УДК 004.93'1

ОПРЕДЕЛЕНИЕ ВЕСОМОСТИ ПРИЗНАКОВ В ЗАДАЧАХ РАСПОЗНАВАНИЯ ОБРАЗОВ И КЛАССИФИКАЦИИ ОБЪЕКТОВ

Е. М. Маматов[1]

1 - Белгородский государственный университет Российская Федерация, 308015, г. Белгород, ул. Победы, 85

Предложен метод определения весомостей признаков в задачах распознавания образов и классификации объектов, позволяющий сократить их размерность.

ВВЕДЕНИЕ

Одним из этапов решения всех видов задач распознавания образов является формирование признакового пространства, то есть его качественного состава и размерности. 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.

При формировании признакового пространства исследователь может столкнуться с некоторыми ограничениями[1]:

1. в словарь включают признаки, относительно которых может быть получена априорная информация, достаточная для описания классов на языке этих признаков;
2. некоторые малоинформативные признаки необходимо включать в основное признаковое пространство;
3. некоторые наиболее информативные признаки не могут быть определены (в виду отсутствия дорогостоящей аппаратуры).

Таким образом, в состав признакового пространства должны входить признаки, которые, с одной стороны, наиболее информативны и , с другой стороны, могут определяться имеющейся аппаратурой.

Другими словами, Задача формирования признакового пространства в общем случае сводиться к тому, чтобы в пределах выделенных ресурсов определить состав аппаратных средств наблюдений, использование которых обеспечит получение наиболее информативных признаков. Построенное таким образом пространство признаков позволяет реализовать максимально возможную эффективность процедуры распознавания.

1. ОПРЕДЕЛЕНИЕ ВЕСОМОСТИ ПРИЗНАКОВ

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.

Для решения задачи определения весов признаков в основном используются статистические методы. Например, разложение Каранунена – Лоэва [2].

Статистические методы и в частности разложение Карунена – Лоэва требуют репрезентативных реализаций сигналов, поэтому на практике не всегда получается их применить. Например, в случаях, когда встречаются классы, которые представляют собой один или два объекта прецедента и само количество классов ограничено.

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.1 , один из которых имеет равномерное распределение центров классов "Признак 1", а другой неравномерное "Признак2".

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.

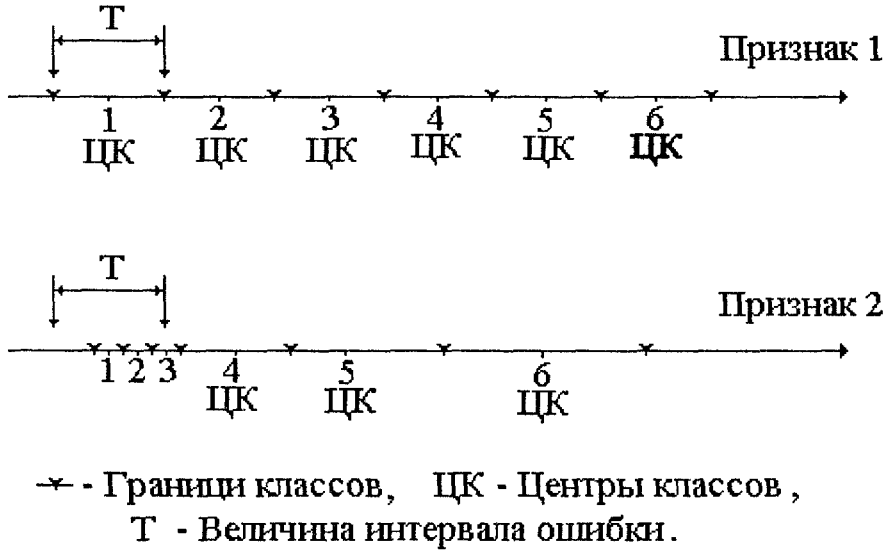

Рис 1 Признаки, характеризующие объекты.

