УДК 519.85 + 681.3

ОБ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ОБЪЕКТОВ И РАСПОЗНАВАНИИ ОБРАЗОВ С ИСПОЛЬЗОВАНИЕМ ВЕСОВ ПРИЗНАКОВ И РЕПРЕЗЕНТАТИВНОСТЕЙ КЛАССОВ

Е.М. МАМАТОВ

Белгородский государственный университет
e-mail: : Mamatov@bsu.edu.ru

В статье рассматривается программно-алгоритмическая информационная технология, использующая созданный в рамках данной работы вариационный алгоритм автоматической классификации объектов с использованием, предложенного авторами работы, нового функционала качества разбиения, основанного на мере однородности. Также в рамках данной работы рассматривается реализация алгоритма вычисления оценок с использованием весов признаков и репрезентативностей классов

Ключевые слова: автоматическая классификация объектов, распознавание образов, алгоритмы вычисления оценок.

При решении задач, связанных с необходимостью проведения начального анализа данных, получаемых в результате проведения вычислительных экспериментов или в результате наблюдения различных процессов и явлений, используют методы и алгоритмы классификации объектов и распознавания образов. Большинство таких методов и алгоритмов реализовано в информационных системах классификации и распознавания объектов, необходимым элементом которых является человек (исследователь). Основная функция исследователя заключается в управлении процессом работы алгоритмов, то есть задание и корректировка свойств классов (границы, максимально возможное количество классов, выделение одного объекта в отдельный класс как прецедент, и.т.д.), что приводит этот вид систем к классу автоматизированных.

На рис.1 отображена схема функциональной структуры программно-алгоритмической системы классификации и распознавания образов, которая решает следующие задачи:

1) задача обработки входной информации;
2) задача классификации объектов;
3) задача распознавания образов;
4) задача оценки устойчивости работы алгоритмов;
5) задача визуализации результатов работы алгоритмов.

Структура выполнена в виде SADT диаграммы.

На рис.1 отражена подсистема, выполняющая функцию оценки устойчивости работы алгоритмов, получая выходные данные которой исследователь имеет возможность принимать решение о дальнейшем использовании результатов работы алгоритмов классификации объектов и распознавания образов.

Подсистема обработки входной информации выполняет следующие функции:

1) обработки информации с клавиатуры и занесение ее в таблицу свойств объектов;
2) приема с клавиатуры параметров классификации объектов и распознавания образов;
3) открытия и обработки таблицы свойств объектов из *dbf* –файла;
4) выбора признакового пространства из наиболее информативных признаков.

Подсистема визуализации результатов работы алгоритмов выполняет функцию подготовки отчета, который содержит информацию о распределении объектов по классам в табличном виде, если количество признаков больше двух. Если количество признаков объектов равняется двум, то отчет может содержать точечный график рас-

пределения объектов по классам, при чем объекты отдельного класса окрашиваются в определенный цвет.

Подсистема классификации объектов содержит автоматическую процедуру, реализующую вариационный алгоритм.

При реализации автоматических процедур классификации возникает необходимость количественной оценки (критерия) качества разбиения [1,2].

Такой критерий по необходимости должен учитывать много факторов, которые описываются на эвристическом уровне с использованием вербальных (качественных) моделей [3].

