УДК 519.67

А.А.Викентьев[1], 2, В.В.Иванов[2]

[1] *Институт математики им. С.Л.Соболева СО РАН, Новосибирск, Россия;*
[2] *Новосибирский государственный университет, Россия*
(E-mail: vikent@math.nsc.ru)

Методы распознавания образов в пространстве знаний

В работе рассмотрены методы анализа данных. В новой постановке рассмотрена задача распознавания образов в пространстве знаний и предложены алгоритмы решения. Вклад Н.Г. Загоруйко и Г.С. Лбова в этом направлении огромен.

Ключевые слова: высказания, гипотеза, выборка, распознавание образов.

Светлой памяти профессоров
Г.С. Лбова и Н.Г. Загоруйко

Введение

Проблема распознавания образов уже давно привлекает внимание психологов, физиологов, инженеров и математиков. Методы распознавания образов находят применение в различных сферах деятельности человека: диагностика заболеваний, сельское хозяйство, добыча полезных ископаемых и многое другое.

Для решения проблемы распознавания образов необходимо проанализировать информацию, поступающую в виде «данных», «знаний» и других структур. Такой анализ включает в себя две процедуры: процедуру обнаружения закономерностей, содержащихся в предоставленной информации, и процедуру использования обнаруженных закономерностей для предсказания значения одной части информации по известным значениям другой её части. Исходная информация, поступаемая для анализа, чаще всего имеет вид числовых таблиц (матриц), состоящих из m строк и n столбцов. Строки $a_1, \ldots, a_i, \ldots, a_m$ отражают информацию об изучаемых объектах или явлениях, а столбцы $x_1, \ldots, x_j, \ldots, x_n$ отражают свойства (признаки, характеристики) этих объектов или явлений.

На пересечении i-ой строки и j-го столбца указывается значение (b_{ij}) j-го признака у i-го объекта. Например, i-ый дом имеет высоту 15 метров. Полные данные об i-ом объекте содержатся в совокупности всех элементов i-ой строки. Информация же обо всех заданных свойствах всех изучаемых объектов, записанная в таблице «объект-свойство», называется таблицей данных.

По мере изучения таблицы данных можно обнаружить некоторые закономерности. Например, в таблице, описывающей некоторые свойства домов (строительный материал, цвет и высоту), все панельные дома серого цвета имеют высоту 10-20 м, оранжевого цвета — 20-30 м, а кирпичные, любого цвета, — высоту меньше 10 м. Теперь, если обозначить признак «вид стройматериала» через x_1 ($_1$=п (панель) или x_1=к (кирпич)), признак «цвет» через x_2 (x_2=серый, или x_2=оранжевый, или x_2=любой) и признак «высота» через x_3 (x_3 принимает значение от 0 до 30 м), то обнаруженные закономерности можно записать в виде логических высказываний:

если (x_1=п) и (x_2=серый), то (x_3 = 10-20);
если (x_1=п) и (x_2=оранжевый), то (x_3 = 20-30);
если (x_1=к) и (x_2=любой), то (x_3<10).

Эти высказывания, не содержащие информации в виде конкретных характеристик каждого отдельного дома, отражают знания о некоторых обобщенных характеристиках всех домов, описанных в таблице данных. Таким образом, описывается переход от данных к знаниям, представляющим собой краткое описание основного содержания информации, представленной в таблице данных. В идеальном случае каждый образ представляется не обучающей выборкой конечного объема, а полным аналитическим описанием распределения всех существующих в природе объектов этого образа («генеральной совокупностью»). Но практически все реальные задачи распознавания отличаются от такого идеального случая самым важным свойством: отсутствием представления о генеральной совокупности изучаемых знаний. Этот недостаток восполняется той или иной эвристической гипотезой. Одной из наиболее известных гипотез является гипотеза компактности.