Таким образом, представляется возможным судить о том, что "Признак1" более информативен чем "Признак2". +

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:

$$\delta_k = \frac{\Delta_k}{\sum_{k=1}^{M-1} \Delta_k}, \text{к=1,...,M-1,} \quad (1)$$

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

$$\Delta_k = x_{k+1}^c - x_k^c. \quad (2)$$

Следует заметить, что выполняется равенство

$$\sum_{k=1}^{M-1} \delta_k = 1. \quad (3)$$

Для вычисления веса признака предлагается использовать следующее выражение

$$V = -\sum_{k=1}^{M-1} \delta_k \ln \delta_k / \ln(M-1). \quad (4)$$

Следует подчеркнуть, что при применении выражения (4) 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 :

$$\delta_k \to 0, \quad k = 1,...,M-1, \quad k \neq m, \quad \delta_m \to 1, \quad (5)$$

где m-любой из номеров интервалов.

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

2. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ ПО СРАВНЕНИЮ ЗНАЧЕНИЯ ВЕСОВ ПРИЗНАКОВ И ОЦЕНКИ ВЕРОЯТНОСТИ ПРАВЕЛЬНОЙ КЛАССИФИКАЦИИ

Возникает вопрос - насколько вычисленное согласно (4) значение веса признака отражает его различающую способность.

Определенный ответ на него дает проведенный в рамках данной работы вычислительный эксперимент , в котором значение веса признака сравнивалось с оценкой вероятности правильной классификации при классификации объектов полученных:

- путем добавления к исходным объектам равномерно распределенной случайной величины в интервале (A,B);
- путем добавления к исходным объектам распределенной по Гауссу случайной величины с математическим ожиданием m и дисперсией σ.

Была использована следующая методика моделирования.

1.В интервале от 0 до 1 генерируем M равномерно распределенных случайных величин, значения которых сортируем по возрастанию.

2.Полученные значения равномерно распределенной случайной величины принимаем за значения признака x_i , i = 1,2,...,M характеризующего M объектов-прецедентов, т е. каждый класс Ω_i представлен одним объектом.

3 Определяем значения границ (левой- E_i^1 и правой- E_i^2, i = 1,2,.. ,M) классов следующим образом.

$$\text{Если i = 1, тогда } \begin{matrix} E_i^1 = x_i - \frac{x_{i+1} - x_i}{2} \\ E_i^2 = x_i + \frac{x_{i+1} - x_i}{2} \end{matrix}, \quad (6)$$

$$E_i^1 = x_i - \frac{x_i - x_{i-1}}{2} \qquad (7)$$

если i = M, тогда

$$E_i^2 = x_i + \frac{x_i - x_{i-1}}{2},$$

$$E_i^1 = x_i - \frac{x_i - x_{i-1}}{2} \qquad (8)$$

если 1<i< M, тогда

$$E_i^2 = x_i + \frac{x_{i+1} - x_i}{2}.$$

4.Для каждого класса Ω_i получаем К объектов x_{ij} путем генерирования К раз равномерно распределенной случайной величины ξ_j (j = 1,2,...,K) в интервале (E_i^1, E_i^2), т.е. $x_{ij} = \xi_j$, i = 1,...,M, j = 1,...,K.

5.Вычисляем оценки вероятностей правильных классификаций P_{ij}^r объектов x_{ij}, полученных путем добавления к исходным объектам x_{ij} псевдослучайной равномерно распределенной в интервале (A,B) случайной величины ξ (рис.2).

Пусть INT = A-B, тогда при условии нормировки $a = \frac{1}{INT}$.

Соответственно можно записать $A = x_{ij} - \frac{INT}{2}$ и $B = x_{ij} + \frac{INT}{2}$.

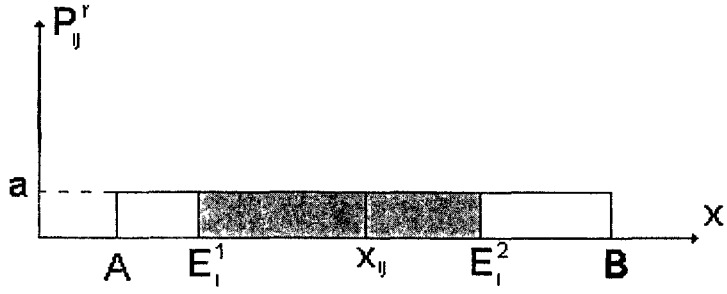

Рис 2 Оценка вероятности правильной классификации.