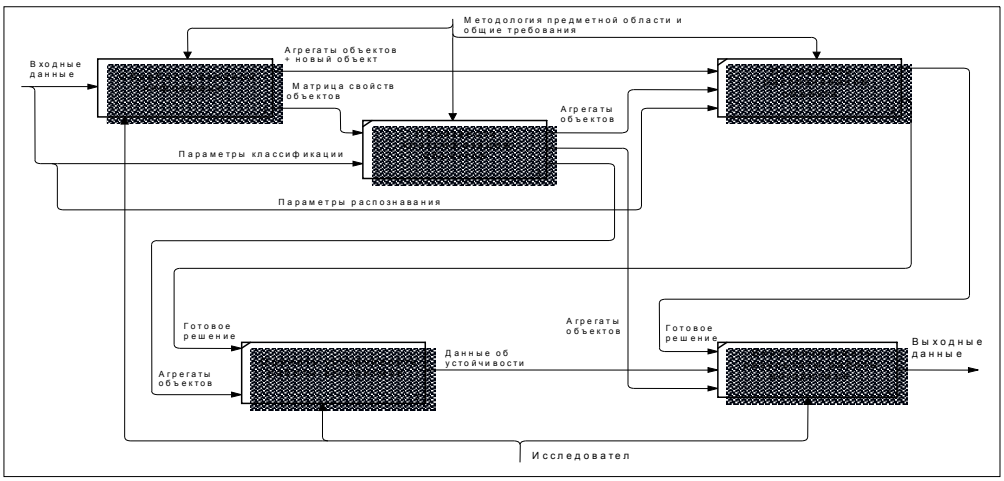

Рис. 1. Схема функциональной структуры программно-алгоритмической системы классификации и распознавания образов

В работе [4] отмечается важность однородности (равномерности) разбиения исходного множества на классы как в смысле отсутствия скачков “плотности” элементов внутри каждого класса, так и в смысле примерного равенства количества элементов в каждом классе.

В рамках настоящей работы при разбиении элементов предлагается кроме этого обеспечивать равномерность в смысле однородности расстояний между геометрическими центрами тяжестей классов и однородности максимальных расстояний между объектами одного и того же класса (размеров классов).

Уточним постановку задачи классификации. Пусть исходное множество содержит M элементов, которые необходимо разбить на K классов. В дальнейшем M_q означает мощность q-ого подмножества, так что

$$\sum_{q=1}^{K} M_q = M \qquad (1)$$

Для характеристики однороднос.и разбиения введем функционал

$$L = \frac{D \cdot H}{G \cdot R}, \qquad (2)$$

где D - мера однородности расстояний между центрами тяжестей классов; H - мера однородности количества элементов в классах; G - мера однородности расстояний между элементами одного и того же класса; R - Мера однородности максимальных расстояний между объектами одного и того же класса.

Максимальное значение функционала (2) будет соответствовать наилучшей степени качества разбиения исходного множества на подмножества.

Поэтому для меры D предлагается на основе работы [5] использовать представление вида

$$D = -\frac{\sum_{q=1}^{K}\sum_{l=q+1}^{K}\mu_{q,l}\cdot \mathrm{Ln}(\mu_{q,l})}{\mathrm{Ln}(K(K-1)/2)}, \quad (3)$$

где

$$\mu_{q,l} = \frac{Y_{q,l}}{\sum_{i=1}^{K}\sum_{j=i+1}^{K} Y_{i,j}}, \quad (4)$$

где $Y_{q,l}$ - расстояние между геометрическими центрами тяжестей q - ого и l - ого классов.

Будет иметь место Dmax = 1 когда все расстояния между геометрическими центрами тяжести классов будут равны, и следовательно будет иметь место Dmin = 0 когда К=1.

Для меры H предлагается использовать представление

$$H = -\frac{\sum_{q=1}^{K} m_q \cdot \mathrm{Ln}(m_q)}{\mathrm{Ln}(K)}, \quad (5)$$

где $m_q = \frac{M_q}{M}$.

Будет иметь место Hmax = 1 когда количества элементов в классах будут равны, и следовательно будет иметь место Hmin = 0 когда в одном классе будут содержаться все элементы исходного множества , а в остальных ни одного.

Мера G определяется соотношением

$$G = 1 + \frac{1}{K}\sum_{q=1}^{K}\left(\frac{\sum_{i=1}^{M_q-1}\rho_{iq}\cdot \mathrm{Ln}(\rho_{iq})}{\mathrm{Ln}(M_q-1)}\right), \quad (6)$$

где

$$\rho_{iq} = \frac{r_{iq}}{R_q}, \quad (7)$$

$$\sum_{i=1}^{M_q-1} r_{iq} = R_q, \quad (8)$$

R_q - общая длина внутренних ребер q-го подмножества, а r_{iq} - длина i-ого ребра в q-ом подмножестве (i=1,...,M_q-1). Ребра получаются путем построения минимального остового дерева для каждого класса.