Гипотеза компактности Загоруйко (H)

Гипотеза компактности состоит в следующем: реализации одного и того же хорошо организованного образа обычно отражаются в признаковом пространстве в геометрически близкие точки, образуя «компактные» сгустки. При всей кажущейся тривиальности и лёгкости опровержения указанная гипотеза лежит в основании большинства алгоритмов не только распознавания, но и многих других задач анализа данных. Конечно, она подтверждается не всегда. Если, например, среди признаков имеется много случайных, не информативных, то такой случай соответствует плохой организации образов, и точки одного и того же образа могут оказаться далекими друг от друга. Но дополнительно предполагается, что работа по организации образов уже проведена и в многомерном признаковом пространстве найдено такое («информативное») подпространство, в котором точки одного класса действительно образуют явно выделяемые компактные сгустки.

Назовем n признаков, входящих в информативное подмножество X, «описывающими», а номинальный $(n+1)$-й признак z, указывающий имя образа, «целевым». Обозначим множество объектов обучающей выборки через A, новый распознаваемый объект через q, а тот факт, что объекты множества A «компактны» («эквивалентны», «похожи» или «близки» друг другу) в пространстве n характеристик X, через $C(X/A)$. Мера «компактности» может быть любой. Например, можно считать, что объекты компактны, если Евклидово расстояние между векторами их признаков не превышает величину r. Фактически гипотеза H равнозначна предположению о наличии закономерной связи между признаками X и z, и ее тестовый алгоритм может быть представлен следующим выражением:

$$if\,[C\,(X, z/A)\,\&C\,(z/A, q)]\,.$$

Т.е. если объекты множества A компактны в совместном пространстве X, z и объекты множества A, q компактны в пространстве описывающих свойств X, то объекты A и q будут компактными и в пространстве целевого признака z.

Сформулированное выше условие компактности для решения задачи распознавания образов является необходимым, но не достаточным. Мало того, чтобы объекты образа A были близкими друг другу, нужно ещё, чтобы объекты образа B не оказались к ним такими же близкими, т.е. нужно, чтобы сгустки объектов разных образов не налагались друг на друга, что будем обозначать так: C – с учетом этого, гипотезу компактности H для распознавания образов можно записать в следующем виде:

$$if\,[C\,(X/A, B)\,\&C\,(A, z/A)\,\&C\,(X/A, q)]\,.$$

Если предполагать, что реализации одного и того же образа образуют один компактный сгусток, то его можно аппроксимировать унимодальным распределением. Этот случай соответствует гипотезе унимодальной компактности (Hu).

Ослабленный вариант обсуждаемой гипотезы (гипотеза полимодальной компактности(H_p) утверждает, что точки одного образа могут образовывать не один, а несколько компактных сгустков. На этом основании можно представлять образ многосвязными областями или смесью нескольких простых распределений.

Полимодальную компактность можно в пределе представить в виде локальной компактности (H_l), она выражает осторожное утверждение о свойстве ближайшего соседства: «обо всём распределении судить не берусь, но в некоторой малой ϵ-окрестности каждой реализации обучающей выборки образа i может появиться только представитель этого же образа». На указанном основании построить общую модель нельзя, но можно построить правило распознавания с опорой на все или часть объектов обучающей выборки (т.е. с опорой на прецеденты).

Остаётся добавить, что описанные выше гипотезы применяются не только при распознавании в пространстве данных, но и в пространстве знаний.

Основная задача алгоритмов распознавания

Пусть даны:

1) элементы алфавита $S = \{s_i\}$ (алфавит — список наименований фиксированных областей, на которые разделено выборочное пространство);

2) конкретные представители этих элементов в виде обучающей выборки

$$D_0 = \{di0\}\ i = 1, \ldots, k; k \geq 2;$$

3) система признаков $X = \{xl\}\ l = 1, \ldots, n;\ n \geq 1;$

4) величина допустимых затрат N_0, складывающихся из стоимости ожидаемых потерь $N(R)$ (R — вероятность ошибок) и стоимости реализации процедурных элементов N(A,X,S).

Требуется найти такую решающую функцию $A = \{ah\}\ h = 1, \ldots, n;\ n \geq$, при которой бы достигался минимум затрат N при ограничениях S, X и $minN \geq N_0$.