Для определения величины интервала INT следует воспользоваться следующим выражением :

$$INT = \frac{x_{max} - x_{min}}{M - 1}. \qquad (9)$$

Оценки вероятностей правильных классификаций P_{ij}^r представляется возможным вычислить исходя из выражения

$$P_{ij}^r = (E_i^1 \le x_{ij}^{'} \le E_i^2) = \frac{1}{INT}\left[\min(B, E_i^2) - \min(A, E_i^1)\right]. \qquad (10)$$

6. Вычисляем среднюю оценку вероятности правильной классификации P^r

$$P^r = \frac{\sum_{i=1}^{M}\sum_{j=1}^{K} P_{ij}^r}{M \cdot K}. \qquad (11)$$

7.Получаем оценки вероятностей правильных классификаций P_{ij}^g объектов $x_{ij}^{'}$, полученных путем добавления к исходным объектам x_{ij} распределенной по Гауссу случайной величины с математическим ожиданием m и дисперсией σ.

$$P_{ij}^g = \frac{1}{\sqrt{2\pi}\sigma} \cdot \int_{E_i^1}^{E_i^2} e^{-\frac{(x_{ij}-M)^2}{2\sigma^2}} dx. \qquad (12)$$

Найти первообразную от подынтегральной функции в правой части выражения (12) в аналитическом виде не представляется возможным, следовательно было принято решение вычислять интеграл с помощью численного метода, а именно с помощью метода трапеций.

Математическое ожидание M в этом случае будет равно нулю, а значение дисперсии σ = S*INT ($S=\frac{1}{5}$, 3 - задается в зависимости от начальных условий эксперимента).

8. Вычисляем среднюю оценку вероятности правильной классификации P^g

$$P^g = \frac{\sum_{i=1}^{M}\sum_{j=1}^{K} P_{ij}^g}{M \cdot K}. \quad (13)$$

9.Определяем вес признака, описывающего объекты в классах, по центрам классов следующим образом. Сначала определяем центры классов по выражению

$$x_i^c = \frac{\sum_{j=1}^{K} x_{ij}}{K}, \ i = 1,2,...,M. \quad (14)$$

Затем согласно (1), (2), (3), (4) определяем вес признака.

10.Повторяем пункты 1-9 еще H раз (H=200) и запоминаем полученные результаты в файле для дальнейшего их сравнения.

На основе результатов эксперимента по указанной методике были построены следующие графические зависимости (при количестве классов M = 200, количестве объектов в классе K=50, количестве раз изменения значений признака H=200):

1. Зависимость оценки вероятности правильной классификации P^r объектов, полученных путем добавления к исходным объектам равномерно распределенной случайной величины в интервале INT от веса признака V (Рис. 3).
2. Зависимость оценки вероятности правильной классификации P^g объектов, полученных путем добавления к исходным объектам распределенной по Гауссу случайной величины с математическим ожиданием m=0 и дисперсией σ = S*Int, где S=3, от веса признака V (Рис. 4).
3. Зависимость оценки вероятности правильной классификации P^g объектов, полученных путем добавления к исходным объектам распределенной по Гауссу случайной величины с математическим ожиданием m=0 и дисперсией σ = S*Int, где $S=\frac{1}{5}$. от веса признака V (Рис. 5).

Сопоставление значений весов признаков V_i и оценок вероятностей правильной классификации показывает, что вычисленные согласно (4) веса признаков в достаточной мере отражают их различающую способность.

Способ вычисления веса признака согласно (4) может быть успешно применен в задачах формирования признакового пространства при классификации объектов и распознавании образов. Например, в работе [3] была предложена информационная мера, на основе которой построен функционал качества разбиения объектов на классы. Данный функционал может быть использован в вариационном алгоритме, работающем с минимальным остовым деревом, при разрезании которого необходимо отыскивать с помощью рекурсивной процедуры поддеревья.

