УДК 004.93`1 + 004.896

В.А. Козловский, А.Ю. Максимова
Институт прикладной математики и механики НАН Украины
Maximova.Alexandra@mail.ru

Нечеткая система распознавания образов для решения задачи классификации жидких нефтепродуктов

Решается задача классификации жидких нефтепродуктов. Строится нечеткая система на правилах с лингвистическими переменными. Настройка базы знаний выполняется на основе функционалов скользящего контроля. Для повышения качества распознавания было выполнено разбиение на кластеры некоторых классов образов, что позволило улучшить качество распознавания на 2% по сравнения с исходным алгоритмом (92% верно распознанных объектов).

Ключевые слова: *распознавание образов, нечеткая логика, анализ данных.*

Введение

В различных областях человеческой деятельности возникают задачи, сводящиеся к классу задач распознавания образов или классификации. Развитие компьютерных технологий позволило накапливать большие объемы информации, требующие обработки и преобразования с целью извлечения из них необходимых данных и знаний. В связи с этим активно разрабатываются методы обработки данных, и, в частности, методы распознавания образов. По [1] целью разведочного анализа данных (exploratory data analysis) является выявление структуры данных и описание ее статистическими моделями. Однако этот подход требует достаточно жестких модельных ограничений, вытекающих из аксиоматики теории вероятностей. Выход за рамки статистических моделей привел к формированию более общего направления интеллектуального анализа данных, в котором используются методы нечеткой логики, эволюционные и генетические алгоритмы, иммунные системы и др., а также гибридные методы, которые в совокупности получили название «мягкие вычисления» (soft computing).

На практике исходные данные зачастую обладают сложно формализуемой, неоднородной структурой с заведомо пересекающимися классами образов. Другой проблемой является неполнота обучающей выборки и ее сильная зашумленность. В дискриминантном анализе в случаях неопределенности ответа о принадлежности образа классу образов ответ может быть получен в виде вероятности принадлежности образа каждому из классов образов. Однако описанные выше особенности данных во многих случаях не позволяют построить адекватные вероятностно-статистические модели, что обуславливает создание эмпирических методов и подходов, при разработке которых удобными становятся методы и модели из области мягких вычислений.

Целью работы является решение прикладной задачи классификации жидких нефтепродуктов (ЖНП), в частности бензинов. Эта задача возникает при контроле качества нефтепродуктов по результатам лабораторных измерений некоторых их физико-химических характеристик. Для ее решения строиться нечеткая система распознавания. При формировании базы знаний за основу взята идея построения интегральных характеристик классов образов в виде нечетких множеств – значений лингвистических переменных, описывающих классы образов по признакам. Принятие решения о принадлежности рассматриваемого образца к определенному виду выполняется методами логического нечеткого вывода.

Особенности рассматриваемой задачи классификации жидких нефтепродуктов

Как отмечалось выше, в работе предлагается метод для решения задачи контроля качества нефтепродуктов, которая сводиться к задаче распознавания образов. Ее суть заключается в определении производителя и вида образца ЖНП в лабораторных условиях. В лаборатории контроля качества нефтепродуктов накапливается информация об образцах топлива, поступающих на экспертизу от разных производителей и потребителей. По каждому образцу разными методами определяется ряд показателей. Такими показателями являются октановое число, содержание ароматических веществ, в том числе олефинов, ароматических бензолов, содержание серы и др. Накопленная информация по всем образцам используется в качестве эталонной для определения производителя рассматриваемого образца. Отсутствие универсальных подходов к измерению параметров ЖНП и неоднородный характер накопленной информации сдерживают развитие методов и приборов для автоматизации процессов управления в лабораториях контроля качества. Существующие разработки обладают рядом

недостатков и ограничений в использовании. Например, в работе [2] предлагается нейросетевой импедансный метод идентификации ЖНП, работающий с данными, полученными с импедансометрического датчика. Эти данные обрабатываются нейросетевым алгоритмом. Однако алгоритм ориентирован на уникальный прибор, предложенный автором указанного метода.

В качестве класса образов в данной задаче рассматривается топливо определенного вида от определенного производителя. Основной особенностью исходных данных является полное или частичное совпадение некоторых классов образов. Весьма условно схема взаимного пересечения классов образов представлена на рис. 1., где обозначения П и В в названиях классов образов являются сокращениями от слов «производитель» и «вид». Области пересечения классов образов заштрихованы.