В процессе обучения алгоритму не сообщается никаких конкретных сведений о контрольной выборке. Предполагается лишь в соответствии с некоторой гипотезой о характере закономерности в структуре обучающей выборки и связи между обучающей и контрольной выборками построить такое решающее правило, которое обеспечивало бы распознавание обучающей выборки с минимальными ошибками. Если эти гипотетические закономерности и связи (обозначим их через H) соответствуют действительным (H_0), то такая решающая функция минимизирует ошибки распознавания и контрольной реализации. Выбор вида решающей функции A осуществляется из конечного набора. Если функциям, входящим в этот набор, присвоить номера (j=1...J), то задача поиска наилучшего вида решающей функции сводится к поиску номера j функции Aj, которая удовлетворяет условию

$$j = arg\ \underset{j \in J}{min} N\left(A_j\right) | S, X, D_0, H, N_0.$$

Описание алгоритмов, используемых в распознавании данных

В распознавании образов применяется огромное количество различных алгоритмов, выбор которых зависит от размера обучающей выборки. Если она велика, то можно опереться на модели, т.е. аппроксимировать эти сгустки распределениями того или иного типа и затем использовать строгие статистические методы. В противном случае единственное, что остается делать, — опереться на прецеденты, т.е. на свойства конкретных объектов из обучающей выборки. При опоре на статистические модели решающие правила могут иметь простую форму плоскостей или поверхностей второго порядка, разделяющих пространство признаков на k непересекающихся

областей. Распознаваемый объект q считается принадлежащим тому образу, в область которого он попадает. Иногда область образа ограничивается замкнутой фигурой простой формы (гиперпараллелепипедом или гиперсферой). Если граница между образами имеет сложную форму, то целесообразно пользоваться алгоритмом «Дробящиеся эталоны».

Дробящиеся эталоны

Этот метод отражает стремление к безошибочному распознаванию обучающей выборки, но с использованием покрытий обучающей выборки каждого образа простыми фигурами, усложняющимися по мере необходимости. Один из вариантов этого метода предусматривает использование в качестве покрывающих фигур набора гиперсфер. Для каждого из k образов строится сфера минимального радиуса, покрывающая все его обучающие реализации. Значения радиусов этих сфер и расстояний между их центрами позволяют определить образы, сферы которых не пересекаются со сферами других образов. Такие сферы считаются эталонными (см. образ 1 на рис.1), а их центры и радиусы запоминаются в качестве «эталонов первого поколения».

Если два образа пересекаются, но в области пересечения не оказалось ни одной реализации обучающей выборки, такое пересечение считается фиктивным, центры и радиусы этих сфер также вносятся в список эталонов первого поколения. При этом область пересечения считается принадлежащей сфере с меньшим радиусом (сфера 2 на рис.1). Если в зоне пересечения оказались точки только одного образа, то эта зона считается принадлежащей этому образу (сфера 3 на рис.1). Точка будет считаться относящейся к образу 4, если она попадает в сферу 4 и не попадает в сферу 3. Если же область пересечения содержит точки разных образов, то для этих точек строятся «эталоны второго поколения» (сферы 4–6). Если и они пересекаются, то для точек из зоны пересечения строятся «эталоны третьего поколения». Процедура дробления эталонов продолжается до получения заданной надежности распознавания обучающей последовательности. При распознавании новых реализаций процесс проверки попадания в ту или иную эталонную сферу начинается со сфер наименьшего диаметра.

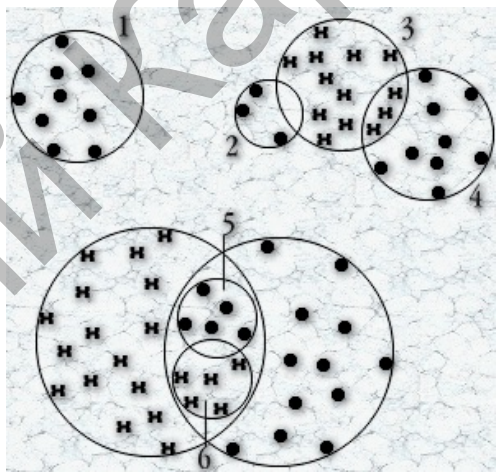

Рисунок 1. Эталоннные сферы

Если число распознаваемых образов велико, то следует применять алгоритмы «МПР» (Метод попутного разделения) или «МПВ» (Метод покоординатного вычеркивания).