Ввиду того, что G находится в знаменателе функционала будет иметь место Gmax = 1 при максимальной неоднородности внутриклассовых расстояний и Gmin =0 когда однородность внутриклассовых расстояний буде наилучшей.

Мера R по аналогии с мерой G имеет представление

$$R = 1 + \frac{\sum_{q=1}^{K}\lambda_q \cdot Ln(\lambda_q)}{Ln(K)}, \quad (9)$$

где

$$\lambda_q = \frac{R_{max,q}}{\sum_{i=1}^{K} R_{max,i}}, \quad (10)$$

где $R_{max,q}$ - максимальное из расстояний между самыми дальними элементами q - ого класса .

Будет иметь место Rmin = 0 когда максимальные расстояния между самыми дальними элементами каждого из классов будут равны, и следовательно будет иметь место Rmax = 1 когда в одном классе будут содержаться все элементы исходного множества, а в остальных ни одного.

На практике предельные случаи мер (3), (5), (6), (9) при большом количестве объектов в исходном множестве встречаются довольно редко.

Так как наилучшему разбиению исходного множества объектов будет соответствовать максимальное значение функционала (2), то для реализации процедуры классификации необходим вариационный алгоритм, основанный на разрезании графа.

Исходные данные данного алгоритма, как и для многих алгоритмов классификации, представляются в виде таблицы «Объекты-свойства» (ТОС).

Выходные данные алгоритма представляют собой структурированную таблицу «Объекты – свойства», то есть с указанием принадлежности каждого объекта к одному из классов.

Таким образом, на основе исходной таблицы «Объекты – свойства» вычисляется матрица расстояний с использованием расстояния Евклида.

На следующем этапе по матрице расстояний осуществляется объединение элементов друг с другом по принципу ближайшего соседа. В этом случае на исходном множестве при помощи алгоритма Р.Прима строится кратчайший незамкнутый путь (КНП) или, по-другому, минимальное остовое дерево. Минимальное остовое дерево представляет собой взвешенный граф без петель, вершинами которого являются агрегируемые элементы, а ребра проведены только между ближайшими относительно друг друга элементами. В результате КНП соединяет все элементы исходного множества, и при этом сумма длин входящих в КНП ребер является минимальной из всех возможных. На рис.2 представлен результат работы алгоритма Р.Прима в двухмерном признаковом пространстве.

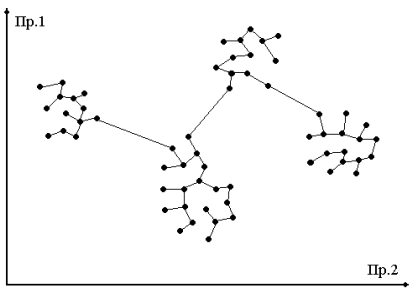

Рис. 2. Результат работыалгоритма Р. Прима

Далее необходимо предусмотреть оценку граничного случая, при котором все объекты принадлежат одному классу, то есть все объекты в заданном признаковом пространстве находятся друг от друга на одинаковых расстояниях.

В таких случаях все ребра КНП равны. Для этого следует вычислить и запомнить значение функционала L (2) при D=H=R=1, а значение меры G вычисляется по выражению

$$G = 1 + \frac{\sum_{i=1}^{M-1} \rho_i \cdot Ln(\rho_i)}{Ln(M-1)}, \quad (11)$$

$$\rho_i = \frac{r_i}{R_q}, \tag{12}$$

$$\sum_{i=1}^{M} r_i = R_q, \tag{13}$$

где R_q - общая длина внутренних ребер КНП, а r_i - длина i-ого ребра в КНП (i=1,..,M-1).

Следующим этапом работы алгоритма будет поиск и «разрезание» в КНП самого длинного ребра. Таким образом, мы получим из исходного КНП два поддерева, соответствующих двум классам (см. рис.3).