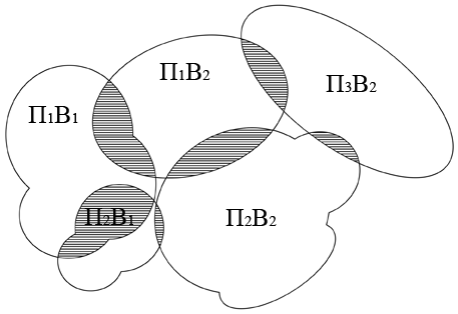


Рисунок 1 – Схема взаимного пересечения классов образов

Наличие пересечений между классами образов связано не столько с зашумлением граничных областей классов образов, сколько с технологическими особенности производства ЖНП, когда разные производители реализуют топливо одного вида и с одинаковыми характеристиками. В ситуациях, когда рассматриваемый образец попадает в области пересечений классов образов и не удается дать однозначный ответ о принадлежности к определенному классу образов, важно получить ответ о степени похожести образца на каждый из предполагаемых классов образов, а, иногда, определить, что это точно не образец определенного класса образов. В связи с этим предлагается ответ получать в виде нечеткого множества, элементы которого - классы образов.

Постановка задачи

Формально постановку прикладной задачи распознавания нефтепродуктов можно рассматривать как нечеткую модификация задачи классификации в многомерном пространстве. Пусть дана обучающая выборка $Y = \{(x^{(i)}, v^{(i)}),\ x^{(i)} \in X, v^{(i)} \in V, i = 1,...,n\}$, где $x^{(i)} \in X \subset R^m$ векторы m-мерного пространства – набор информативных признаков, $V = \{v_i \mid i = 1,...,k\}$, $v_i \in \mathrm{N}$ - множество номеров классов образов. Пары $(x^{(i)}, v^{(i)})$ определяют, представителем какого класса образов $v^{(i)}$ является образ $x^{(i)}$. В общем случае, необходимо определить степень принадлежности образца $\overline{x'}$ рассматриваемым классам, т. е. построить нечеткое множество $\tilde{y}_X = \sum_{i=1}^{k} \mu_X(v_i)/v_i$, где $\mu_X(v_i)$ - степень принадлежности образа $\overline{x'}$ классу v_i. Здесь и далее обозначения соответствуют введенным в классической теории нечетких множеств [3].

Основная идея алгоритма распознавания образов

Как отмечалось ранее, задача классификации нефтепродуктов сводиться к нечеткой модификации задачи распознавания образов. Авторами в [4] был предложен алгоритм распознавания образов, основанный на нечетких портретах классов образов. Основная идея алгоритма заключается в представлении исходной информации о классах в виде их нечетких портретов, которые формируются в результате анализа обучающей выборки. Такие портреты описываются совокупностью лингвистических переменных, соответствующих информативным признакам. Терм-множества этих лингвистических переменных описывают значение признака для каждого из классов образов и строятся в результате анализа частоты встречаемости значений признака в каждом классе образов. В отличие от метрических алгоритмов распознавания образов, таких, например, как алгоритма k-ближайших соседей или метод потенциальных функций [5], где в рассмотрение берется каждая точка обучающей выборки, в предлагаемом подходе знания о выборке прецедентов обобщаются. В результате нет необходимость хранить в памяти сведения о всех реализациях обучающей выборки, что дает выигрыш по объему памяти.

На основе анализа выборки прецедентов строиться нечеткая система, основанная на лингвистических переменных. База знаний формируется по построенным нечетким портретам. Решение принимается алгоритмом нечеткого логического вывода. Результат работы алгоритма представляется нечетким множеством.

Извлечение знаний в виде интегральных характеристик классов образов

Рассмотрим механизм анализа и извлечения знаний по выборке прецедентов.

Результатом такого анализа являются нечеткие портреты классов образов. Остановимся более подробно на алгоритме построения таких «портретов».

В рассмотрение берутся признаки с низкой попарной корреляцией, поэтому данные по каждому показателю рассматриваются независимо друг от друга. Первоначальный выбор информативных признаков в задаче классификации ЖНП был осуществлен экспертом.