МПР

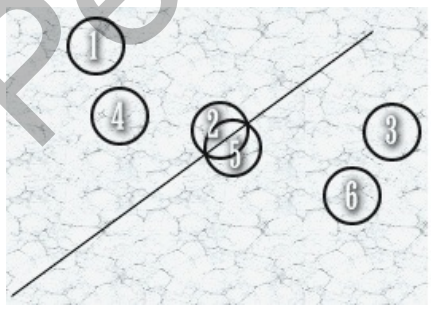

Рисунок 2. Разделение образов

Если построить оптимальную разделяющую границу между образами i и j, то можно обнаружить, что эта граница попутно разделяет и другие некоторые пары образов с ошибкой R, не превышающей допустимого значения. Так, решающая граница (см. рис.2), построенная для разделения образов 2 и 5, может использоваться также для разделения образов 1 и 4 от образов 3 и 6.

Процесс нужно начинать с разделения двух наиболее трудно разделяемых образов. Отмечаются все пары

образов, успешно разделяемые построенной границей. Эти пары из списков пар исключаются. Среди пар, оставшихся неразделенными, снова выбирается пара самых трудно разделяемых, для которой строится следующая разделяющая граница. Такие шаги повторяются до полного исчерпания списка неразделенных пар. При удачном расположении образов число h разделяющих границ может оказаться значительно меньше, чем число распознаваемых образов. Для каждой границы фиксируется список тех k_1 образов, которые находятся по одну сторону от нее, и тех k_2 образов, которые находятся по другую сторону. Будем характеризовать «информативность» границы величиной $I = (1 - |k_1 - k_2|/k)$, которая принимает значения в диапазоне от 0 до 1 и равна 1, если граница делит множество образов на две равные части. Процесс распознавания с помощью построенных границ состоит в следующем. Начиная с самой информативной границы, определяется, по какую сторону от нее находится контрольная точка q, и вычеркиваются из списка претендентов все образы, находящиеся с противоположной стороны от границы. С помощью следующей по информативности границы список оставшихся претендентов сокращается снова. Так продолжается до тех пор, пока в списке не останется один-единственный образ, к которому и относится распознаваемая точка q.

МПВ

Представим себе, что мы считаем допустимыми потери типа «пропуск цели» равными R_0. Такие потери могут возникать тогда, когда разделяющая граница проходит на расстоянии d между контрольной точкой q и математическим ожиданием $*i$ образа i. Если контрольная точка q удалена от математического ожидания i-го образа на расстояние, большее d, то можно считать, что она не принадлежит образу i. Свой вклад в это расстояние вносит каждая координата пространства признаков. Если хотя бы по одной l-й координате расстояние $r(q_l, *i_l)$ будет для образа i равно или больше d, то i-тый образ из списка конкурентов можно вычеркнуть. На этом соображении основан метод покоординатного вычеркивания, который состоит в следующем. Рассматриваются проекции точки q и распределений всех образов на каждую координату в отдельности. По первой (l-той) координате определяются расстояния r между точкой q_l и математическими ожиданиями $*il$ всех k образов. Те образы, для которых выполняется условие $r(q_l, *i_l) >= d$, из списка претендентов на включение точки q в свой состав исключаются. Для оставшихся образов та же процедура повторяется с использованием проекции на вторую координату, и это продолжается до тех пор, пока в списке претендентов не останется заданное число k_1 образов. Для этих самых сильных претендентов вычисление расстояний и оценка ожидаемых потерь осуществляются в исходном n-мерном пространстве с использованием оптимальных решающих правил.

При использовании прецедентов(т.е. типичных представителей) объект q сравнивается с каждым из прецедентов и относится к тому образу, чей прецедент (представитель) (или прецеденты (в алгоритме к-ближайших соседей)) оказался самым похожим на знание q.

