера, и такой компьютер можно считать интеллектуальной машиной (т.е. системой искусственного интеллекта). А. Тьюринг полагал, что уже к 2000 году компьютер с объемом памяти в 10^9 бит и скоростью обработки (10^6- 10^7) оп/с сможет пройти этот тест в течение 5 мин. с вероятностью 70%. Заметим, что в настоящее время компьютеры имеют память и быстродействие на несколько порядков выше, чем предполагал А. Тьюринг, однако пока ни один компьютер не в состоянии пройти тест Тьюринга в полном объеме, поскольку выполнение теста Тьюринга связано с рядом сложных проблем, а именно пониманием смысла текстов на естественном языке, автоматизацией поиска решений задач в различных предметных областях и рядом других, пока еще далеких от своего решения.

С момента появления первых работ в области искусственного интеллекта прошло более 50 лет. За прошедшие годы данная отрасль науки прошла сложный и поучительный путь развития. В ней сформировался *ряд направлений*, таких, как *системы основанные на знаниях, логический вывод, поиск решений, системы распознавания образов, системы машинного перевода, обучение и самообучение,*

планирование действий, агенты и мультиагентные системы, самоорганизация и самоорганизующиеся системы, нейронные сети, системы с нечеткой логикой и нечеткие нейронные сети, моделирование эмоций и психики, интеллектуальные игры, роботы и робототехнические системы.

В настоящее время существует достаточно много определений понятия искусственный интеллект. Многие из них приведены в монографиях [Рассел Стюарт, 2007], а также Ф. Люгера [Люгер Ф., 2006], в которых дается фундаментальное изложение основных направлений искусственного интеллекта. Нет смысла приводить их в данной работе. На наш взгляд, более важным является выделить **основные особенности и свойства**, **систем искусственного интеллекта,** отличающие их от обычных систем автоматизации. Эти свойства таковы [Згуровский М.З., 2013]:

1) *наличие цели или группы целей функционирования;*
2) *способность планирования своих действий и поиск решений задач;*
3) *способность к обучению и адаптации поведения в процессе работы,*
4) *способность работать в плохо формализованной среде, в условиях неопределенности; работать с нечеткими инструкциями;*
5) *способность к самоорганизации и саморазвитию,*
6) *способность понимать тексты на естественном языке*
7) *способность к обобщению и абстрагированию накопленной информации.*

Для того, чтобы создать машины, которые бы приближались по своим возможностям человеческому мозгу, необходимо прежде всего понять сущность интеллекта человека, раскрыть механизмы человеческого мышления. За истекшие десятилетия этой проблеме было посвящено много работ. Среди монографий, появившихся в последнее время, необходимо выделить монографию Джеффа Хокинса и Сандры Блейксли [Хокинс Джефф, 2007], в которой авторы, на наш взгляд, ближе всего подошли к пониманию основы человеческого интеллекта. В ней авторы, на стыке нейробиологии, психологии и кибернетики разработали пионерскую теорию, в которой построена модель мозга человека, главными функциями которой являются запоминание прошлого опыта и прогнозирование мозгом результатов восприятия окружающей действительности и своих действий. Авторы приводят множество убедительных примеров поведения человека в различной обстановке, подтверждающих эту идею. Дж. Хоукинс отмечает: «… Прогнозирование по моему мнению- это не просто одна из функций коры головного мозга, Это *первичная функция неокортекса и основа интеллекта*. Кора головного мозга является органом предвидения. Если мы хотим понять, что такое разум, что такое творчество, как работает наш мозг, и как научиться создавать разумные машины, нам нужно постичь природу прогнозов и понять, каким образом кора головного мозга их формирует. Даже поведение можно лучше всего представить как промежуточный продукт процесса прогнозирования» [Хокинс Джефф, 2007].

Целью настоящей работы является обзор и анализ современного направления в области ИИ, получившего название вычислительный интеллект.

Основные компоненты вычислительного интеллекта

В ходе развития работ в области ИИ в начале 90-х гг. путем интеграции ряда интеллектуальных технологий и методов сформировалось новое направление, получившее название *вычислительный интеллект (*computational intelligence).

Существует несколько определений термина вычислительный интеллект. Впервые термин «вычислительный интеллект» (ВИ - computational intelligence) был введен Бездеком [Bezdek J., 1994], который определил его так: «система является интеллектуальной вычислительно, если она: оперирует только с цифровыми данными; имеет компоненты распознавания образов; не использует знания в смысле искусственного интеллекта и вдобавок когда она проявляет:

а) вычислительную адаптивность;

б) вычислительную отказоустойчивость;

в) уровень ошибок, аппроксимирующий характеристики человека.

В дальнейшем, это определение уточнялось и расширялось. Так, Маркс в определении ВИ делает акцент на составляющие технологии ВИ [Marks R., 1993]: «…нейронные сети, генетические алгоритмы, нечеткие системы, эволюционное программирование и искусственная жизнь являются строительными блоками ВИ».

Другая попытка определения ВИ была сделана Фогелем [Fogel D.,1995]: «эти технологии нейронных сетей, нечетких и эволюционных систем были интегрированы под вывеской «вычислительный интеллект»-сравнительно новый термин, предложенный для обобщенного описания методов вычислений, которые могут быть использованы, чтобы адаптировать решения к новым проблемам и не базироваться на явных человеческих знаниях.

За прошедшие годы было выполнено большое число работ, посвященных различным направлениям в области ВИ, регулярно проводятся международные конференции и конгрессы по вычислительному интеллекту, в Международном институте IEEE издается специализированный журнал, посвященный проблематике вычислительного интеллекта IEEE Transactions on Computational Intelligence.

Анализ этих работ позволяет дать следующее определение ВИ [Згуровский М.З., 2013].

Под вычислительным интеллектом (ВИ, computational intelligence) будем понимать совокупность технологий, моделей, методов и программных средств, предназначенных для решения неформальных, творческих задач в различных сферах человеческой деятельности с использованием аппарата и логики, в определенной степени отождествляющих мыслительную деятельность человека (нечеткость рассуждений, качественный и интуитивный подходы, креативность, логический вывод, самообучение и т.д.) в частности принятии решений, классификации, распознавании образов и т.д.

Следует отметить взаимосвязь между искусственным интеллектом (ИИ) и вычислительным интеллектом. ВИ – это составная часть направлений (разделов) современного ИИ, использующих специальные модели, методы и технологии и ориентированных на решение определенных классов задач.

Структура направлений и методов ВИ приведена на рис.1.

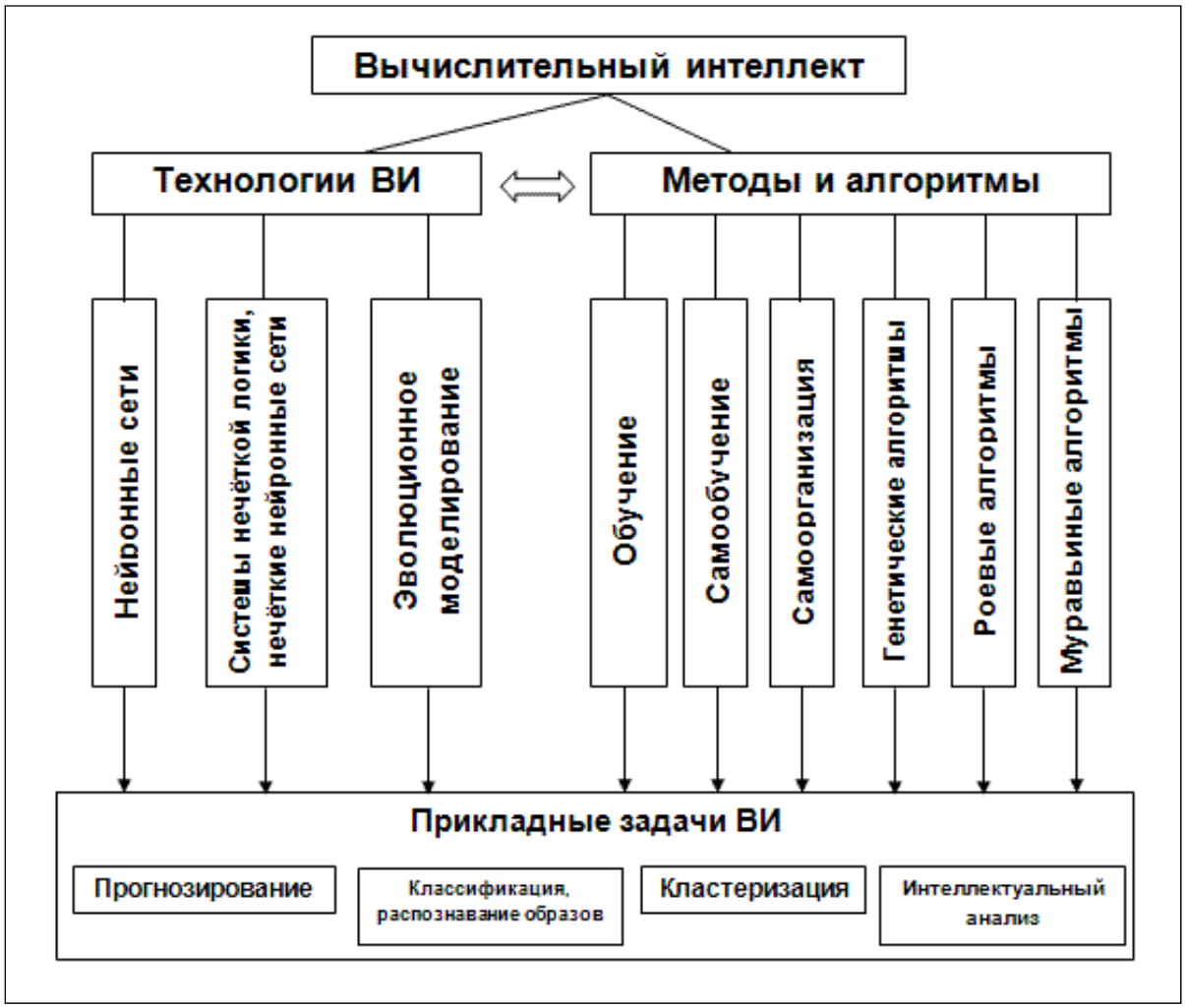


Рис. 1 Структура вычислительного интеллекта

В структуре ВИ можно выделить следующие компоненты [Згуровский М.З., 2013]:

- технологии;
- модели, методы и алгоритмы;
- прикладные задачи.

Технологии ВИ включают:

- нейронные сети (НС);
- системы нечеткой логики (СНЛ) и нечеткие нейросети (ННС);
- эволюционное моделирование (ЭМ).

К методам и алгоритмам ВИ можно отнести:

- методы обучения;
- методы самообучения;
- методы самоорганизации;

- генетические алгоритмы (ГА);
- роевые алгоритмы;
- муравьиные алгоритмы.

Технологии и методы ВИ используются при решении соответствующих задач ИИ. Логично выделить следующие основные классы задач ВИ, характерных для мыслительной деятельности человека:

- прогнозирование и предвидение;
- классификация и распознавание образов;
- кластеризация, самопроизвольное разбиение множества объектов на классы однородных объектов;
- интеллектуальный анализ данных;
- логический вывод
- принятие решений.

Взаимосвязь вычислительного интеллекта и «мягких вычислений»

Термин ВИ близок по своему значению, широко используемому в зарубежной литературе термину «soft computing» [Zadeh L.A., 1984] (т.е. «мягкие» вычисления), под которым понимается совокупность моделей, методов и алгоритмов, базирующихся на применении нечеткой математики (нечетких множеств и нечеткой логики).

Понятие мягких вычислений (soft computing) впервые было упомянуто в работе Л. Заде (Lotfi A. Zadeh) по анализу мягких (soft) данных в 1981 году. Мягкие вычисления (МВ) - это сложная компьютерная методология, основанная на нечеткой логике (НЛ), генетических вычислениях, нейрокомпьютинге и вероятностных вычислениях [Zadeh L.A., 1984]. Составные части не конкурируют, но создают эффект синергизма. Ведущий принцип МВ - это учет неточности, неопределенности, частичной истины и аппроксимации для достижения робастности, низкой цены решения, большего соответствия с реальностью.

Четыре составные части мягких вычислений включают в себя:

- нечеткую логику - приближенные вычисления, грануляция информации, вычисление на словах;
- нейрокомпьютинг - обучение, адаптация, классификация, системное моделирование и идентификация;
- генетические вычисления - синтез, настройка и оптимизация с помощью систематизированного случайного поиска и эволюции;
- вероятностные вычисления - управление неопределенностью, сети доверия, хаотические системы, предсказание.

Традиционные компьютерные вычисления (hard computing) слишком точные для реального мира. Имеется два класса задач для мягких вычислений: во-первых, существуют проблемы, для решения которых полная и точная информация не может быть получена и, во-вторых, проблемы, определение которых является не достаточно полным.

Каждая из отдельных компонент МВ обладает рядом внутренних проблем, взаимный учет которых и создает синергетический эффект. Например, нейронные сети решают задачи аппроксимации или

классификации, причем настройка весов происходит за счет обучения. Но знания распределены по многим связям и результаты работы нейронной сети кажутся пользователю необъяснимыми, то есть необходима экспликация знаний. В тоже время, системы нечеткого вывода, наоборот, обладают явными знаниями в виде продукций и позволяют легко построить протокол объяснений. Однако такая ясность достигается долгим и творческим по своей сути предшествующим процессом извлечения знаний и последующей отладкой совокупности правил.

Отсутствие возможности извлечения знаний из данных, настройки параметров функций принадлежности в ходе обучения, автоматизированной редукции правил – это характерные проблемы систем логического вывода. Применение генетических алгоритмов (ГА) требует предварительного исследования проблемы с целью выбора вероятностной модели, формулировки целевой функции.

Выбор отдельной компоненты МВ в качестве базовой обычно определяет общую архитектуру гибридной системы (под гибридной системой в данном случае понимается “мягкая” система).

Таким образом, между вычислительным интеллектом (ВИ) и мягкими вычислениями много общего: общие парадигмы, принципы, методы, технологии и прикладные задачи. Отличия их, на наш взгляд, состоят в подходах, МВ ориентированы на методологические, философские и математические проблемы, а ВИ ориентирован в основном на вычислительные алгоритмы и технологии и практическую реализации соответствующих моделей и методов.

Общая характеристика технологий и методов вычислительного интеллекта

Между технологиями и методами ВИ с позиций системного анализа можно выделить следующие взаимосвязи:

1) методы обучения широко используются для настройки весов связей нейронных сетей и нечетких нейронных сетей. В качестве критерия обучения здесь используются следующие критерии:

– СКО (MSE):

$$\overline{q^2} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i}(w))^2 \to \min_w \tag{1}$$

– средняя модульная процентная ошибка (MAPE):

$$\varepsilon_{cano}^2 = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \widehat{y_i}(w)\right| \Big/ \left|y_i\right| \to \min_w \tag{2}$$

где y_i - желаемый выход НС для i -го входного сигнала; $\widehat{y_i}$ - фактический выход НС для i -го входа; W - матрица весов, $W = \left|w_{kj}\right|_{k=\overline{1,n},\, j=\overline{1,m}}$.

2) при обучении, как правило, задается обучающая выборка $L = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, где x_i - входной вектор $(x \in R^{(n)})$, y_i - выходной сигнал (желаемый выход – цель обучения).

3) для обучения используются методы или алгоритмы оптимизации, в частности: градиентный метод и его модификации [16, 20, 44], метод сопряженных градиентов (СГ), методы второго порядка (метод Ньютона), а также широко используются ГА [15].

В том случае, когда учитель отсутствует (т.е. цель y_i неизвестна или не задана) используются методы и алгоритмы самообучения. В этом случае задача состоит в автоматическом разбиении выборки L на ряд подмножеств, называемых «кластерами». При этом задаются соответствующие критерии и метрики разбиения.

Одной из первых технологий ВИ явились искусственные нейронные сети (ИНС). Они являются моделями нейронной системы (мозга) человека и высокоорганизованных животных. Первой известной ИНС явился разработанный Фр. Розенблаттом в 1962-65 гг. перцептрон (от слова percertio – восприятие) [Розенблатт Ф., 1965].

Первый вариант перцептрона – трехслойный перцептрон (элементарный) состоял из 3 типов нейронов: S - нейронов – сенсорных элементов, A - нейронов – ассоциативных, и R - нейронов – реагирующих, представляющих выходную реакцию нейросети на входной сигнал.

Перцептрон является моделью зрительного механизма мозга человека.

При проектировании изображения на сенсорное поле из S - нейронов, представляющих чувствительные фотоэлементы, S - нейроны (i, j), на которые попадают элементы изображения, срабатывают и выдают сигнал $S_{ij}(t) = 1$. Сенсорные элементы соединены случайными связями с A - элементами, аналогично тому как в зрительном механизме человека. A - связи (i, j) бывают двух типов: возбуждающие, их вес $w_{ij} = 1$, и тормозящие $w_{ij} = -1$. Число входных возбуждающих связей x, а тормозящих y - одинаково для всех A-нейронов.

Суммарный сигнал g_j на входе ассоциативного нейрона A_j определяется так:

$$g_j = \sum_{i \in P_j} w_{ij} s_i = \sum_{i \in P_j^{в}} s_i - \sum_{i \in P_j^{т}} s_i \qquad (3)$$

где P_j - множество входных связей (сигналов нейрона A_j), $P_j^{в}$ - множество возбуждающих связей у нейрона A_j, P_j^{m} - множество тормозящих связей. Очевидно, $\left|P_j^{в}\right| = x$, $\left|P_j^{m}\right| = y$.

Выходной сигнал ассоциативного нейрона A_j определяется

$$a_j = f_1(g_j)$$

где $f_1(\cdot)$ - функция активации A-нейронов.

Здесь чаще всего используется 2 вида функций:

а) релейная $a_j = \begin{cases} 1, \textit{если } g_j \geq \theta, \textit{где } \theta - \textit{величина порога}; \\ 0, \textit{в противном случае.} \end{cases}$

б) сигмоидальная $f_1(x) = \frac{1}{1+e^{-x}}$.

A-нейроны связаны с R-нейронами «каждый с каждым». Обозначим через w^o_{jk} - вес связи нейрона A_j с нейроном R_k. Суммарный сигнал S_k на входе R_k определяется из условия:

$$S_k = \sum_{j=1}^{J} a_j w^o_{jk} \tag{4}$$

Срабатывает тот нейрон R_k, на входе которого сигнал S_k максимален. Сработавший нейрон определяет соответствующий класс объектов (образ). Если обозначить через r_k выходную реакцию нейрона R_k, то

$$r_k = \begin{cases} 1, \textit{ если } S_k = \max\limits_l S_l; \\ 0, \textit{ в противном случае.} \end{cases} \tag{5}$$

Веса связей $\left\| w^o_{jk} \right\|$ настраиваются так, чтоб обеспечить требуемую классификацию изображений. Для этого используются соответствующие алгоритмы обучения. Фр. Розенблатт разработал несколько алгоритмов обучения элементарного перцептрона, α - алгоритм обучения и γ-алгоритм обучения. Это так называемые алгоритмы «с поощрением – наказанием» [Розенблатт Ф., 1965].

В случае правильной реакции перцептрона веса активных входных связей срабатывающего нейрона R_k увеличиваются, поощряются так:

$$w_{jk}(t+1) = w_{jk}(t) + \triangle w. \tag{6}$$

В случае неправильной реакции, наоборот, наказываются:

$$w_{jk}(t+1) = w_{jk}(t) - \triangle w.$$

Это так называемый « α - алгоритм обучения».

γ-алгоритм обучения отличается тем, что в нем веса связей изменяются так, что общая сумма весов у каждого нейрона остается прежней.

Фр. Розенблатт провел множество экспериментальных исследований 3-х слойного перцептрона по распознаванию простейших изображений: цифр, букв и геометрических фигур. Эти эксперименты показали способность перцептрона обучаться правильной классификации образов. Тем самым были убедительно продемонстрированы интеллектуальные возможности перцептрона как модели зрительного механизма мозга.

Опыты Розенблатта стимулировали интерес ученых к нейронным сетям. В конце 60-х – начале 70-х годов возник «бум» работ по исследованию НС. Появились новые архитектуры нейронных сетей, в частности работы Б. Уидроу предложившего нейрон Adaline и НС Madaline, обучающие матрицы Штейнбуха, распознающая система Альфа А.Г. Ивахненко.

Усилия многих ученых были направлены на выяснение потенциальных возможностей перцептрона. Математик П. Новиков доказал ряд теорем о сходимости процесса обучения перцептрона к искомой реакции. Однако ожидания ученых относительно чрезвычайно широких возможностей перцептрона как модели механизма мозга не оправдались. В процессе дальнейших исследований было выяснено, что элементарный трехслойный перцептрон не обладает такими неотъемлемыми свойствами интеллекта человека, как способность к экстраполяции (обобщению), т.е. перцептрон обученный распознавать некий объект в одной области сенсорного поля не способен правильно распознать его в другой области (без дополнительного обучения). Кроме того, перцептрон не обладает инвариантностью к изменению масштаба и повороту изображения.

Наконец, в 1971 г. профессор из МТИ М. Минский и С. Пайперт опубликовали книгу [Минский М., 1971], в которой показали, что трехслойный перцептрон не способен решать даже простую логическую задачу – реализовать логическую функцию XOR – исключительное «Или», в которой бинарные объекты $X_1 = [0;0]$ и $X_2 = [1;1]$ должны быть отнесены к одному классу (к=1), а $X_3 = [1;0]$ и $X_4 = [0;1]$ - к другому (к=2) (см. рис.3).

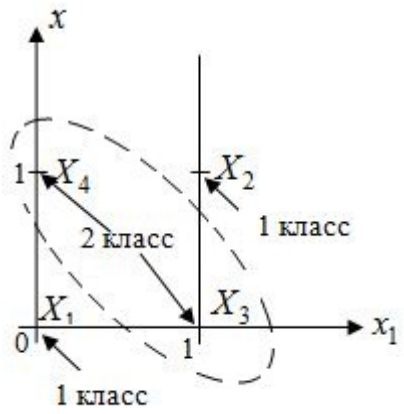


Рис. 3. Задача XOR для размерности n=2

Правда, Фр. Розенблатт в ответ на эту критику предложил усложнить структуру перцептрона, перейдя к 4-х слойному перцептрону с двумя настраиваемыми слоями связей. Такой перцептрон способен легко решить задачу XOR. Однако в нем возникает новая проблема – непонятно, как настраивать 2 слоя весов, чтобы обеспечить сходимость обучения к требуемой реакции. В связи с этой критикой математиков работы по НС были незаслуженно приостановлены на целых 10 лет, и был нанесен урон целому направлению в области ВИ.

Лишь спустя 10 лет, в результате работ Андерсона, Хинтона, Уильямса была предложена НС Back Propagation (обратного распространения) [Хайкин Саймон, 2006], представляющая по структуре многослойный перцептрон (MLP). Для этой сети был предложен алгоритм обратного распространения (ошибки), который позволяет настраивать сеть с произвольным числом слоев и обеспечивать сходимость к требуемой реакции (классификации или прогнозу).

Была доказана теорема об универсальной аппроксимации с помощью НС, в которой утверждается, что «существует НС ВР, позволяющая реализовать любое непрерывное отображение: $y = F(x)$ при некоторых достаточно обобщенных предположениях».

Эта теорема обосновывает широчайшие потенциальные возможности НС ВР, и стимулировала резкий рост исследований в области НС.

Кроме классических ИНС с последовательными связями между слоями нейронов, в последние годы появились нейронные сети с обратными связями (рекуррентные НС [Хайкин Саймон, 2006]). Например, НС Хопфилда и Хемминга. Такая сеть за счет использования положительной ОС обладает удивительным свойством: при определенных условиях поведение системы в процессе ее работы стремится к устойчивому состоянию (неподвижной точке), которое называется аттрактором. Аттракторы представляют собой память системы; а нейронная сеть Хопфилда – модель ассоциативной памяти человека. Когда на вход НС Хопфилда подается зашумленное или искажённое изображение объекта, она как бы по ассоциации «узнает»его и выдаёт на выходе неискаженное изображение – эталон, который является одним из аттракторов.

В настоящее время НС широко применяются в разнообразных сферах, в частности, прогнозировании, классификации и распознавании образов, аппроксимации функций, управлении технологическими процессами.

Достоинством НС является способность к обучению и настройке весов связей между нейронами с целью получения требуемой реакции. Вторым достоинством НС является параллельная обработка входной информации, что обеспечивает высокую производительность НС в целом, несмотря на медленную скорость передачи информации по нейронным связям и низкую скорость работы самих нейронов.

Важным свойством интеллекта является *способность к самоорганизации, самообучению*, когда «учитель» в системе отсутствует, и «правильная» классификация объекта неизвестна (если речь идет о распознавании образов). В таком случае система сама, рассматривая различные объекты (изображения) в пространстве признаков, разбивает их на подмножества, называемые кластерами, по степени «похожести», близости объектов в пространстве признаков. При этом могут использоваться самые различные метрики близости.

Нейронные сети с самоорганизацией (в частности, НС Кохонена), используя соревновательное обучение, проявляют способность к автоматической классификации объектов самой различной природы [Хайкин Саймон, 2006].

В задачах обучения и самообучения нейронных сетей настраиваются веса связей $w = w'$ с учителем или без него. При этом структура сети не изменяется.

Однако зачастую встречаются задачи, когда в процессе настройки необходимо не только настраивать веса, но и менять структуру сети. В таких случаях необходимо использовать модели и методы самоорганизации.

К числу методов самоорганизации относится, в частности, метод индуктивного моделирования, так называемый метод группового учета аргументов (МГУА). Метод используется для восстановления неизвестной функции прогноза или классификации $y = f(x_1, x_2, ..., x_n)$ по экспериментальным

данным. Метод был предложен в конце 60-х годов 20-го века академиком А.Г. Ивахненко [Ивахненко А.Г., 1976].

Он использует основные идеи эволюции скрещивания (кроссинг-овер) родительских пар и генерацию потомков, селекцию (отбор лучших потомков), проверку условий останова.

Для восстановления неизвестной функции n переменных в классе полиномов (т.н. полиномов Колмогорова-Габора) вида

$$y = f(x_1, x_2, ..., x_n) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{j \geq i}^{n} \sum_{i=1}^{n} a_{ij} x_i x_j + \sum_{k \geq j} \sum_{j \geq i} \sum_{i=1}^{n} a_{ijk} x_i x_j x_k + ... \quad (7)$$

используются элементарные функции от каждой пары переменных, называемые частичными описаниями (11) следующего вида:

$$y = \varphi(x_i, x_j) = \begin{cases} a_0 + a_i x_i + a_j x_j - \textit{линейные}; \\ a_0 + a_i x_i + a_j x_j + a_{ii} x_i^2 + a_{jj} x_j^2 + a_{ij} x_i x_j - \textit{квадратичные}. \end{cases} \quad (8)$$

Для каждого такого частичного описания (8) по обучающей выборке $L_{обуч}$ методом МНК, находятся оценки неизвестных коэффициентов $\{a_i\}$, $\{a_{ij}\}$, а по проверочной $L_{пров}$ находим наилучшую модель.

Для этого используются внешние критерии селекции:

а) регулярности

$$\overline{\varepsilon}_{пр}^2 = \frac{1}{N_{пр}} \sum_{i=1}^{N_{пр}} (y_i - \widehat{y_i}(x))^2 , \quad (9)$$

где y_i - реальный выход для i-ой точки, $\widehat{y_i}(x)$ - выход модели, $N_{пр}$ - объем проверочной выборки;

б) несмещенности

$$N_{см} = \frac{1}{N_{пр}} \sum_{i=1}^{N_{пр}} (y_i^* - y_i^{**})^2 \quad (10)$$

Далее происходит процедура селекции – отбор F лучших частичных описаний по $\min$ критерия $\overline{\varepsilon}_{пр_i}^2$.

Число F называется «свободой выбора». Отобранные лучшие частичные описания являются входами следующего ряда синтеза. На этом 1 итерация метода заканчивается.

Процесс синтеза продолжается до тех пор, пока не будет достигнут минимум критерия регулярности, т.е. $\min \overline{\varepsilon}_{пр}^2(k)$, где $\overline{\varepsilon}_{пр}^2(k) = \min_i \overline{\varepsilon}_{пр_i}^2(k)$, $\overline{\varepsilon}_{пр_i}^2(k)$ - значения критерия $y\ i$-ой модели на k-ой итерации.

Таким образом, находится оптимальная модель минимальной сложности. Найдя искомую (оптимальную) модель, далее двигаемся в обратном направлении, по её связям с предыдущим рядом, и делая замену переменных, приходим в конце к искомой модели в исходных переменных $x_1, x_2, ..., x_n$. Метод в

отличии от других методов идентификации моделей сложных систем, позволяет автоматически восстановить структуру искомой модели.

Это обеспечивается за счет использования *принципа самоорганизации* [1, 3]: с ростом сложности модели S, значения критерия регулярности $\overline{\varepsilon}^2_{пр}$ сначала падает, достигает минимума, а затем остается постоянным (при отсутствии шумов) или начинает расти. Здесь S – число членов полинома Колмогорова-Габора.