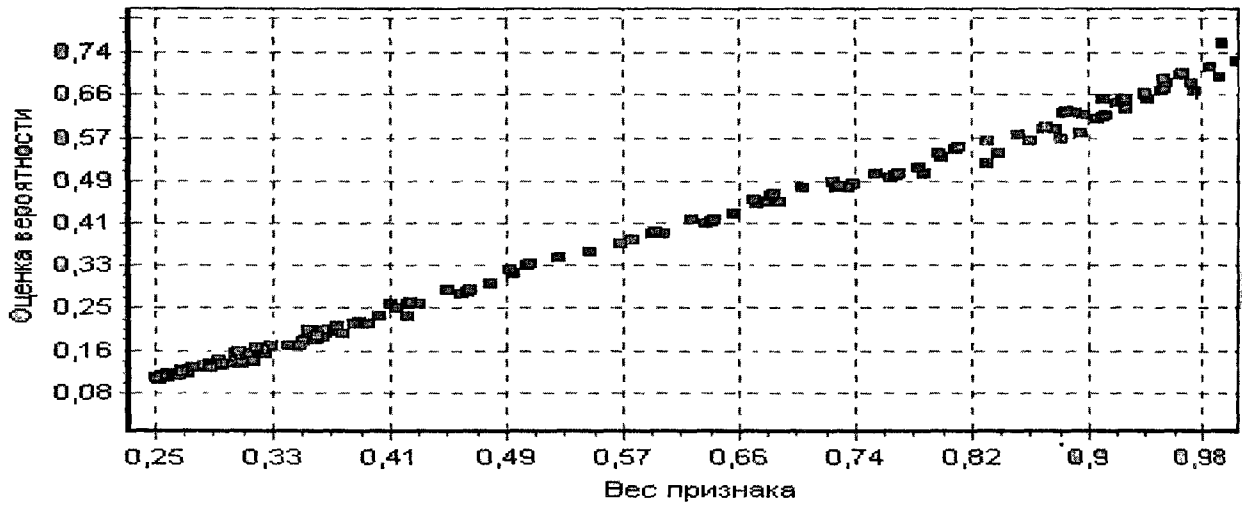

Рис.3 Зависимость оценки вероятности правильной классификации P^r объектов от веса признака V для первого случая.

