

УДК 004.94

05.13.10 - Управление в социальных и экономических системах (технические науки)

**АНАЛИЗ МЕТОДОВ И СРЕДСТВ ИЗВЛЕЧЕНИЯ ЗНАНИЙ И ОЦЕНИВАНИЕ КАЧЕСТВА МОДЕЛЕЙ В СППР ДЛЯ ИНДУСТРИАЛЬНО-ТЕХНОЛОГИЧЕСКИХ И СТРОИТЕЛЬНЫХ ПРОИЗВОДСТВ**Лаптев Владимир Николаевич  
к.т.н, доцентЛипин Константин Михайлович  
аспирантСтепанов Владимир Васильевич  
д.т.н., профессор

Степанова Марина Валерьевна  
старший преподаватель  
e-mail: faruelone@gmail.com  
e-mail: vvs04367@mail.ru  
*Кубанский государственный аграрный университет, Россия г. Краснодар*  
*Кубанский государственный технологический университет, Россия г. Краснодар, ул. Красная 135, к.123*  
*Краснодарское высшее военное авиационное училище летчиков им. А.К. Серова, г. Краснодар, Россия*

В работе рассмотрены наиболее используемые методы и средства извлечения знаний с учетом оценки качества моделей в системах поддержки принятия решений. В научном и практическом плане исследованы возможности совместного эффективного использования экспертных систем, интеллектуального анализа данных (ИАД) и машинного логического вывода (МЛВ), что обеспечивает более глубокую обработку данных, учитывая существенные отличия баз данных (БД) и баз знаний (БЗ). БД является единицей информации несвязанной друг с другом сведениями, в то время как БЗ – не только связаны между собой, но и с понятиями окружающего мира, что дает возможность решать сложные многокритериальные задачи в различных предметных областях. В настоящее время все возрастающее внимание уделяется нейросетевым технологиям, обладающие способностью моделирования нелинейных процессов, работе с зашумленными данными, а также способностью к обучению и самообучению, извлекая из поступающей информации существенные особенности. В тоже время интеграция нейросетевых технологий и моделей искусственного интеллекта в единую гибридную систему вместе с методами логического вывода в виде иерархической последовательности структуры

UDC 004.94

05.13.10 - Management in social and economic systems (technical sciences)

**ANALYSIS OF METHODS AND MEANS OF EXTRACTING KNOWLEDGE AND ASSESSING THE QUALITY OF MODELS IN THE DSS FOR INDUSTRIAL-TECHNOLOGICAL AND CONSTRUCTION INDUSTRIES**Laptev Vladimir Nikolaevich  
Cand.Tech.Sci., associate ProfessorLipin Konstantin Mikhailovich  
postgraduate studentStepanov Vladimir Vasilievich  
Doctor of technical sciences, Professor

Stepanova Marina Valerievna  
senior lecturer  
e-mail: faruelone@gmail.com  
e-mail: vvs04367@mail.ru  
*Kuban state agrarian University, Russia, Krasnodar*  
*Kuban state technological University, Russia*  
*Krasnodar, Krasnaya, 135*  
*Krasnodar higher military aviation school of pilots named after A. K. Serov, Krasnodar, Russia*

The article considers the most used methods and means of knowledge extraction taking into account the quality assessment of models in decision support systems. In scientific and practical terms, the possibilities of joint effective use of expert systems, data mining (IAD) and machine logical inference (MLV), which provides deeper data processing, taking into account the significant differences between databases (DB) and knowledge bases (BZ). DB is a unit of information unrelated to each other information, while BZ – not only related to each other, but also with the concepts of the world, which makes it possible to solve complex multi-criteria problems in various subject areas. Currently, increasing attention is paid to non-network technologies that have the ability to simulate nonlinear processes, work with noisy data, as well as the ability to learn and self-study, extracting essential features from the incoming information. At the same time, the integration of neural network technologies and artificial intelligence models into a single hybrid system together with the methods of logical inference in the form of a hierarchical sequence of the "If-then" rules structure significantly improves the understanding of the studied process and the quality of presentation of the result. Nevertheless, these methods and means of knowledge extraction are insufficient if the fuzzy linguistic inference