Одним из существенных свойств интеллекта человека является его *способность работать с неполной нечеткой информацией*, в условиях неопределенности. Возникла потребность в создании аппарата формализованного описания нечеткой и качественной информации для принятия решений в условиях неопределенности. Такая проблема была решена Л. Заде, который ввел понятие нечеткого множества (НМ) (Fuzzy set) и разработал аппарат операций над НМ [Zadeh L.A., 1965]. Основным атрибутом НМ является его *функция принадлежности* $\mu_a(x)$, такая что $\mu_a(x) \in [0;1]$.

Величина $\mu_a(x_i)$ - есть степень принадлежности элемента x_i к НМ A, лежащая между 0 и 1. Здесь принципиальное отличие от обычных множеств, для которых степень принадлежности элементов равна либо 1 (элемент x принадлежит), либо 0 (элемент x не принадлежит множеству A).

Следует подчеркнуть, что ФП не является плотностью распределения случайной величины A, а степень принадлежности $\mu_a(x_i)$ - это не вероятность появления значения x_i, а шансы появления такого значения по оценке ЛПР, и отображает его субъективную оценку.

Дальнейшим развитием теории НМ явилось введение Л.Заде понятия лингвистической переменной её использование для принятия решений в условиях качественной и нечеткой информации.

По определению [Zadeh L.A., 1975], лингвистическая переменная (ЛП) – это переменная (количественная или качественная), задаваемая пятеркой кортежей:

$$\text{ЛП} = < N, X, T, \mu, G >, \tag{11}$$

где N - имя переменной;

$T = \{T_j\}$ - множество базовых термов или значений ЛП;

$X = [x_{\min}, x_{\max}]$ - универсальная шкала;

μ - оператор, ставящий в соответствие каждому значению T_j его ФП $\mu_{T_j}(x)$;

G - генератор новых термов из базовых с использованием предикатов: «НЕ», «ОЧЕНЬ», «БОЛЕЕ-МЕНЕЕ».

При этом ФП новых термов определяются по соотношениям:

$$\mu_{неT_j}(x) = 1 - \mu_{T_j}(x);\ \mu_{оченьT_j}(x) = (\mu_{T_j}(x))^2;\ \mu_{более-менееT_j}(x) = \sqrt{\mu_{T_j}(x)}. \tag{12}$$

Приведем примеры:

1) Пусть количественная переменная «Доход фирмы» рассматривается как ЛП, возможный доход фирмы изменяется в диапазоне $[1;10^6]$ грн. Введем базовые термы: T_1 - низкий, T_2 - низкий - средний, T_3 - средний, T_4 - средний - высокий, T_5 - высокий. Тогда ЛП Доход фирмы задается так:

$$\text{ЛП Доход фирмы} = < N, X, \{T_j\}, \{\mu_{T_j}(x)\}, G >$$

где N - Доход фирмы;

$X = [1;10^6]$; $T = \{T_j\}$ - набор базовых термов; а их ФП $\mu_{T_j}(x)$ приводятся на рис.4

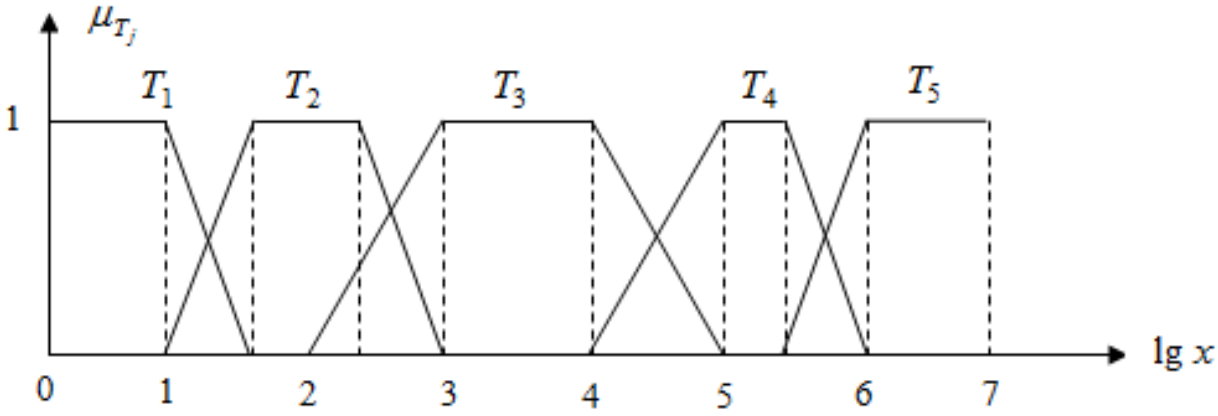


Рис. 4. ФП $\mu_{T_j}(x)$

2) Рассмотрим теперь качественную переменную. Например, внешность. Такую переменную задаем на универсальной шкале $X = [0;1]$. Введем набор базовых термов для ЛП «внешность»: T_1 - безобразная, T_2 - несимпатичная, T_3 - приятной наружности (симпатичная), T_4 - красивая, T_5 - идеал красоты (например, Сикстинская Мадонна Рафаэля). ФП базовых термов представим на универсальной шкале (см. рис. 5).

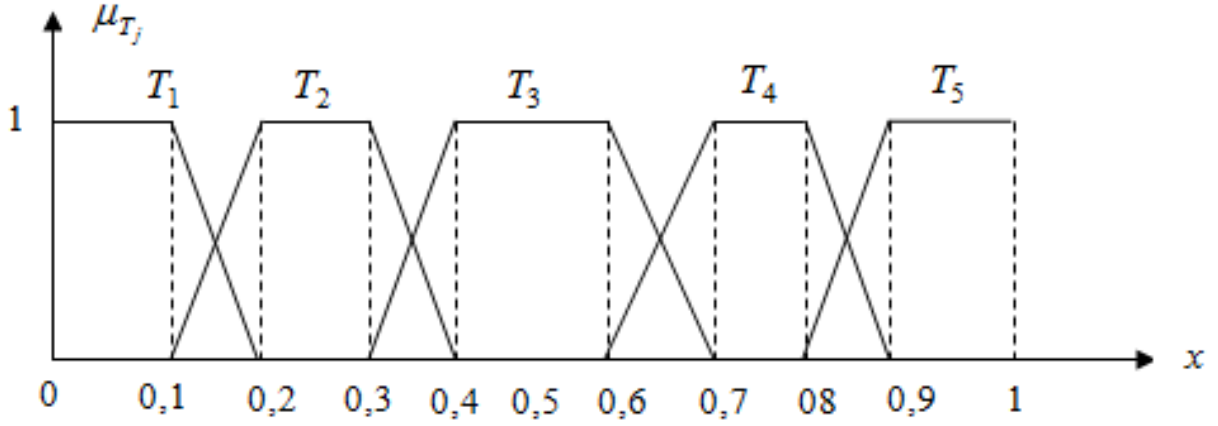


Рис. 5. ФП $\mu_{T_j}(x)$

Эти ФП отражают информацию эксперта о качественной переменной «красота», его субъективную оценку соответствия этой переменной степени наличия данного признака у конкретного лица.

Лингвистические переменные используются для принятия решений в условиях качественной и недостоверной информации в системах нечеткой логики (СНЛ). Главным элементом таких систем является нечеткая база правил (НБП), отображающая нечеткие знания эксперта о соответствующей предметной области. Нечеткие правила вывода R_k имеют вид:

R_1: если x_1 есть A_{11}, x_2 есть A_{21}, ..., x_n есть A_{n1}, то z есть C_1;

.....

R_k: если x_1 есть A_{1k}, x_2 есть A_{2k}, ..., x_n есть A_{nk}, то z есть C_k;

где x_1, x_2 ,..., x_n - входные переменные; z - выходная переменная; $\{A_{ij}\}$ и $\{C_k\}$ - значения лингвистических переменных (термов).

Рассмотрим в качестве примера нечеткие правила в задаче анализа риска банкротства корпораций в условиях неопределенности [Згуровский М.З., 2013].

Пусть для оценки финансового состояния корпораций используются следующие финансовые показатели:

x_1 - коэффициент автономности;

x_2 - коэффициент обеспеченности активов собственными средствами;

x_3 - коэффициент промежуточной ликвидности;

x_4 - коэффициент абсолютной ликвидности;

x_5 - коэффициент оборачиваемости активов;

x_6 - рентабельность всего капитала.

Все эти показатели рассматриваются как лингвистические переменные со значениями: очень низкий (ОН), низкий (Н), средний (Ср), высокий (В) и очень высокий (ОВ).

Пусть выходная переменная «риск банкротства» принимает аналогичные значения: очень высокий (ОВ), высокий (В), средний (Ср), низкий (Н) и очень низкий (ОН).

Тогда можно записать следующие правила:

R_1: если x_1 - «ОН», x_2 - «ОН», x_3 - «ОН», x_4 - «ОН», x_5 - «ОН», x_6 - «ОН», то риск банкротства «В»;

R_k: если x_1 - «Ср», x_2 - «Ср», x_3 - «Ср», x_4 - «Н», x_5 - «ОН», x_6 - «ОН», то риск банкротства «В»;

R_M: если x_1 - «ОВ», x_2 - «ОВ», x_3 - «ОВ», x_4 - «ОВ», x_5 - «ОВ», x_6 - «ОВ», то риск банкротства «ОН».

Логический вывод осуществляется за следующие этапы [Згуровский М.З., 2013]:

1) *Фаззификация* (введение нечеткости).

2) *Логический вывод*, состоящий из двух подэтапов – определение степени выполнения условий правил и нахождение выхода каждого правила.
3) *Композиция* выходов правил.
4) *Дефаззификация*. Приведение к четкости (т.е. нахождение четкого выхода).

Имеется несколько алгоритмов нечеткого логического вывода, основными из которых являются: алгоритм Мамдани; Цукамото; Сугено; Ларсена.

Отметим основные достоинства систем с нечеткой логикой СНЛ [Згуровский М.З., 2013].

1) Они позволяют учитывать качественную и нечеткую информацию, работать в условиях неполноты и неопределенности исходной информации;
2) Они позволяют учитывать нечеткие знания экспертов в виде базы нечетких правил.

Вместе с тем СНЛ имеют следующие недостатки.

Для работы СНЛ необходимо задать ФП лингвистических переменных в нечетких правилах.

1) Эту работу выполняет человек-эксперт. Для упрощения проблемы предполагается, что все ЛП имеют один класс ФП (например, гауссовские, треугольные и трапецеидальные, и ФП разных термов отличаются лишь параметрами, которые должен указать эксперт). Однако эксперт может не знать их;
2) База нечетких правил, формулируемая экспертом, может оказаться неполной или противоречивой.

Для устранения указанных недостатков необходимо сделать СНЛ аддитивными и обучать по входной выборке.

Чтобы использовать арсенал накопленных методов и алгоритмов обучения, СНЛ реализуют структурно в виде нечетких нейронных сетей, в которых имеется слой входных нейронов X_1, X_2,..., X_n, слой нейронов правил R_1, R_2,..., R_k и слой выходных нейронов C_1, C_2,..., C_m.

При этом связи между слоями w_{ij} являются нечеткими, задаваемыми в виде некоторых ФП с неопределенными параметрами, настраиваемыми в процессе обучения. Для этого используются алгоритмы обучения, разработанные для обычных НС.

Таким образом, интеграция двух технологий: нейронных сетей и систем с нечеткой логикой позволила позволила создать новую *гибридную технологию ВИ – нечеткие нейронные сети*, позволяющие использовать достоинства обоих исходных технологий. Подробно системы нечеткой логики и ННС рассмотрены в главах 3, 4, 5 монографии.

Системы нечеткой логики и ННС находят широкие применения во многих задачах ВИ. Их широкое распространение базируется на теореме FAT – Fuzzy Approximation Theorem об универсальной аппроксимации систем нечеткой логики.

Ванг доказал, что СНЛ с Гауссовскими ФП при увеличении числа правил $n \to \infty$ является универсальным аппроксиматором [Wang F., 1992]. Несколько позднее Коско доказал аналогичную теорему, если ФП нечетких множеств (термов) являются треугольными [Kosko B., 1994].

Заметим, что данная теорема является полным аналогом соответствующей теоремы об универсальной аппроксимации для обычных нейронных сетей. Таким образом, между свойствами обычных НС и нечетких НС имеются глубокие взаимосвязи.

Одним из свойств живых организмов биосистем в том числе человеческого мозга является способность к саморазвитию, приспособлению к применяющимся условиям внешней среды, самосовершенствованию в процессе эволюции. Как известно, биологическая эволюция базируется на следующих механизмах:

а) скрещивание родительских пар и генерация потомков, берущих гены у обоих родителей;

б) мутации, воздействие случайных факторов;

в) селекция – отбор лучших особей, переходящих в следующее поколение.

Этот процесс продолжается многократно, в результате идет постепенное целенаправленное улучшение особей.

Именно эти механизмы были использованы в эволюционном моделировании (ЭМ) – важной технологии ВИ. Основная идея ЭМ – это заменить процесс построения сложной системы (распознавание образов, технической диагностики) процессом его эволюции [Fogel L. J., 1966]. Этот процесс проводится естественно в ускоренном масштабе времени, требуются тысячи или десятки тысяч поколений, чтобы достичь желаемого результата, желаемого качества функционирования.

ЭМ широко используется в задачах структурного синтеза сложных систем РО и классификации, медицинской диагностики, прогнозирования, систем управления сложными объектами и процессами.

Следует отметить, что крайне важным в ЭМ является использование механизма мутаций, который обеспечивает необходимое разнообразие особей и исключает возможность вырождения популяций.

Вышеперечисленные технологии ВИ базируются и широко используют соответствующие методы и алгоритмы. В частности, НС и ННС используют обучения и адаптации, такие как градиентный, сопряженных градиентов, стохастическая аппроксимация, РМНК, современные алгоритмы обучения Качмаржа, Марквардта, Гульвина-Рамаджи-Кейнса и др.

Кроме того, они также используют для своей работы методы и алгоритмы самообучения и самоорганизации: Кохонена, методы кластерного анализа, к-средних (четкий), дисперсионный алгоритм, метод Уорда, «ближайшего соседа», нечеткие методы кластерного анализа: К-средних, Густавссона Кесселя, на основе нечеткого отношения квазитолерантности, адаптивные нечеткие методы кластерного анализа [Згуровский М.З., 2013].

Наконец, технология ЭМ базируется на использовании генетических алгоритмов (ГА), использующих те же механизмы эволюции. ГА отличаются способами выбора родительских пар, кроссовера, механизмов мутаций способом селекций [Gen M, 1996].

Технологии и методы ВИ широко используются для решения прикладных задач ВИ. Так, нейронные сети и нечеткие НС, метод МГУА используются для прогнозирования нестационарных временных процессов, в частности в экономике и финансовой сфере.

НС и МГУА широко используются в задачах классификации и распознавания образов, в задачах диагностики, в том числе технической и медицинской [Згуровский М.З., 2013].

Нейронные сети с самоорганизацией, методы кластерного анализа (четкие и нечеткие) применяются в задачах автоматической классификации объектов по их признакам сходства – различия.

Эволюционное моделирование, генетические алгоритмы используются для структурного синтеза сложных систем распознавания образов, классификации, а также оптимизации структуры информационно-компьютерных сетей. Кроме того, ГА находят широкое применение в задачах комбинаторной оптимизации на графах и сетях.

Системы с нечеткой логикой и ННС эффективно используются в задачах анализа финансового состояния корпорации, прогнозирования риска банкротства корпораций и банков, оценке кредитоспособности заемщиков банковских капиталов в условиях неопределенности [Згуровский М.З., 2013].

Таким образом, подводя итого следует сказать, что современные технологии и методы ВИ тесно взаимодействуют друг с другом, имеется глубокое взаимопроникновение методов и алгоритмов вычислительного интеллекта в соответствующие технологии, и наоборот. А в целом технологии и методы ВИ являются отображением и технической реализацией свойств и способностей интеллекта человека в различных областях практической деятельности человека.

Заключение

В работе дано определение термина вычислительный интеллект и определены его основные компоненты. указаны взаимосвязи между вычислительным интеллектом и мягкими вычислениями.

Дан обзор современных направлений вычислительного интеллекта, проанализированы их основные свойства, взаимосвязи и возможности применения.

Отмечены взаимосвязи между направлениями и методами вычислительного интеллекта. Дальнейшее развитие вычислительного интеллекта по-видимому будет идти в нескольких направлениях.

Во-первых – это расширение сфер применения задачах ВИ, новые приложения в различных задачах и предметных областях, например в экономике, финансовой сфере, телекоммуникационных системах, управлении технологическими процессами и т.д.

Во-вторых – это развитие и совершенствование самих методов ВИ, в частности, генетических (ГА) и эволюционных алгоритмов (ЭА), роевых алгоритмов оптимизации, имунных алгоритмов, муравьиных алгоритмов. Одним из перспективных направлений здесь является адаптация и самообучение параметров ГА и ЭА с целью ускорения сходимости и повышения точности в задачах оптимизации. Актуальным является также развитие и совершенствование параллельных генетических алгоритмов.

В – третьих, это дальнейшая интеграция различных технологий ВИ, например, интеграция нечеткой логики и генетических и эволюционных алгоритмов в задачах принятия решений и распознавания образов в условиях неполноты информации, роевых алгоритмов и алгоритмов их обучения и самообучения.

Acknowledgement

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Fogel D., 1995] D. Fogel, (1995) Review of “Computational intelligence: imitating life,” IEEE Trans.Neural Networks, 6, 1995.-pp. 1562–1565

[Fogel L.J., 1966] L.J. Fogel, A. Owens, and M. Walsh. Artificial Intelligence through Simulated Evolution. John Wiley & Sons, 1966.

[Bezdek J., 1994] J.C. Bezdek, “What is computational intelligence?” in Computational Intelligence Imitating Life,. Zurada, J. M., R. J. M. II, and C. J. Robinson, Eds., IEEE Press, New York, 1–12, 1994.

[Gen M, 1996] Gen M. and Cheng R.: Genetic Algorithms and Engineering Design, John Wiley $ Sons, Inc., 1996, 410 p.

[Holland J.H., 1990] J.H. Holland. ECHO: Explorations of Evolution in a Miniature World. In J.D. Farmer and J. Doyne, editors, Proceedings of the Second Conference on Artificial Life, 1990.

[Kosko B., 1994] Kosko B.: Fuzzy Systems as Universal Approximators, // IEEE Transaction on Computers, 1994, N.11, pp. 1329-1333

[Marks R., 1993] Marks R., (1993) “Intelligence: computational versus artificial,” IEEE Trans. Neural Networks, 1993. 4, 737–739.

[Wang F., 1992] Wang F.: “Neural Networks Genetic Algorithms, and Fuzzy Logic for Forecasting,” Proceedings, International Conference on Advanced Trading Technologies, New York, July 1992, pp. 504-532.

[Zadeh L.A., 1965] Zadeh L.A.: Fuzzy sets // Information and Control, 1965, V.8, № 3, p. 338-353.

[Zadeh L.A., 1975] Zadeh L.A.: The Concept of a Linguistic variable and its application to approximate reasoning, // Part 1 and 2, Information Sciences, 1975, v.8-p. 199-249, 301-357.

[Zadeh L.A., 1984] Zadeh L.A. Theory of commonsense knowledge: aspects of vagueness / Dordrecht: D.Reidel, 1984, p. 257-296.

[Згуровский М.З., 2011] Згуровский М.З. Зайченко Ю.П. Модели и методы принятия решений в нечетких условиях.-К.: Изд «Наукова думка, 2011.-275 с.

[Згуровский М.З., 2013] Згуровский М.З. Зайченко Ю.П. Основы вычислительного интеллекта.-К.: Изд. «Наукова думка», 2013.- 406 с.

[Ивахненко А.Г., 1976] Ивахненко А.Г., Зайченко Ю.П. Димитров В.Д.: Принятие решений на основе самоорганизации. – Москва, Сов. Радио. – 1976. – 363стр.

[Минский М., 1971] Минский М., Пайперт С. Перцептроны. Пер. с англ. Под ред. В.А. Ковалевского. Киев. «Наукова Думка», 1971.

[Рассел Стюарт, 2007] Рассел Стюарт и Норвиг Питер. Искусственный интеллект: современный подход, Второе издание.: Пер. с англ. М. : Изд. Дом «Вильямс» , 2007- 1408 с.

[Розенблатт Ф., 1965] Розенблатт Ф. Принципы нейродинамики. Перцептроны и теория механизмов мозга.- М.: Мир, 1965.

[Ф. Люгер, 2006] Люгер Ф. Искусственный интеллект. Пер. с англ. М.: Изд. Дом « Вильямс», 2006.

[Хайкин Саймон, 2006] Хайкин Саймон. Нейронные сети: полный курс, 2-е изд. ,испр. Пер. сангл. М.: Изд. Дом Вильямс, 2006.- 1104 с.

[Хокинс Джефф, 2007] Хокинс Джефф, Блейксли Сандра. Об интеллекте. Пер. с англ. –М.: ООО «ИД. Вильямс», 2007, 240с.

Authors' Information

Згуровский Михаил *– ректор НТУУ «Киевский политехнический институт», директор Института прикладного системного анализа», академик НАН Украины, д.т.н., профессор. 03056, Киев-56, проспект Победы, 37, Украина. phone: 38044 -2366913 e-mail: mzgurovsky@gmail.com*

Области научных исследований*: системный анализ, модели и методы устойчивого развития, принятие решений*

Зайченко Юрий *- д.т.н., профессор ННК «Институт прикладного системного анализа», 03056, Киев-56, проспект Победы, 37, Украина phone: 38044 -4068393; e-mail: baskervil@voliacable.com,*

Области научных исследований*: теория принятия решений в условиях неопределенности, модели и методы вычислительного интеллекта в задачах прогнозирования и анализа в экономике и финансовой сфере, моделирование и оптимизация компьютерных сетей*

System Analysis of Computational Intelligence Main Trends

Michael Zgurovsky, Yuriy Zaychenko

Abstract*: The survey and analysis of new field in AI -computational intelligence (CI) is performed. The main components of CI- technologies, methods and applications are outlined, their interconnections are determined and analyzed. The interrelations between CI and soft-computing are established and discussed.*

Keywords*: computational intelligence, soft computing, neural networks, fuzzy logic, evolutionary modeling.*

РЕШЕНИЕ ПРОБЛЕМЫ ФОРМАЛЬНОЙ ОЦЕНКИ ЭФФЕКТИВНОСТИ ТЕХНОЛОГИЙ ИДЕНТИФИКАЦИИ ЗНАНИЙ В СЛАБОСТРУКТУРИРОВАННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

Нина Хайрова, Наталья Шаронова, Дмитрий Узлов

***Аннотация**: В работе показана возможность использования интегральных количественных показателей полноты, точности и меры Ван-Ризбергена для оценки эффективности информационно-лингвистических технологий идентификации знаний в текстах. Обосновывается возможность использования метода тестовых коллекций для экспериментального подтверждения достоверности получаемых коэффициентов эффективности. В работе исследуется проблема максимизации надежности использования результатов, полученных по репрезентативной выборке, для выводов о генеральной совокупности текстовой коллекции. Рассмотрены процедуры использования выборочной доли признака как статистической характеристики для оценивания доли релевантных документов в генеральной совокупности. Предложен метод определения доверительного интервала для доли признака, основанный на подходе Вилсона, и метод определения необходимого объема релевантной выборки. Приведены примеры реализации предложенного подхода средствами Microsoft Excel.*

***Ключевые слова**: полнота, точность, релевантность, эффективность идентификации знаний, доверительный интервал, объем тестовой коллекции.*

***ACM Classification Keywords**: H.3.3 .Information Search and Retrieval, I.2.4. Knowledge Representation Formalisms and Methods, G.3. Probability and statistics – Statistical computing*

Введение

Для оценки эффективности информационно-лингвистической технологии экстракции и идентификации знаний из слабоструктурированной текстовой информации необходимо определить метрики - совокупность объективно измеряемых показателей, характеризующих деятельность пользователей до и после внедрения данной технологии. К таким метрикам обычно относят время поиска пользователем информации по тому или иному вопросу и уровень знаний, извлеченных пользователями данной системы.

При этом, в отличие от временных показателей, характеризующих длительность выполнения тех или иных процессов и достаточно просто поддающихся объективному измерению, метрика уровня знаний поддается измерению достаточно сложно. В то же время ясно, что основную ценность для социально-экономических организационных систем представляют, новые, скрытые, неявные и неформализованные знания, извлеченные, в том числе, из текстовых информационных потоков. Именно они позволяют принимать новые нетрадиционные решения.

На сегодняшний день не существует стандартных бенчмарков для измерения качества и эффективности технологии идентификации знаний, извлечённых из текстовых массивов. Обычно для определения эффективности технологий лингвистического процессора, используемых различными системами

семантической классификации и информационного поиска (Text Mining, Opinion Mining, Web Mining), используют метод тестовых коллекций [Шабанов, 2003; Cormack, 1998]. Суть данного метода заключается в сравнении результатов работы исследуемой системы на заранее определенных данных с оценками экспертов на тех же данных. В результате сравнения получается одно-двух-критериальная оценка эффективности. Но, поскольку для задач, связанных с экстракцией и идентификацией знаний, понятия „эффективного излечения знаний", „качества знаний" не имеют общепринятого определения, количественная оценка результатов работы системы не тривиальна. Традиционный подход в подобных случаях — сравнение с „эталонным" результатом — плохо применим из-за необходимости создания эталонного ответа для каждого конкретного набора электронных документов. Поэтому для оценки работы системы используется алгоритм, для которого выводы, сделанные системой, согласуются с мнением экспертов. Две основные, возникающие при этом проблемы заключаются в субъективности эксперта и в необходимом размере текстовой коллекции, позволяющем получить достоверный результат.

Под достоверностью понимается доказанная правильность того, что полученные в результате проведения эксперимента значения выполняются в определенных условиях для определенного класса объектов. Достоверность должна быть подтверждена верификацией, то есть повторением результатов в одних и тех же условиях при большом количестве проверок на разных объектах.

Общая постановка задачи

При использовании метода тестовых коллекций для оценки эффективности технологий идентификации знаний в слабоструктурированной текстовой информации возникают проблемы выделения интегральных количественных показателей оценки, учета реальных условий и количества исследований, а также проблема обоснования вывода о свойствах всех текстов коллекции (всей совокупности) по результатам выборочного метода экспериментального исследования. Иными словами при проведении проверки результатов информационно-лингвистических исследований на контрольных примерах возникает вопрос учета реальных условий и количества исследований.

Таким образом, прежде всего, необходимо обоснование использования в качестве усредненных метрик эффективности идентификации знаний в слабоструктурированной текстовой информации принятых показателей количественной оценки качества обработки текстовой информации, а именно – полноты, точности, шума и аккуратности, а также меры Ван Ризбергена.

Кроме того, в связи с тем, что при проведении экспериментальной проверки достоверности информационно-лингвистических технологий нецелесообразно или, скорее, практически невозможно в силу объективных причин, исследовать все тексты совокупности, то необходимо исследовать некоторую выборочную репрезентативную их часть. При этом возникает новый ряд проблем, связанный с необходимостью максимальной надежности использования результатов, полученных по выборке, для выводов о генеральной совокупности. Предлагается рассмотреть применение подхода, основанного на методах математической статистики (в частности, математической теории выборки), для определения объема экспериментальной выборки, необходимой для подтверждения достоверности разработанных информационно-лингвистических технологий, а также для нахождения погрешности оценивания выбранных показателей качества идентификации знаний в слабоструктурированной текстовой информации.

Интегральные показатели эффективности работы системы идентификации знаний

Для получения интегральных показателей эффективности работы системы идентификации знаний в слабоструктурированных текстовых информационных потоках применяем методики усредненных метрик.

Будем использовать показатели количественной оценки эффективности поиска и классификации, утвержденные межгосударственным стандартом по информации, библиотечному и издательскому делу [ISO 12620:2009]. Такими показателями являются: коэффициент точности — precision, коэффициент полноты — recall и коэффициент аккуратности accuracy, базирующиеся на субъективно определяемом понятии релевантности. Понятие релевантности является сложно определяемым и имеет, скорее, психологическую природу. Мы используем определение релевантности [Mizzaro, 1997], в котором релевантность зависит от четырех понятий Relevance (IR, IN, C, T), где IR — информационный ресурс, IN — информационные потребности, C — контекст и T — время. Информационный ресурс представлен множеством текстов коллекции, поступающим на обработку IR = D. Наибольшая субъективность при этом заключается в понятии информационной потребности, которую можно разделить на неосознанную (истинную потребность) эксперта в знаниях, оперируя которыми эксперт решает некоторую информационную проблему, стоящую перед ним, и осознанную (внутреннее понимание реальной потребности). Переход между двумя составляющими потребности вносит дополнительную погрешность в вычисление эффективности работы систем, основанных на знаниях, но только осознанная потребность эксперта в знании определяет полноту и точность работы системы. Дело в том, что именно осознанная потребность эксперта в знаниях, необходимых для решения некоторых задач, формируется в сфере мышления, и, сформировавшись в реальном контексте предметной области C и времени T, информационная потребность IN затем уже описывается средствами естественного языка.