Рассмотренные алгоритмы позволяют производить распознавание образов в пространстве данных. Новая постановка задачи

Как говорилось выше, при обработке таблиц данных можно выделить некоторые закономерности, которые можно записать в виде логических высказываний далее называемых знаниями. При увеличении числа знаний возникает потребность в анализе этих знаний. В частности, допустим, задана некоторая структурированная база знаний, на вход которой подаётся некоторое новое знание q. Требуется определить, к какому из имеющихся k таксонов (поименованных областей, содержащих элементы, похожие друг на друга по каким-то характеристикам) следует отнести это новое знание, т.е. получаем задачу распознавания образов.

Постановка задачи. Пусть в пространстве знаний заданы:

1. Набор характеристик X.

2. Список наименований фиксированных областей (таксонов, называемых также образами), на которые разделено выборочное пространство $S = s_i$, $i = 1 \dots I$.

3. Обучающая выборка в виде знаний экспертов D_{0i} (в пространстве X) для каждого S_i.

4. Контрольное знание q.

Требуется определить номер i S_i : $q \geq S_i$, используя алгоритм k-ближайших соседей по прецедентам (т.е. типичным представителям каждого образа)

$$i = arg \min_{i \in I} \sum_{k=1}^{K} Rik/K|S, Do, X, k, R,$$

где Rik — k-минимальные расстояния от q до M знаний для каждого таксона; R — ошибка распознавания. Т.е. находятся расстояния от контрольного знания до каждой реализации каждого образа, выбираются k-минимальные расстояния, определяются средние (для каждого образа), среди которых находится минимальное, и таким образом восстанавливается номер таксона, которому принадлежит контрольное знание.

Для решения поставленной задачи была написана компьютерная программа. Кроме того, в программе рассмотрен алгоритм, реализованный ранее, отличие которого от рассмотренного здесь заключается в использовании для определения i эталонных знаний, создаваемых для каждого образа:

$$i = arg \min_{i \in I} R_i,$$

где R_i — расстояние от q до Eti (эталонного знания i-го образа).

При создании искусственного эталона для каждого предиката в отдельности находится его «среднее распределение», т.е. такое, расстояние от которого до соответствующих предикатов всех знаний данного образа было бы минимальным. С этой целью суммируются плотности вероятностей для каждой из T градаций данного предиката всех mi знаний I-го образа и затем полученные суммы делятся на mi. В результате получится нормированное распределение предиката. Его плотность в каждой градации представляет собой среднеарифметическое значение плотностей этой градации у всех знаний образа, что обеспечивает минимум суммы расстояний от этого «центрального» предиката до соответствующих предикатов всех знаний этого образа.

Метрика в пространстве знаний

При решении задачи проблемой являлось сравнение знаний. Для решения данной проблемы необходимо измерить меру похожести знаний или ввести метрику для измерения расстояния в пространстве знаний. Такая метрика рассмотрена в [1],[2, 10], что было использовано при вводе метрики данной работы.

Можно считать, что каждая характеристика отражает знание эксперта о распределении возможных значений данной характеристики. Если два эксперта одной и той же характеристике xi приписывают одинаковый диапазон значений, расстояние между мнениями экспертов равно нулю. Если же один из них считает, что $x_j = x_{j,max}$, а второй — что $x_j = x_{j,min}$, то расстояние между их мнениями максимально и можно считать равным 1. Эксперты могут указывать диапазон значений x_j и вообще любые распределения допустимых значений в пределах от $x_{j,min}$ до $x_{j,max}$. Следовательно, задача оценки расстояния между мнениями экспертов сводится к задаче поиска расстояния между двумя распределениями.

Предложенная мера для измерения этого расстояния $R = f(r*h*w)$ учитывает расстояние r от всех элементов одного распределения до всех элементов другого, энтропийную меру h, близкую

по смыслу k дисперсии распределений, и степень w пересечения распределений (величину области «консенсуса»). Эти аргументы вычисляются так.