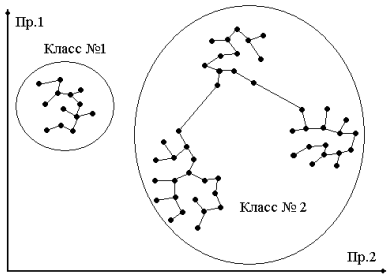

Рис. 3. Два поддерева, соответствующие двум классам

На основе полученного разбиения вычислим функционал L (2), содержащий компоненты D, H, R и G.

Мера D вычисляется согласно выражений (3) и (4), где $Y_{1,2}$ определяется как расстояние между геометрическими центрами тяжестей между 1 и 2 классами (см. рис. 4). На данном шаге мера D будет равна нулю, так как расстояние между классами будет одно (приграничный случай). Поэтому в данном случае положим D=1 для исключения влияния этой меры на значение функционала.

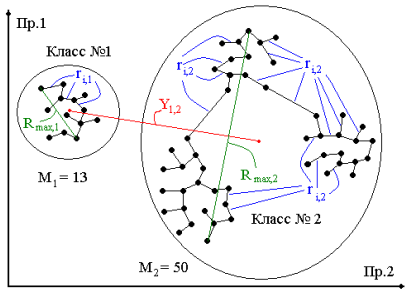

Рис. 4. Данные для вычисления функционала

Мера H вычисляется с помощью выражения (5) при известных M_1 и M_2 (см. рис. 4).

Мера R вычисляется согласно выражения (9) при известных $R_{\max,1}$ и $R_{\max,2}$.

Мера G определяется согласно выражений (6), (7) и (8).

Таким образом, вычисленное значение функционала L сравнивается с предыдущем значением, и если настоящее значение больше предыдущего то принимается разбиение на два поддерева, а если нет, то происходит перебор максимальных ребер и разбиение на поддеревья до тех пор, пока не увеличиться значение функционала качества разбиения. Если увеличение функционала не произошло, то происходит отказ от разбиения и алгоритм прекращает свою работу.

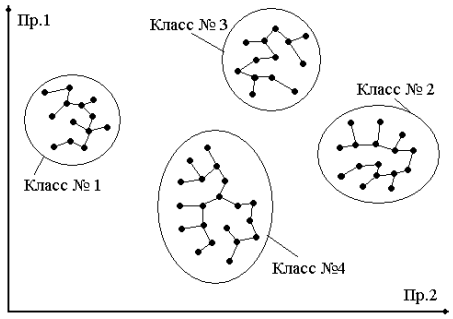

Рис. 5. Результат работы алгоритма

Следующим шагом алгоритма будет поиск среди поддеревьев максимального ребра и его последующее «разрезание» с оценкой функционала качества разбиения.

Алгоритм прекратит свою работу после того как будет найден глобальный максимум функционала и последующие попытки разрезания ребер поддеревьев не приведут к увеличению значения L. Для КНП, представленного в качестве примера, максимальное значение функционала качества разбиения будет соответствовать ситуации изображенной на рис. 5

Разработанные в рамках настоящей работы вариационный алгоритм автоматической классификации и новый функционал качества классификации объектов, как показали проведенные эксперименты, наилучшим образом классифицируют исходные множества объектов с точки зрения человеческих предпочтений.

Данный алгоритм был реализован с помощью среды программирования DELPHI 7 и зарегистрирован в Отраслевом фонде алгоритмов и программ.

Подсистема распознавания образов. В общем случае задачей распознавания образов является задача отнесения объекта исследования, характеризующегося вектором значений признаков, к одному из априорно заданных классов объектов, существующих в некотором признаковом пространстве.

Одним из этапов решения всех видов задач распознавания образов является формирование признакового пространства, то есть его качественного состава и размерности. О необходимости формирования достаточно информативного словаря признаков излагается в работе [6]. Признаковое пространство должно подбираться таким образом, чтобы каждый признак обладал достаточной для решения задачи разделительной способностью при как можно меньшей размерности данного пространства. Уменьшение размерности признакового пространства при сохранении его различительной способности в целом необходимо для осуществления реализации алгоритмов распознавания образов на вычислительных машинах. В некоторых случаях размерность пространства признаков является критичной при машинной реализации процедур распознавания. Например, при реализации всевозможных вариационных методов или при реализации алгоритмов основанных на разрезании графов.