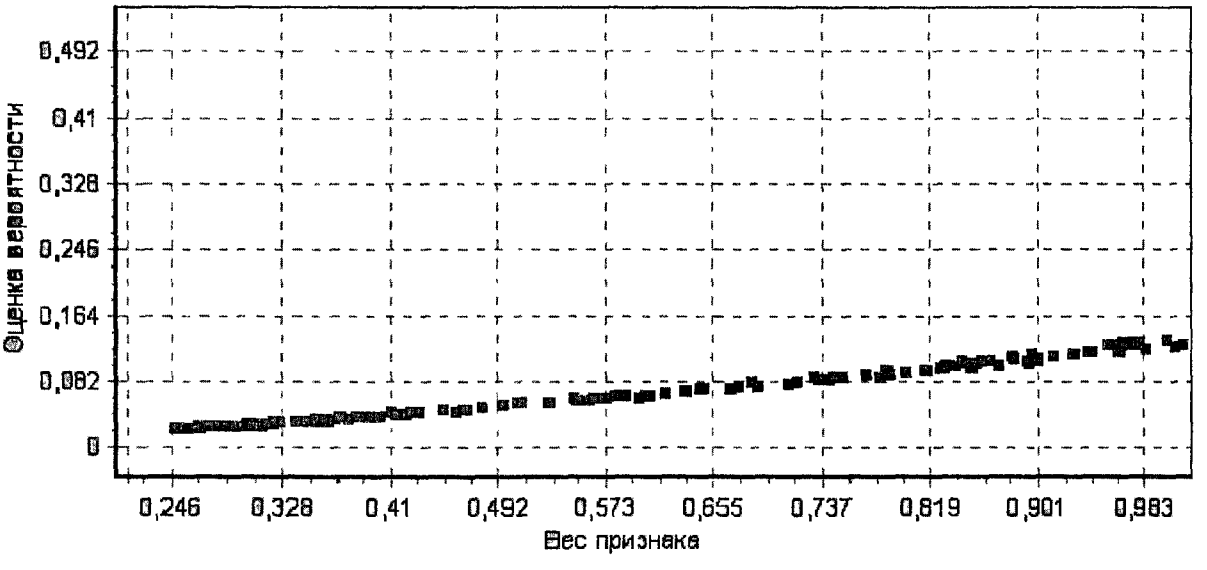

Рис.4 Зависимость оценки вероятности правильной классификации P^r объектов от веса признака V для второго случая.

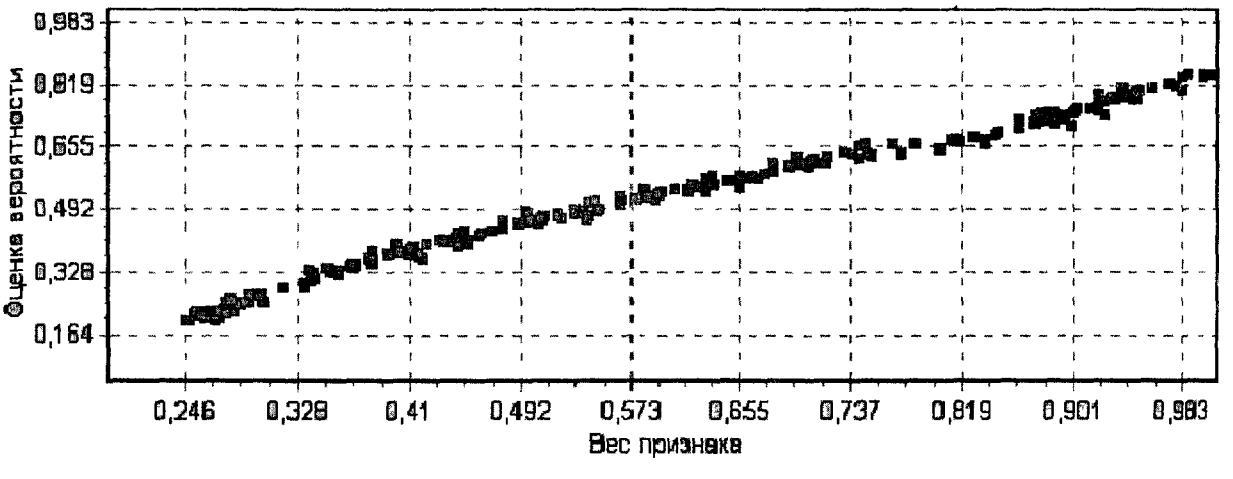

Рис.5 Зависимость оценки вероятности правильной классификации P^r объектов от веса признака V для третьего случая.

При увеличении признакового пространства значительно усложняется сама рекурсия из за необходимости увеличения количества вызовов процедуры из самого тела данной процедуры. Таким образом, при формировании признакового пространства в данном случае просто необходима предварительная оценка информативности признаков, для исключения необходимости усложнять рекурсивную процедуру.

3. ОПРЕДЕЛЕНИЕ ВЕСОВ ПРИЗНАКОВ И РЕПРЕЗЕНТАТИВНРОСТЕЙ КЛАССОВ В АЛГОРИТМАХ ВЫЧИСЛЕНИЯ ОЦЕНОК

В алгоритмах, реализующих методы основанные на принципе полной или частичной прецедентности (реализующих способ принятия решения по аналогии) возникает необходимость в вычислении разделительной способности признаков. Например, в алгоритмах вычисления оценок (АВО) [4]. Одним из этапов процедуры распознавания неизвестного объекта исследования (ОИ) в рамках Г-модели является вычисление индивидуальных (полных или частичных) оценок близости распознаваемого объекта ω к каждому объекту ω_i , $i = 1,2,...,G$, по опорному множеству $S_к$,

$$\Gamma_{S_k}(\omega,\omega_i) = W_i \cdot B_{S_k}(\omega,\omega_i) \cdot \sum_{j \in S_k} V_j , \qquad (15)$$

где V_j - вес, учитывающий различающую способность признака, W_i - вес, учитывающий репрезентативность, $B_{S_k}(\omega,\omega_i)$ - близость полных и частичных описаний объектов ω и ω_i.