Для каждого информативного признака P_i, $i = 1,...,m$, определим лингвистическую переменную L_i: $L_i = \{имя(P_i), T_i, U_i, G, M\}$, где U_i – множество значений признака P_i, $T_i = \{\mu_{ij} \mid j = 1,...,k\}$ – терм-множество лингвистической переменной, $\mu_{ij}(pr_i\,\bar{x}) \in [0,1]$ – функция принадлежности, определяющая степень уверенности, с которой образ $\bar{x}$ относиться к классу образов v_j, причем $pr_i\,\bar{x}$ определяет значение признака P_i для образа $\bar{x}$. Синтаксическое правило G, порождающее названия переменных, в данном случае тривиально, т.к. все термы атомарные, и заключается в присвоении функции принадлежности имени класса, который она представляет. Семантическое правило M представлено в виде алгоритма формирования функций принадлежности, который основан на концепции скользящего окна и является расширением подхода, используемого при построении гистограмм в статистике [6]. Вид функций принадлежности зависит от выбора коэффициентов α и β, которые фактически определяют ширину скользящего окна и шаг скольжения и являются настраиваемыми параметрами алгоритма распознавания.

Нечеткий портрет первого порядка S_j класса v_j определяется как совокупность значений лингвистических переменных, соответствующих классу v_j: $S_j = \{\mu_{ji} \mid i = 1,...,m\}$. Для каждого нечеткого портрета S_j строится правило нечеткого вывода по определенным ранее лингвистическим переменным.

Построение нечеткой системы распознавания образов

Нечеткие системы весьма успешно показали себя для решения задач управления объектами, для которых получить модель в рамках классической теории управления невозможно или нецелесообразно, в связи с неполнотой данных и высокой сложностью моделей [7]. Системы распознавания образов также включают в себя нечеткие модели, однако в этих случаях чаще используются гибридные модели [8].

В работе строиться нечеткая система, основанная на правилах с лингвистическими переменных, схема которой представлена на рис. 2. На вход системы подается m-мерный вектор $\bar{x}'$. Для каждой компоненты вектора x_i', $i = 1,...,m$ в блоке «фаззификатор» строиться синглетон – одноточечное нечеткое множество. На втором этапе выполняется обработка данных с помощью механизма нечеткого вывода, который состоит из базы знаний и нечеткого процессора. База знаний строиться по нечетким портретам, полученным на этапе анализа выборки прецедентов. Следует отметить, что стандартный для систем данного типа блок «дефаззификации» отсутствует и заменен «анализатором», где строиться модифицированное нечеткое множество $\tilde{y}$.

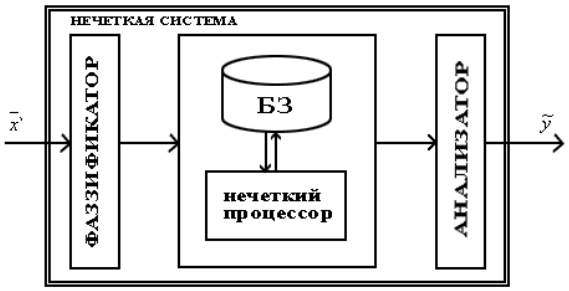


Рисунок 2 – схема нечеткой системы распознавания с нечетким выходом

Рассмотрим способ формирования базы знаний. Каждое ее правило соответствует нечеткому портрету класса образов.

ПРАВИЛО «S_1»:

ЕСЛИ «L_1 *есть* v_1» И… И «L_i *есть* v_1» И … И «L_m *есть* v_1» ТО «$\tilde{v}_1$ *есть* V_1»;

…

ЕСЛИ «L_1 *есть* v_j» И… И «L_i *есть* v_j» И … И «L_m *есть* v_j» ТО «$\tilde{v}_j$ *есть* V_j»;

…

ПРАВИЛО «S_k»:

ЕСЛИ «L_1 *есть* v_k» И… И «L_i *есть* v_k» И … И «L_m *есть* v_k» ТО «$\tilde{v}_k$ *есть* V_k».

В нечетком предикате «L_i *есть* v_j» лингвистическая переменная соответствует построенной на этапе анализа выборки прецедентов, а v_j - имя класса образов, соответствующее значениям лингвистической

переменной.

Количество нечетких предикатов в поле ЕСЛИ правила соответствует m информативным признакам. В поле ТО нечеткое множество V_j является монотонной функцией, что используется в алгоритме нечеткого вывода Цукамото [9]. Принятие решения осуществляется на основе механизма нечеткого вывода. Следует отметить, что в качестве операции «И» на этапе агрегирования используется m-местная логарифмическая функция:

$$f(a_1,a_2,...,a_m)=\begin{cases}0, \textit{если } \exists\, a_i=0\\ \log_2((a_1+1)\cdot...\cdot(a_m+1))/m,\\ \quad \textit{если } a_i>0, i=1,...,m\\ a_1, m=1\end{cases} \quad (1)$$

Результатом работы алгоритма нечеткого вывода является совокупность синглетонов $\tilde{v}_j$, $j=1,...,k$.