На практике встречаются случаи, когда априорный словарь признаков неизвестен, а представляется возможным получить только некоторую совокупность реализаций сигналов, характеризующих явления или процессы. В данных случаях возникает следующая задача: на основе совокупности сигналов, характеризующих некие классы объектов, определить и упорядочить признаки, приписывая больший вес признаку, несущему больше информации при различении объектов. Таким образом, зная информативность каждого признака можно сформировать словарь признаков, включая в него только признаки с наибольшим весом.

Таким образом, в рамках настоящей работы предлагается определять информационные веса количественных признаков исходя из следующих соображений.

Признак будет наиболее информативен в том случае, когда для классов (каждый из которых представлен одним объектом-прецедентом) все его значения будут

отстоять друг от друга на равных расстояниях. Информативность признака будет уменьшаться по ходу нарушения равномерного распределения значений признака. Если признак описывает классы, в каждом из которых будет больше чем один объект, то следует обратить внимание на расстояния между центрами классов относительно этого признака. Информационный вес признака будет наибольшим при одинаковых расстояниях между центрами классов, и будет уменьшаться при нарушении равномерного расположения центров классов (под центром класса, вычисленного относительно конкретного признака, следует понимать среднее значение признака по всем объектам данного класса).

Такое суждение об информативности признака можно обосновать следующим образом.

Рассмотрим два признака на рис. 6, один из которых имеет равномерное распределение центров классов “Признак 1”, а другой неравномерное “Признак2”.

Рис. 6. Признаки, характеризующие объекты

Предположим, что для объекта 1-ого класса были получены значения признаков P_1 и P_2 (близкие к центрам классов) с некоторой ошибкой ζ ($T=(P_1+\zeta) - (P_1- \zeta) =(P_2+\zeta) - (P_2- \zeta) = 2\zeta$), тогда по 1-му признаку объект будет правильно отнесен к 1-му классу, а по 2-му признаку он может быть отнесен как к 1-му, так и ко 2-му, и к3-му классам.

Таким образом, представляется возможным судить о том, что “Признак1” более информативен чем “Признак2”, и его информативность напрямую связана с равномерным распределением значений.

В связи с этим, предлагается использовать следующий подход для определения весов информативности количественных признаков.

Пусть $x_1^c, x_2^c, ..., x_M^c$ значения центров классов признака, которые изменяются при переходе от одного класса к другому, тогда можно вычислить следующие величины:

$$\delta_k = \frac{\Delta_k}{\sum_{k=1}^{M-1} \Delta_k}, \quad к=1,...,M-1, \tag{14}$$

где Δ_{kj} - расстояние между соседними значениями центров классов признака

$$\Delta_k = x_{k+1}^c - x_k^c . \tag{15}$$

Следует заметить, что выполняется равенство

$$\sum_{k=1}^{M-1} \delta_k = 1 . \tag{16}$$

Для вычисления веса признака предлагается использовать следующее выражение

$$V = -\sum_{k=1}^{M-1} \delta_k \ln \delta_k / \ln(M-1) . \tag{17}$$

Следует подчеркнуть, что при применении выражения (17) значение V будет максимальным и равным 1 только тогда, когда $\Delta_k = const$, т.е. значения центров классов признака распределены равномерно, соответственно V→0 при выполнении условия :

$$\delta_k \to 0, \quad k = 1,...,M-1, \quad k \neq m, \quad \delta_m \to 1, \tag{18}$$

где m-любой из номеров интервалов.

Такое поведение V соответствует интуитивному представлению об информационной различающей силе признаков.