правил «Если-то» значительно улучшает понимание изучаемого процесса и качества представления получаемого результата. Тем не менее эти методы и средства извлечения знаний являются недостаточными если не использовать механизм нечеткого лингвистического вывода. Базовой характеристикой нечетких множеств является функция принадлежности, которая представляет собой обобщенную характеристику обычного множества. Для задания этой функции используются три типа форм – треугольная, трапецидальная и гауссова типа и две основных процедуры – фазификация и дефазификация, что и рассмотрено на примере метода Мамдани. Наряду с изложенным наиболее перспективным направлением в этой области является алгоритм адаптивного усиления AdaBoost, где ограничение усиления за счет фильтрации состоит в применении подвыборки, контур которой есть обычный контур пакетного обучения, допускающий повторное использование данных обучения. Это и обеспечивает возможность работы со слабыми моделями, а в условиях гибридизации обуславливает повышение эффективности, усиливает классификаторы, объединенные в «комитет». Каждая следующая совокупность классификаторов строится по объектам неверно классифицированных предыдущими совокупностями. AdaBoost чувствителен к шуму в данных и выбросам и менее подвержен переобучению, что позволяет значительно уменьшить количество примеров и получать более качественный вывод в СППР

Ключевые слова: ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ, ОБРАБОТКА ДАННЫХ, МНОГОКРИТЕРИАЛЬНЫЕ ЗАДАЧИ, НЕЙРОСЕТЕВЫЕ ТЕХНОЛОГИИ, ФАЗИФИКАЦИЯ, ДЕФАЗИФИКАЦИЯ, АЛГОРИТМ АДАПТИВНОГО УСИЛЕНИЯ

DOI: <http://dx.doi.org/10.21515/1990-4665-157-013>

mechanism is not used. The basic characteristic of fuzzy sets is the membership function, which is a generalized characteristic of a normal set. To set this feature, we use three types of shapes – triangular, trapezoidal and Gaussian type and two main procedures – phasefication and de-phasefication which is considered by the example of the method of Mamdani. Along with the stated most promising direction in this area is the adaptive gain algorithm called AdaBoost, where the limitation of the gain due to the filtering is to apply the subsampling circuit which has the normal contour of batch training, reusable training data. This provides an opportunity to work with weak models, and in the conditions of hybridization causes efficiency increase, strengthens the classifiers united in the "Committee". Each next set of classifiers is built on objects incorrectly classified by previous sets. AdaBoost is sensitive to data noise and emissions and is less susceptible to retraining, which can significantly reduce the number of examples and obtain better output in the DSS

Keyword. DATA MINING, DATA PROCESSING, MULTICRITERIA PROBLEM, NEURAL NETWORK TECHNOLOGY, PHASEFICATION, DE-PHASEFICATION, ADAPTIVE ALGORITHM GAIN

Информация и разнородные данные – это основной строительный материал не только для построения обычных вычислительных, но и нейронных сетей, а также деревьев решений в интеллектуальных системах поддержки принятия решений, обеспечивающих осуществление высококачественного анализа, выявления скрытых закономерностей и выработки стратегии принятия решений. Интеллектуальный анализ данных (ИАД) – процесс поддержки принятия решений, основанный на

поиске в данных скрытых закономерностей, шаблонов, сведений – автоматически обобщенных до информации, которая может быть отождествлена как знания. Особенности алгоритмов, методов и средств, применяемых в ИАД – отсутствие ограничительных рамок в процессе выявления закономерностей и предположений в структуре выбора анализируемых показателей с необработанными (сырыми) данными, сокращения объема которых не производится за счет некоторых значений величинами или производными от них, например, средних значений.

ИАД основан на методологии «обнаружения знаний» в базе данных (БД), как мультидисциплинарная область, возникшая из теории БД (DW, OLAP), теории искусственного интеллекта распознавания образов, экспертных систем, нейросетевых технологий, математической статистики, деревьев решений, нечетких множеств, иерархических моделей данных и т.д.

Методы ИАД позволяют выделить основные типы закономерностей:

- ассоциация – выявление связи (совместимость) нескольких событий с учетом их временной составляющей (последовательность);
- классификация – выделение признаков классов (сгруппированных данных);
- кластеризация – формирование классов.

Как следствие, имеет место прогнозирование, чему наилучшим образом соответствуют нейронные технологии, обладающие способностью к моделированию нелинейных процессов, работе зашумленными данными, способностью обучаться и самообучаться, извлекать из поступающей информации существенные особенности [1].

Интеграция нейронных сетей и моделей искусственного интеллекта (ИИ) в единую гибридную систему позволяет значительно повысить эффективность работы в области поддержки принятия решений.