При определении эффективности работы системы релевантность, т.е. соответствие связного текста крупной смысловой парадигме, определяется экспертом по шкале “Relevance/irrelevant/undefined” и показывает соответствие или несоответствие электронного текста некой локальной области знаний (или крупной смысловой парадигме). Для определения коэффициентов полноты, точности и аккуратности необходимо для каждой области знаний эксперта определить: n_{yy} — число идентифицированных элементов (связных текстов или фрагментов связных текстов), релевантных области знаний эксперта, с его точки зрения, n_{yn} — число идентифицированных элементов, не релевантных области знаний, с точки зрения эксперта, n_{ny} — количество релевантных элементов, не идентифицированных системой, и n_{nn} — количество нерелевантных элементов, не идентифицированных системой (рис.1).

При этом, если нет элементов, получивших с точки зрения эксперта определение undefined, сумма значений метрик равна количеству элементов, поступивших на обработку: $D=n_{nn}+n_{ny}+n_{yn}+n_{yy}$, где D — множество элементов, поступивших в систему на обработку

Коэффициент точности определяется как:

$$precision = \frac{n_{yy}}{n_{yy} + n_{yn}}, \quad (1)$$

коэффициент полноты определяем как:

$$recall = \frac{n_{yy}}{n_{yy} + n_{ny}}. \quad (2)$$

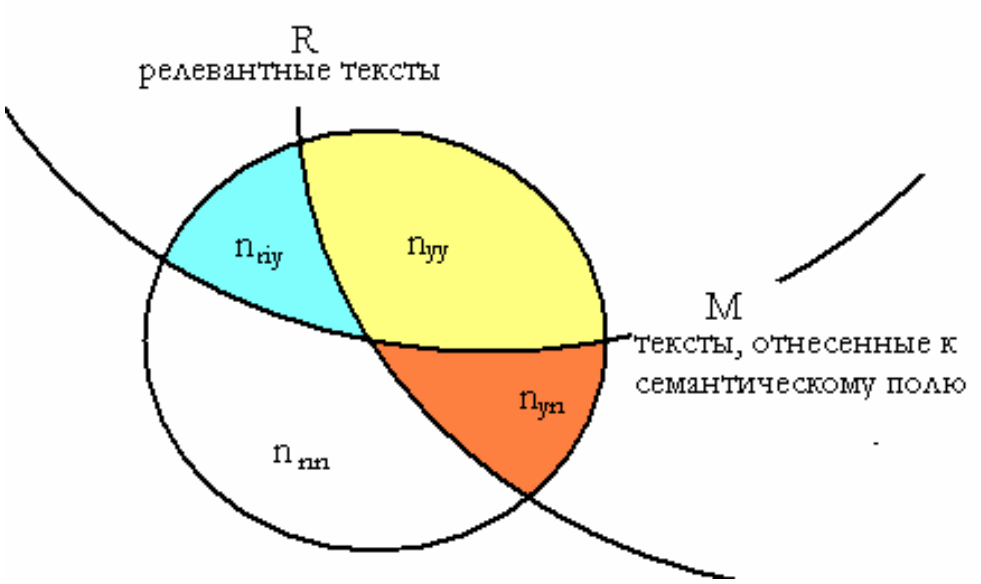


Рис. 1. Метрики оценки эффективности работы системы идентификации знаний

Для одновременного учета полноты и точности в одной усреднённой величине, с учетом различных весов α, можно использовать меру Ван Ризбергена или F-measure:

$$F_\beta - measure = \frac{1}{\alpha \frac{1}{precision} + (1-\alpha)\frac{1}{recall}} = \frac{(\beta^2+1)\times precision \times recall}{\beta^2 \times precision + recall} \quad (3)$$

где $\alpha \in [0,1]$, $\beta^2 = \frac{(1-\alpha)}{\alpha}$, $\beta \in [0,\infty]$. При значении коэффициентов α = 1/2 или β = 1 в F-мере полнота и точность имеют одинаковый вес и получаемая мера называется сбалансированной F_1-мерой:

$$F_1 - measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

Если $0<\beta<1$ — большее значение при расчете уделяется точности, а при $\beta>1$— больший вес приобретает полнота.

Выборочный метод исследования изучаемой совокупности

В связи с тем, что определяемые показатели эффективности работы системы идентификации знаний используют субъективно определяемое понятие релевантности той или иной смысловой парадигме или области знаний эксперта, достоверность значений показателей полноты, точности и меры Ван Ризбергена той или иной модели требует экспериментального подтверждения на совокупности текстов. Так как совокупность текстов, на которых реализуются модели, практически бесконечна, имеет смысл исследовать лишь часть объектов из изучаемой совокупности, т.е. осуществить так называемый выборочный метод исследования совокупности текстов и сделать обоснованные выводы о свойствах всей совокупности.

Одной из основных проблем выборочного исследования является получение репрезентативной выборки. Под репрезентативной выборкой понимается такая выборка, которая позволяет получить наиболее точную информацию о генеральной совокупности, а также не противоречит требованиям применимости вероятностных методов к обработке выборочных данных. Такая выборка образуется методом случайного отбора, то есть методом, при котором каждый элемент совокупности попадает в выборку случайным образом. Так как строгая реализация требований случайности отбора объектов в выборку требует разработки и применения специальных процедур и не всегда осуществима, на практике предполагается, что рассматриваемая выборка получена случайным отбором.

Так как в нашем случае отношение объема выборки к объему генеральной совокупности значительно меньше 5 – 10% (генеральной совокупностью любой модели обработки естественно-языковых текстов является стремящийся к бесконечному объему информационный текстовый поток), можно использовать математический аппарат теории возвратной выборки [Четыркин, 1982]. Кроме того, полученные результаты для возвратной выборки в ряде случаев можно перенести на соответствующий безвозвратный случай, вводя необходимый поправочный коэффициент.

Вторая проблема выборочного исследования заключается в проблеме оценки, связанной с тем, что выводы, делаемые на основе данных выборки, адекватно характеризуют только свойства выборки, а их перенос на свойства генеральной совокупности будет приводить к некоторой погрешности. Проблема оценки заключается в необходимости использования с максимальной возможной надежностью результатов, полученных по выборке для выводов о генеральной совокупности.

Для задачи оценки эффективности технологии идентификации знаний в слабоструктурированной текстовой информации будем оценивать долю признака (долю релевантных документов) в генеральной совокупности по соответствующей выборочной характеристике. При этом допускается, что любой рассматриваемый метод идентификации, как правило, не безошибочен: он может отнести к числу релевантных реально не релевантные документы, а также счесть не релевантными в действительности релевантные объекты.

Пусть доля признака в генеральной совокупности, которая показывает отношение числа релевантных документов к общему числу документов в генеральной совокупности, равна R. Выборочная оценка R_S доли R равна $R_S = M/N$, где, N – объем исследуемой возвратной выборки, а M – количество выявленных в ней рассматриваемым методом идентификации релевантных документов. Можно показать, что эта оценка удовлетворяет всем требованиям, предъявляемым к статистическим оценкам (состоятельность, несмещенность, достаточность и эффективность) [Четыркин, 1982]. Так как объекты случайно отбираются в выборку, то выборочная доля R_S может принимать любые значения в интервале [0;1], причем $R_S=0$, когда ни один релевантный документ не попал в выборку и $R_S=1$, если все документы в выборке релевантны.

Так как выборочная оценка R_S является точечной оценкой доли признака, то для нахождения погрешности приближения R оценкой R_S необходимо обратиться к интервальной оценке последней, то есть установить ошибку выборки. Поскольку R_S, а, следовательно, и ошибка выборки являются случайными величинами с одним и тем же распределением вероятностей, введем порождаемую конкретной точечной оценкой R_S интервальную оценку, в пределах которой с некоторой доверительной вероятностью P будет лежать доля признака генеральной совокупности.

Таким образом, доверительная вероятность P будет показывать вероятность того, что интервальная оценка содержит в себе неизвестную долю признака генеральной совокупности. Дополняющей вероятность P до 1 вероятностью α будем измерять риск выхода доли признака генеральной совокупности R за пределы интервальной оценки.

Поскольку в общем случае распределение величины R несимметрично, то интервальная оценка, или доверительный интервал, случайной величины R имеет вид:

$$P(R_S - E_1 < R < R_S + E_2) = 1 - \alpha, \quad (5)$$

где $[R_S - E_1; R_S + E_2]$ – доверительный интервал, $R_S - E_1$; $R_S + E_2$ – доверительные границы, $P = 1 - \alpha$ – доверительная вероятность, α – уровень значимости, или существенности.

В случае симметричного распределения R доверительный интервал также симметричен относительно величины R и имеет вид: $P(|R - R_S| < E) = 1 - \alpha$, где величина E – предельная ошибка, характеризующая точность выборки.

Обычно при таком подходе возникает несколько типов задач:

1) Определение доверительной вероятности по заданному доверительному интервалу и объему выборки;
2) Определение доверительного интервала по заданной доверительной вероятности и объему выборки;
3) Определение необходимого объема выборки по заданной доверительной вероятности и предельной ошибке.

В нашей задаче оценки эффективности технологий идентификации знаний в слабоструктурированной текстовой информации более актуальными являются определение доверительного интервала и определение необходимого объема выборки.

Определение доверительного интервала по заданной доверительной вероятности и объему выборки

Доверительный интервал для доли признака надо определять, строго говоря, базируясь на биноминальном законе распределения [Clopper, 1934]. Начиная с выборок объемом не менее 20, биномиальное распределение симметризуется и хорошо аппроксимируется нормальным распределением с параметрами: среднее $\langle R_S \rangle = R$, дисперсия $D(R_S) = R(1 - R)/N$, стандартное отклонение $\sigma(R_S) = [D(R_S)]^{1/2}$.При этом доверительный интервал может быть рассчитан по формуле $P(|R - R_S| < E_\alpha) = 2\Phi(Z_\alpha) = 1 - \alpha$, где $\Phi(Z_\alpha)$ – функция Лапласа. Предельная ошибка выборки находится при этом из равенства $E_\alpha = Z_\alpha\sigma(R_S)$.

В качестве величины доверительной вероятности обычно выбирают значение 0,95 (тогда уровень значимости α = 0,05). При этом $Z_{0,05} = 1,96$. Величина Z_α просто связана со статистической функцией Excel НОРМСТОБР(вероятность): Z_α = НОРМСТОБР(1 – α). С помощью этой функции может быть найдена величина Z_α при любой доверительной вероятности.

Теперь отталкиваясь от соотношения

$$|R - R_S| < Z_\alpha[R(1 - R)/N]^{1/2} \quad (6)$$

можно получить выражения для левой и правой границ доверительного интервала R, решая соответствующее квадратное уравнение относительно R [Wilson, 1927]. Использование подобного соотношения для адекватной оценки доверительных интервалов доли признака на малых выборках доказано статистиками [Brown, 2001; Garcia-Perez, 2005]. Тогда для левой границы доверительного интервала R_L имеем:

$$R_L = \frac{R_S + \frac{Z^2}{2N} - Z\left[\frac{R_S(1-R_S)}{N} + \frac{Z^2}{4N^2}\right]^{1/2}}{1+\frac{Z^2}{N}} \quad (7)$$

Для правой границы R_R получаем соответственно:

$$R_R = \frac{R_S + \frac{Z^2}{2N} + Z\left[\frac{R_S(1-R_S)}{N} + \frac{Z^2}{4N^2}\right]^{1/2}}{1+\frac{Z^2}{N}} \quad (8)$$

Опираясь на рекомендации нахождения доверительных границ, имеющиеся в работе [Agresti, 1998], получим значения доверительных границ более простым образом. Именно, переписав (6) в виде

$$Z_\alpha = |R - R_S|/[R(1 - R)/N]^{1/2}, \quad (6A)$$

и поместив формулу для Z_α в некоторую ячейку электронной таблицы, а какое-то, скажем 0,1, значение *R* в другую ячейку, воспользоваться сервисом ПОДБОР ПАРАМЕТРА MS-Excel, потребовав, чтобы в ячейке для Z_α подбиралось значение -1,96 (что соответствует доверительной вероятности 0,95), меняя параметр ячейки, содержащей *R*. Таким образом, будет найдена левая граница R_L доверительного интервала для *R*, а требуя подбора значения +1,96 в ячейке для Z_α, найдем соответствующую правую границу R_R.

Определение необходимого объема выборки

Для того чтобы определить объем нужной нам выборки при заданной доверительной вероятности и предельной ошибке, заменим в (6A) |R – RS| на E и разрешим получившееся уравнение относительно N. Тогда

$$N = [Z^2 R_S(1 - R_S)]/E^2 \quad (6B)$$

В соотношение (6B) входит выборочная доля R_S для определяемого еще неизвестного объема выборки. Поскольку эта доля неизвестна, разумно определить ее так, чтобы объем выборки *N* был максимальным (то есть годился при всех допустимых R_S). Нетрудно видеть, что максимум *N* как функции R_S достигается при R_S = ½, то есть $N_{MAX} = Z^2/4E^2$. Надо подчеркнуть, что если при исследовании возникают или априори имеются (по аналогии, из опыта) некоторые предположения о величине доли признака, то надо использовать эту величину в формуле (6B). Необходимый объем выборки будет при этом меньше максимального, что целесообразно.

Достаточно часто при нахождении долей признаков используется величина предельной ошибки E = 0,05. Используя электронные таблицы MS-Excel, рассмотрим следующий иллюстративный пример.

Пусть у нас имеется возвратная выборка объемом N = 10 объектов и доля признака в ней R_S = 0,9. Используя сервис „Подбор параметра”, получим значения левой и правой доверительных границ для доли признака в генеральной совокупности R при доверительной вероятности, равной 0,95: 0,59 < R < 0,98. Размах полученного доверительного интервала представляется излишне широким. Найдем максимальный объем выборки N_{MAX} для той же доверительной вероятности 0,95 (напомним, что при этом Z = 1,96) и предельной ошибке E = 0,05. Округляя рассчитанный требуемый объем выборки до целого вверх, имеем: N_{MAX} = 385. Вновь, как и ранее, используя сервис „Подбор параметра”, получаем новый более узкий доверительный интервал: 0,87 < R < 0,93. Это достигнуто ценой значительного роста требуемого объема выборки.

Изложенная выше схема расчета требуемого объема для возвратной выборки может быть перенесена на случай безвозвратной. При этом, правда, генеральная совокупность должна быть конечной и нужно знать ее объем $N_{ГС}$. Для указанного переноса необходимо лишь в использованную ранее формулу для выборочного стандартного отклонения возвратной выборки ввести корректирующий множитель $[(N_{ГС} - N)/(N_{ГС} - 1)]^{1/2}$: $\sigma_{ББ}(R_S) = \sigma_{ВВ}(R_S)[(N_{ГС} - N)/(N_{ГС} - 1)]^{1/2}$. Здесь $\sigma_{ВВ}$ и $\sigma_{ББ}$ стандартные отклонения для возвратной и безвозвратной выборок соответственно.

Действуя, как и ранее, получим соотношение для требуемого при заданной погрешности и доверительной вероятности объема безвозвратной выборки:

$$N_{ББ} = [Z^2 R_S(1 - R_S)N_{ГС}]/[E^2(N_{ГС} - 1) + Z^2 R_S(1 - R_S)] \quad (6C)$$

Вновь наибольшее значение объема выборки получается при R_S = ½, и $N_{ББМАХ} = (Z^2 N_{ГС})/[4E^2(N_{ГС} - 1) + Z^2]$. Это выражение может быть переписано в виде $N_{ББМАХ} = (N_{ВВМАХ}N_{ГС})/[N_{ГС} - 1 + Z^2/4E^2]$. Отсюда следует, что $N_{ББМАХ} < N_{ВВМАХ}$, если $Z^2/4E^2 - 1$ положительно, что заведомо имеет место при используемых обычно Z = =1,96 и E = 0,05. Так при объеме генеральной совокупности в 1000 объектов мы получим для максимального объема безвозвратной выборки (при тех же требованиях к точности и надежности) $N_{ББМАХ}$ = 278.

Формулу для доверительного интервала в случае безвозвратной выборки точно так же можно получить, модифицируя вышеуказанным образом выражение для стандартной ошибки. Проще всего это сделать, отправляясь от (1A), с использованием сервиса „Поиск решения” MS-Excel. При этом получаем:

$$Z_\alpha = |R - R_S|/\{[R(1 - R)(N_{ГС} - N)]/[N(N_{ГС} - 1)]\}^{1/2} \quad (6D)$$

Для прежних исходных данных, $N_{ГС}$ = 1000, N = 10, и доверительной вероятности 0,95 с точностью до второго знака после запятой получим прежний доверительный интервал: 0,59 < R < 0,98. При большей точности доверительный интервал для безвозвратной выборки незначительно уже, чем для возвратной, поскольку объем выборки составляет всего 1% от объема генеральной совокупности. Для безвозвратной выборки максимального размера, обеспечивающей заданную точность и надежность, доверительный интервал с точностью до второго знака после запятой совпадает с аналогичным результатом для возвратной выборки: 0,87 < R < 0,93. Подчеркнем, что требуемый объем безвозвратной выборки при этом на 28% меньше.

Выводы

Таким образом, в работе обосновывается использование в качестве усредненных метрик эффективности идентификации знаний в слабоструктурированной текстовой информации принятых показателей

количественной оценки качества обработки текстовой информации, а именно – полноты, точности, шума и аккуратности, а также меры Ван Ризбергена. Для подтверждения достоверности разрабатываемых технологий применяется метод тестовых коллекций, при использовании которого необходимо решать проблему максимизации надежности использования результатов, полученных по тестовой коллекции, для выводов о генеральной совокупности всех исследуемых текстов. В работе рассмотрено применение подходов математической статистики для определения погрешности оценивания выбранных показателей качества идентификации знаний, а именно, использование методов определения доверительного интервала для доли признака и методов определения необходимого объема релевантной выборки в зависимости от заданной погрешности и доверительной вероятности.

Литература:

[Agresti, 1998] Agresti A., B. Coull A. Approximate is better than exact for interval estimation of binomial proportions // American statistician. – 1998. – N 52. – C. 119–126.

[Brown, 2001] L. D. Brown, T. T. Cai, A. Dasgupta. D. Interval estimation for a binomial proportion // Statistical science. – 2001. – N 2. – P. 101–133.

[Clopper, 1934] Clopper C. J., E. S. Pearson. The use of confidence or fiducially limits illustrated in the case of the binomial // Biometrika. – 1934. – N 26. – P. 404–413.

[Cormack, 1998] Cormack G.V. A Efficient construction of large test collections // G. V. Cormack , C. R. Palmer , C. L. Clarke // Proc. of the SIGIR'98 — P. 282—289.

[Garcia-Perez, 2005] Garcia-Perez M. A. On the confidence interval for the binomial parameter // Quality and quantity. – 2005. – N 39. – P. 467–481.

[ISO 12620:2009] "ISO 12620:2009 - Terminology and other language and content resources -- Specification of data categories and management of a Data Category Registry for language resources".iso.org. 2011. Retrieved 9 November 2011.

[Mizzaro, 1997] Mizzaro S. Relevance: The whole history. Journal of American Society for Information Science. — 1997. — V.48. — Is. 9 — P. 810-832

[Wilson, 1927] Wilson E. B. Probable inference, the law of succession, and statistical inference // Journal of American Statistical Association. – 1927. – N 22. – P. 209–212.

[Медик, 2007] Медик В. А., Токмачев М. С. Математическая статистика в медицине. – М.: Финансы и статистика. 2007.

[Четыркин, 1982] Четыркин Е.М., Калихман И.Л. Вероятность и статистика. М.: Финансы и статистика. 1982

[Шабанов, 2003] Шабанов В.И. Метод классификации текстовых документов, основанный на полнотекстовом поиске / В.И. Шабанов, А.М. Андреев // Труды РОМИП'2003. — СПб. : НИИ Химии СПб гос. ун-та, 2003. — С.52—71.

Сведения об авторах

Нина Хайрова *– профессор кафедры интеллектуальных компьютерных систем Национального технического университета „Харьковский политехнический институт", ул. Фрунзе, 21, Харьков, 61002, Украина; e-mail: nina_khajrova@yahoo.com*

Научные интересы: искусственный интеллект, идентификация знаний из текстов, Text Mining, Opinion Mining, Web Mining, Natural language processing, искусственный интеллект

Наталья Шаронова *- профессор, заведующий кафедрой интеллектуальных компьютерных систем Национального технического университета „Харьковский политехнический институт", ул. Фрунзе, 21, Харьков, 61002, Украина; e-mail: nvsharonova@mail.ru*

Научные интересы: искусственный интеллект, математическое моделирование, автоматизированные библиотечные системы

Узлов Дмитрий *– соискатель кафедры интеллектуальных компьютерных систем Национального технического университета „Харьковский политехнический институт", ул. Фрунзе, 21, Харьков, 61002, Украина; e-mail: poputcik@mail.ru*

Научные интересы: системы автоматической обработке текстов на естественном языке, экстракция и идентификация знаний, искусственный интеллект

Solution of the Problem of Formal Evaluation of Effectiveness of the Technology Knowledge Identification in Semistructured Text Information

Nina Khairova, Nataliya Sharonova, Dmytro Uzlov

Abstract: *The traditional approach (the comparison with a "reference" result) for evaluating quality of the technology to identify knowledge extracted from text arrays is badly applicable out of a need to create the reference answer for each specific set of electronic documents. In this paper we show that integral quantitative coefficients of recall, precision and F-measure can be used to assess effectiveness of linguistic technologies of knowledge identification in texts. Justifying the possibility of using the test collections method for the experimental validation of obtained efficiency coefficients, we propose the use of the approach based on mathematical statistics methods. The procedures of using sampling fraction of the indicator as a characteristic of evaluating the proportion of relevant documents in the general population are reviewed. The paper shows the argumentation to the fact that, in important practical cases of text collection samples, asymmetry of a confidence interval at the binomial distribution can be overcome by approximated transition to the normal distribution. We also propose the methods of determining the confidence interval for the indicator fraction that are based on Wilson approach, and the method of determining the required size of the relevant sample depending on the specified error and confidence probability as well.*

Key worlds: *evaluation of effectiveness, semistructured text information, test collections method, size sample*

OWL КАК СТАНДАРТНАЯ МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ТРАНСДИСЦИПЛИНАРНЫХ ЗНАНИЙ В SEMANTIC WEB

Андрей Михайлюк

***Аннотация**: Рассматривается причины появления OWL как стандарта для представления онтологий в Semantic Web, его структура и возможности для представления онтологий трансдисциплинарных знаний. Предложены рекомендации по использованию OWL на междисциплинарном уровне.*

***Ключевые слова**: OWL, междисциплинарные исследования, онтология.*

***ACM Classification Keywords**: 1.2.4. Knowledge Representation Formalisms and Methods*

Введение

Не вызывает сомнений тот факт, что подавляющее большинство знаний, которыми владеет человечество, представлено в цифровом виде в сети Интернет. Простота доступа, надежность хранения и широкий диапазон вариантов представления (включая интерактивные) делают Интернет неотъемлемым источником информации в любых научных исследованиях. Особенно важным данный вид представления и доступа к знаниям выступает в междисциплинарных исследованиях, где рабочие группы территориально и инфраструктурно отдалены. Однако, невзирая на объективные преимущества цифрового представления знания в компьютерных сетях для трансдисциплинарных исследований, данная область требует дальнейшего усовершенствования.

Постановка задачи

Одним из перспективных направлений исследования являются исследования моделей представления онтологий, предназначенных для использования в различных областях науки и хозяйства, в частности, модель и язык OWL, рекомендованный концерном W3C и де-факто ставший стандартом представления онтологий в развивающимся пространстве Semantic Web. Несмотря на то, что назначением OWL является представление онтологических знаний в глобальном информационном пространстве, текущие исследования, исходя из существующих онтологий, онтолого-ориентированных проектов и систем, ограничиваются узкоспециализированными областями: медицина, биология, незначительная часть социальных и технических наук [W3C Wiki, Semantic Web Wiki]. Несомненно, внимания требует **слабо исследованная область применения OWL на трансдисциплинарном уровне**, что в перспективе послужит интегрирующим звеном в задаче построения Semantic Web. Целью данной работы является анализ структуры и возможностей OWL в рамках применения на междисциплинарном уровне и **выработка рекомендаций для представления в OWL трансдисциплинарных онтологических знаний**.

Предпосылки создания OWL

Первым шагом в создании глобальной межпредметной коммуникации, которая на порядок повысила эффективность и возможности взаимодействия исследователей, стало развитие Интернет и его основного способа представления информации (в т.ч. и знаний) – языка гипертекстовой разметки **HTML**. Основной функцией HTML является **оформление** текстовой и графической информации в структурированном виде (абзацы, списки, таблицы и т.п.). Его структура призвана упростить и унифицировать **способ визуального отображения** информации в окне браузера. Данные возможности позволяют авторам наглядно представлять результаты своих исследований, соискателям – легко понимать содержание изложенной информации, а компьютерным системам – упростить алгоритмы ее отображения.

Следующим вопросом, возникшим с ростом глобальной сети, стал поиск информации. Очевидно, что исследователь при поиске необходимой информации не в состоянии пересмотреть (а тем более понять) содержание миллионов источников. Для этого были созданы **автоматические системы поиска** ресурсов, релевантных запросам пользователя, а язык HTML был дополнен **метаинформацией** (тегами), **идентифицирующей общее содержание** или область, к которой относится ресурс. Отметим, что такая метаинформация уже описывает не способ отображения основной информации ресурса, а его **общую семантику**. Поисковые системы значительно упрощают поиск ресурса, необходимого исследователю. Однако с дальнейшим ростом количества информации в глобальной сети отфильтрованный список возможных релевантный ресурсов сам стал настолько объемным, что обработка исследователем такого количества источников с целью выделения знаний из информационных ресурсов стала невозможной. Эволюция поисковых систем лишь количественно решала данную проблему, сужая круг поиска, в то время как выделение знаний все еще оставалось исключительно задачей человека. Кроме того, подход обособленной идентификации общей семантики ресурса по специальным тегам не дает гарантии, что ресурс действительно отвечает заявленной семантике. Так, множество ресурсов для поднятия собственного рейтинга в поисковых системах помечаются наиболее популярными тегами вместо наиболее релевантных содержанию ресурса.

С целью преодоления указанных проблем возникло два основных направления: попытка **автоматического выделения семантики из самого ресурса** (а не из искусственно прописанных тегов) и **внесение в структуру ресурса семантической составляющей** наряду (а иногда и вместо) визуальной разметки. Первый путь включает обработку естественного языка, распознавание графический образов, аудио-сигналов и т.п. Данный подход дает возможность получать знания из уже созданных и значительно упрощает выделение знаний из новых ресурсов, но реализация такого подхода на должном уровне выступает научно-технической проблемой, надежные способы решения которой не определены и по сегодня. Следует, однако, отметить существование методов и систем, которые реализуют этот подход поверхностно (как правило, средствами математической статистики) но достаточном для ряда прикладных задач уровне (реферирование естественно-языкового текста, поиск людей по фотографии или фотороботу и т.д.). Другой подход заключается в создании информационных ресурсов, структура которых изначально описывала бы семантику информационных составляющих, таких как термины,

определения, графические изображения и др. Такая метаинформация должна быть простой для обработки существующими инструментами работы с Web-ресурсами в автоматическом режиме. Так, на основе языка XML как расширения HTML был предложен **фреймворк описания ресурсов RDF** (Resource Description Framework) и определена **схема описания RDFS** (RDF Schema). Концептуально RDF всего лишь определяет широко известный принцип формирования фактов при помощи бинарных отношений между двумя ресурсами [Hoekstra 2009]. В дополнение, RDFS вносит семантические структуры, такие как таксономические отношения, возможность типизации (определение некоторых ресурсов как типов для других ресурсов), определения классов ресурсов, их свойств и т.п. RDFS как модель является достаточно абстрактной для выражения любых фактов, включая правила, ограничения и мета-утверждения. Однако, такая абстрактность вынуждает разработчиков онтологий вводить собственную мета-семантику и семантику верхнего уровня в каждую онтологию, что значительно снижает степень определенности при анализе RDF/RDFS онтологии автоматическими машинами вывода. Очевидной стала необходимость внесения общепринятой семантики (логические и множественные отношения, транзитивные, симметричные и др. свойства отношений, арифметические и списковые примитивы, эксплицитные понятия тождественной истинности и ложности и др.) в спецификацию языка представления онтологий.