Помимо значений весов признаков в некоторых алгоритмах распознавания используются значения репрезентативностей классов, например в алгоритмах вычисления оценок. Следует заметить то, что выбранная для исследования случайным образом из генеральной совокупности группа величин будет называться репрезентативной, если она наилучшим образом представляет всю генеральную совокупность в смысле соответствия выборочных параметров параметрам генеральной совокупности. В основном для определения численности репрезентативной выборки используются параметры генеральной совокупности, но при решении задачи определения репрезентативностей классов в алгоритмах вычисления оценок (АВО) информация о генеральной совокупности практически отсутствует. Поэтому, определить веса W_i репрезентативности (представительности) классов в рамках данной работы предлагается следующим образом.

Репрезентативность класса будет тем выше, чем больше объектов он содержит и при этом расстояния между ближайшими объектами в классе должны быть наиболее однородными. Например, такое утверждение справедливо для твердых тел неорганической природы. Действительно, каждое тело (класс) имеет свою кристаллическую решетку в узлах которых находятся атомы (объекты). Наличие структуры – кристаллической решетки говорит о том, что атомы находятся на равномерном расстоянии друг от друга. Естественно чем больше атомов в теле, расположенных в определенной последовательности, тем больше вес самого тела.

Для оценки равномерности расстояний между объектами в классе следует построить в выбранном признаковом пространстве конечный незамкнутый путь (КНП) или по другому минимальное остовое дерево. Зная расстояния между объектами, то есть длины ребер КНП можно определить репрезентативность класса объектов.

Таким образом, репрезентативность i-ого класса будет равна

$$W_i = -\sum_{r=1}^{K_i-1} \eta_r \ln \eta_r \tag{19}$$

где K_i - количество объектов в i – ом классе, а

$$\eta_r = \frac{R_r}{\sum_{r=1}^{K-1} R_r}, \quad k = 1,...,К-1, \tag{20}$$

где R_r - ребро КНП i-ого класса.

Следует отметить в (19) отсутствие нормировочного знаменателя Ln(К-1), что дает возможность учесть не только равномерность (однородность) ребер КНП, но и их количество.

Таким образом, использование выражений (17) и (19) позволит реализовать АВО с определением весов признаков и репрезентативностей классов, что в свою очередь придаст данным алгоритмам определенную гибкость и позволит реализовать автоматические процедуры распознавания образов, так как присутствие экспертов для определения весов признаков и репрезентативностей классов будет исключено.

В рамках настоящей работы был проведен вычислительный эксперимент, целью которого являлась демонстрация работы алгоритмов вычисления оценок с использованием весов признаков и репрезентативностей классов, вычисленных по вы-

ражениям (17) и (19), и без их использования. Работа алгоритмов оценивалась относительно критерия, который можно назвать ошибкой распознавания.

Работа алгоритмов вычисления оценок с использованием весов признаков и репрезентативностей классов, вычисленных по выражениям (17) и (19), и без их использования при коэффициенте корреляции R_k=0,7 и с разными значениями дисперсии $\delta_k^x = \delta_k^y$ показана в таблице №1.

Большинство алгоритмов и методов классификации объектов и распознавания образов основаны на эвристических принципах. Например, результат классификации будет лучше, если будет достигнута максимально возможная компактность объектов внутри классов, и классы будут находиться на максимально возможном расстоянии друг от друга. Для многих частных задач данные принципы могут быть и другими, в виду наличия всевозможных априорных ограничений в исходных данных.

Таблица 1

Результат работы алгоритмов вычисления оценок с использованием весов признаков и репрезентативностей классов (1) и без их использования (2)

Значения дисперсии $\delta_k^x = \delta_k^y$, при R_k=0,7	1	2	3	4	5	6	7	8	9	10
Кол. правильно расп. объектов (1)	350	350	350	350	350	350	349	347	344	338
Кол. правильно расп. объектов (2)	350	350	350	350	348	346	344	340	337	333

Представляется возможным говорить о том, что применять алгоритмы классификации объектов необходимо для последующего решения задач распознавания образов, т.е. определять классы похожих друг на друга объектов, затем описывать обобщенные характеристики классов и, наконец, нераспознанный объект относить к тому или иному классу.

На основе информационных технологий, использующих алгоритмы классификации и распознавания, строятся информационные системы, которые называют системами распознавания с обучением.