Рассмотрим несколько примеров интеграции в единую систему указанных выше технологий, обеспечивающих ожидаемый эффект.

Экспертные системы (ЭС) моделируют рассуждение эксперта в некоторой предметной области, и использует базу знаний, содержащей факты и правила об этой области, а также специальную процедуру логического вывода и является составной частью исследований по ИИ, цель которых использовать алгоритмы обработки примеров. Ориентация на обработку данных с помощью некоторых правил вывода, которые могут извлекать знания из баз знаний (БЗ) или же непосредственно получать их от эксперта, последнее свойственно именно экспертным системам, когда возникает необходимость разработки моделей и выбора метода для решения определенной задачи предметной области (ПрО) [1].

Существенно важным в ЭС является их взаимодействие с пользователем, структурно-технологической частью системы, что обусловило появление нового технологического приема основанного на знаниях, как составной части ИИ, и имеет целью создание компьютерных методов решения задач (проблем) требующих привлечение экспертов. Основой экспертных систем является БЗ, отличающаяся от базы данных (БД) единицами информации несвязанных друг с другом сведений, в то время как в БЗ все элементы не только связаны между собой, но и с понятиями окружающего мира, что необходимо для понимания, формулирования и решения задач различных предметных областей. Основным элементом ЭС является машина логического вывода (МЛВ), которая манипулирует информацией из БЗ, и указывает, в каком порядке следует выявлять взаимосвязи и делать выводы. Кроме того, она используется для обработки вопросов, моделирования рассуждений и подготовки ответов [3]. В тоже время в рамках интеграции с другими системами этот подход имеет ряд недостатков. Например, в случае использования нейросетевых технологий и построения деревьев решений

на основе экспертных оценок с помощью некоторых правил вывода, он реализует цепочку рассуждений эксперта при выборе метода решения задачи в конкретных условиях предметной области и не предназначен для решения универсальных задач.

Поскольку ЭС – компьютерная программа, и как конечный продукт зависит от качества создаваемой базы знаний и решает большой круг задач исключительно в узкоспециализированной хорошо изученной предметной области, располагает четкими стратегиями принятия решений, то передача ЭС более глубоких знаний о ПрО весьма сложной проблемой. Это связано со сложностью формализации эвристических знаний экспертов, не способностью к самообучению, отсутствием возможностей осмысленно объяснить собственные рассуждения.

Способность нейронных сетей к моделированию нелинейных процессов, а также к работе с зашумленными данными в процессе решения задач ИАД априорных предположений о структуре выборки и виде распределенных знаний анализируемых показателей, вызывает необходимость применения нейросетевых технологий.

Интеграция нейросетевых технологий с методами логического вывода в виде иерархической последовательности структуры правил «Если-то» значительно улучшает понимание изучаемого явления за счет более качественного представления результата полученного в ходе обучения нейронной сети [2].

Основой для использования механизма нечеткого логического вывод, как известно, является база правил, основанная на нечетких высказываниях в форме «Если-то» для соответствующих лингвистических термов и функций принадлежности удовлетворяющая двум условиям:

- входная переменная для любого терма должна иметь хотя бы одно правило, в котором этот терм используется в качестве предпосылки (левая часть правила);

- существует хотя бы одно правило для каждого лингвистического термина выходной переменной.

Базовой характеристикой нечеткого множества является функция принадлежности, которую чаще всего обозначают через  $MF_c(x)$  - обобщенное понятие характеристической функции обычного множества.

При этом  $MF_c(x) = 0$  означает отсутствие принадлежности множеству, а  $MF_c(x) = 1$  – полная принадлежность [2], где  $c$  - нечеткое множество, при этом является множеством упорядоченных пар, то есть  $c = MF_c(x' / x)$ ;  $MF_c(x) [0; 1]$ .

Для задания функции принадлежности существует достаточно много типовых форм, из которых приемлемыми явились – треугольная, трапециевидальная и гауссова.