Внесением общей онтологической семантики в модель представления знаний одновременно занимались две организации: DARPA (US Defense Advanced Research Projects Agency) и European Research Area в рамках проекта Information Society Technologies. Независимо были разработаны две соответствующие модели – **DAML** (DARPA Agent Markup Language) [DARPA, McIlraith, 2001] и **OIL** (Ontology Inference Layer или Ontology Interchange Language) [Fensel, 2001]. Обе базируются на RDF/RDFS и в значительной мере пересекаются. В дальнейшем две модели были объединены в одну **DAML+OIL** [Connolly, 2001]. Как отмечается в [Hoekstra 2009], она получила семантику дискриптивной логики от OIL, что дало возможность построения эффективных машин вывода, и совместимость с RDF/RDFS от DAML, что позволило использовать DAML+OIL в рамках Semantic Web. Приняв во внимание все преимущества DAML+OIL, концерн W3C создал аналогичную модель и язык с большей выразительной мощностью и точностью описания онтологий – **OWL** (Ontology Web Language). Последним на сегодня является стандарт **OWL2** [OWL2 Overview, 2012] – в дальнейшем изложении, если иное не указано явно, будем использовать аббревиатуру “OWL” в значении последнего доступного стандарта OWL, то есть OWL2.

Язык и модель OWL

Язык OWL можно рассматривать в разных аспектах. Так, “горизонтально” можно выделить **синтаксис** и **семантику** языка OWL,а также **структуру** OWL онтологии. Семантика и абстрактная структура онтологии определяют **модель OWL**.

Синтаксисы и структура OWL

Единственным необходимым для OWL **синтаксисом** является **RDF/XML**. Он гарантирует поддержку обмена онтологиями между гетерогенными системами, которые могут даже не поддерживать сам язык OWL як таковой (чистые RDF/RDFS системы) – сериализация в данный синтаксис определяет правила

приведения элементов языка OWL к абстрактным конструкциям RDF, записанным в виде иерархии XML тегов.

Другим синтаксисом, который использует XML как низкоуровневый язык структуризации элементов OWL онтологии является синтаксис **OWL/XML**. Его отличие от RDF/XML заключается в том, что сериализатор не описывает встроенные элементы языка OWL в терминах RDF, полагаясь на то, что система-получатель должна корректно распознать и проинтерпретировать специфические элементы OWL. Целью введения данного синтаксиса является предоставление возможности оптимального представления онтологий для систем, ориентированных на OWL как внутренний язык и XML як формат обмена.

Важным, но не обязательным для поддержки, является **Functional Syntax**. Его особенность заключается в компактной структурной записи элементов OWL в виде, подобном записи предикатов математической логики или функциональным языкам (например LISP) – отсюда и возникло название синтаксиса. Как отмечено в [OWL2 Overview, 2012], данный синтаксис наилучшим образом отображает структуру онтологии. Стандарт [OWL2 Structure, 2012], который описывает структуру абстрактной онтологии, использует именно Functional Syntax.

Manchester Syntax также оперирует встроенными элементами OWL, которые должны распознаваться системой, и направлен на представление, оптимизированное под сохранение и загрузку онтологий, базирующихся на дискриптивной логике.

Последний опциональный синтаксис OWL – это **Turtle** (Terse RDF Triple Language) – компактный способ записи графов бинарных отношений RDF.

На общем уровне OWL как языка стандарт [OWL2 Structure, 2012] выделяет три составляющих:

- **Сущности**, такие как классы, свойства и экземпляры;
- **Выражения** – сложные записи в описываемом домене, семантика которых соответствует множествам определенных анонимных сущностей;
- **Аксиомы** – утверждения, которые считаются тождественно истинными в универсуме дискурса.

Уровни использования OWL

Для целей практического использования OWL, целесообразно разделять его элементы также по уровню использования:

- **Мета-уровень** онтологии, где определяется, из каких частей состоит онтология (импорт других онтологий), какие словари доступны в онтологии, какие сокращения будут использоваться в онтологии, название, версия, аннотация к онтологии и т.п. Структуры и элементы данного уровня значительно расширяют возможности удаленного и разделяемого пользования онтологиями, их классификации, обмена, оптимизации и т.д.;
- Уровень **структур данных** – часть словаря и правил OWL, которые определяют доступные типы литералов (например, текстовые строки, целые числа, даты и время), уникальные и универсальные способы идентификации сущностей (IRI [Duerst, 2005], Unicode [Yergeau, 2003]), методы обращения к бинарным ресурсам и др. Данный уровень обеспечивает максимальную универсальность относительно локации и области использования OWL онтологий;

- **Онтологический уровень** – структуры и элементы OWL, которые используются для непосредственного описания содержательной части онтологии. Составляющие данного уровня определяют выразительную мощность, семантику, противоречивость и возможности анализа OWL онтологий.

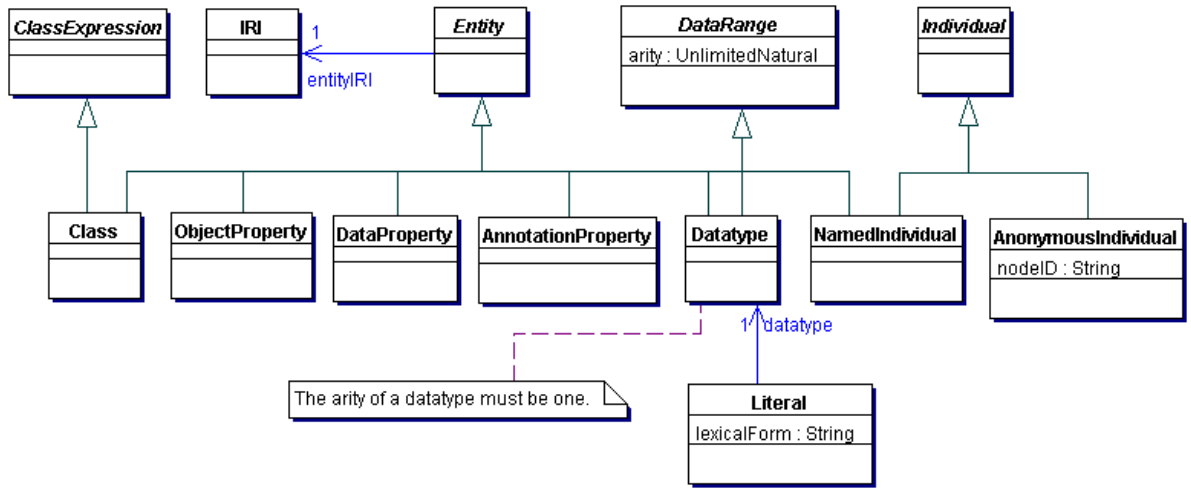


Рис. 1. Сущности OWL и их зависимости

Языковые составляющие по уровням использования OWL взаимно пересекаются, то есть на каждом из трех уровней OWL предоставляет набор языковых средств всех трех типов. Поскольку в цели данной работы входит анализ прежде всего онтологических (а не технических) возможностей OWL, рассмотрим лишь последний уровень использования OWL – онтологический уровень.

Базовыми элементами OWL являются **сущности**. Общую структуру зависимостей между типами сущностей и их место в OWL отражает рис. 1 [OWL2 Structure, 2012].

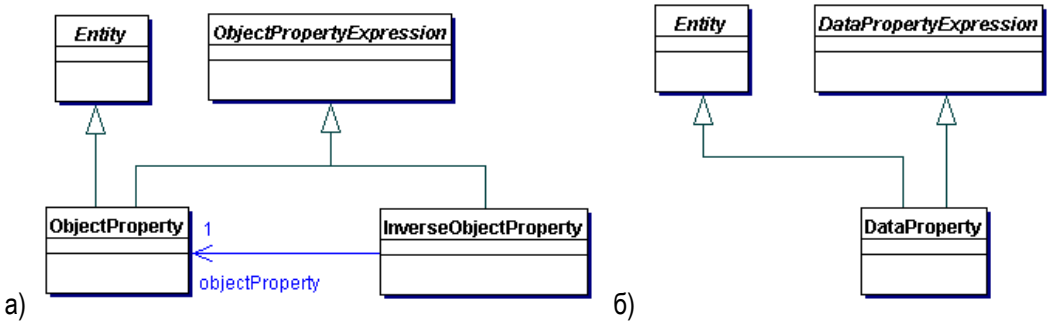
Сущности онтологического уровня OWL включают:

- **Классы** (*Class*). Общепринятое понятие в онтологическом инжиниринге, которое описывает множество элементов (объектов, действий, явлений) объективного мира, собранных по наличию некоторого общего признака. Другие источники могут также использовать аналогичные термины "**понятие**", "**фрейм**" и др. Аналогично многим другим онтологическим высокоуровневым языкам [Genesereth, Sowa, 2000 (1)], OWL содержит как минимум два специальных имплицитно определенных класса: *owl:Thing* и *owl:Nothing*;
- **Типы данных** (*Datatype*). Определяет естественное (качественное) ограничение на множество литеральных значений и их специфическую семантику (например, тип данных *integer* ограничивает связанную переменную множеством целых чисел Z и определяет набор математических операций над такой переменной как множество математических операций над полем Z);
- **Объектные свойства** (*ObjectProperty*) описывают бинарные отношения между экземплярами классов. В отличие от широко известного разделения характеристик объектов на отношения (N-

арные отношения) и атрибуты (унарные отношения), в OWL свойством объекта называется исключительно бинарное отношение. В соответствии с отображением на семантику RDF [Patel-Schneider, 2012], объектные отношения OWL являются RDF классами, а поэтому могут формировать иерархии, иметь экземпляры и т.п.;

- **Свойства данных** (*DataProperty*) – аналогично объектным свойствам характеризуют (вводят ограничения на) экземпляр класса путем определения бинарного отношения, однако вторым операндом здесь выступает не другой экземпляр, а литерал. Свойства данных, как и объектные свойства, также являются RDF классами со всеми вытекающими возможностями;
- **Экземпляры классов** (*Individual*), которые описывают элементы объективного мира, типом которых являются некоторые классы. Важным отличием экземпляров от всех других сущностей является то, что их имена локальны для данной онтологии, то есть если имена сущностей в данной и импортированной онтологиях совпадают, такие сущности считаются различными.

Выражения OWL подразделяются на выражения свойств (объектных и свойств данных) и выражения классов. **Выражения объектных свойств**, согласно [OWL2 Structure, 2012] (см. рис. 2), помимо самих свойств, может описывать свойство, обратное данному. Данная часть OWL, несмотря на столь незначительную выразительность, на данный момент не нуждается в расширении, т.к. основные описательные возможности в отношении свойств обеспечиваются многообразием выражений классов (т.к. любые свойства OWL сами являются классами).



а) б)

Рис. 2. Структура выражений свойств OWL:

а) выражения объектных свойств; б) выражения свойств данных.

Выражения классов, как говорилось ранее, имплицитно определяют новый класс, удовлетворяющий указанным характеристикам. Такими характеристиками и соответствующими типами выражений являются:

- **Теоретико-множественные** операции, соответствующие основным способам задания множеств, а именно **пересечение** (*ObjectIntersectionOf*), **объединение** (*ObjectUnionOf*), **дополнение** (*ObjectComplementOf*) и **перечисление элементов** множества (*ObjectOneOf*). В терминах понятий как аналогов OWL классов, данная группа выражений описывает новое понятие через его объем, или, иначе говоря, количественно (экстенсионально) определяет класс;

- **Ограничения по объектным свойствам**, в отличие от теоретико-множественных операций, определяют класс качественно, интенсионально. Семантика этого типа выражений заключается в определении класса, экземпляры которого неким образом связанны заданным бинарным отношением (объектным свойством) с указанными экземплярами. Объектное свойство и экземпляры, формирующие его диапазон значений, так же могут быть заданы эксплицитно. Типы ограничения включают ограничения с **квантором существования** (*ObjectSomeValuesFrom*), **универсально квантифицированные** (*ObjectAllValuesFrom*), **индивидуальные** ограничения (*ObjectHasValue*), **самоограничения** (*ObjectHasSelf*), ограничения с **минимальной** (*ObjectMinCardinality*), **максимальной** (*ObjectMaxCardinality*) и **точной** (*ObjectExactCardinality*) **кардинальностью**. Отметим, что подобное многообразие вариантов интенсионального определения классов наилучшим образом соотносится с применением OWL к представлению онтологий в междисциплинарных исследованиях, т.к. данный способ определения понятий является, с одной стороны, одним из основных в научном мире, а с другой – описывает определяемое понятие безотносительно к конкретным, специфическим для данной дисциплины, частным его (понятия) проявлениям или экземплярам;
- **Ограничения по свойствам данных** аналогичны предыдущему типу выражений, с тем лишь основным отличием, что операндом свойства является ограничение экземпляром данных (открытым или закрытым диапазоном, множеством, единичным или анонимным значением). Т.к. данный тип выражений, как и другие, определяет класс, а ограничивающим операндом свойства является не экземпляр класса, в нем отсутствует самоограничение, доступное в выражениях ограничения по объектным свойствам.

Аксиомы можно назвать наиболее важными элементами OWL, т.к. их многообразие во многом определяет богатство языка и его удобство в составлении онтологий, что особенно важно в описании сложных онтологий, представляющих домены научных дисциплин. С другой стороны, определение закономерностей в описываемом онтологией универсуме при помощи заранее предопределенного набора аксиом упрощает разработку машин вывода и других систем анализа онтологий, т.к. семантика каждой предопределенной аксиомы известна, а способ идентификации в онтологии сводится к сравнению с шаблонной строкой. В противовес этому, анализаторы “чистых” RDF онтологий должны определять, например, свойство транзитивности некого отношения анализом структуры данного отношения, его родовых отношений и логических отношений, в которые данное входит как операнд. Как и выражения, аксиомы целесообразно разделять по типу описываемых сущностей:

- **Аксиомы выражения классов** определяют взаимоотношения между OWL классами, тождественно истинные в данной онтологии. Сюда входят отношения **родовидовой зависимости** (*SubClassOf*), **эквивалентный** класс (*EquivalentClasses*), **несовместный** (непересекающийся, отличный, дизъюнктный) класс (*DisjointClasses*), **объединение несовместных** классов (*DisjointUnion*). Данные аксиомы являются мощным инструментом в определении родовидовой иерархии онтологии;

- **Аксиомы объектных свойств** можно подразделить на две группы согласно двойственности природы самих OWL свойств:
  - Ввиду того, что объектные свойства являются классами, над ними определен ряд аксиом, аналогичных аксиомам выражения классов. Это аксиомы утверждения **родовидовой зависимости** (*SubObjectPropertyOf*, *ObjectPropertyChain*), **эквивалентности** (*EquivalentObjectProperties*), **несовместности** свойств (*DisjointObjectProperties*) и, согласно рис. 2 (а), аксиомы утверждения **инверсности** двух объектных свойств (*InverseObjectProperties*);
  - С другой стороны, свойства в OWL являются бинарными отношениями, поэтому вторым подмножеством аксиом над ними являются утверждения о типе бинарного отношения, а именно **функциональность** (*FunctionalObjectProperty*) и **обратная функциональность** (*InverseFunctionalObjectProperty*), **рефлексивность** (*ReflexiveObjectProperty*) и **иррефлексивность** (*IrreflexiveObjectProperty*), **симметричность** (*SymmetricObjectProperty*) и **асимметричность** (*AsymmetricObjectProperty*), **транзитивность** (*TransitiveObjectProperty*). Так же, будучи бинарным отношением, объектное свойство имеет **область определения** и **область значения**, что так же может быть описано соответствующими аксиомами (*ObjectPropertyDomain* и *ObjectPropertyRange* соответственно).
- **Аксиомы свойств данных** включают утверждения, аналогичные аксиомам объектных свойств (*SubDataPropertyOf*, *EquivalentDataProperties*, *DisjointDataProperties*, *FunctionalDataProperty*, *DataPropertyDomain*, *DataPropertyRange*), за тем лишь исключением, что имея областью определения экземпляр класса, а областью значения – литеральный тип, не могут быть обратно-функциональными, рефлексивными и иррефлексивными, симметричными и асимметричными, транзитивными;
- **Определения типов данных** (*DatatypeDefinition*) служит инструментом для введения специфических (пользовательских) типов в онтологию ограничением существующих. Такие аксиомы, помимо удобства использования специфического для данной предметной области типа и автоматической поддержки операций над ним (унаследованной от родительского предустановленного типа), дает возможность анализировать специфические типы различных предметных областей в междисциплинарных исследованиях при интеграции онтологий с пользовательскими типами. Иначе говоря, пользовательские типы – это не только удобный технический прием, но и элемент семантики предметной области;
- **Определение ключей** (*HasKey*) помечает подмножество свойств экземпляров некого класса как уникально идентифицирующих каждый экземпляр. Как и в случае с определением пользовательских типов, данная возможность, помимо очевидного влияния на логический вывод над онтологией, вносит семантику, специфическую для данной предметной области, что также может быть важно при анализе онтологий в рамках междисциплинарных исследований;

- **Утверждения** об экземплярах классов (также известные как **суждения** или **факты**) определяют **эквивалентность** (*SameIndividual*) или **различность** (*DifferentIndividuals*) каждого из перечисленных экземпляров, **принадлежность экземпляра к** некому **классу** (*ClassAssertion*), **присутствие** или **отсутствие свойства** у данного экземпляра (*ObjectPropertyAssertion* и *NegativeObjectPropertyAssertion*, *DataPropertyAssertion* и *NegativeDataPropertyAssertion*).

Следует также отметить, что такой богатый набор всевозможных аксиом и возможность явного указания истинности или ложности некоторого выражения в значительной степени упрощает создание онтологии верхнего уровня – основы для построения общей междисциплинарной онтологии из онтологий различных предметных областей.

Использование OWL в онтологиях трансдисциплинарного уровня

Основной особенностью междисциплинарных исследований, по сравнению с исследованиями в рамках некой отдельной предметной области, является очерчивание нерешенных проблем, слабо исследованных вопросов и поиск новых знаний, находящихся на стыке дисциплин. Говоря об онтологическим инжиниринге, где каждая предметная область, ее знания и задачи описываются онтологиями, можно заметить, что основной задачей при проведении междисциплинарных исследований является **обобщение** онтологий. Обобщение включает в себя две основных составляющих: **построение онтологии верхнего уровня**, определяющей общенаучную, или хотя бы общую для данных двух онтологию, а так же эффективные механизмы непротиворечивой **интеграции онтологий** разных предметных областей.

Обсуждения подходов к **построению онтологий верхнего уровня** неоднократно проводились в многочисленных работах [Палагін, 2006, Найханова, 2005, Schneider, 2003]. В рамках данной статьи важно отметить, что OWL предоставляет широкий спектр возможностей для построения верхнеуровневых онтологий. Так, помимо общеиспользуемых аксиом описания таксономических иерархий и раскрытия содержания понятий категориального уровня, **особо полезными элементами OWL** в данном случае являются:

- Аксиомы, устанавливающие высокоуровневые **отношения эквивалентности** между классами (*EquivalentClasses*, *DisjointClasses* и *DisjointUnion*), позволяющие на категориальном уровне определить как точки соприкосновения различных предметных областей или теорий (что необходимо при интеграции доменных онтологий), так и потенциальные конфликты (например, при анализе онтологий из разных дисциплин легко установить недостаточность или противоречивость некоторых их частей посредством проверки на несовместность надклассов категориального уровня);
- Аксиомы объектных свойств, предназначенные для **построения иерархий отношений**, что является важной частью онтологической картины мира [Sowa, 2000 (2)]. К ним относятся аксиомы *SubObjectPropertyOf*, *ObjectPropertyChain*, *EquivalentObjectProperties*, *DisjointObjectProperties* и *InverseObjectProperties* и аналогичные отношения над литеральными типами;

- Аксиомы типизации объектных свойств как бинарных отношений, что позволяет на верхнем уровне определить **схему взаимозависимости неизвестных** на данный момент **классов и отношений** конкретных предметных областей приписыванием верхнеуровневому отношению некого типа. Это аксиомы *FunctionalObjectProperty*, *InverseFunctionalObjectProperty*, *ReflexiveObjectProperty*, *IrreflexiveObjectProperty*, *SymmetricObjectProperty*, *AsymmetricObjectProperty*, *TransitiveObjectProperty*, *ObjectPropertyDomain* и *ObjectPropertyRange*;
- Возможность оперирования **анонимными сущностями** и введения **мета-описаний** позволяют более лаконично и абстрактно устанавливать категориальные правила и основные объективные законы (например, в OWL возможно определить понятия “тяжелый” и “легкий”, оперируя анонимными массами анонимных предметов, имеющими данные характеристики).

Для задачи **интеграции онтологий**, принадлежащих разным доменам знаний, основными рекомендациями являются рекомендации (а в перспективе – методики) по созданию OWL онтологий, пригодных для автоматического анализа на междисциплинарном уровне. Одной из очевидных рекомендаций в данном контексте является искусственное ограничение используемого в предметной онтологии синтаксиса OWL. Данная идея, в общем, не является оригинальной. Как известно, возможность использования логических и мета-онтологических конструкций на одном уровне приводит к неразрешимости OWL в общем смысле [Motik, 2007], а богатая выразительность значительно усложняет разработку эффективных алгоритмов анализа онтологий. Поэтому, наряду с полным языком OWL (который иногда называют OWL Full [Hoekstra 2009]), стандартом [Motik, 2012] определено также три его подмножества (профиля): OWL EL, OWL QL и OWL RL, предназначенных соответственно для оптимизированного вывода за полиномиальное время на онтологиях значительного объема; ускорения выполнения конъюнктивных запросов (в частности, на SQL-подобных языках) к базам со значительным количеством сравнительно небольших онтологий; оптимизации выполнения (вплоть до полиномиальной сложности) вывода на правило-ориентированных системах анализа RDF-подобных графов. Детальное рассмотрение указанных подмножеств выходит за рамки данной работы, поэтому, приняв искусственное ограничение синтаксиса языка за широко используемую практику, можно предложить следующие рекомендации:

1. Использование **интенсиональных определений** понятий предметной области. Относительно OWL, это означает преимущественный способ введения классов через **наследование** и **ограничение по объектным свойствам**. Так, для описания нового класса, отличающегося от надкласса некоторым набором отношений (собственное содержание понятия), следует ввести в онтологию две записи: (1) аксиому родовидовой зависимости *SubClassOf* от надкласса и (2) аксиому родовидовой зависимости от анонимного класса (или классов), описанного выражением ограничения по объектному свойству *ObjectSomeValuesFrom* – от этого анонимного класса будут унаследованы новые свойства;
2. Максимальное описание **содержания понятий**, специфических данной дисциплине. Данная необходимость мотивирована тем, что формальное обобщение на междисциплинарном уровне возможно через имена понятий, либо через их содержание [Палагин, 2010]. Известными

проблемами инженерии знаний, особо остро стоящей в междисциплинарных исследованиях, являются синонимия и омонимия понятий. Поэтому, для качественного анализа онтологии некой предметной области каждое понятие рекомендовано раскрывать содержательно. Со стороны OWL, это должно выражаться в явном описании максимального количества объектных свойств классов и их значимых свойств данных. Как говорилось ранее, для описания объектных свойств класса необходимо ввести в онтологию аксиому родовидовой зависимости от анонимного класса (или классов), описанного выражением ограничения по объектному свойству *ObjectSomeValuesFrom*;

3. **Минимизация используемого словаря** OWL. Данная рекомендация означает описание предметной области наиболее простыми конструкциями языка. Причиной этому служит необходимость обеспечения однозначного и максимально прозрачного анализа онтологий на стыке дисциплин. Машины вывода способны обработать значительное количество сложных логических и теоретико-множественных структур, сводя их к проверке истинности или ложности десятков (или даже сотен) сгенерированных атомарных выражений (например, дизъюнктов в методе резолюции), семантика которых сложно ассоциируется с какой-либо предметной областью или классом. Для междисциплинарного обобщения же необходим более наглядный метода анализа, результатом которого должны быть понятия или правила, принадлежащие одной из исходных онтологий (а значит и дисциплин), или легко воспринимаемые пользователем системы (например, экспертом в одной из исследуемых предметных областей) как сущности на границе дисциплин. Так, целесообразно ограничивать синтаксис OWL к одному из подмножеств описательной логики [Baader, 2010]. Например, поскольку основным назначением научных онтологии предметных областей является представление понятий, но не экземпляров, возможно использование только элементов TBox [Baader, 2010], т.е. ограничиться минимальным онтографом <X,R>.

Данный список рекомендаций не является полным, однако следования им значительно упрощают анализ онтологий предметных областей на междисциплинарном уровне, что в свою очередь, делает возможным его автоматическую реализацию, простейшим и первым необходимым вариантом которого является системная интеграция онтологий. С другой стороны, указанные рекомендации предоставляют достаточный инструментарий для создания онтологий (как в ручном, так и автоматическом режиме), используя лишь небольшое подмножество элементов OWL, что делает их совместимыми с существующими системами визуализации и редактирования онтологий [Semantic Web Wiki].

Выводы

Эволюция моделей и языков представления знаний в глобальных масштабах привела к объективной необходимости исследований в области онтологического инжиниринга на междисциплинарном уровне. В результате анализа причин возникновения той или иной технологии в этом направлении установлено, что перспективным является использование языка OWL как стандартного в глобальной сети нового поколения – Semantic Web. Однако неисследованная область применения данного языка в онтолого-

ориентированных междисциплинарных исследованиях значительно затрудняет практические шаги в этом направлении.

Первым необходимым на данном пути этапом, проделанным в работе, является краткое описание модели и возможностей языка OWL с акцентом на его использовании на междисциплинарном уровне. Помимо этого, опыт в области системной интеграции онтологических знаний дал возможность предложить общие и практические рекомендации по созданию онтологий предметных областей, подходящих для автоматического анализа в рамках трансдисциплинарных исследований. Указанные рекомендации также могут быть полезны при разработке систем автоматического построения онтологий (например, из естественно-языкового текста), так как упрощают модель онтологии без значительного ущерба в отношении выразительной способности. Важным является и вытекающий из анализа OWL факт совместимости таких онтологий с наиболее общей моделью представления знаний в Semantic Web – моделью RDF/RDFS, обеспечивающий перспективу их использования в глобальной сети третьего поколения.

Перспективное развитие вопроса использования OWL в трансдисциплинарных исследованиях включает три основных направления:

1. Эмпирические исследования по созданию и экспертному анализу онтологий, принадлежащих разным областям знаний;
2. Адаптация существующих программных систем генерации и интеграции онтологий к междисциплинарному уровню, следуя обозначенным рекомендациям;
3. Разработка методики (возможно, не полностью формальной) создания и анализа онтологий, ориентированных на использование в трансдисциплинарных исследованиях.

Литература

[Baader, 2010] F.Baader, D.Calvanese, D.L.McGuinness, D.Nardi, P.F.Patel-Schneider. The Description Logic Handbook. Theory, Implementation and Applications. 2nd Edition. Cambridge University Press, 2010. – 624 p.

[Connolly, 2001] D.Connolly, F.van Harmelen, I.Horrocks, D.L. McGuinness, L.A.Stein, Lucent Technologies, Inc. DAML+OIL (March 2001) Reference Description. 2001. http://www.w3.org/TR/daml+oil-reference

[DARPA] The DARPA Agent Markup Language Homepage. http://www.daml.org/index.html

[Duerst, 2005] M.Duerst, M.Suignard. RFC 3987: Internationalized Resource Identifiers (IRIs). IETF, January 2005. http://www.ietf.org/rfc/rfc3987.txt

[Fensel, 2001] D.Fensel, F.van Harmelen, I.Horrocks, D.L.McGuinness, P.F.Patel-Schneider. OIL: An Ontology Infrastructure for the Semantic Web. // IEEE INTELLIGENT SYSTEMS, The Semantic Web 2001, 2001. – pp. 38-45

[Genesereth] M.R.Genesereth. Knowledge Interchange Format (draft proposed American National Standard). http://logic.stanford.edu/kif/dpans.html

[Hoekstra 2009. R.Hoekstra. Ontology Representation: Design Patterns and Ontologies that Make Sense. IOS Press, 2009. – 236p.

[Lacy 2005] L.W.Lacy. OWL: Representing Information Using the Web Ontology Language. Trafford Publishing, 2005. – 205p.

[McIlraith, 2001] S.A.McIlraith, T.Cao Son, H.Zeng. Mobilizing the Semantic Web with DAML-Enabled Web Services, 2001. http://www.csd.abdn.ac.uk/ebiweb/papers/mcilraith.doc

[Motik, 2007] B.Motik. On the Properties of Metamodeling in OWL. // Journal of Logic and Computation, 17(4), 2007. – pp 617-637

[Motik, 2012] B.Motik, B.C.Grau, I.Horrocks, Z.Wu, A.Fokoue, C.Lutz. OWL 2 Web Ontology Language Profiles (Second Edition). W3C Recommendation 11 December 2012, 2012. http://www.w3.org/TR/2012/REC-owl2-profiles-20121211/

[OWL2 Overview, 2012] OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation 11 December 2012, 2012. http://www.w3.org/TR/2012/REC-owl2-overview-20121211/

[OWL2 Structure, 2012] OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition). W3C Recommendation 11 December 2012, 2012. http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/

[Patel-Schneider, 2012] P.F.Patel-Schneider, B.Motik. OWL 2 Web Ontology Language RDF-Based Semantics (Second Edition). W3C Recommendation 11 December 2012, 2012. http://www.w3.org/TR/2012/REC-owl2-mapping-to-rdf-20121211/

[Schneider, 2003] L.Schneider. How to Build a Foundational Ontology: The Object-Centered High-level Reference Ontology OCHRE // KI 2003: Advances in Artificial Intelligence. Lecture Notes in Computer Science. Volume 2821, 2003. – pp 120-134

[Semantic Web Wiki] Semantic Web Wiki. Tools. http://semanticweb.org/index.php?title=Tools&oldid=53928

[Sowa, 2000 (1)] J.F.Sowa. Conceptual Graph Standard (updated version from 6 December 2000). NCITS.T2 Committee on Information Interchange and Interpretation. http://users.bestweb.net/~sowa/cg/cgdpans.htm

[Sowa, 2000 (2). J.F.Sowa] Knowledge Representation: Logical, Philosophical and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA, 2000. – 594 p.