Если рассматривать такие системы с точки зрения стратифицированного подхода, то их структуру можно представить в виде рис. 7.

В страте № 1 решается вопрос обучения системы, а в страте №2 непосредственно решается задача распознавания. Эти страты выделены по функциональному признаку.

Следует отметить то, что на каждой страте может использоваться свое описание, свои алгоритмы, свои модели, но система будет обладать эммерджентростью до тех пор, пока не изменятся ее свойства, принципы и концепция на верхней страте.

В настоящей работе предлагается использовать свойство однородности (равномерности), в обоих стратах рассматриваемой системы, для поддержания целостности.

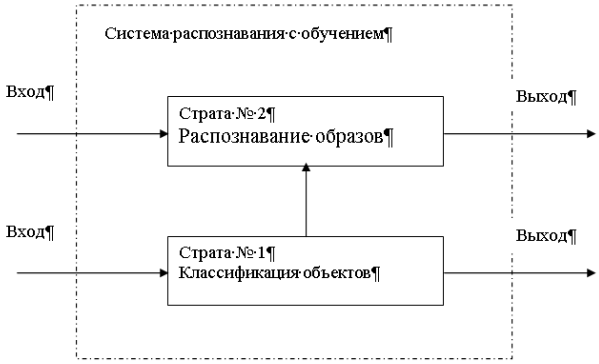

Рис. 7. Стратифицированное представление системы распознавания образов с обучением

То есть в рамках данной работы построено аналитическое выражение функционала качества разбиения, позволяющее количественно оценивать свойство однородности. Данное выражение применено в программно-алгоритмической информационной технологии, использующей созданный в рамках данной работы вариационный алгоритм автоматической классификации объектов с использованием, предложенного авторами работы нового функционала качества разбиения, основанного на мере однородности. Также в рамках данной работы рассмотрен вопрос реализации алгоритма вычисления оценок с использованием весов признаков и репрезентативностей классов.

Литература

1. Дюран, Н. Кластерный анализ [Текст] / Н. Дюран, П Оделл; под общ. ред. Н.Дюрана. – М.:Статистика, 1977. – 128 с.

2. Кропотов, Д.В, Метод группировки объектов, основанный на оптимальных разбиениях [Текст] / Д.В Кропотов, О.В.Сенько // Доклады Всероссийской конференции «Математические методы распознавания образов», ММРО – 10: Изд-во ВЦ РАН, Москва, 2001. с. 77 – 79.

3. Мандель, И.Д. Кластерный анализ [Текст] / И.Д. Мандель, Москва: Финансы и статистика, 1988.

4. Загоруйко, Н.Г. Алгоритмы обнаружения эмпирических закономерностей. [Текст]/ Н.Г.Загоруйко, В.Н.Елкина, Г.С. Лбов. – Новосибирск : Наука, 1985 – 111с..

5. Жиляков, Е.Г. Об Автоматической классификации объектов [Текст] /Е.Г. Жиляков, Е.М.Маматов // Математическое моделирование в научных исследованиях. / Материалы Всероссийской научной конференции. Ч.1. – Ставрополь: Изд-во СГУ, 2000. с. 36-38.

6. Ветров, Д.П., О минимизации признакового пространства в задачах распознавания [Текст] / Д.П. Ветров, В.В. Рязанов // Доклады Всероссийской конференции «Математические методы распознавания образов», ММРО – 10: Изд-во ВЦ РАН, Москва, 2001. с. 22 – 25.

AUTOMATIC CLASSIFICATION OF OBJECTS AND PATTERN RECOGNITION WITH THE USAGE OF FEATURES WEIGHTS AND CLASSES REPRESENTATIVES

E. M. Mamatov

[2] Belgorod state university
e-mail: Mamatov@bsu.edu.ru

In the present article a software information technology which employs a variational algorithm for automatic classification of objects with the use of a new functional of quality partitioning based on uniformity measure is described. The realization of algorithm for calculation of estimation using feature weights and representatives of classes is also implemented.

Key words: Automatic classification of objects, Pattern recognition, Algorithms for calculation of estimation.