Треугольная функция, принадлежность которой строится на основе тройки чисел  $(a, b, c)$  и ее значение в точке  $x$  вычисляется исходя из соотношения (рис.1)

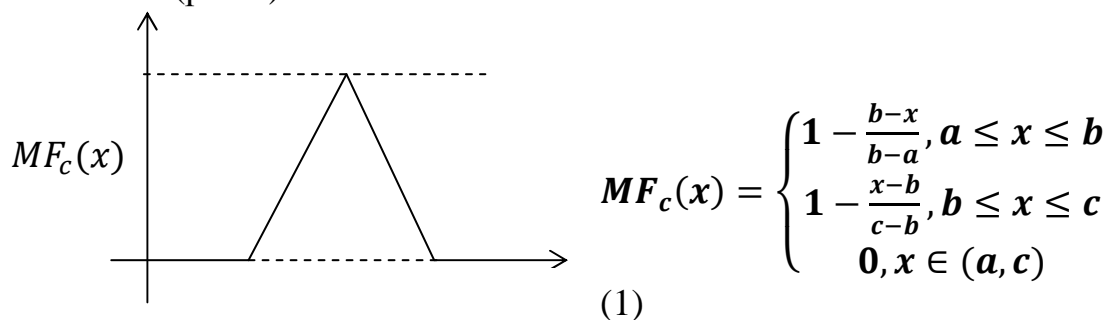
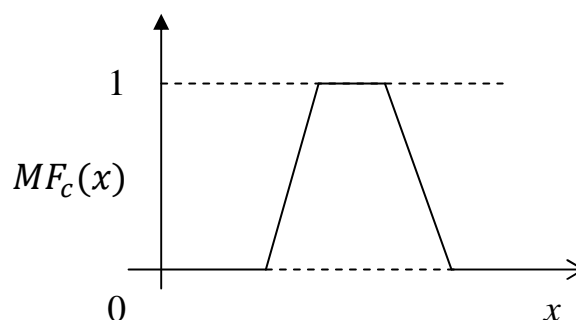


Рисунок 1. Кусочно-линейная типовая функция треугольной функции принадлежности

Как видно из соотношения [1] в случае  $(b - a) = (c - b)$  имеет место случай симметричной треугольной функции принадлежности, которая может быть однозначно задана двумя параметрами из тройки чисел  $(a, b, c)$ .



Для более сложного случая имеет место трапецеидальная функция принадлежности, ее можно записать с помощью четверки чисел  $(a, b, c, d)$  (рис.2).

$$MF_c(x) = \begin{cases} 1 - \frac{b-x}{b-a}, & a \leq x \leq b \\ 1, & b \leq x \leq c \\ 1 - \frac{x-c}{d-c}, & c \leq x \leq d \\ 0, & x \notin (a, d) \end{cases} \quad (2)$$

Рисунок 2. Кусочно-линейная трапецеидальная функция принадлежности

Функция принадлежности гауссового типа может быть представлена следующей формулой:

$$MF(x) = \exp\left[-\left(\frac{x-c}{\sigma}\right)^2\right] \quad (3),$$

где  $c$  – центр нечеткого множества;

$\sigma$  - параметр, отвечающий за крутизну функции (рис.3).

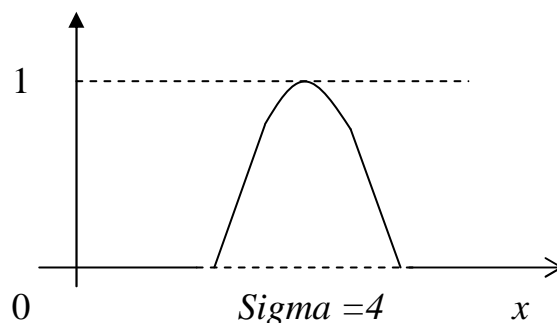


Рисунок 3. Гауссова функция принадлежности.

Обычно функции принадлежности для каждого термина из базового множества  $T$  изображают на одном графике лингвистической переменной, учитывая, что в ней количество термов редко превышает 7.

Для углубленного понимания рассмотрим механизм нечеткого вывода на примере алгоритма Мамдани, в основу которого положены виды используемых правил, логических операций и методы дефазификации, а

также минимаксная композиция нечетких множеств. Полученный таким образом алгоритм включает следующую последовательность действий [2]:

1. процедура дефазификации: определяется истинности, то есть значение функции принадлежности для левых частей соответствующего правила (предпосылки). Степень истинности обозначается как  $A_{ik}(x_k)$ , где  $i=1, \dots, m$ ;  $k=1, \dots, n$ .
2. Нечеткий вывод - вначале определяются уровни «отсечения» для левой части каждого из правила:  $\alpha_i = \min_i(A_{ik}(x_k))$ . Затем находят «усеченную» функцию принадлежности:  $B_i^*(y) = \min_i(\alpha_i, B_i(y))$ .
3. Композиция или объединение полученных усеченных функций, в основе чего используется максимальная композиция нечетких множеств:  $MF(y) = \max_i(B_i^*(y))$ , где  $MF(y)$ - функция принадлежности итогового нечеткого множества.
4. Дефазификация или приведение к четкости - методом среднего центра (центроидный метод). Геометрический смысл которого заключается в том, что центр тяжести для кривой:  $MF(y)$  определяется на примере двух выходных переменных, что является распространенным способом логического вывода в нечетких системах. Последнее может найти широкое применение, особенно в инженерии.