[W3C Wiki] Ontology repositories. W3C Wiki. http://www.w3.org/wiki/index.php?title=Ontology_repositories&oldid=56154

[Yergeau, 2003] F.Yergeau. RFC 3629: UTF-8, a transformation format of ISO 10646. IETF, November 2003. http://www.ietf.org/rfc/rfc3629.txt

[Найханова, 2005] Л.В.Найханова. Основные аспекты построения онтологий верхнего уровня и предметной области. // В сборнике научных статей "Интернет-порталы: содержание и технологии". Выпуск 3. / Редкол.: А.Н. Тихонов (пред.) и др.; ФГУ ГНИИ ИТТ "Информика". - М.: Просвещение, 2005. - С. 452-479

[Палагин, 2010] А.Палагин, А.Михайлюк, В.Величко, Н.Петренко. К интеграции онтологий предметных областей. // Information Models of Knowledge. ITHEA, Kiev, Ukraine – Sofia, Bulgaria, 2010. – 470 с. С. 69-85

[Палагін, 2006] О.В.Палагін, М.Г.Петренко. Модель категоріального рівня мовно-онтологічної картини світу // Математичні машини і системи. – 2006. - №3. – С. 91-104

Информация об авторах

Андрей Васильевич Михайлюк *– Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40; e-mail: fruler@ukr.net*

Основные области научных исследований: формальные модели представления знаний, логико-онтологические системы обработки знаний

OWL as a Standard Model for Transdisciplinary Knowledje Representation in Semantic Web

Andrey Mihailyuk

Abstract*: The reasons for emerging OWL as a standard for ontology representation in the Semanic Web are discussed in the paper. OWL structre and possibilities for representing ontologies of transdisciplinary knowledge are outlined. Recommendations for using OWL at the interdisciplinary level are proposed.*

Keywords*: OWL, interdisciplinary research, ontology*

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ РАСПОЗНАВАНИЯ РУКОПИСНЫХ МАТЕМАТИЧЕСКИХ ВЫРАЖЕНИЙ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ НА ОСНОВЕ НЕЧЕТКИХ НЕЙРОННЫХ СЕТЕЙ

Эдрис Надеран

Аннотация: *в статье рассматривается разработанная информационная технология распознавания рукописных математических выражений, вводимых в ЭВМ в режиме реального времени, в основе которой лежат предложенные подходы к распознаванию рукописных символов и структурному анализу математических выражений.*

Ключевые слова: *распознавание рукописных математических выражений, нечеткая логика, нечеткие нейронные сети, структурный анализ.*

ACM Classification Keywords: *I.5.2 Computing Methodologies - Pattern Recognition - Design Methodology - Pattern analysis. G.4 Mathematics of Computing – Mathematical Software - Algorithm design and analysis. I.5.1 Computing Methodologies - Pattern Recognition - Models - Structural.*

Вступление

Математические выражения составляют основную часть в большинстве научных и технических дисциплин, однако на сегодняшний день ввод математических выражений в ЭВМ является сложным и неудобным для пользователя ПК, поскольку осуществляется с помощью традиционных устройств ввода, таких как клавиатура и мышь, к тому же отнимает большое количество времени.

Активное развитие планшетных ПК, наблюдаемое сегодня, приводит к необходимости ввода данных в ЭВМ без использования клавиатуры. По прогнозам компании DisplaySearch к 2016 году объем мировых продаж планшетные ПК превысит объем мировых продаж ноутбуков (Рис. 1). Возможность взаимодействия пользователя с ПК с помощью сенсорной функциональности станет безусловно самой естественной, удобной и быстрой альтернативой для ввода математических выражений в ЭВМ.

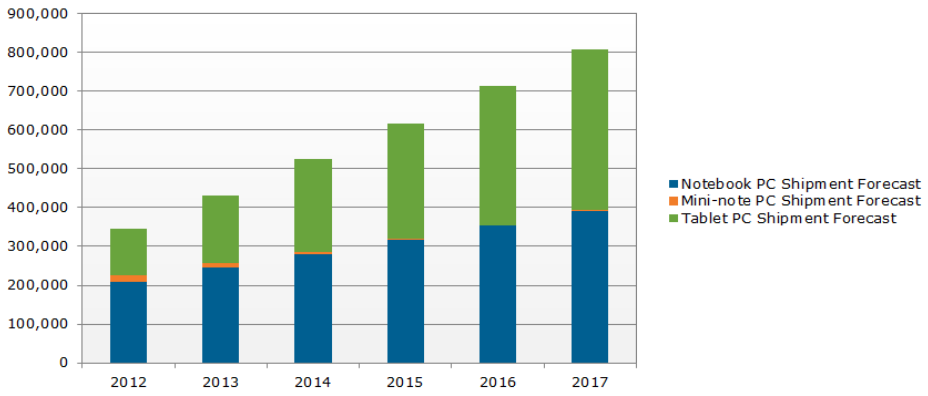


Рис. 1 Прогноз мировых поставок мобильных ПК (источник NPD DisplaySearch)

Целью данной статьи является рассмотрение разработанной информационной технологии, обеспечивающей эффективное решение задачи распознавания рукописных математических выражений, вводимых в ЭВМ в режиме реального времени.

Описание и ключевые особенности информационной технологии

Рассмотрим общую схему работы информационной технологии распознавания рукописных математических выражений, вводимых в режиме реального времени. Процесс распознавания математических выражений состоит из двух основных этапов: распознавания рукописных символов и структурного анализа.

При распознавании рукописных символов в режиме реального времени можно проследить траекторию написания символа и определить точку касания пера и точку отрыва пера при написании. Для получения информативных признаков, характеризующих символы математического выражения, использован предложенный в [Надеран Э., Зайченко Ю.П., 2012] подход, основанный на использовании точек ломаной аппроксимирующей кривую символа. Опытным путем был подобран оптимальный коэффициент точности сглаживания, при котором получено оптимальное соотношение между количеством оставляемых алгоритмом точек и максимальным сохранением очертания символа. В качестве классификатора используется модифицированная ННС NEFCLASS, отличительной особенностью которой является возможность формирования логического вывода на основе аппарата нечеткой логики, что позволяет заложить знания эксперта в систему, при этом параметры функций принадлежности настраиваются с использованием алгоритмов обучения нейронных сетей. Для повышения качества распознавания символов применяется генетический алгоритм обучения параметров функции принадлежности на этапе первичного обучения системы и алгоритм сопряженных градиентов на этапе дообучения системы для улучшения временных показателей [Надеран Э., Зайченко Ю.П., 2012].

Этап структурного анализа позволяет определить пространственные отношения между составляющими математического выражения и включает в себя следующие этапы: размещение, реконструкция и группировка символов [Надеран Э., Зайченко Ю.П., 2013]. Символы разделены на группы, в зависимости от допустимых и обязательных позиций размещения других символов относительно них. На этапе размещения определяется наилучшая позиция размещения для символа по формуле:

$$NP = P * k$$

где P – процент попадания символа

k – коэффициент позиции, принимающий значения:

0 - для недопустимых позиций;

1 - для допустимых позиций;

1,5 - для обязательных позиций.

На этапе реконструкции символов применяется динамическая база эвристических правил, основанная на знаниях о порядке записи и пространственных отношениях между символами, и позволяющая проводить реконструкцию символов и коррекцию неправильно распознанных символов в последовательности путем нахождения ее семантического значения. На этапе группировки некоторые символы позволяют сгруппировать несколько отдельных символов в одну группу. К таким символам относятся: различного

вида скобки, математические аббревиатуры, дробная черта, сумма Σ, произведение $\prod$, интеграл, арифметический корень, точка и запятая. Написанное пользователем математическое выражение просматривается после каждого внесенного изменения, что дает возможность записывать составляющие математическог о выражения в любом порядке, а также вносить изменения в уже написанное выражение.

Основные требования, которым должна соответствовать информационная технология:

1. Гибкость, способность к адаптации и дальнейшему развитию.
2. Надежность.
3. Минимальное время отклика на запрос.
4. Возможность адаптации к различным почеркам.
5. Возможность работы независимо от платформы.
6. Возможность использования результата распознавания в популярных текстовых редакторах.
7. Удобство и простота интерфейса.
8. Эффективность.

Информационная технология распознавания рукописных математических выражений состоит из следующих компонентов (Рис.2):

1. Event Processing, который обеспечивает многозадачность и синхронную обработку событий. К событиям относятся: распознавание символов, структурный анализ математического выражения, удаление символов, обучение нейронной сети.
2. Training, который включает в себя модуль GA Training, реализующий генетический алгоритм обучения, и модуль CGA Training, реализующий алгоритм сопряженных градиентов обучения нейронной сети.
3. Recognition, который включает в себя модуль Classifier NEFCLASS, реализующий нечеткий классификатор NEFCLASS и модуль Features Processing, вычисляющий значения информативных признаков.
4. Structural Analysis, который отвечает за решение задачи структурирования последовательности введенных пользователем символов и состоящий из модуля Expression Analyzer, отвечающего за подготовку входных и выходных данных, модуля Symbol Arrangement, реализующего этап размещения символов, модуля Reconstruction, отвечающего за реконструкцию символов, и модуля Grouping, обеспечивающего этап группировки символов.

Модуль Database Layer обеспечивает взаимодействие системы с базой данных. В базе данных хранятся параметры обученного с помощью генетического алгоритма обучения нечеткого классификатора NEFCLASS, необходимые для этапа распознавания рукописных символов, написанных с помощью одного штриха. Таким образом, в базе данных хранятся параметры функции принадлежности Гаусса, количество нейронов входного и выходного слоя, а также база нечетких правил. Для проведения структурного анализа в базе данных хранится база эвристических правил, а также для каждого символа хранится информация о допустимых, обязательных и недопустимых позициях расположения других символов относительно них.

При коррекции пользователем результата распознавания символов будет дообучаться нейронная сеть и обновляться параметры классификатора в базе данных, а также при коррекции пользователем результатов структурного анализа будут обновляться данные эвристической базы правил для данного

пользователя. Таким образом, предлагаемое решение позволяет персонализировать систему распознавания рукописных математических выражений, повысить эффективность ее работы и подстроить систему под почерк пользователя.

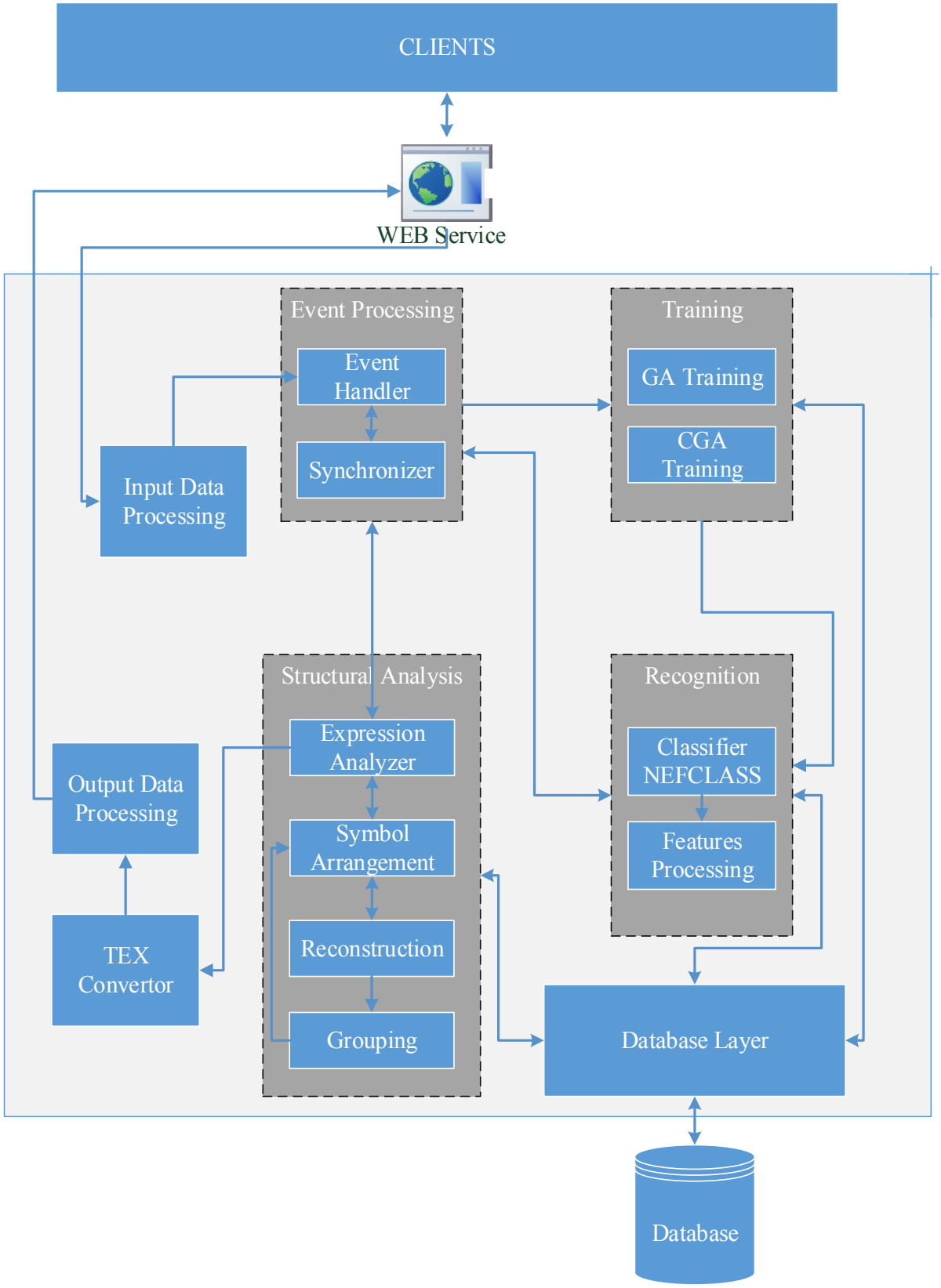


Рис.2 Структурная схема

Экспериментальные исследования и результаты тестирования

Программная часть системы реализована на языке C# на базе платформы Microsoft .NET. На Рис.3 представлена диаграмма классов ИТ. Преимущество использования платформы .NET Framework заключается в возможности реализации исполняемого кода независимого от аппаратной части, поскольку исходный код переводится компилятором в промежуточный байт-код Common Intermediate Language (CIL) и затем код исполняется виртуальной машиной Common Language Runtime (CLR), которая с помощью встроенного в неё JIT-компилятор преобразует промежуточный байт-код в машинные коды нужного процессора. Кроме того, CLR заботится о базовой безопасности, управлении памятью и системе исключений. Необходимо отметить, что использование современной технологии динамической компиляции позволяет достигнуть высокого уровня быстродействия.

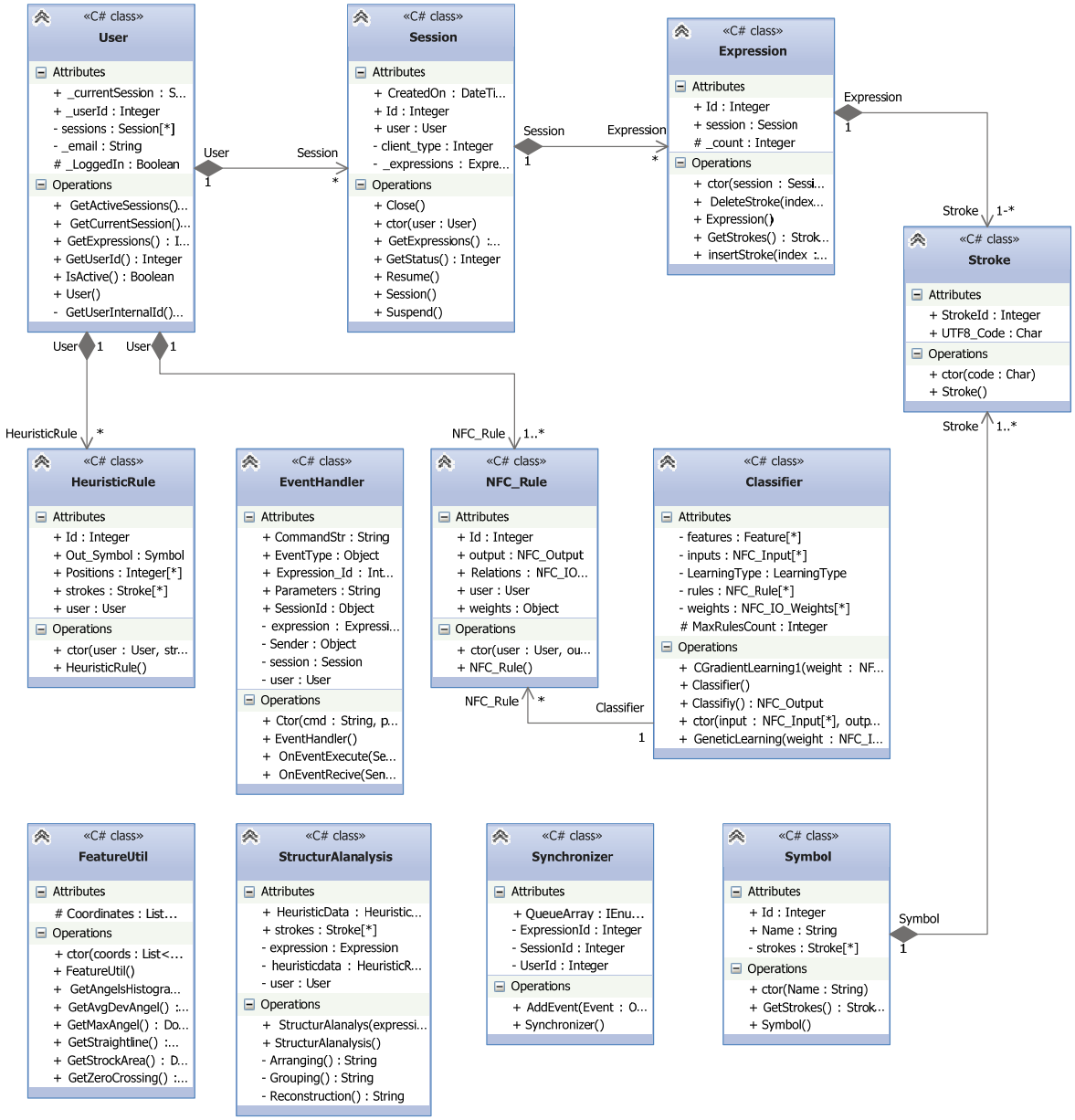


Рис.3 Диаграмма классов ИТ

Для проверки эффективности работы, предложенной информационной технологии были проведены экспериментальные исследования, позволяющие оценить качество распознавания рукописных математических выражений. Для обучения нечеткого классификатора NEFCLASS была использована выборка, состоящая из 260 рукописных математических выражений. Для генерации базы правил каждому образцу ставился в соответствие подходящий рукописный символ. После чего нейронная сеть была обучена с помощью генетического алгоритма обучения. Таким образом, были найдены значения весовых коэффициентов необходимые для инициализации нейронной сети.

Тестирование проводилось путем рукописного ввода математических выражений в режиме реального времени в планшетный ПК с помощью стилуса. Общая выборка для тестирования составила 140 математических выражений. Среднее время распознавания каждого символа составило 150 миллисекунд на компьютере со следующими характеристиками: CPU Intel Core 2 Quad с частотой 2.4 ГГц, RAM 4 Гб. Результаты тестирования отражены на Рис.4 и отражают качество распознавания и структурирования математических выражений.

Распознавание отрезков	Реконструкция символов (для правильно распознанных отрезков)	Структурный анализ математических выражений
91,54%	95,32%	71,29%

Рис.4 Результаты тестирования

Поле “Распознавание отрезков” показывает процент правильно классифицированных символов, написанных без отрыва пера с помощью одного штриха, и позволяет оценить эффективность распознавания. Поле “Реконструкция символов” отображает процент верно реконструированных символов из распознанных правильно отрезков. Поле “Структурный анализ математических выражений” показывает процент правильно структурированных выражений.

Заключение

В статье рассмотрена разработанная информационная технология распознавания рукописных математических выражений, вводимых в ЭВМ в режиме реального времени, в основе которой лежат предложенные подходы к распознаванию рукописных символов и структурному анализу математических выражений. Описаны основные требования, которым должна соответствовать информационная технология и рассмотрены компонентов из которых она состоит. Описаны проведенные экспериментальные исследования рассмотренной информационной технологии. Среднее время распознавания каждого символа составило 150 миллисекунд, что удовлетворяет требованиям к использованию системы в режиме реального времени. Точность распознавания составила 91,54%, точность структурного анализа составила 71,29%.

Благодарности

Статья частично финансирована из проекта ITHEA XXI Института Информационных теорий и Приложений FOI ITHEA и консорциума FOI Bulgaria (www.ithea.org, www.foibg.com)

Литература

[Надеран Э., Зайченко Ю.П., 2012] Надеран Э., Зайченко Ю.П. Выделение информативных признаков рукописных математических символов. - Вісник НТУУ “КПІ” Інформатика, управління та обчислювальна техніка, №57, 2012. – с. 124-128.

[Надеран Э., Зайченко Ю.П., 2013] Надеран Э., Зайченко Ю.П. Подход к разработке системы распознавания рукописных математических выражений вводимых в эвм в режиме реального времени. - Вісник НТУУ “КПІ” Інформатика, управління та обчислювальна техніка, №58, 2013. – с.56-60.

Информация про автора

Надеран Эдрис *– аспирант Национального технического университета Украины “КПИ”, адрес электронной почты: e.naderan@gmail.com*

Основные сферы научных исследований автора*: применение нечеткого классификатора NEFCLASS к задаче распознавания рукописных математических выражений.*

Online Handwritten Mathematical Expressions Recognition System Using Fussy Neural Network

Edris Naderan

Abstract: *The paper is devoted to the development of the new online handwritten mathematical expressions recognition system. The paper presents the recognition method to the handwritten symbols using fussy neural network NEFCLASS as a means for classification*.

Keywords: *online handwriting recognition, mathematical expressions, fuzzy logic, fuzzy neural networks, structural analysis, artificial neural network, genetic algorithm, conjugate gradient algorithm.*

THE INTELLIGENT DECISION SUPPORT SYSTEM FOR DIAGNOSTIC OF DIFFICULT DISEASES OF VISION

Aleksandr Eremeev, Ruslan Khaziev, Irina Tcapenko, Marina Zueva

***Abstract**: The work is devoted to methods and software tools of designing intelligent decision support systems (IDSS), which helps professionals (decision making persons) helping to diagnose complex problem situations on the example of complicated pathologies of view. Unlike traditional Bayesian belief networks, the proposed application of advanced multilevel (difficult-structured) networks, more convenient for complex research of the problem and providing expert data. Integration of Bayesian belief networks and Dempster-Shafer method allows using at diagnostics both expert data, and numerical (probabilistic) data obtained in the result of measurements. The proposed approach is implemented in the prototype of the intelligent decision support system for diagnostics of difficult diseases of vision.*

***Keywords**: intelligent system, decision support, diagnostics, problem situation, Bayesian belief network, Demster-Shafer method.*

***ACM Classification Keywords**: H.4.2 [Information systems applications]: Types of systems – Decision support; I.2.3 [Artificial intelligence]: Deduction and Theorem Proving – Uncertainty, "fuzzy," and probabilistic reasoning; I.2.4 [Artificial intelligence]: Knowledge Representation Formalisms and Methods – Bayesian belief network.*

Introduction

At the Appiled Mathematics Department of National Research University “Moscow Power Engineering Institute, for more than twenty years researches on the development of mathematical methods and software for intelligent decision support systems (IDSS) intended for the help to experts (DMP – decision making person) in diagnostics and monitoring of complex problematic situations of different types [Vagin et al., 2001; Eremeev et al., 2009] have been actively conducted. Together with the Laboratory of Clinical Physiology of Vision of Moscow Helmholtz Research Institute of Eye Diseases the studies on creation of IDSS for diagnostics of complicated pathologies of vision have been carried out [Eremeev et al., 2013].

The joint use of the apparatus of Bayesian belief networks (BBN) [Bidyuk, et al., 2005] and the Dempster-Shafer method (DSM) [Lyuger, 2003], oriented to help the DMP in diagnostics of complex problematic situations is examined in this study. Unlike the classic BBN-system, a layered architecture, which combines methods and allows us to explore the problem in complex, is used. Split in tiers allows also reducing the scope of the search of different situations to find the most probable outcome. In the system, not just the reasoning of experts but also the analysis of numerical data in the various diagnostics with the use of probabilistic methods is used.

It is known [Eremeev, et al., 2009] that in applying the BBN in IDSS, especially in IDSS of real time (IDSS RT), is not enough to define the main components of the event in the network, and their relationships. One also needs to take into account all possible situations that the expert can foresee. It is recommended to use the hard-structured BBN. This type of BBN has been adapted and applied in the prototype of IDSS for diagnostics of complex eye

diseases on the example of retinal pathologies when it is impossible to say with absolute certainty what kind of disease a patient has and at what stage.

Even after detection of symptoms, with the help of special computer diagnostic tool and conclusions of experienced professionals, patients had to re-take examination due to inaccuracies and incomplete of available information. When creation the model and experimental approbation there was found that the use of the BBN allows us to analyze only one of the possible outcomes of (the existence of one of the diseases), ignoring the other options. Joint application of BBN and DSM would allow also assessing relationships between the possible outcomes (situations) and specify their probabilities.

The based architecture of the intelligent decision support system

IDSS (including IDSS RT) is based on the integration of knowledge representation and knowledge operation models that are capable to adaptation, modification, and learning. Such models are oriented to specific problem areas and respective uncertainty types, what reflects the ability to develop and modify their states.

The generalized structure of an IDSS (IDSS RT) is shown in Fig. 1.

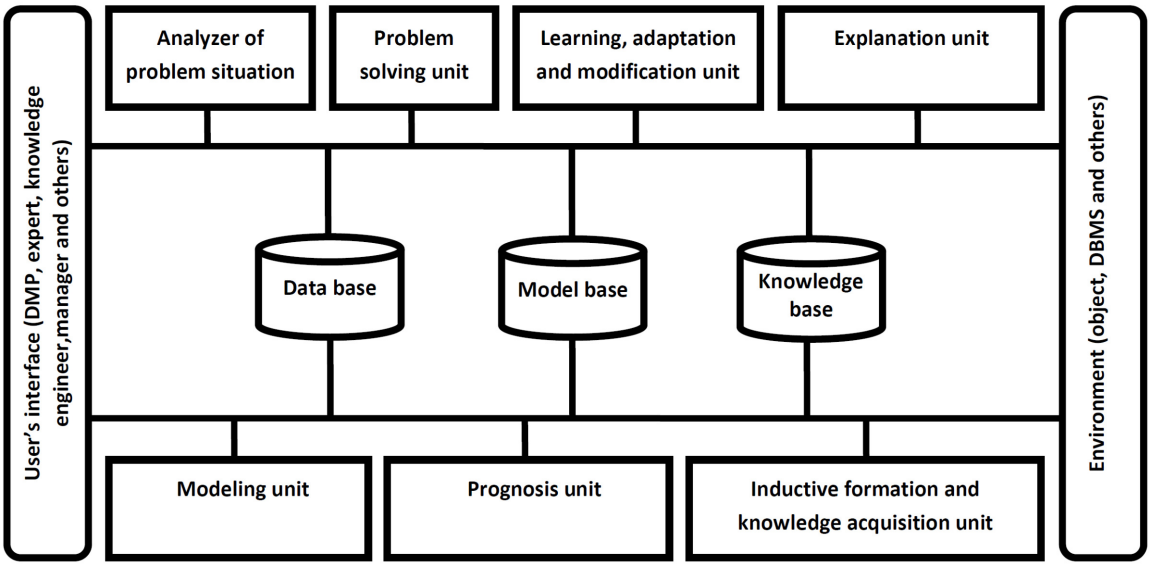


Figure 1. *General architecture of an IDSS*

By realizing methods of reasoning modeling in an IDSS one should take into consideration the features of these systems:

- The necessity to take a decision under time constrains defined by an actually controlled process;
- The need to consider a time factor in the description of a problem situation in a process of finding the solution;
- The impossibility of obtaining all the objective information necessary for decision making and in this connection the usage of subjective expert information;
- The multivariate character of search;

– The necessity of applying methods of plausible reasoning and the active participation of DMPs in decision making;
– The presence of incomplete, fuzzy and even inconsistent data for the description of situations.