Как было отмечено ранее, часть классов образов априори пересекается, поэтому в качестве результата работы классификатора важно получить список классов, в область пересечения которых попадает рассматриваемый образ $\bar{x}'$. Сохранить информацию о степени похожести на каждый из возможных классов образов позволяет блок «анализатор», работающий по следующему принципу. На его вход поступает дискретное нечеткое множество $\{v_j \mid j=1,...,k\}$, где каждый элемент несет информацию о соответствии рассматриваемого образца j-ому классу образов. Отсутствие элементов с ненулевыми значениями функции принадлежности сигнализируют об отсутствие в обучающей выборке соответствующего рассматриваемому образцу класса образов. В противном случае элементы множества сортируются по возрастанию степеней принадлежности. Результирующее нечеткое множество $\tilde{y}$ (выход нечеткой системы) формируется по трем лидирующим классам образов, если таковые существуют, либо по меньшему их числу.

Для решения задачи классификации нефтепродуктов обучающая выборка состояла из 870 элементов, шести классов $\{v_1,v_2,v_3,v_4,v_5,v_6\}$ и описывалась шестью признаками, такими как содержание ароматических углеводородов (Aroma), олефинов (Olf), бензолов (ArmBnz), ксилолов (ArmKs), октановое число (RON) и массовая доля МТБЭ (MTBE). На рис. 3 представлена лингвистическая переменная «Olf», соответствующей признаку, отвечающему за содержание олефинов, и ее значения для некоторых классов, полученные при значениях параметров $\alpha=0.1$ и $\beta=4$.

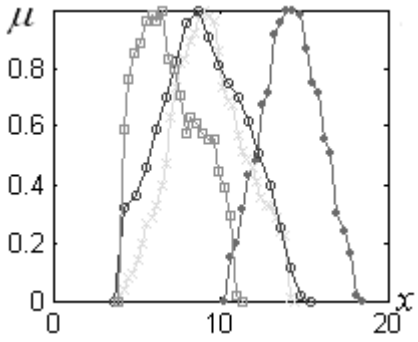
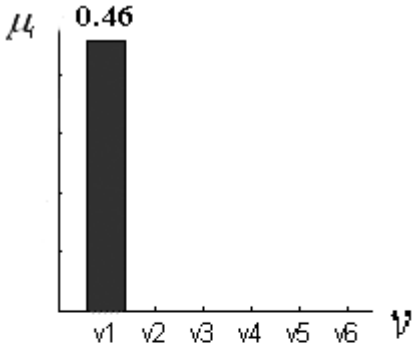
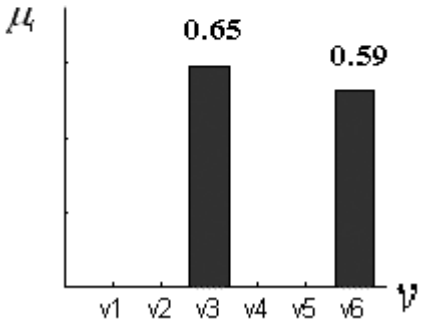


Рисунок 3 – Значения лингвистической переменной для признака Olf для классов образов v_1,v_2,v_3,v_4

Результат работы алгоритма удобно представлять в виде столбиковых диаграмм. Для образцов $\bar{x}^{(1)}=(20,25,1.3,0,93,6)$ и $\bar{x}^{(2)}=(9,40,2,1,97,10)$ получены множества $\tilde{y}_1=(0.46/v_1)$ и $\tilde{y}_2=(0.65/v_3+0.59/v_6)$. На рис. 4 представлен результат работы алгоритма в виде столбиковых диаграмм



а)



б)

Рисунок 4 – Результат нечеткого вывода для образцов а) $\bar{x}^{(1)}$, б) $\bar{x}^{(2)}$

Следует отметить, что образец $\bar{x}^{(1)}$ однозначно является представителем классам образов v_1, а класс образов, к которому относиться $\bar{x}^{(2)}$, определяется неоднозначно - v_3 или v_6.