Рассматриваемый алгоритм работает следующим образом. На итерации  $n$  алгоритм усиления реализует слабую модель обучения с распределением  $D_n$  множества примеров обучения  $T$ , затем для нее вычисляется гипотеза  $F_n: X \rightarrow Y$ , которая осуществляет классификацию примеров обучения  $F_1, F_2, \dots, F_T$  в единую гипотезу  $F_{fin}$ , для вычисления распределения  $D_n$  на итерации  $n$ ,  $F_{fin}$ .

После окончания обучения третьего эксперта, на отфильтрованном множестве примеров процесс обучения завершается. Таким образом, алгоритм концентрирует основной объем веса на трех примерах,



классифицировать которые наиболее сложно [1] и где ограничение усиления за счет фильтрации состоит в использовании предвыборки, контур которой является обычным контуром пакетного обучения, допускающей повторное использование данных обучения, что позволяет работать со слабыми моделями обучения.

Отличие *Ada Boost* от других алгоритмов состоит в следующем:

- адаптивно настраивается на ошибки слабых гипотез, возвращаемых слабыми моделями;
- ограничение производительности алгоритма зависит от производительности слабой модели обучения, только на трех распределениях, которые фактически формируются в процессе обучения.

Алгоритм *Ada Boost* в сочетании с другими алгоритмами классификации обеспечивает повышение эффективности, усиливает классификаторы, объединяя их в «комитет». Он является адаптивным и строится по следующему правилу – каждая следующая совокупность классификаторов строится по объектам, неверно классифицированным предыдущими совокупностями. Кроме того, *Ada Boost* чувствителен к шуму в данных и выбросам, менее подвержен переобучению по сравнению с другими алгоритмами.

Рассмотрим на достаточно простом примере решения задачи построения бинарного классификатора [4].

Дано  $(x_1, y_1), \dots, (x_m, y_m)$ , где  $x_i \in X$ ,  $y_i \in Y = \{-1; 1\}$ .

Инициуем  $D_1(i) = \frac{1}{m}$ ,  $i = 1, \dots, m$  для каждого  $t=1, \dots, T$ .

Находим классификатор  $h_t: X \rightarrow \{-1; +1\}$ , который минимизирует взвешенную ошибку классификации  $h_t = \operatorname{argmin}_j \varepsilon_j$

$$h_j = H$$

где  $\varepsilon_j = \sum_{i=1}^m D(i)[y_i \neq h_j(x_i)]$ .

При этом,

- если величина  $\varepsilon_t > 0,5$ , то останавливаемся.

- выбираем  $\alpha_t \in R$ . Обычно  $\alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$ , где  $\varepsilon_t$  - взвешенная ошибка классификатора  $h_t$ .

$$\text{Обновляем: } D_{t+1}(i) = \frac{D_t(i)e^{-\alpha_t y_i h_t(x_i)}}{z_t},$$

где  $z_t$  - нормализующий параметр, выбираемый так, чтобы  $D_{t+1}$  представлял собой распределение вероятностей  $\sum_{i=1}^n D_{t+1}(i) = 1$ .

Строим результирующий классификатор:

$$H(x) = \text{sign}(\sum_{t=1}^T \alpha_t h_t(x)).$$

Соотношения для последующего обновления  $D_t$  должны удовлетворять условиям:

$$e^{-\alpha_t y_i h_t(x_i)} \begin{cases} < 1, y_i = h_t(x_i) \\ > 1, y_i \neq h_t(x_i) \end{cases}$$

После выполнения этих условий и выбора оптимального классификатора  $h_z$  для распределения  $D_t$  объекты распределения  $x_i$  ( $i=1, \dots, m$ ), характеризующие  $h_z$  идентифицируют корректно и имеют веса меньше, чем те, которые идентифицируют некорректно. При этом на каждой последующей итерации веса каждого неверно классифицируемого объекта увеличиваются, вызывая у нового комитета необходимость, концентрировать внимание именно на этих объектах.