Diagnostics of the pathology of vision

Glaucoma had chosen as an object of research. The main symptom of this disease is the increase in intraocular pressure, and decreased vision. A violation of outflow and circulation of intraocular fluid lays in the basis of the disease [Neroev, et al., 2010]. Statistics data and patients' parameters were obtained in Moscow Helmholtz Research Institute of Eye Diseases. The created IDSS prototype should help DMP (doctor-ophthalmologist) in identifying early signs of glaucoma and clarifying the type of the disease in order the earliest adequate treatment.

Let us consider the main stages of research in the diagnosis of glaucoma. First of all, an ophthalmologist must carry out a complex of clinical examinations in the patient. At this stage, a specific set of psychophysical tests are applied for detection of pathologies. If there is reliable evidence that the patient has a disease, then further investigation is not reasonable; otherwise, the examination may continue due to the lack of available information. In particular, further examination is often necessary in the early stages of the disease or in difficult differentiable cases, e.g. low pressure glaucoma, when the data of clinical examination of the patient can almost do not differ from the normal state of a healthy person.

Then one should do the complex of electrophysiological tests, which analyze the electrical responses of the retina in different light stimuli. It uses various types of electroretinogram (ERG) in computer graphics, representing changes of retina electric potentials that occur in response to light flashes, and the amplitude and peak latency of the ERG waves (components) are estimated. Based on the type and degree of ERG change, the existing pathology is detected. At the final stage the expert-specialist (physiologist) or group of experts, on the basis of the obtained resulting data of the patient examination must confirm or deny the diagnosis of ophthalmologist (in the latter case it is necessary to clarify the diagnosis), after that the appropriate treatment can be recommended.

In Fig. 2 BBN is presented, which characterizes the process of glaucoma diagnosis in the form of multiple-defined events: T – results of ophthalmological examinations (suspicion for this disease); Z - is the conclusion of physiologist about the existence of the disease based on the ERG parameters and clinical data of ophthalmologist; I - is the result of ERG testing, D – is the resulting event, which characterizes the solution of the problem (the exact diagnosis, the current stage, recommendations for treatment or further examination etc.). Figures are events related to the stages of patient research.

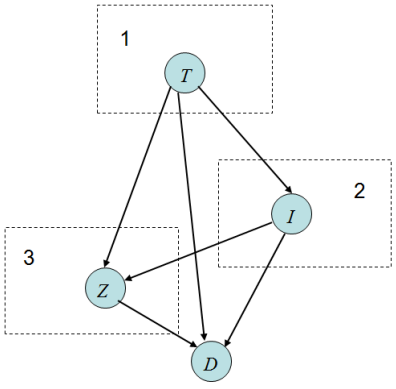


Figure 2. BBN-model for glaucoma diagnosis

Consider the event determining the result of the ERG study on the example of primary open angle glaucoma (POAG). For diagnosis of this disease (condition) and its stage it is necessary to analyze several kinds of ERG, to make conclusions about the detected signs (symptoms) of the disease, and to determine the degree of diagnostic confidence (probability).

In Fig. 3 the model of the second level is presented for definition the value of the event "The ERG testing results", where *I*- is the target event (diagnosis), I_K,..., I_{RT} - the results of studies; I_K - the result obtained in the study of cone ERG, I_M – maximal ERG, I_P – rod-ERG, I_{OP} – oscillatory potentials (OP), I_F – photopic negative response (PhNR), I_{PT} – pattern ERG (PERG).

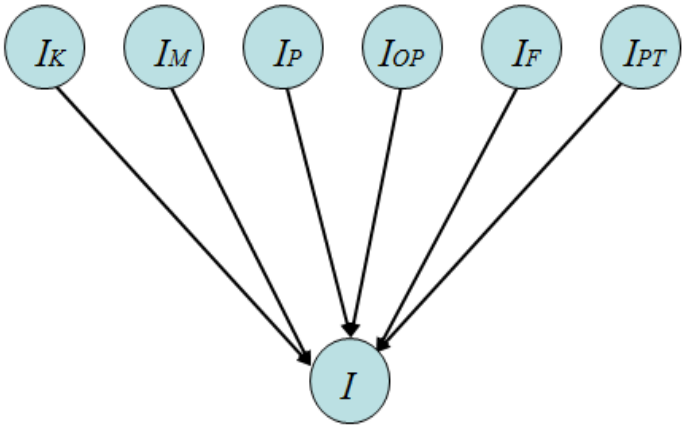


Figure 3. BBN-model for definition the value of the event "The ERG testing results"

In the model of the second level, the data of main types of ERG are presented, each of which has its own waveform, and varies depending on the type of the disease, i.e. the key values that characterize a particular disease are identifying. Rod, cone and maximal ERG are used to test the ability of the eye to the dark and light adaptation. Each of the ERG components is generated by different structures in the retina. In Fig. 4 the classic analysis of ERG components is represented (by Ragnar Granit) [Neroev, et al., 2010].

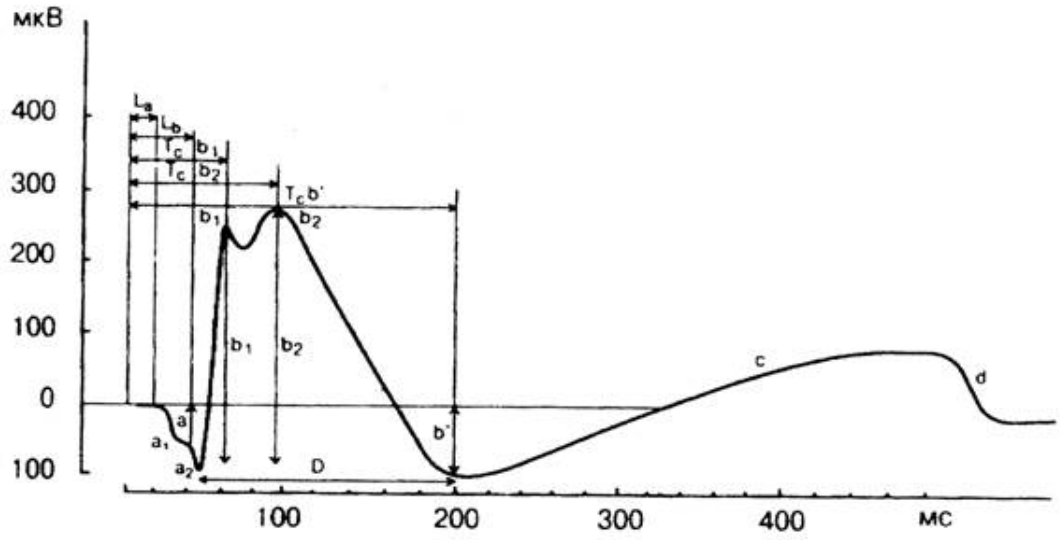


Figure 4. The Granit component analysis of ERG

Each of the ERG components is generated by different structures of the retina. The main indicators are the a- and b-waves. In the course of diagnostics, values of amplitude and peak latency of the components of various types of ERG are analyzing. In the Table 1 the statistical data of ERG amplitudes in the examined patients with POAG stage I and II are presented.

Table 1. Amplitudes of the *a- и b*-waves in the I and II stage of POAG (M±m)

Kinds of standard ERG		**Norm, uV M±2SD**	**POAG I stage**		**POAG II stage**	
			μV	**% of Norm**	**μV**	**% of Norm**
Rod ERG	*b*-wave	58,1±21,7	44,3±5,9	76,2%	35,1±27,2	60,4%
Maximal ERG	*a*-wave	103,7±37,3	87,8±7,9	84,6%	70,3±30,4	67,8%
	b-wave	203,2±55,0	176,5±13,3	86,8%	196,6±46,1	96,7%
Cone ERG	*a*-wave	18,2±7,5	16,7±2,7	91,7%	15,1±6,1	82,5%
	b-wave	91,2±21,0	49,0±6,0	53,7%	49,6±21,8	54,4%

According to the data of the amplitudes it is possible to analyze the changes in value of the signal, passed through the retina. In the presence of disease, there is the reduction of the a- and b-waves. In addition, one can change their culmination time (peak latency): it usually is delayed in the advanced stages of diseases, while latency remains unchanged in the initial stages. The progression of the disease contributes to a significant reduction in the amplitude and to a prolongation in the peak time of the ERG waves, resulting to the reduction of visual functions.

Consider the model (BBN) for the POAG diagnosis on the basis of the analysis of maximal ERG (Fig. 5), where *Aa, Ab* - the maximum amplitude of a- and b-waves; *Ta, Tb* – their peak time; *Va, Vb* – the conclusion on the basis of input data, defining the disease; I_M - target event that corresponds to the common result of analysis of the ERG (diagnosis).

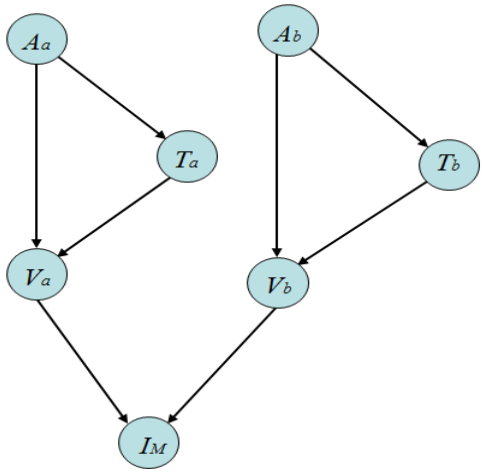


Figure 5. BBN-model for the diagnosis on the base of maximal ERG

The proposed model takes into account all the key values of standard ERG for POAG. At this step of POAG diagnosis, only possible stage of POAG are the I and II stages, therefore, we can note a number of peculiarities: A_a, A_b - can take on different meanings depending on the type and stage of the disease in some range relative to the average values of the amplitudes listed in the table. 1, for example, for the POAG stage I, each event has three states: “more typical for the POAG stage I”, “typical for the POAG stage I”, “less typical for the POAG stage I”. The zones of states can intersect, i.e. it is necessary to analyze all of the states, if the values of maximal amplitudes of a- and b-waves are in the area of intersection; T_a, T_b - change only when the POAG of II stage, therefore, it is possible to allocate two states: “more than normal for a healthy person”, “normal for a healthy person”; V_a, V_b - depending on the key values it is possible to ascertain the most appropriate state (diagnosis): “The POAG stage I”, “The POAG stage II” or “Healthy”; I_M - this is the event, which result in the analysis of all conclusions according to maximal ERG, has states similar to V_a, V_b.

Consider a patient with the suspected POAG stage I. In Tables 2 and 3 the key values of the amplitude and peak time of the a- and b-waves in the analysis of maximal ERG are set, as well as relevant event states and the coefficients of confidence.

Table 2. Amplitudes of the *a*- и *b*-waves in the analysis of maximal ERG

Parameters	**A_a – *a*-wave amplitude**	**A_b – *b*-wave amplitude**
μV	85,1	177,1
% of Norm	82 %	87,1 %
The most appropriate event status	Typical for the POAG stage I Less typical for the POAG stage I	Typical for the POAG stage I
The coefficient of confidence, P(x), x=A_a, A_b	0,8 (typical for the POAG stage I); 0,75 (less typical for the POAG stage I)	0,8

Table 3. Peak latency (the culmination time) of the *a*- и *b*-waves in the analysis of maximal ERG

Parameters	**T_a – the *a*-wave peak latency**	**T_b – the *b*-wave peak latency**
μV	24	46
% of Norm	135 %	124,1 %
The most appropriate event status	exceeds the normal value	exceeds the normal value
The coefficient of confidence, P(x), x=T_a, T_b	0,65 (A_a= “typical for the POAG stage I”); 0,8 (A_a= “less typical for the POAG stage I”)	0,65

The value of a-wave amplitude corresponds to the two states of events – “typical for the POAG stage I” and “less typical for the POAG stage I”, one need to calculate the probability of execution of both events. The most likely value of events states V_a and V_b is “The POAG stage I”. Set the probability of these events: $P(V_a/A_a,T_a)=0,68$

when A_a = a "typical for the POAG stage I'; $P(V_a/A_a,T_a)=0,62$ when A_a = a "less typical for the POAG stage I'; $P(V_b/A_b,T_b)=0,68$.

The final event for data analysis of maximal ERG is the event of I_M, which is the most probable consequence of the patient disease or a denial in suspicion during the clinical examination. In this case it is necessary to confirm or deny that the patient has the POAG stage I. The coefficient of confidence in this situation is: $P(I_M/V_a,V_b)= 0,95.$ Calculate the joint probability of the event, provided that A_a = a "typical for the POAG stage I"

$$P(A_a,T_a,A_b,T_b,V_a,V_b,I_M)=P(A_a)\cdot P(T_a)\cdot P(A_b)\cdot P(T_b)\cdot P(V_a/A_a,T_a)\cdot P(V_b/A_b,T_b)\cdot P(I_M/V_a,V_b)\approx 0,119$$

In order to ensure that the result of the analysis of maximal ERG is the presence of the disease one need to determine the highest value of the joint probability of events. When calculate the joint probability of the event, provided that A_a = a "less typical for the POAG stage I", we'll get $P(A_a,T_a,A_b,T_b,V_a,V_b,I_M)\approx 0,125$

Thus, the result (diagnosis) A_a = a "less typical for the POAG stage I" has the highest value of the joint probability of events. In order to ensure that the result of the analysis of maximal ERG is that the patient has the POAG stage I, it is necessary to calculate the corresponding conditional probability (when the given events), which determine the value of the key points on the ERG.

The calculation is made according to the formula

$$P(I_M \mid A_a,T_a,A_b,T_b) = P(I_M,A_a,T_a,A_b,T_b) / P(A_a,T_a,A_b,T_b) = (P(I_M,,A_a,T_a,A_b,T_b,V_a,V_b,) + P(I_M,A_a,T_a,A_b,T_b,V_a,\neg V_b,) + P(I_M,A_a,T_a,A_b,T_b,\neg V_a,V_b) + P(I_M,A_a,T_a,A_b,T_b,\neg V_a,\neg V_b)) / ((P(A_a,T_a,A_b,T_b,V_a,V_b,I_M) + P(A_a,T_a,A_b,T_b,V_a,\neg V_b,I_M) + P(A_a,T_a,A_b,T_b,\neg V_a,V_b,I_M) + P(A_a,T_a,A_b,T_b,\neg V_a,\neg V_b,I_M) + P(A_a,T_a,A_b,T_b,V_a,V_b,\neg I_M) + P(A_a,T_a,A_b,T_b,V_a,\neg V_b,\neg I_M) + P(A_a,T_a,A_b,T_b,\neg V_a,V_b,\neg I_M) + P(A_a,T_a,A_b,T_b,\neg V_a,\neg V_b,\neg I_M)).$$

After appropriate calculations we obtain the result 0.67 (67 %), meaning that in the analysis of maximal ERG of the patient the most likely diagnosis is the presence of the POAG stage I, but this information is not sufficient to confirm (or refute) the conformity of a given diagnosis to the reality, because, as it was already noted, the use of BBN allows us to analyze just one of many possible diagnoses, in particular, the presence or absence of the POAG stage I. To address this shortcoming, i.e. to conduct the comprehensive analysis with accounting and other possible diagnoses, one can offer in addition to BBN also use the DSM, which allows to compare all the available values, as a set of elements that characterize the disease.

For example, the value of V_a can have three possible values (states). Depending on the entered values of the amplitude and peak time of the ERG waves, one can set the probability of facilities for all possible states of the event V_a. Let each of V_a states be considered as the opinion of one expert, who argues that when entered data his conclusion is the most probable. Proceeding from this, one can determine intervals of the belief and plausibility (likelihood) for each of the diseases. In Table 4 the corresponding values of measure, coefficients of belie and plausibility, used for calculations are shown.

Selected values in Table 4 can be interpreted as the lower and upper boundaries of the intervals at which the values of the corresponding probabilities are contains. According to the obtained data the values of the probabilities for BBN can be then corrected. The combined use of the BBN and DSM allows improving the effectiveness of diagnostics when increase of disease databases and methods of their analysis.

Table 4. The values of the coefficients of belief and plausibility for the event V_a

Events	Measure	Belie	Plausibility
a value NULL	0	0	0
Healthy	0,05	**0,1**	**0,15**
POAG I stage	0,55	**0,62**	**0,65**
POAG II stage	0,3	**0,35**	**0,4**
One of the (healthy, POAG I stage, POAG II stage)	0,1	1,0	1,0

The prototype of the intelligent decision support system for glaucoma diagnostics

For realization of the prototype of diagnostic IDSS, promising artificial intelligent language Clips was selected [Giarratano, et al., 2007]. Clinical trials of patients, data of various kinds of ERG and the conclusion of ophthalmologists and physiologists are used as the input information. The output information is the diagnosis with corresponding probability (confidence) and recommendation for treatments or for further examinations of the patient. According to the data obtained in the analysis of different kinds of ERG, the database (evidence base) is constructed that stores the original description of the task (information about the patient) in the form of verbal descriptions. To build BBN and to operate with database of facts, the knowledge base (rules) is used, which contains the methods for obtaining results.

Below there is a fragment of the program in language Clips with the sample of rule of adding a new element - the maximum amplitude of a wave when it gets into the interval for the POAG 1 stage:

```
(defglobal ?*Vvod-Aa-m* = 0)
(defrule opredelenie-Aa-max-ERG-2 ""
(declare (salience 98))
(not (varAa ?))
(AaPOUG 1 st.)
=>
(printoutt "Select the interval at which gets the value of the maximum a-wave amplitude:" crlf)
(printoutt "Response options" crlf)
(printoutt "[1] more typical of the disease POAG I stage" crlf)
(printoutt "[2] typical of the disease POAG I stage" crlf)
(printoutt "[3] less typical for disease POAG I stage" crlf)
(bind ?*Vvod-Aa-m* (ask-question " Response: " 1 2 3))
(if (eq ?*Vvod-Aa-m* 1)
then
  (assert (varAa more for POUG 1 st.))
(printoutt "Aa = more typical of the disease POAG I stage" crlf)
else
  (if (eq ?*Vvod-Aa-m* 2)
then
```

```
   (assert (varAa norm for POUG 1 st.))
(printoutt "Aa = typical of the disease POAG I stage" crlf)
else
   (if (eq ?*Vvod-Aa-m* 3)
then
     (assert (varAa less for POUG 1 st.))
(printoutt "Aa = less typical for disease POAG I stage" crlf))))
(assert (labelAaSingularlabel
Plurallabels)))
```

After we set in the system many facts to describe events A_a, A_b, T_a, T_b based on the coefficients of confidence, the search of rules is performed for the determination of status of events V_a and V_b. In the search process one can be a situation when several rules are followed that form the certain set of facts under the same conditions. This is because the extrema of the a- and b-waves can correspond to several states. In this case the highest value of the conditional probability of an event I_m is used.

In Fig. 6 an example of the dialog box for DMP is shown when specifying the patient data at the maximal ERG. Note that this example demonstrates only use one of the methods of diagnosis (BBN), and does not show the final result.

Conclusion

In conclusion we note that the proposed formal apparatus and implementing it software tools are included in the basic tools of constructing of modern IDSS, including IDSS RT, on the basis of integrated methods and models of knowledge representation and decision making in the conditions of different types of uncertainty (incomplete, incorrect, unclear, contradictory) in available information (data and knowledge).

Acknowledgement

The work was supported by RFBR, projects No 14-01-00427.

Bibliography

[Bidyuk P.I., et al., 2005] Bidyuk P.I., Terentev A.N., Gasanov A.S. Building and methods of learning Bayesian networks. - Cybernetics and System Analysis, 2005;(4):133–147 (in Russian).

[Eremeev et al., 2013] Eremeev A.P.,Khaziev R.R., Tsapenko I.V., Zueva M.V. Prototype of the diagnostic decision support system on the basis of integration Bayesian belief networks and the Demster-Shafer method. Software & Systems. Tver, 2013;1(101):11-16. ISSN 0236-235X (in Russian).

[Eremeev, et al., 2009] Eremeev A.P., Vagin V.N. Method and tools for modeling reasoning in diagnostic systems. - ICEIS 2009. Proc. of the 11th International Conference on Enterprise Information Systems. Vol. AIDSS. Milan, Italy, May 6-10, 2009. INSTICC, 2009. pp. 271-276.

[Giarratano, et al., 2007] Giarratano J., Riley G., Expert Systems: principles and programming, Moscow, Giarratano J., Riley G. Expert Systems: principles and programming. Fourth Edition. Moscow: OOO «I.D. Vilyams». 2007. 1152p (in Russian).

[Neroev, et al., 2010] Neroev V.V., Zueva M.V., Tsapenko I.V., Ryabina M.V., Kiseleva O.A., Kalamkarov G.R. Ischemic aspects in pathogenesis of retinal diseases. Rossiyskiy oftal'mologicheskiy zhurnal [Russian Ophthalmologic Journal]. 2010;3(1):42–49 (in Russian).

[Vagin, et al., 2001] Vagin, V.N., Eremeev, A.P. Some Basic Principles of Design of Intelligent Systems for Supporting Real-Time Decision Making. Journal of Computer and Systems Sciences International. 2001;40(6):953-961.

[Zueva, et al., 1992] Zueva M.V., Tsapenko I.V. Electrophysiological characteristics of glia-neuronal interaction in retinal pathology. Sensornye sistemy [Sensory Systems]. 1992;3:58-63 (in Russian).

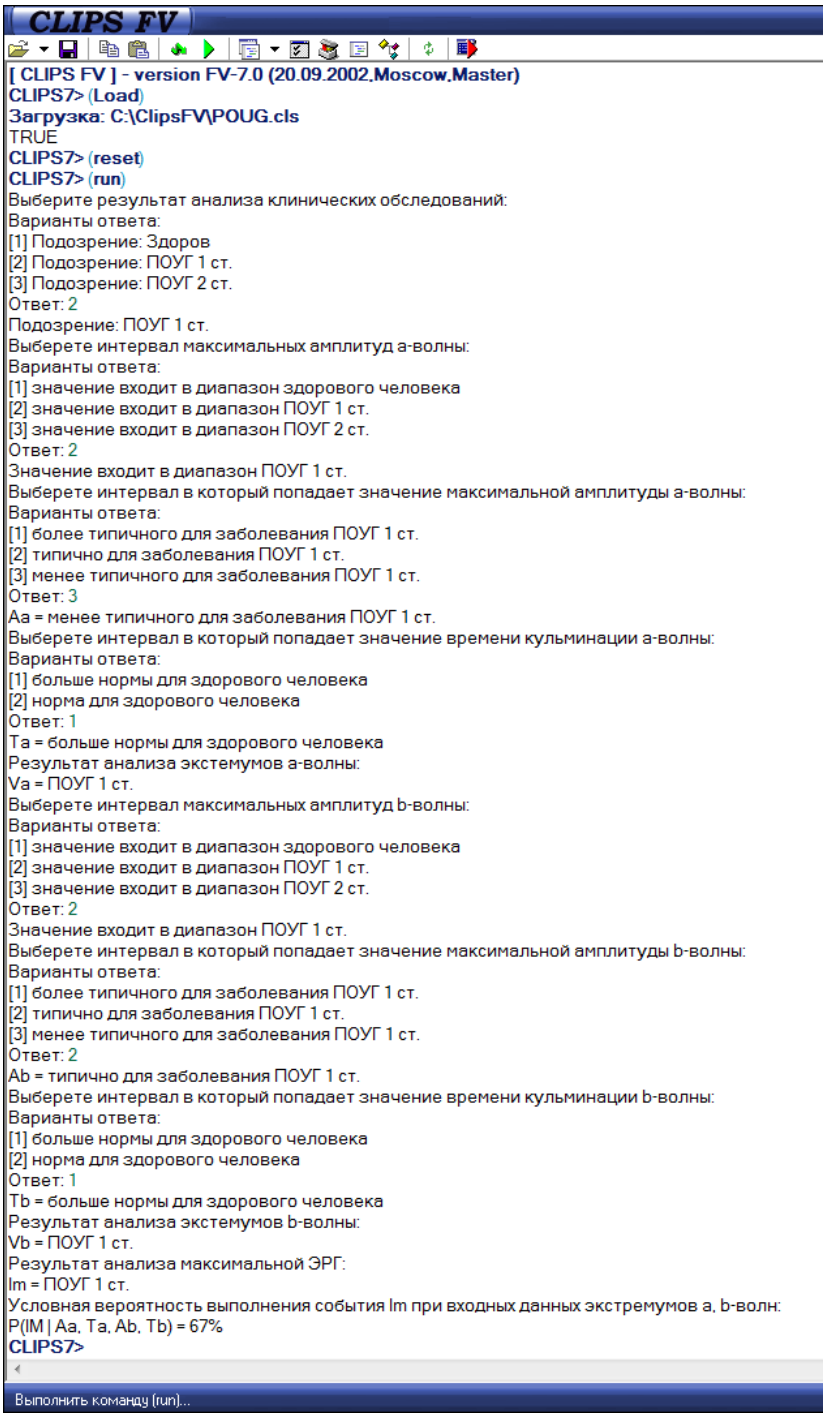


Figure 6. Example of a dialog box for DMP

Authors' Information

Eremeev Aleksandr Pavlovich *– Ph.D., Professor, Head of the Applied Mathematics Department of National Research University "Moscow Power Engineering Institute", 14, Krasnokazarmennaya Str., Moscow, 111250, Russia, Moscow, e-mail: eremeev@appmat.ru*

Area of scientific interests: artificial intelligence, decision making, decision support system, expert system

Khasiev Ruslan Robertovich *– postgraduate student of the Applied Mathematics Department of National Research University "Moscow Power Engineering Institute", 14, Krasnokazarmennaya Str., Moscow, 111250, Russia, Moscow, e-mail: ruslan.haziev@gmail.com*

Area of scientific interests: artificial intelligence, decision support system, Bayesian belief network

Tcapenko Irina Vladimirovna *– Ph.D., senior researcher of the laboratory of Clinical Physiology of Vision of Moscow Helmholtz Research Institute of Eye Diseases, 14/19, Sadovaya-Chernogriazskaya Str.,Moscow, 105062, Russia, e-mail: sunvision@mail.ru*

Area of scientific interests: artificial intelligence, decision making, visual physiology, retinal diseases

Zueva Marina Vladimirovna *– Ph.D, Dr. Biol. Sci, Professor of Pathophysiology, Head of the laboratory of Clinical Physiology of Vision of Moscow Helmholtz Research Institute of Eye Diseases, 14/19, Sadovaya-Chernogriazskaya Str.,Moscow, 105062, Russia, e-mail: visionlab@yandex.ru*

Area of scientific interests: artificial intelligence, decision making, visual physiology, retinal diseases, nonlinear physiological processes

NOVEL METHOD FOR ANALYSIS OF FINGERPRINT POROSCOPICAL MAPS

David Asatryan, Grigor Sazhumyan, Bagrat Sakanyan

***Abstract**: During last decades grows the interest of investigators to the methods of using the third level features for increasing identification accuracy in automated fingerprint recognition systems. It was noticed that the poroscopical maps carry important information about fingerprint pores, namely, their number, sizes, coordinates etc. However, at creation of automated identification technique, there arises a problem in connection with impossibility of using directly the pixel-by-pixel comparison methods of corresponding images. In this paper, we propose technique for comparison of poroscopical maps, based on using investigated earlier structural proximity assessment measure, which was determined using the gradient field of images. The poroscopical map is determined using some known algorithms of image processing, namely binarization, segmentation etc. The results of poroscopical maps processing and comparative analysis of some items from the database of Hong Kong Polytechnical Institute are given. It is shown that the proposed technique for poroscopical maps processing can be used in AFRS for increasing the accuracy of fingerprint identification.*

***Keywords**: Fingerprint, third level, pores, poroscopical map, gradient field, image proximity assessment.*

***ACM Classification Keywords**: Image Processing and Computer Vision*

Introduction

Fingerprint recognition is widely popular but is a complex pattern recognition problem. The information contained in a fingerprint can be categorized into three different levels, namely, Level 1 (pattern), Level 2 (minutia points), and Level 3 (pores and ridge contours) [Jain, 2007]. Most existing automated fingerprint recognition systems (AFRS) utilizes only level one and level two fingerprint features for personal identification [Jain, 2003]. Level-three fingerprint features like pores, are also very distinctive, though they are seldom used [Stosz, 1994]. During last decades more and more researchers are exploring how to extract and use level-three features in AFRS.