Пусть характеристика x имеет конечное число (n) градаций. Если первый эксперт указывает в качестве градаций $1, 2, \ldots, \alpha, \ldots, A$ с вероятностями $P\alpha(\Sigma P\alpha = 1\alpha = 1, \ldots, A)$, а второй эксперт допускает градации $1, 2, \ldots, \beta, \ldots, (\Sigma P\beta = 1\beta = 1, \ldots, B)$, то расстояние между этими распределениями можно определить следующим образом [2]:

$$r = \frac{\sum\limits_{\alpha=1}^{A}\sum\limits_{\beta=1}^{B}|X_\alpha - X_\beta| \cdot (P_\alpha + P_\beta)}{\sum\limits_{\alpha=1}^{A}\sum\limits_{\beta=1}^{B}(P_\alpha + P_\beta) \cdot (X_{\min} - X_{\max})}.$$

Принимается, что чем больше область пересечения двух распределений (область консенсуса), тем меньше расстояние R. Определяется w величина, связанная с областью консенсуса [1]:

$$w = \sum_{t=1}^{T}|P_{1t} - P_{2t}|,$$

где T — число делений, равномерно распределённых вдоль оси X (ось X отображает мнение эксперта о распределении характеристики, в программе T равно разбиению); а P_{1t} и P_{2t} — указанные экспертами вероятности попадания оценок в t–ю градацию.

Расстояние между суждениями экспертов зависит и от категоричности (h) их оценок. При одних и тех же расстояниях r и w R считается тем больше, чем больше распределения отличаются от равномерного распределения по всему диапазону значений от x min до x max . Величина h находится следующим образом [1]:

$$h = 0.5\,(h_1 + h_2)\,, h_1 = 0.5\sum_{t=1}^{T}|P_{1t} - 1/T|,$$

$$h_2 = 0.5\sum_{t=1}^{T}|P_{2t} - 1/T|.$$

Общая мера расстояния между двумя распределениями $R = r \times h \times w$. Общее же расстояние между двумя знаниями вычисляется по формуле

$$R_0 = \frac{\sum\limits_{j=1}^{m}\gamma_j \cdot R_j}{\sum\limits_{j=1}^{m}\gamma_j},$$

где γ_j — весовой коэффициент, характеризующий относительную «информативность», «важность» характеристики X_j.

Получена машинная реализация алгоритмов, что будет рассмотрено в следующей работе. Другие методы получения метрик и подходы по информативности рассмотрены в [4–14].

Заключение

В данной работе в новой постановке рассмотрена задача распознавания образов в пространстве знаний, в виде программы реализованы алгоритм к-ближайших соседей, позволяющий решить

данную задачу, и ранее рассмотренный алгоритм сравнения по эталонам. Заданы обучающие выборки, проведено распознавание знаний и подтверждена связь между характером распределений и правильностью работы алгоритмов, в случае унимодальных распределений оба алгоритма распознают практически одинаково, а в случае же полимодальных сравнение по эталонам даёт больше ошибок. Проведённое моделирование показывает возможность дальнейшего использования работающей программы.

Работа осуществлена при финасовой поддержке грантов РФФИ 14-07-00851а, 14-07-002490а.

Список литературы

1 *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Изд-во Ин-та мат., 1999.

2 *Загоруйко Н.Г., Бушуев М.В.* Меры расстояния в пространстве знаний // Анализ данных в экспертных системах. — 1986. — Вып. 117. Вычислительные системы. — С.24–35.

3 *Загоруйко Н.Г., Ёлкина В.Н., Лбов Г.С.* Алгоритмы обнаружения эмпирических закономерностей. — Новосибирск: Наука, 1985.

4 *Викентьев А.А.* Мера опровержимости высказываний экспертов, расстояния в многозначной логике и процессы адаптации // XIV International Conference «Knowledge-Dialogue-Solution» KDS 2008. — Varna, Bulgaria, 2008. — С. 179– 188.

5 *Миркин Б.Г.* Методы кластер-анализа для поддержки принятия решений: обзор. — М.: Изд. Дом ВШЭ, 2011. — 88 с.

6 *Викентьев А.А., Кабанова Е.С.* Расстояние между формулами пятизначной логики Лукасевича и мера недостоверности высказываний экспертов в кластеризации // Материалы междунар. науч. конф., посвящ. памяти и 70-летию проф. Т. Г. Мустафина. — Караганда, 2012. — С. 28—29.