В ситуации, соответствующей второму примеру, когда ответ неоднозначен, эксперт, обладающий дополнительными знания может принят окончательное решения, опираясь на имеющиеся данные.

Адаптация и контроль качества алгоритма

В предлагаемом алгоритме этап адаптации осуществляется за счет выбора параметров α и β по результатам серии экспериментов. Поиск наилучшего решения затруднен отсутствием общепризнанных универсальных критериев качества решений. На практике для оптимизации небольшого числа параметров используют функционалы скользящего контроля [10]. Фактически методами скользящего контроля измеряется обобщающая способность метода обучения на заданной конечной выборке.

В полной мере оценить обобщающую способность алгоритма позволяет комбинаторный функционал полного скользящего контроля:

$$Q_c(\mu, X^L) = \frac{1}{N}\sum_{n=1}^{N} \nu(\mu(X_n^l), X_n^k), \qquad (2)$$

где μ - алгоритм распознавания, настраиваемый по конечной совокупности объектов X^L, где (X_n^l, X_n^k), $n = 1,...,N$ - всевозможные разбиения выборки X^L на обучающую и контрольную, $L = l + k$; $\nu(\mu, X^L)$ - частота ошибок алгоритма μ на обучающей выборке X^L. Для некоторых алгоритмов получены эффективные формулы вычисления функционала полного скользяещго контроля, как, например, для алгоритма k ближайших соседей, где он вычисляется через профиль компактности. Однако с увеличением объема обучающей выборки его вычисления становиться ресурсоемким и затратным по времени. Такими же недостатками обладает функционал среднего отклонения частоты ошибок на контроле от частоты ошибок на обучении:

$$Q_d(\mu, X^L) = \frac{1}{N}\sum_{n=1}^{N} (\nu(\mu(X_n^l), X_n^k) - \nu(\mu(X_n^l), X_n^l)).$$

Решить данную проблему возможно используя механизм *k*-кратного скользящего контроля (k-fold cross validation), вычислительная сложность которого уменьшается за счет способа формирования разбиения выборки прецедентов. В работе [11] экспериментально показано, что функционал *k*-кратного скользящий контроля не менее эффективен, чем дорогостоящий функционал полного скользящего контроля, при $k = 10$.

Для поиска оптимальных значений параметров α и β был разработан экспериментальный комплекс в системе Matlab, схема которого представлена на рис. 5.

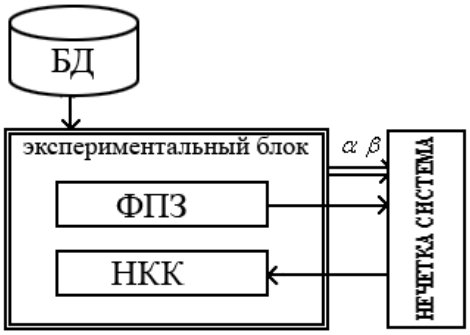


Рисунок 5 – Экспериментальная схема поиска оптимального решения

Экспериментальный блок состоит из двух базовых компонентов: ФПЗ и НКК. Первый отвечает за формирование всевозможных разбиений выборки прецедентов на обучающую и тестовую, а второй осуществляет вычисление функционалов контроля качество по результатам работы нечеткой системы на сформированных заданиях. В результате определяются оптимальные α и β по критерию:

$$(\alpha, \beta) = \arg\min_{\alpha \in A, \beta \in B} Q_c^{k-fold}, \qquad (3)$$

где множества A и B определяются диапазоны возможных значений параметров алгоритма. Для данной задачи $A = [0.01, 0.5]$ и $B = [0.5, 5]$.

В результате серии экспериментов для задачи распознавания ЖНП в качестве оптимальных параметров были определены $\alpha = 0.1$ и $\beta = 4$, что обеспечило 92% верно распознанных образцов.

Кластеризация и повышение уровня распознавания

Как уже отмечалось ранее, структура обучающей выборки может быть достаточно сложной. В рассматриваемой задаче в результате визуального анализа диаграмм рассеивания, функций принадлежности нечетких портретов, а также консультаций с экспертом было определено наличие кластеров в некоторых классах образов, в частности для класса v_4. Это обусловлено технологическими особенностями производства ЖНП. Для класса v_4 были выделены два кластера и в базе знаний, соответственно, было определено два правила для данного класса. На рис. 6 изображена ROC-кривая, для случая, когда в качестве позитивных примеров рассматриваются элементы класса v_4, а в качестве отрицательных – все остальные примеры.