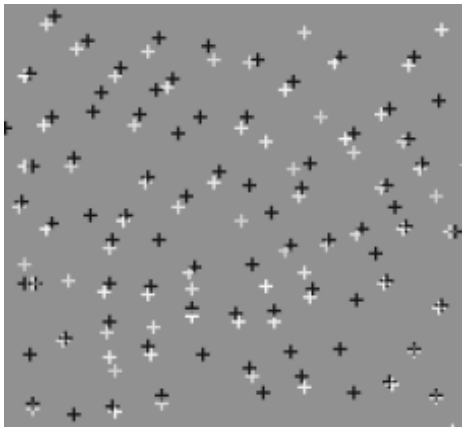


Figure 1. Overlapped positions of two fingerprint pores

Many methods have been proposed for fingerprint pore detection, extraction and matching. Large class of algorithms for pore extracting directly from a Gray scale image has been proposed in [Jain, 2007]. A review of pore extraction methods are considered in [Zhao, 2010]. Usually a pore is considered as extracted when the pore center coordinates becomes known. Then the coordinates of two fingerprint pores are compared by overlapping corresponding maps of pores positions as it is shown in Figure 1 (a fragment from the map of [Parsons, 2008]). Then the proximity of corresponding coordinates is estimated by using visual analysis.

The next important problem is poroscopy, which means study of the configuration, size, and relative positions of extracted pores. Poroscopy also involves a comparative study of the pores visible in impressions left at the site of a crime and in the fingerprints of an identified person.

The comparative study usually is performed by visual comparison of the positions of the pores on the poroscopical maps which are overlie in an appropriate manner. Using measures based on the mean-squared error doesn't solve the problem due to uncertainty of the coordinates of centers of pores. This fact is especially important in AFRS. Unlike to such measures the human visual system (HVS) can solve the poroscopical maps comparing problem. In several papers it is noted that HVS extracts substantive and/or structural information from an image (see, for example, [Bovik, 2002]). Therefore in many cases HVS can have certain advantages vs. formal methods. During last decades there proposed some formal methods in the scientific literature which give the results comparable with HVS results [Bovik, 2002; Bovik, 2004].

An approach to image quality assessment based on structural properties, is investigated in [Asatryan, 2009]. A measure using the information of two image gradient magnitude distributions is offered, which gives the results similar to the HVS perception. Some applications of the mentioned measure show its low sensitivity to image scaling or/and rotations [Asatryan, 2010]. This paper is devoted to testing the ability of the specified measure for poroscopical investigations.

In Section 2 the closed pore extraction method and a technique for poroscopical map creating is described which is based on segmentation procedure using also Otsu method for image binarization. Section 3 devoted to application of method [Asatryan, 2009] to the poroscopical maps comparison problem. Section 4 includes some experimental results obtained by proposed technique.

Proposed Technique for Fingerprint Pore Extraction

We use the technique for closed pore extraction, which is described in [Asatryan, 2012]. A closed pore assumed to be as a segment, i.e. it is a set of pixels which satisfy the following requirements:

- It consists of connected pixels, i.e. every pixel from the set has neighbors only from the same set;
- The intensity of the pixels from the set and the intensity of the pixels from outside of that set belong to different value intervals.

The extracting algorithm consists of following steps.

Step 1. The Gray Scale image is binarized by Otsu method [Otsu, 1979]. This method allows good enough binarization of fingerprint and is applied in many papers on image processing. As a result of binarization we have new image with pixels of intensity "0" or "255".

Step 2. The binarized image is inverted. This means that the black pixels of binarized image turned into white and vice-versa. This operation is performed to have the extracted pores of black color in the white background (it is more convenient for printing processes).

Step 3. The binarized and inverted image is fully segmented. As a result we get K segments with pixels of two intensities. Thus the closed pore will look as a black segment.

Step 4. Let n_k be the number of pixels of k-th segment, $k = 1,2,...,K$. Let $(T_{\min}, T_{\max})$ be the interval of acceptable sizes for closed pores. This interval must be determined by prior consideration of fingerprint scanner resolution and known sizes of closed pores from special investigations. For example according to [Zhang, 2010] $T_{\min} = 3, T_{\max} = 30$ for scanner with resolution of 1200 dpi. More detailed analysis of fingerprint scanner resolution requirements and corresponding pores sizes is given in [Busselaar, 2010].

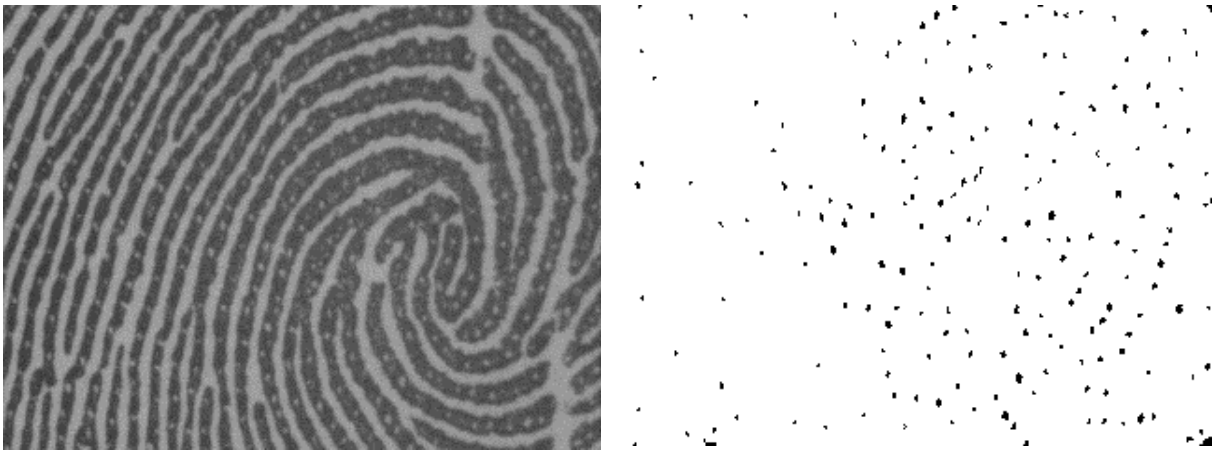


Figure 2. Fingerprint 5-1-1 and its poroscopical map ($T_{\min} = 3, T_{\max} = 30$)

Step 5. Extracting pores, i.e. all segments with $T_{\min} \le n_k \le T_{\max}$, $k = 1,2,...,K$. Creating the poroscopical map by locating the images of pores on the map of fingerprint sizes. Figure 2 shows an example of fingerprint and corresponding poroscopical map.

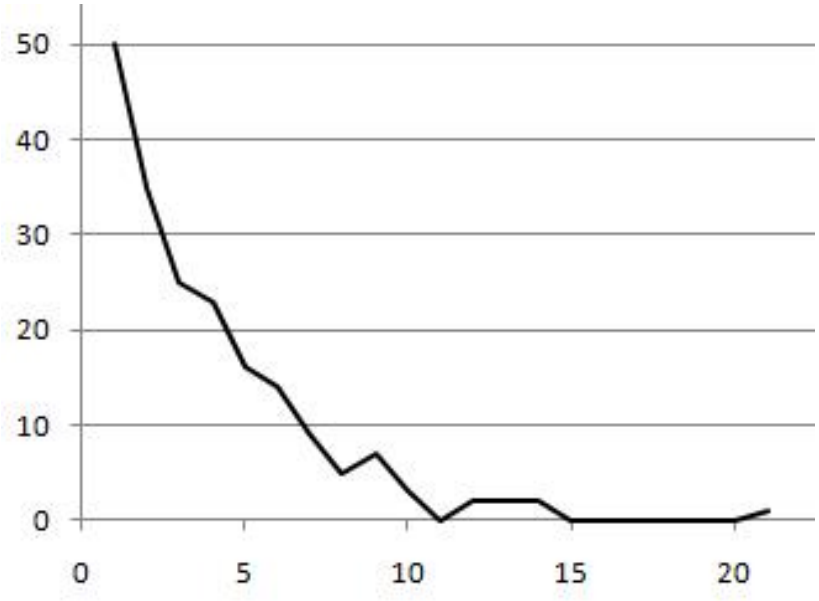


Figure 3. Frequency polygon of sizes of extracted pores

We check the size of each pore, i.e. the number of pixels it contains. The size of chosen pores varies within the pre-specified range of valid pore size (from 3 to 30 pixels in our experiments). General number of extracted pores is 194, average size is $\bar{x} = 5.85$, which corresponds to pore diameter of about 60 micron. If $T_{\min} = 7$, the average pore diameter increases up to 74 micron. The frequency polygon of sizes of 194 extracted pores is shown in Figure 3.

Comparison of Poroscopical Maps of Two Fingerprints

The method originates in a new approach to the problem of image quality assessment, described in [Asatryan, 2009]. According to that approach the set of gradient magnitudes is considered as a feature of an image. A measure for quality assessment is created to use the information contained in two image gradient magnitude distribution. Such measure gives the results similar to the HVS.

Let's describe briefly above mentioned quality assessment measure. The related algorithm consists of the following steps.

Step 1. Calculate the gradient magnitudes $\|\Delta_j(m,n)\|$ for both considered images, $j = 1,2$, $m = 0,1,...,M_j$, $n = 0,1,...,N_j - 1$.

Step 2. Assume that the gradient magnitudes $\|\Delta_j(m,n)\|$ are two-dimensional independent random variables with Weibull distributions $F_1(x;\eta_1,\sigma_1)$ and $F_2(x;\eta_2,\sigma_2)$. Parameters $\eta_1,\sigma_1,\eta_2,\sigma_2$ are estimated by the gradient magnitude samples of two images.

Step 3. Calculate the proximity of the images by formula

$$W^2 = \frac{\min(\eta_1,\eta_2)\min(\sigma_1,\sigma_2)}{\max(\eta_1,\eta_2)\max(\sigma_1,\sigma_2)},\ 0 < W^2 \le 1. \tag{1}$$

The measure (1) is invariant to sizes and rotations [Asatryan, 2010]. It has tested for certain images and showed the results more corresponding to HVS, than usual mean-square measure.

The main purpose of this paper is to demonstrate a method of comparison of two poroscopical maps to estimate its proximity.

Experimental Results

Fingerprints with pores in the series of experiments were chosen from High-Resolution-Fingerprint (HRF) Database of The Hong Kong Polytechnic University (PolyU [Hong Kong]). The fingerprints in this database were captured by resolution of 1200 dpi.

We have performed two series of experiments. In the first series we have chosen five transformed issues of the same fingerprint. The poroscopical maps of these fingerprints are determined by method described in Section 2 and shown in Table 3. The pixel-by-pixel comparison of these maps does not give any reason to identify the

corresponding fingerprints, while the proximity measure W^2, described in Section 3, shows high proximity between the mentioned items (see Table 1).

Table 1. Values of proximity measure W^2 between different maps for transformed items of the same fingerprint.

	2-1-1	2-1-2	2-1-3	2-1-4	2-1.5
2-1-1	1	0.859	0.81	0.876	0.874
2-1-2		1	0.942	0.753	0.984
2-1-3			1	0.71	0.927
2-1-4				1	0.766
2-1-5					1

Table 2. Values of proximity measure W^2 between maps of different fingerprints.

	2-1-1	4-1-2	69-1-5	54-2-5	58-2-2
2-1-1	1	0.078	0.392	0.026	0.179
4-1-2		1	0.199	0.334	0.435
69-1-5			1	0.067	0.457
54-2-5				1	0.146
58-2-2					1

Then, different fingerprints from the same database are collected in Table 4. The proximity measure between poroscopical maps of these fingerprints are given in Table 2. The values of proximity measure W^2 are significantly less than presented in Table 1. Though in Tables 1 and 2 there are presented only a few samples from the specified database, the considered results show that the poroscopical maps can be processed by proposed technique and being used in AFRS will increase the recognition accuracy.

Conclusion

In this paper, a technique for poroscopical maps creating and analyzing is proposed. It is assumed that a closed pore is a segment in the fingerprint image which can be extracted and simplified after binarization and inversion of the fingerprint. Binarization threshold can be estimated by Otsu method. The size of the segments are preliminary determined with a glance of fingerprint scanner resolution that must be enough high. Thus all the extracted segments are depicted in the map which is considered as a poroscopical map. The proximity of two chosen map is estimated by technique described in Section 2. The results obtained by experiments show that the information containing in the poroscopical map can be used for AFRS investigations being processed with fingerprint features of Levels 2 and 3 and have a resource for increasing the accuracy of fingerprint recognition.

Table 3. Transformed items of the same fingerprint and corresponding poroscopical maps

$$T_{\min} = 3, T_{\max} = 30$$

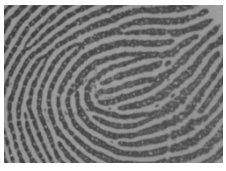
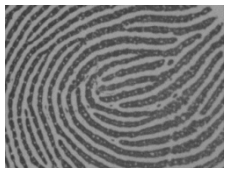
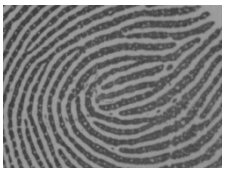
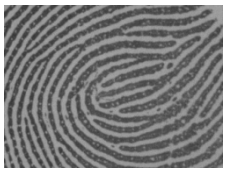
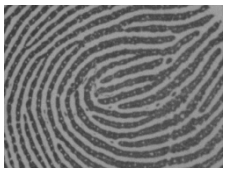
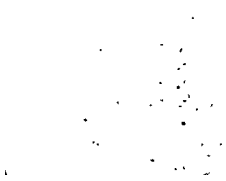
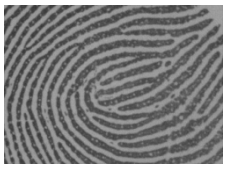
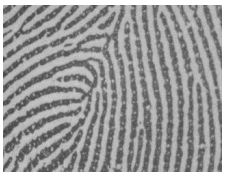
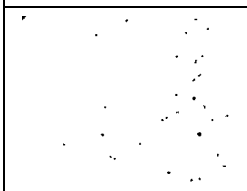
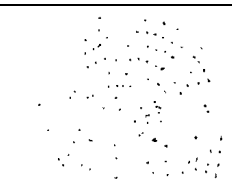
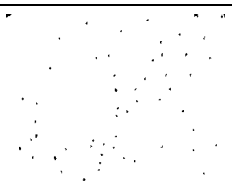
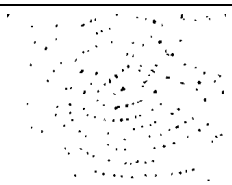
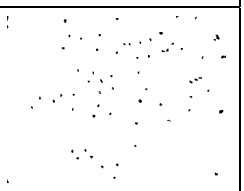
2-1-1	2-1-2	2-1-3	2-1-4	2-1-5
				
				

Table 4. Different fingerprints and corresponding poroscopical maps ($T_{\min} = 3, T_{\max} = 30$)

2-1-1	4-1-2	69-1-5	54-2-5	58-2-2
				
				

Bibliography

[Asatryan, 2009] D. Asatryan, K. Egiazarian. Quality Assessment Measure Based on Image Structural Properties. Proc. of the International Workshop on Local and Non-Local Approximation in Image Processing, Finland, Helsinki, pp. 70-73, 2009

[Asatryan, 2010] D. Asatryan, K. Egiazarian, V. Kurkchiyan. Orientation Estimation with Applications to Image Analysis and Registration. International Journal “Information Theories and Applications”, Vol. 17, Number 4, pp. 303-311, 2010

[Asatryan, 2012] D. Asatryan, G. Sazhumyan. Segmentation Based Fingerprint Pore Extraction Algorithm. International Journal «Information models and Analysis», Vol. 1, pp. 134-138, 2012

[Bovik, 2002] Z. Wang, A.C. Bovik. A universal image quality index. IEEE Signal Processing Letters, vol. 9, no. 3, pp. 81-84, 2002.

[Bovik, 2004] Z. Wang, A.C. Bovik, H.R. Seikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004

[Busselaar, 2010] E.J. Busselaar. Improved pores detection in fingerprints by applying ring led’s (525 nm). Optica Applicata, Vol. XL, No. 4, pp. 843-861, 2010.

[Hong Kong] http://www4.comp.polyu.edu.hk/~biometrics/HRF/HRF.htm

[Jain, 2003] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar. Handbook of Fingerprint Recognition. Springer, New York, 2003

[Jain, 2007] A. K. Jain, Y. Chen, M. Demirkus. Pores and Ridges: High-Resolution Fingerprint Matching Using Level 3 Features. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 15-27, 2007

[Otsu, 1979] N. Otsu (1979). A threshold selection method from gray-level histograms. IEEE Trans. Syst. Manage. Cybern. (SMC) 9: 62; pp. 377–393.

[Parsons, 2008] N. R. Parsons, J.Q. Smith, E. Thönnes, L. Wang, and R.G. Wilson. Rotationally Invariant Statistics for Examining the Evidence from the Pores in Fingerprints (March 2008). Law, Probability & Risk, Vol. 7, Issue 1, pp. 1-14, 2008

[Stosz, 1994] J. Stosz and L. Alyea. Automated system for fingerprint authentication using pores and ridge structure. Proc. SPIE Conference on Automatic Systems for the Identification and Inspection of Humans, 2277:210–223, 1994.

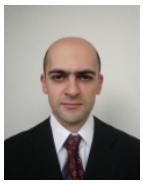
[Zhang, 2010] Q. Zhao, D. Zhang, L. Zhang, and N. Luo. High Resolution Partial Fingerprint Alignment Using Pore-Valley Descriptors. Pattern Recognition, vol. 43(3), pp. 1050-1061, 2010

[Zhao, 2010] Q. Zhao, D. Zhang, L. Zhang, N. Luo. Adaptive fingerprint pore modeling and extraction. Pattern Recognition, 43, pp. 2833–2844, 2010

Authors' Information

David Asatryan *– Professor, Doctor of Sciences (Engineering), Head of Research Center for Critical Technologies of Russian-Armenian (Slavonic) University, Head of group of the Institute for Informatics and Automation Problems of NAS Armenia, e-mail: dasat@ipia.sci.am .*

Major Fields of Scientific Research: Digital signal and image processing



Grigor Sazhumyan *– Candidate of Technical Sciences, Software Engineer, Institute for Informatics and Automation Problems of NAS Armenia, e-mail: grigorsazhumyan@gmail.com .*

Major Fields of Scientific Research: Digital signal and image processing, Software developing.

Bagrat Sakanyan *– PhD student, Russian-Armenian (Slavonic) University, e-mail:* mr.org@mail.ru.

Major Fields of Scientific Research: Digital signal and image processing

JOINT STUDY OF VISUAL PERCEPTION MECHANISM AND COMPUTER VISION SYSTEMS THAT USE COARSE-TO-FINE APPROACH FOR DATA PROCESSING

Anton Sharypanov, Alexandra Antoniouk, Vladimir Kalmykov

***Abstract**: Aspects of visual perception mechanism and pattern recognition methods are examined jointly. Latest results from neurophysiology in studying the visual system of living beings are discussed. Another view on coarse-to-fine approach for technical vision tasks is presented. On the basis of systemological analysis of neurophysiology sources a new hypothesis about visual neuron's functioning is proposed. This hypothesis explains the mechanism and takes into account receptive fields excitatory zones resizing during visual act.*

***Keywords**: coarse-to-fine, neurons of visual system, intercellular processes, pattern recognition, image processing, variable resolution*

***ACM Classification Keywords**: I.4.1 Digitization and Image Capture, I.5.1 Models*

Introduction

State of the art in technical vision, various approaches to pattern recognition, image processing and object detection tasks shows that researchers frequently meet the problem of great computational complexity, particularly – the problem of great time and machine resource consuming. One of the strategies that solves these problems is so-called "coarse-to-fine approach", i.e. the technique of refining initial data that exclude inappropriate objects or irrelevant ranges of the image on earlier stages of processing in order to apply computationally intensive part of the algorithm to reduced volume of data. One of the variants of this approach is to use a coarse-to-fine modification of known algorithm, suitable for solving the particular class of problems.

For example, in [1] proposed a multi-resolution part based model and a corresponding coarse-to-fine inference algorithm which is extremely efficient. The method is based on the observation that matching of each part of the image is the most expensive computational operation in comparison to detection of significant parts and computation of their optimal configuration, so the minimization of number of part-to-image comparisons implies detection acceleration. Starting from matching the lowest resolution part the method selects only the best placement in each image neighborhood. These locally optimal placements are then propagated recursively to the parts at higher resolution. By recursive elimination of unlikely part placements from the search space, the set of possible locations is narrowed so that the computation of only few part-to-image comparisons is performed. This method gives a ten-fold speed-up over the standard dynamic programming approach.

The algorithm in [2] follows the same idea of discarding of large regions from the hypothesis space at an early stage of the recognition however the object detector for each resolution is not the same. Each detector uses inexpensive tests on the image features and narrows down the set of possible variants to a smaller number, which are explored in greater detail by the next detector. Also each detector computes a quantity for each region of hypothesis space; this region is accepted for consideration at the next level of resolution if the quantity exceeds a given threshold. All thresholds are set automatically based on probabilistic measurements.

Coarse-to-fine strategy application to vehicle motion trajectories clustering is discussed in [3]. Raw trajectories are clustered into “coarse clusters”. Each “coarse cluster” consists of trajectories having very similar directions, but with different positional characteristics. The feature for further fine clustering is trajectory resampling point set, and the Euclidean distance is used as the distance measure between two trajectories.

For face recognition, a coarse-to-fine procedure can be implemented by sequential application of different face recognition methods in order to reduce the candidate examples on each step. In [4] the decision process is developed through consecutive stages, such as “one-against-all (OAA) of SVM (support vector machine)”, “one-against-one (OAO) of SVM”, “Eigenface”, and “RANSAC”. The stage 1 “OAA of SVM” and stage 2 “OAO of SVM” uses the discrete cosine transform features extracted from the entire face image. On stage 3 “Eigenface” face images are projected onto a feature space (face space). The face space is defined by the “Eigenfaces”, which are the eigenvectors of the set of faces and based on intensity information of the face image. “RANSAC” is applied in the last stage, in which the epipolar geometry method with space information of the testing image is matched with the two training images, and then the image with the greatest match numbers of and the shortest distance to corresponding feature points is selected.

The task of establishing the correspondence between pixels in two images (finding a markup) with human faces, addressed in [5], is effectively solved by building “cascades” of markups. The resolution is decreased two times in both initial images per cascade and new markup for them is built. After that the starting approximation for initial markup is defined based on the new markup and the field of motion is searched but with less quantity of markings. The algorithm that solve the task utilizing one “cascade” runs eight times faster while preserving accuracy in finding the field of motion for two images.

Examples of dynamic programming (DP) application to recognition problems are numerous and include speech recognition, character recognition, deformable template matching, soft decoding and road tracking. However such problems often lead to enormous state spaces and the computations can be infeasible, even with DP. To overcome this obstacle, in [6] proposed a variation on DP - coarse-to-fine dynamic programming (CFDP). The essential idea of this algorithm is to form a series of coarse approximations to the original DP trellis by aggregating trellis states into “super states”. For each coarse approximation, the optimal path is found using DP with “optimistic” arc costs between the super states. The super states along this optimal path are refined and the process is iterated until a demonstrably globally optimal path is found. In many cases this global optimum is achieved with considerably less computational expenditure than straight DP. This CFDP algorithm is particularly well-suited to DP problems with large state spaces. According to [7], the speed of CFDP depends on the structure of the grouping and the nature of the problem. In the best case, CFDP gives a large computational savings over standard DP; in the worst case, it will actually be slower.

The goal of coarse-to-fine approach in mentioned cases is to reduce intensive processing only to some regions of starting image or to some parts of initial dataset containing the information that seems to be useful and no mater what the particular coarse-to-fine mechanism is used.

But many of the image recognition problems that result in NP-complete tasks or even can not be addressed with traditional methods are solved by human visual system in less than no moment. It definitely looks like one can benefit from studying the processes taking place there to reach the level of performance comparable to that of the visual system. In previous decades researchers have already tried to examine some aspects of image processing in conjunction with contemporary results in neurophysiology.

Investigations on changes in spatial sizes of direction-selective cells in the primate visual system [8] has inspired Battiti et al. to study how integrating motion information across different spatial scales could help improving the estimate of the optical flow. An adaptive multiscale method, where the discretization scale is chosen locally according to an estimate of the relative error in the velocity estimation, based on image properties was proposed in [9]. This coarse-to-fine method provided substantially better estimates of optical flow than did conventional algorithms, while adding little computational cost.

It will be shown in the following paragraphs that studying the processes taking place before V1 visual cortex seems to be useful for developing new efficient methods in technical vision systems construction.

On mechanism of visual perception

Objects of concern are retinal ganglion cells and LGN neurons found in visual system (visual neurons) and their receptive fields. It was discovered [10, 11] that the sizes of receptive fields' excitatory zones change during the visual act, which eventually mean dynamical changes in visual system's resolution. It is known that the receptive field of a visual neuron consists of many receptors that send signals to it through one or more synapses [12]. Receptive field is circular in simplest case. The functioning of such visual neuron is studied for its action potentials in response to stimulus – a circle on image, contrasting relative to the background, which is projected onto the retina at a given time interval.

Human visual system operates in a sequence of visual acts that last for ~150ms each. After that period a saccade (oculomotorius muscles' twitching) happens, the image on retina shifts and the next visual act begins.

The fact that the size of receptive fields' excitatory zones does not stay stationary during visual act was established in the course of research work in [10]. Excitatory zone diameter is the minimum diameter of stimulus at which the maximum number of neuron responses is achieved. In order to get quantitative description of changes in receptive field's excitatory zone the time slices method was proposed. This method is based on the assumption that if the diameter of excitatory zone changes during visual act, then the maximum number of spikes from this neuron matches the time interval when diameter of excitatory zone meets the stimulus diameter or their diameter values varies insignificantly.

The time slices method for obtaining the diameter of excitatory zone as a function of time is used as follows. Post-stimulus histograms (PSH) that represent the responses of neuron on circular stimuli of different sizes were divided into a series of sequential temporal intervals (7.5 or 15ms each) and the number of spikes in each interval was determined (Figure 1 [10]). The time slices method also allows determining the serial number of temporal interval with maximal neuron responses. The bigger the size of stimulus, the earlier maximal response is achieved, so one can conclude that the diameter of round receptive field excitatory zone of a neuron changes in the course of visual act, namely it shrinks from maximal to minimal, up to 1-2 receptors in the case of ganglion cell. Thus there exists maximal resolution for visual system defined as the number of receptors in the field of view center and the variable resolution that changes during visual act. Variable resolution is determined by the size of neuron's receptive field excitatory zone.

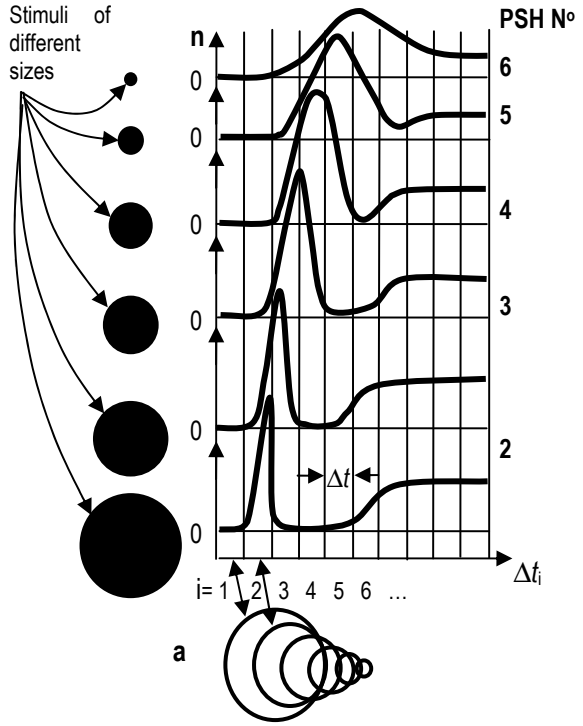


Figure 1 [10]. Neuron responses on stimuli of 6 different sizes
n axes represents the number of spikes in corresponding time slice Δt_i.
Maximum number of spikes corresponds to interval where the size of stimulus meets the excitatory zone size.
a – decreasing of receptive field's excitatory zone area during visual act.

Temporal response pattern and dynamics of receptive field structure in single lateral geniculate nucleus (LGN) neuron of a cat using static spot stimuli flashed on the receptive field for 400–500 ms were studied in [11]. Spatial receptive field parameters from spatial summation curves, determined for successive 5-ms intervals throughout the stimulus period, were estimated (Figure 2 [11]). Thereby it was possible to study dynamics of the response properties during periods similar to those in natural fixations, and with a method that did not presuppose a linear system. The results showed pronounced changes in the receptive field structure during the spot presentation.

Initially, the neurons had wide receptive field centers. The center rapidly shrank to a minimum that occurred on average ~70 ms after stimulus onset whereupon the center widened slightly [11]. Thus the maximum spatial resolution occurred in a brief time window after onset of stimulation. In parallel, the center-surround antagonism increased. The changes in spatial resolution did not follow the changes of firing rate. The initial strong burst of action potentials appeared earlier than the maximal spatial resolution. The authors state that these results are consistent with the hypotheses that the firing pattern of the neurons during brief static stimulation initially mediates a strong but spatially coarse message to cortex that gradually changes into a weaker, but spatially more precise message.

For all the nonlagged cells, there was a pronounced change in the selectivity for spot size during the spot-on period. At the beginning of the period, the neurons responded well to a broad range of spot sizes, but subsequently, the response was restricted to a gradually narrower range of the smaller spots.