Таким образом, когда алгоритм *Ada Boost* тестирует классификаторы на рассматриваемом распределении  $D_{t+1}$ , он ориентируется на классификатор, который лучше идентифицирует объекты неверно распознаваемые предыдущим классификатором.

Обобщая изложенное в данной работе, следует отметить, что проведенный анализ методов и средств извлечения знаний и оценки качества моделей в СППР показал, что широко известные методы

автоматического анализа данных, в основе которого лежит обучение на примерах, позволяют решать более сложные задачи различных сложных предметных областей.

Чаще всего в сложных ПрО может возникнуть ситуация когда точно классифицировать исследуемый объект весьма трудоемко, а использование даже экспертных систем, ориентированных на обработку данных с помощью некоторых правил вывода, которые могут извлекать знания из БЗ, ограничены хорошо изученными предметными областями. Возникает необходимость использования нейронных сетей, поскольку они хорошо адаптивны к классификации образов, формированию классов, моделирование нелинейных процессов, работа с зашумленными данными, обладают способностью к обучению и самообучению.

Интеграция нейронных сетей и методов искусственного интеллекта (ИИ), а также машины логического вывода (МЛВ) в единую гибридную систему значительно повышают эффективность работы в области поддержки принятия решений. Наиболее эффективным в этой ситуации является то, что когда в единую гибридную систему к МЛВ добавляется метод логического вывода в виде иерархической последовательности структуры правил «Если-то», значительно улучшает качество разрабатываемых моделей, а использование механизма нечеткого логического вывода для этого же случая, на основе соответствующей базы правил и лингвистических термов, функций принадлежности ограниченной двумя известными условиями.

На основе использования нечеткого вывода и алгоритма, имеющего четыре последовательных действия, последним из которых является дефазификация или приведение к четкости (центроидный метод) получаем наиболее обоснованный четкий вывод.

Достаточно перспективным направлением в этой области является алгоритм адаптивного усиления *Ada Boost*, который адаптивно

настраивается на ошибки слабых моделей, при этом производительность его зависит от производительности слабой модели только на трех распределениях, которые фактически формализуются в процессе обучения. То есть, когда алгоритм *Ada Boost* тестирует классификаторы на рассматриваемом распределении  $D_{t+1}$ , он ориентируется на классификатор, который лучше идентифицирует объекты неверно распознаваемые предыдущим. Учитывая, что в процессе обработки БЗ имеет место огромное количество примеров, последнее обеспечивает значительные уменьшения и более качественный вывод в области СППР.

### Литература:

1. Саймон Хаки. Нейронные сети. Полный курс Второе издание. [Текст] / Саймон Хаки. –М.: Издательский дом Вильямс, 2006.-1104 с.
2. Солодовников И.В., Солодовников В.И. Подход к созданию подсистем автоматизации проектирования нейросетевых структур данных с использованием деревьев решений. [Текст]/ Солодовников И.В., Солодовников В.И.// / Информационные технологии в проектировании и производстве. 2006. № 2.С. 62-66.
3. Рассел С., Норвиг П. Искусственный интеллект: современный подход [Текст]/ Рассел С., Норвиг П. М.: Издательский дом «Вильямс», 2006.
4. Алгоритм AdaBoost/ [Электронный ресурс] Режим доступа: <http://www.machinelearning.ru>, свободный.

### References

1. Sajmon Xaki. Nejronny`e seti. Polny`j kurs Vtoroe izdanie. [Tekst] / Sajmon Xaki. –M.: Izdatel`skij dom Vil`yams, 2006.-1104 s.
2. Solodovnikov I.V., Solodovnikov V.I. Podxod k sozdaniyu podsystem avtomatizacii proektirovaniya nejrosetevy`x struktur danny`x s ispol`zovaniem derev`ev reshenij. [Tekst]/ Solodovnikov I.V., Solodovnikov V.I.// / Informacionny`e texnologii v proektirovanii i proizvodstve. 2006. № 2.S. 62-66.
3. Rassel S., Norvig P. Iskusstvenny`j intellekt: sovremenny`j podxod [Tekst]/ Rassel S., Norvig P. M.: Izdatel`skij dom «Vil`yams», 2006.
4. Algoritm AdaBoost/ [E`lektronny`j resurs] Rezhim dostupa: <http://www.machinelearning.ru>, svobodny`j.