7 *Викентьев А.А., Викентьев Р.А.* Расстояния и меры недостоверности на высказываниях n-значной логики // Вестн. НГУ. Сер. Мат., механика, информатика. — Новосибирск: Изд-во НГУ, 2011. — Т. 11. — Вып. 2. — С. 51–64.

8 *Викентьев А.А., Кабанова Е.С.* Расстояние между формулами пятизначной логики Лукасевича и мера недостоверности высказываний экспертов // Вестн. Караганд. ун-та. Сер. Математика. — 2013. — №1 (69). — С. 18–27.

9 *Викентьев А.А.* О возможных расстояниях и степенях недостоверности в многозначных высказываниях экспертов и приложение этих понятий в проблемах кластеризации и распознавания // Проблемы информатики. — Новосибирск: СО РАН, 2011. — №3 (11). — С. 33–45.

10 *Vikent'ev A. A., Lbov G. S.* Setting the metric and informativeness on statements of experts // Pattern Recognition and Image Analysis. — 1997. — Vol. 7. — No. 2. — P. 175–183.

11 *Викентьев А.А.* О возможных расстояниях и степенях недостоверности в многозначных высказываниях экспертов и приложение этих понятий в проблемах кластеризации и распознавания // Проблемы информатики. — Новосибирск: СО РАН, 2011. — №3 (11). — С. 33–45.

12 *Лбов Г.С.* Анализ данных и знаний. — Новосибирск: Изд. НГУ, 2013.

А.А.Викентьев, В.В.Иванов

Білімдер кеңістігінде бейнелерді тану әдістері

Мақалада берілгендердің талдау әдістері қарастырылған. Білім кеңістігінде бейнелерді тану есебі жаңа қойылымда қарастырылды және шешімдер алгоритмі ұсынылды. Осы бағытта Н.Г.Загоруйко мен Г.С. Лбовтың ықпалы зор.

A.A.Vikentiev, V.V.Ivanov

Methods of recognition in the space of knowledge

The paper discusses methods of data analysis. In the new statement of the problem of image recognition in the area of knowledge and solutions offered algoritmy. Effect of N.G. Zagoruiko and G.S.Lbov in this impact of huge.

References

1 Zagoruiko N.G. *Applied methods of data analysis and knowledge*, Novosibirsk: Publ. House of the Institute of Mathematics, 1999.

2 Zagoruiko N.G., Bushuyev M.V. *Data analysis expert systems*, 1986, 117: Computer systems,p.24–35.

3 Zagoruiko N.G., Yolkina V.N., Lbov G.S. *Detection algorithms empirical regularities*, Novosibirsk: Nauka, 1985.

4 Vikent'ev A.A. *XIV International Conference «Knowledge-Dialogue-Solution» KDS 2008*, Varna, Bulgaria, 2008, p. 179–188.

5 Mirkin B.G. *Methods of cluster analysis for decision support: a review*, Moscow: Ed. Home Higher School of Economics, 2011, 88 p.

6 Vikent'ev A.A., Kabanova E.S. *Materials Intern. scientific. conf., is dedicated. memory and 70th anniversary of prof. T.G.Mustafin*, Karaganda, 2012, p. 28—29.

7 Vikent'ev A.A., Vikent'ev R.A. *Bull. of University of Novosibirsk, Ser. Mathematics, Mechanics and Computer Science*, Novosibirsk: Publ. NSU, 2011, 11, 2, p. 51–64.

8 Vikent'ev A.A., Kabanova E.S. *Bull. of University of Karaganda, Ser. Mathematics*, 2013, 1(69), p. 18–27.

9 Vikent'ev A.A. *Informatics Problems*, Novosibirsk: SB RAS, 2011, 3 (11), p. 33–45.

10 Vikent'ev A.A., Lbov G.S. *Pattern Recognition and Image Analysis*, 1997, 7, 2, p. 175–183.

11 Vikent'ev A.A. *Informatics Problems*, Novosibirsk: SB RAS, 2011, 3 (11), p. 33–45.

12 Lbov G.S. *Analysis of data and knowledge*, Novosibirsk: Publ. NSU, 2013.