The quantitative estimates of the receptive field center width revealed pronounced changes during the stimulus presentation in all the nonlagged cells, in particular during the first 150 ms after spot onset. The receptive field center was initially wide, but rapidly shrank to a minimum. In the majority of neurons, the center thereafter widened again such that the minimal size occurred only briefly.

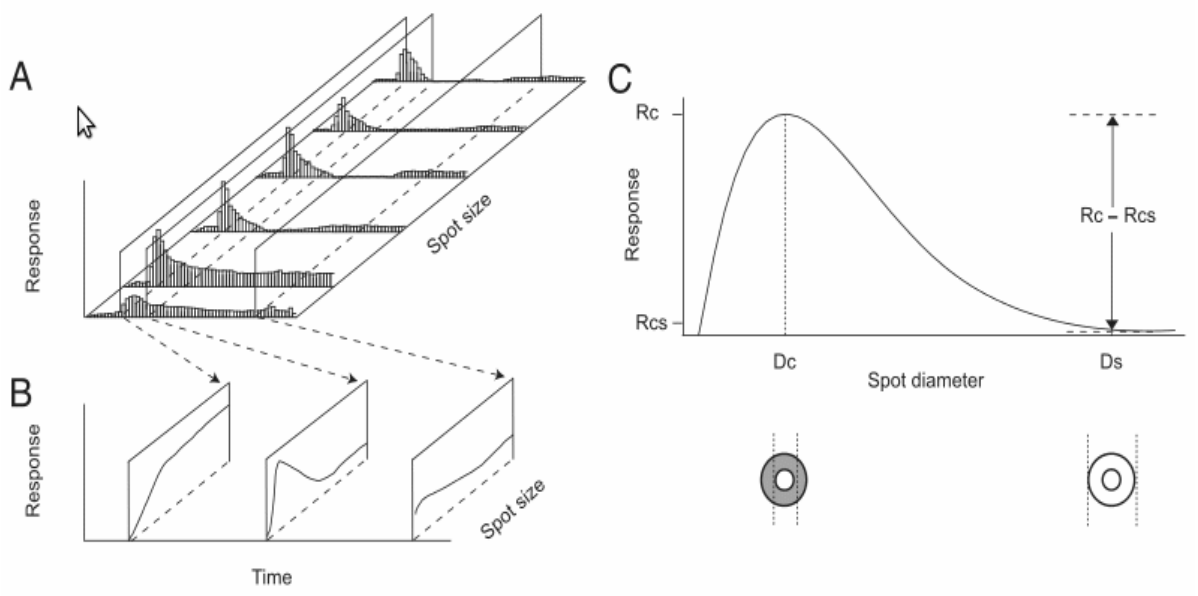


Figure 2 [11]. Schematic illustration of method for studies of dynamics of receptive field properties

A: peristimulus time histogram (PSTH) with 5-ms bin width was determined for response to each of a series circular spot stimuli presented on receptive field for 400-500 ms. Spots differed in width from smaller than receptive field center to larger than whole receptive field. PSTHs for different spot sizes were used (shown by 6 schematic PSTHs). Time slices through whole set of histograms were made for each 5-ms bin, and set of response vs. spot width values in each time slice was used to make a spot width tuning curve for respective time after stimulus onset , shown by 3 schematic tuning curves in B. C: in each tuning curve, width of spot that gave maximum response (R_c) was taken as estimate of width of receptive field center (D_c), width of smallest spot that gave minimum response (R_{cs}) was taken as estimate of width of receptive field surround (D_s), and reduction of response from maximum to minimum divided by response maximum was taken as estimate of center-surround antagonism at respective time.

Contrary to that the lagged cells showed no clear dynamic changes in the spatial structure of the receptive field. When the response appeared, the neuron already had a small receptive field center [11].

Also the hypothesis that a small stimulus spot will initially activate many neurons, most of them only transiently was checked (Figure 3 [11]). The conditions when the stimulus falls outside of the receptive fields centers during their shrinkage were modeled. For all three positions, there was an initial transient response, but the subsequent sustained response that occurs to a small spot in the minimum receptive field center was lacking in these cases.

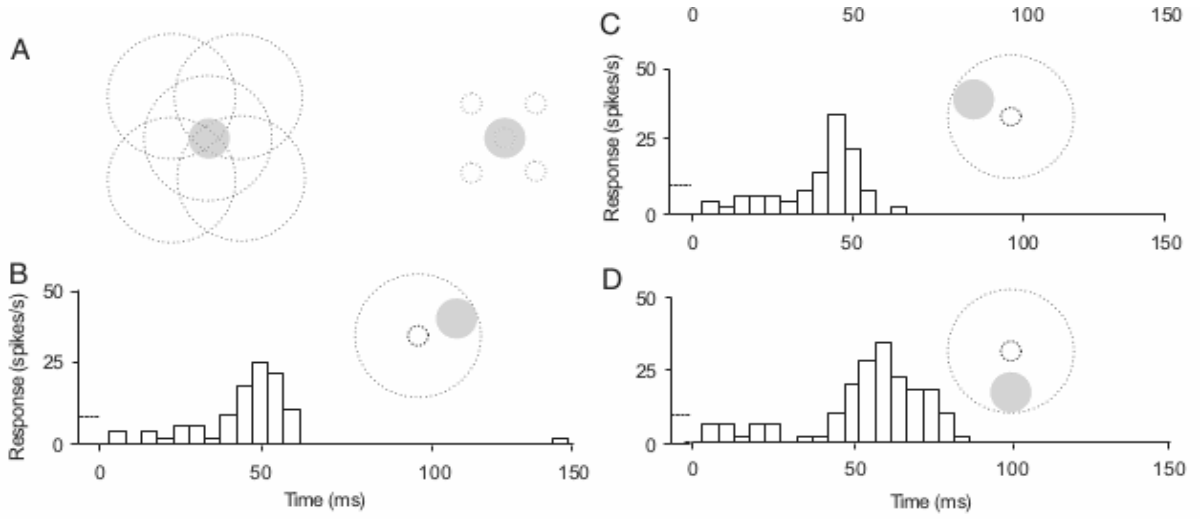


Figure 3 [11]. Initially wide RFCs suggest that number of activated neurons decrease during stimulation

A: schematic illustration of hypothesis. B–D: eccentric stimulus spots presented outside minimum field center (small center circle) but inside maximum field center (large circle) gave a fast initial response that disappeared as center shrank toward minimum. Only responses over the 1st 150 ms of a 500-ms stimulation period are plotted. Dotted line before stimulus onset shows level of spontaneous activity. Nonlagged cell. Bin width in histograms, 5 ms. Number of spot presentations at each location (interleaved) was 100.

The conducted experiments with combined S-potential and action potential recordings showed that the initial fast shrinkage of the receptive field center was present already in the retinal input to LGN neurons. The degree of shrinkage was similar for the retinal input and the LGN neuron, and apart from the faster shrinkage in the LGN neurons, the temporal pattern of the shrinkage was also similar.

However the authors mention that mechanisms for generation of center width dynamics in retinal ganglion neurons are unclear, so further studying of intracellular processes is needed.

The fact that an optimal visual stimulus flashed on receptive field center in retinal ganglion cells typically evokes a strong transient response followed by weaker sustained firing and so happens in thalamocortical (TC) neurons of LGN in a state-dependent manner is well-known and described in literature, but the mechanisms by which transient firing changes to sustained are less well known. One of the results obtained in [13] shows that constant frequency pulse train stimulation of retinal afferents causes depolarization through temporal summation of excitatory post-synaptic potentials (EPSPs) in TC neurons (Figure 4 [13]) which occurs no matter what the holding potential of the cell membrane was set but the value of the holding potential influenced the ability of TC neurons to generate spikes in later part of the train.

In this study the holding potential of a cell was adjusted to different steady state values by direct current injection through recording electrode at the beginning of each experiment and didn't change during particular measurement. The authors suggest that regulation of the sustained response through the level of the membrane potential is a key mechanism for regulation of the strength of input to cortex depending on states like arousal, attention etc.

Another aspect of coarse-to-fine approach in computer vision

Researches in the field of visual perception physiology and creation of information technologies for automatic processing of visual information (technical or computer vision in other words) are fairly interconnected domains of human activity. Indeed, the subject in both disciplines is the study of visual perception. For physiology of vision the subject is the visual perception of humans and animals, while one of the subjects for computer sciences is creation and testing of technical vision means. The progress in one of these domains would initiate the progress in the other.

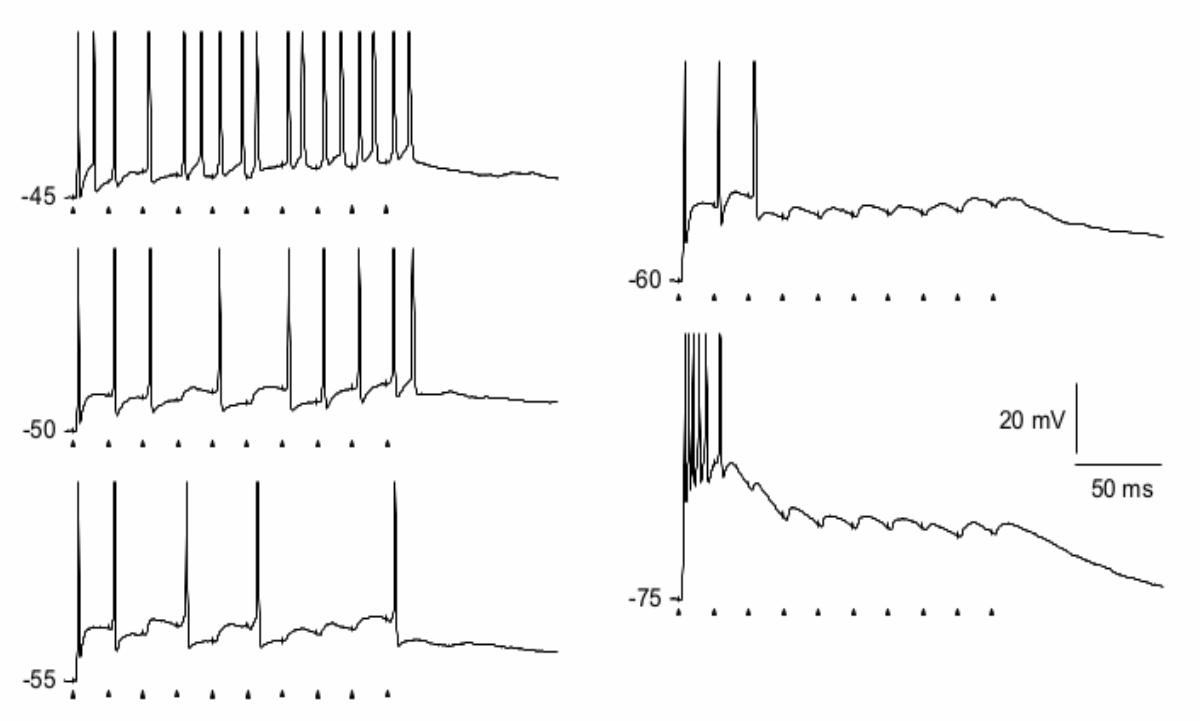


Figure 4 [13]. Firing pattern of a TC neuron at different holding potentials evoked by pulse train stimulation of retinal afferents

The frequency of pulse train was 50 Hz. Holding potentials are indicated to the left of the trace. Timing of stimulus pulses is indicated by arrow-heads below the traces, and can also be deduced from the stimulation artifacts on the traces (truncated). Spike amplitudes were truncated at 0 mV. At -75mV, as expected, the pattern was dominated by a short-latency low-threshold calcium potential, and the elicited action potentials lacked precise timing with respect to the single pulses in the stimulus train.

Unfortunately mutual results' interchange between these disciplines doesn't happen, maybe due to weak interaction among specialists representing different sciences and different methods that are typically used by them.

In the field of image recognition and image processing it wasn't paid much attention to prototyping computer vision from neurophysiological aspects of visual perception (see, for example [14, 15]) or simplified models were considered [16].

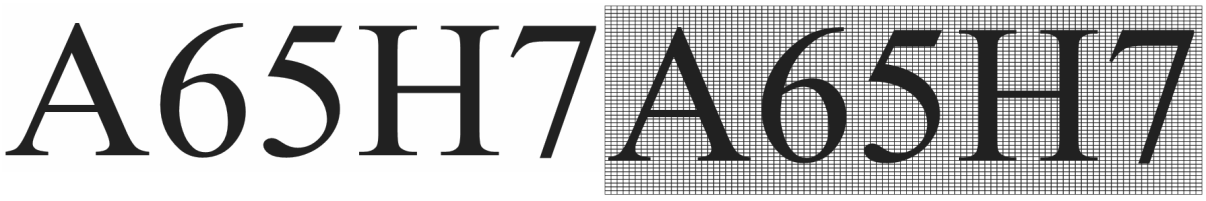
It is used to suggest [15] that initial image for processing is presented in analogous form. As a rule the image is bounded by rectangle – field of view – with dimensions that are suitable for processing. The initial image should always fit this rectangle and fill it if possible.

The first operation one do with image being processing is field of view discretization and brightness quantization for colors in image palette. The following parameters of computer vision system are chosen according to practical considerations: resolution – the number of discrete, usually squared, elements of image – pixels – that fit in one measurement unit (inch, centimeter); the size of image in pixels; the set of values of brightness that each pixel can take value from for each of basic colors (grayscale, RGB, etc.).

It is suggested that the resolution of computer vision system is best suited for specified class of images being processed. This will not lead to unnecessary details being arising with extremely high resolution (like object's contour distortion) and to essential details of the objects being disappearing when the resolution is extremely low. The form of quantization function and the number of brightness levels for colors should conform to images being processing in terms of ability to display the essential details of the object in exactly the same way.

Thus the image can be presented as a two-dimensional array $V(N, M)$ having width N and height M. Each element $v(n, m)$ of this array corresponds to either the brightness of pixel with coordinates n, m for grayscale image or to brightness values of basic colors (e.g. red, green, blue) for color image. This array can be treated as a matrix or a vector depending of what the mathematical methods are chosen.

A lot of methods and algorithms for dealing with image which is presented as a matrix or a vector are developed and applied successfully. Most of them are successfully used for grayscale image processing if discretization and quantization are chosen adequately for that class of images. At the same time the following should be mentioned. For example if the image is processed by means of statistical recognition methods, the brightness values of all pixels in the image are used for determining the measure of similarity of two images $f(V_1, V_2)$. But the pixels in the image usually belong to either object or background. In this case the result of recognition not only depends on pixels brightness of the object but also from brightness values of background pixels that in major cases is not acceptable. Consider an image consisting of a line of arbitrary text on a one-color background and another one where the same line of text is placed over an arbitrary grating (Figure 5). The text on (Figure 5a) can be processed by both statistical and structural methods of recognition.



a) Neutral background; b) Square grating in background

Figure 5. Examples of image with arbitrary text

The text on (Figure 5b) is a far more complex task for recognition. If statistical methods for similarity computing with reference image are applied, the result will be distorted due to pixels belonging to background, but representing an arbitrary placed grating. It is not guaranteed that mutual placement of grating lines and text in the field of view of the image will be the same for arbitrary imposition and subsequent digitization. If attempting to use

structural methods of recognition for images like (Figure 5b) it will not be possible to get the contours of the objects. In this case the contours of grating cells will be selected instead of object contours.

At present time the attention of specialists is drawn to tasks dealing with recognition of textures. Grating being placed over text on image is a particular case of them. But methods for texture recognition generally far beyond the complexity of text recognition, face recognition etc. That fact hampers the application of such methods for mentioned tasks significantly or even eliminates the possibility to use them. Similar situation with some differences happens when the grating being placed over the text is of same color as background (Figure 6). For statistical methods computed similarity with reference image will be distorted due to pixels belonging to objects but representing the arbitrary placed grating. The mutual placement of text and grating, again, may be different from image to image after digitization. Attempting to apply the structural methods to images on Figure 6 will give the same result as for image on Figure 5b – the contours of grating cells.

At the same time visual perception cope with similar tasks insensibly, seemingly on subconscious level.

The method of time slices for examining the resizing of receptive field's excitatory zone (Figure 1 [10], Figure 2 [11]) shows that the time of getting the maximum number of spikes on the neuron's output counting from the beginning of visual act depends on the size of stimulus. It is possible to suggest that making a decision versus image in visual system happens when the maximum number of spikes on the neuron's output is reached, in other words – decision can be made for images with different resolution. It is naturally to conjecture that the best resolution for decision making is selected in visual system in the sense of image processing, when the unnecessary details don't arise and the essential parts of objects don't disappear.

a) b)

Figure 6. Grating of the same color as background being placed over the text. Images on (a) and (b) have different width of grating lines

It is possible to suggest that it is the processing of observed low resolution images at the beginning of visual act that makes possible consistent visual perception of symbols on different background. Therefore if some optical character recognition (OCR) program would successfully process the image on Figure 5a digitized with low resolution, then it should also successfully process the images from Figure 5b, Figure 6a and Figure 6b digitized with the same value of resolution. A simple experiment with a well-known OCR program FineReader can be conducted in order to check this statement.

The initial dataset consists of four images (Figures 5a, b; 6a, b) 900x280 pixels each, having resolution of 72x72 pixels per inch. The meaningless combination of symbols "A65H7" was chosen intentionally, in order to eliminate the influence of dictionaries on the result of recognition.

The text on Figure 5a was recognized successfully. Processing of the rest three images gave denial of recognition because of inability to find (determine) an object on image. On Figure 7 the same four images as Figures 5 and 6 but digitized with six times lower resolution are presented. Then the text string was recognized successfully on all of them.

Figure 7. Images having 6 times lower resolution:

a) and b) are from Figure 5 (a, b); and c) and d) are from Figure 6 (a, b)

This experiment clearly shows that the concept of optimal resolution in the sense of processing results may be applied not only to whole image but to each object on it. In this case coarse-to-fine approach was used for different purpose: not to decrease the number of calculation-intensive operations but to solve the problem that can not be addressed at all by traditional methods.

Hypothesis about visual neuron's functioning at the time of action potential generation

Proposed model of visual neuron's functioning at the time of action potential generation is developed on the basis of systemological analysis of known ideas [17] about neurons' functioning and results of other researchers, presented above. It is an attempt to explain the mechanism and to take into account receptive fields excitatory zones resizing during visual act.

It is known that maintenance of stable resting potential of cell membrane is achieved due to chemical and electrical gradients equilibrium of potassium and chlorine ions. Being changed randomly, the value of membrane potential is restored in the course of potassium and chlorine ions' transfer through membrane channels. To be exact, chlorine ions transport inward the cell and potassium ions transport outward the cell during membrane depolarization. Considering the neuron as a system, the transport of chlorine ions inward the cell should take place not only in dormancy state but especially during depolarization at the time of action potential generation. Total restoration of chlorine ions' chemical gradient presumably happens at the end of visual act.

Thus the key concept in our approach is to take into consideration the influence of cell membrane chlorine conductance at the time of action potential generation in the context of whole visual act.

Under the influence of excitatory receptors post-synaptic potential and also with no impact of inhibitory receptors the potential-dependent sodium channels start to open. Sodium flow increases and depolarization increases too, resulting in opening more sodium channels, so the intracellular membrane potential increases up to the value of sodium equilibrium potential.

While sodium flow increments and depolarization increases, chlorine ions enter cell through membrane pores due to concentration gradient while counteractive electrical potential decreases. Furthermore, a point is reached

(perhaps the point of maximum potential value) when potential-dependent chlorine channels that also let in chlorine ions cannot but open. Reaching depolarization maximum starts the drift of sodium and potassium ions outward the cell. That returns the cell membrane potential to its initial value but due to chlorine ions ingress the resting potential shifts for some value toward hyperpolarization (Figure 8) at the time of each following pulse generation.

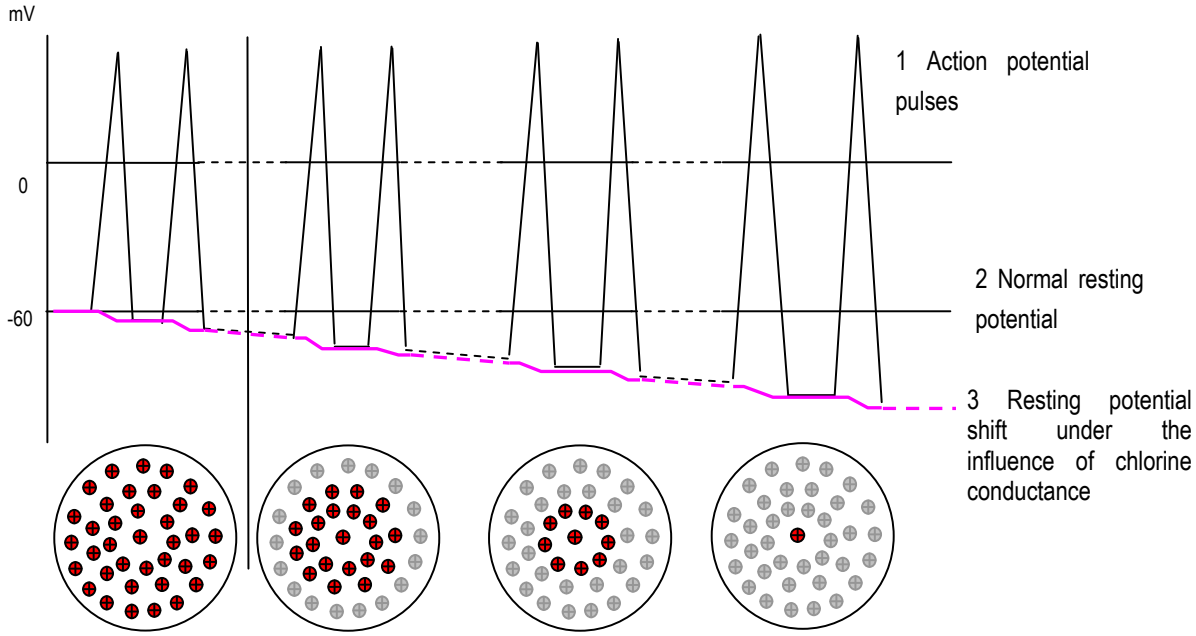


Figure 8. Changes in receptive field excitatory zone

It is known that excitability of different membrane areas in big (afferent and efferent) neurons is not distributed evenly [18]. There exists a low-threshold zone in the area of neuron starting segment (axon hillock and unmyelinated axon initial segment) where membrane possesses several times higher excitation comparing to other cell areas. The opening threshold of potential-dependent sodium channels increases as the distance from axon initial segment grows [19].

So after the *n*-th action potential pulse the resting potential shifts toward hyperpolarization. At the beginning of next (*n+1*-th) pulse this results in potential-dependent sodium channels that have maximal opening threshold (outermost from axon hillock) stops to open under the post-synaptic potential impact of excitatory receptors, located in the same area as this sodium channels. This means that potential of opening threshold for mentioned sodium channels is greater than initial resting potential plus shift value. In other words a number of potential-dependent sodium channels fail to participate in charge accumulation for the *n+1*-th pulse generation.

It is also possible to assume that the distance from excitatory receptor to axon hillock matches distance in the field of view from the point of given receptor to receptive field center. So the “nonparticipation” of some receptors in pulse generation matches lessening of receptive field excitatory zone. On the other hand the less the potential-dependent sodium channels opened in the time of next pulse generation, the greater the time needed for charge accumulation that sufficient to form this pulse. It seems like the reason of decrementing pulse generation

frequency over time. This assumption is consistent with experiment on action potentials generation for different holding potentials (artificially set resting potentials) of a cell membrane found in [13].

Even this mechanism is not examined experimentally its presence for excitation zone decreasing of visual system neurons' at the time of visual act fully complies to, explains and confirms the results obtained in [10, 11].

So, during the optical perception, namely one visual act, there exists image data with different resolution for image being viewed and the resolution changes serially from lowest to highest possible value up to the end of visual act. This means that technical vision systems will also have different image elements over time – the pixels of variable size that are changing from maximal to minimum possible size.

Conclusion

As for now we can state that a coarse-to-fine approach is used spontaneously by researchers in different fields of technical sciences. At the same time some studies are carried out in the domain of neurophysiology, showing the presence of such mechanism in living beings' visual system. A hypothesis explaining the functioning of that mechanism in retinal ganglion cells and LGN neurons was presented in this paper.

This hypothesis approval perhaps will enable systematization of coarse-to-fine approach in the field of technical vision; developing of general recommendations and best practices for its application to recognition tasks that can not be resolved at all by traditional methods.

For neurophysiology proof of this hypothesis will mean the possibility and necessity for combined consideration of visual system neurons' intracellular processes and intercellular interaction, refinement of visual neuron's functioning model at the time of action potential generation and explanation of coarse-to-fine mechanism in visual system of living beings.

Bibliography

[1] M. Pedersoli, A. Vedaldi, J. Gonz`alez. A Coarse-to-fine approach for fast deformable object detection. In CVPR, june 2011

[2] P. Moreels, P. Perona. Probabilistic Coarse-To-Fine Object Recognition. Technical report, 2005. California Institute of Technology

[3] Xi Li, Weiming Hu, Wei Hu. A Coarse-to-Fine Strategy for Vehicle Motion Trajectory Clustering. ICPR, 2006

[4] Jiann-Der Lee and Chen-Hui Kuo (2009). A Multi-Stage Classifier for Face Recognition Undertaken by Coarse-to-fine Strategy, State of the Art in Face Recognition, Julio Ponce and Adem Karahoca (Ed.), ISBN: 978-3-902613-42-4, InTech, Available from: http://www.intechopen.com/books/state_of_the_art_in_face_recognition/a_multistage_classifier_for_face_recognition_undertaken_by_coarse-to-fine_strategy

[5] Tyshchenko M. A. 3D reconstruction of human face in person identification problems. PhD thesis. International Research and Training Center for Information Technologies and Systems, Kyiv, 2012

[6] Christopher Raphael. Coarse-to-Fine Dynamic Programming. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001. vol. 23. p. 1379-1390.

[7] Brian O. Lucena. Dynamic Programming, Tree-width and Computation on Graphical Models. PhD thesis. Division of Applied Mathematics. Brown University 2002

[8] Maunsell, J.H.R. And Van Essen, D.C. Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed and orientation. J. Neurophysiol. 49: p. 1127-1147, 1983.

[9] R. Battiti, E. Amaldi, C. Koch. Computing Optical Flow Across Multiple Scales: An Adaptive Coarse-to-Fine Strategy. International Journal of Computer Vision, 6:2, p. 133-145, 1991

[10] Н.Ф. Подвигин Динамические свойства нейронных структур зрительной системы Ленинград: Наука, 1979, 158 с.

[11] Ruksenas O, Bulatov A, Heggelund P. Dynamics of Spatial Resolution of Single Units in the Lateral Geniculate Nucleus of Cat During Brief Visual Stimulation. J Neurophysiol 97:1445-1456, 2007.

[12] David H. Hubel. Eye, brain, and vision New York: Scientific American Library: Distributed by W.H. Freeman, 1988 240 p.

[13] S. Augustinaite and P. Heggelund. Changes in firing pattern of lateral geniculate neurons caused by membrane potential dependent modulation of retinal input through NMDA receptors. J Physiol 582.1 pp 297–315, 2007

[14] М.Шлезингер, В.Главач Десять лекций по статистическому и структурному распознаванию. – Київ: Наукова думка, 2004. – 535с.

[15] Т.Павлидис Алгоритмы машинной графики и обработки изображений. М Радио и связь 1986г. 400с

[16] У.Прэтт Цифровая обработка изображений: Пер. с англ.—М.: Мир, 1982.— Кн.1—312 с, ил.

[17] Nicholls J. G., Martin A. R., Wallace B. G., Fuchs P. A. 2001. From Neuron to Brain. 4th Ed. Sinauer Associates, Inc. Sunderland, MA USA.

[18] Wollner, D.A., and Catterall, W.A. 1986. Localization of sodium channels in axon hillocks and initial segments of retinal ganglion cells. Proc. Natl. Acad. Sci. USA 83, 8424–8428.

[19] Fried S. I., Lasker A. C. W., Desai N. J., Eddington D. K., Rizzo J. F. 3rd. Axonal sodium-channel bands shape the response to electric stimulation in retinal ganglion cells. J. Neurophysiol. 101: 1972–1987, 2009.

Authors' Information

Anton Sharypanov *– leading software engineer, master of science in flexible cybernated systems and robotechnics, Institute of cybernetics, prosp. akad. Glushkova 40, 03680, Kiev, Ukraine;*
e-mail: _sha_@ukr.net

Alexandra Antoniouk *– senior researcher, PhD (candidate of sciences in mathematics and physics), Institute of Mathematics NAS Ukraine, Tereshchenkivska, 3, 01 601, Kyiv, Ukraine; e-mail: antoniouk@imath.kiev.ua*

Vladimir Kalmykov *– senior researcher, candidate of engineering sciences, Institute of problems of mathematical machines and systems, prosp. akad. Glushkova 42, 03680, Kiev 187, Ukraine; e-mail: vl.kalmykov@gmail.com, kvg@immsp.kiev.ua*

TABLE OF CONTENTS