Данная кривая ошибок показывает зависимость количества верно

классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров. В результате такой модификации алгоритма повысилось качество распознавания за счет верно распознанных элементов класса v_4 на 2%.

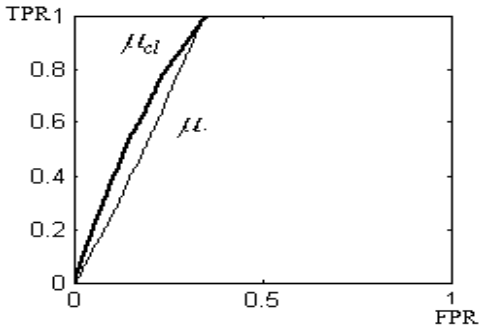


Рисунок 6 – ROC- кривая для класса v_4

Выводы

В работе была решена задача классификации ЖНП алгоритмом распознавания на базе нечетких портретов. На данном этапе в лаборатории контроля качества решение принимается профильными специалистами, однако предлагаемые подход позволит автоматизировать процесс и усовершенствовать работу лаборатории. Достоинством предложенного алгоритма является возможность интерпретации результата непосредственно экспертом в ситуациях, когда однозначного ответа не существует. Многие задачи в химической, пищевой промышленности могут быть сведены к описной выше модификации задачи распознавания образов.

Стабильность алгоритма обусловлена подходом к формирования нечетких портретов как интегральных характеристик классов образов.

В ситуации, когда возникает новый класс образов, которая может быть обусловлена появлением нового производителя, нечеткая система достаточно легко модифицируется за счет добавления нового правила.

В качестве нечетких портретов второго порядка и более высоких порядков рассматриваются нечеткие портреты построенные по парам признаков или группам признаков большей размерности. Их использование в сложных случаях позволит повысить качество распознавания. Также для повышения качества в дальнейших исследованиях планируется получить эффективную алгоритм для вычисления функционала полного скользящего контроля.

В качестве модификации основного алгоритма было выполнено разбиение на кластеры некоторых классов образов, что позволило улучшить качество распознавания на 2% по сравнения с исходным алгоритмом (92% верно распознанных объектов).

Литература

1. Прикладная статистика: классификация и снижение размерности [Справ. изд.] / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, А. Д. Мешалкин; ред. С.А. Айвязяна. – М.: Финансы и статистика, 1989. – 607 с.: ил.

2. Никифоров И.К. Нейросетевой импедансный метод и устройства идентификации и определения параметров жидких нефтепродуктов: автореф. на соиск. уч. степени канд. тех. наук / И.К. Никифоров. – Казань, 2005. – 19 с.

3. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений / Л. Заде ; ред. Н.Н. Моисеева, С.А. Орловского. – М.: Мир, 1976. – 168 с.

4. Козловский В.А. Решение задачи распознавания по нечетким портретам классов / В.А. Козловский, А.Ю. Максимова // Искусственный интеллект. – 2010. – №4. – С. 221-228.

5. Загоруйко Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Ин-та математики, 1999. – 270 с.

6. Афифи А. Статистический анализ. Подход с использованием ЭВМ / А. Афифи, С. Эйзен. – М.: Мир, 1982. – 289 с.

7. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов / К.В. Воронцов // Математические вопросы кибернетики. – 2004. – Вып. 13. – С.5-36.

8. Зайченко Ю.П. Нечеткие модели и методы в интеллектуальных системах: учебное пособие для студентов высших учебных заведений / Ю.П. Зайченко. – К.: «Издательский дом «Слово», 2008. – 344 с.

9. Каргин А.А. Введение в интеллектуальные машины. Интеллектуальные регуляторы / А.А. Каргин. – Донецк: Норд-Пресс, 2010. – К. 1. – 526 с.

10. Рутковская Д. Нейронные сети, генетические алгоритмы и нечеткие системы / Д. Рутковская, М. Пилиньский, Л. Рутковский ; перевод с польского И. Д. Рудинского. – М. : Горячая линия - Телеком, 2006. – 383 с.: ил.

11. Kohavi R.A. A study of cross-validation and bootstrap for accuracy estimation and model selection // IJCAI. – 1995. – Элек. ресурс http://citeseer.isu.psu.edu/kahavi95study.html .

Надійшла до редакції 20.03.2011