Методы Г-модели отличаются высокой гибкостью ввиду возможности введения весов $\{V_j\}_{j=1}^{N}$ для учета различающей способности (информативности) каждого признака из состава признаков априорного описания объектов в обучающей выборке.

В данном случае для определения весов $\{V_j\}_{j=1}^{N}$ признаков, также приемлемо выражение (4).

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 гвердых 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 одинаковом 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 по аналогии с выражениями (1-4) можно определить репрезентативность класса объектов.

Таким образом, репрезентативность i-ого класса будет равна

$$W_i = -\sum_{r=1}^{K_i-1} \eta_r \ln \eta_r , \qquad (16)$$

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

$$\eta_r = \frac{R_r}{\sum_{r=1}^{K-1} R_r} , \; k = 1,...,K-1, \qquad (17)$$

где R_r - ребро КНП i-ого класса.

Таким образом, использование выражений (4) и (16) позволит реализовать АВО с использованием весов признаков и репрезентативностей классов.

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 выражениям (4) и (16), и без их использования. Работа алгоритмов оценивалась относительно критерия, который можно назвать ошибкой распознавания.

При проведении вычислительного эксперимента генерировалось пять классов с различным количеством объектов в каждом из классов. Способ генерации был следующим.

Задались точками, которые должны будут определять условные центры классов, то есть математическими ожиданиями M_k^x и M_k^y. Также задались дисперсиями, характеризующими классы, δ_k^x и δ_k^y и коэффициентами корелляции R_k, где k=1…5.

Для получения двумерных объектов A_i (i=1…M), имеющих нормальный закон распределения по классам, сначала относительно каждого объекта вычислили две независимые случайные величины с нормальным законом распределения по выражениям :

$$\xi^1 = \sum_{j=1}^{12} S_j - 6; \quad (18)$$

$$\xi^1 = \sum_{i=1}^{12} S_j - 6, \quad (19)$$

где S_j - случайная величина из интервала [0..1].

Далее получили X_i - значение координаты объекта по оси X, с учетом принадлежности к k-му классу, посредством следующей формулы

$$X_i = \delta_k^x \xi_i^1 + M_k^x, \quad (20)$$

После чего вычислили математическое ожидание M_k^{yx} и дисперсию δ_k^{yx} по выражениям :

$$M_k^{yx} = M_k^y + R_k \frac{\delta_k^y}{\delta_k^x}(X_i - M_k^x); \quad (21)$$

$$\delta_k^{yx} = \delta_k^y \sqrt{1-(R_k)^2}. \quad (22)$$

В итоге получили Y_i - значение координаты объекта по оси Y, с учетом принадлежности к k-му классу, посредством следующей формулы

$$Y_i = \delta_k^{yx} \xi_i^2 + M_k^{yx}. \quad (23)$$

Таким образом, было сгенерировано исходное множество M объектов (M=350), представленное на рис.6.

Сгенерировав исходное множество M объектов (M=350) произвели их классификацию с помощью алгоритма, предназначенного для агрегирования элементов, заданных матрицей расстояний между ними. Данный вариационный алгоритм автоматической классификации объектов изложен в работе [5]. Результат работы алгоритма классификации представлен на рис. 7.

В общем случае выходные данные алгоритма классификации являются входными данными для алгоритмов вычисления оценок. Так же в качестве входной информации

должны быть координаты нового (неизвестного) объекта, который необходимо отнести к тому или иному классу известных объектов, представленных на рис. 7.

Далее было сгенерировано 350 неизвестных объектов с использованием выражений (18-23) на базе уже известных объектов, задавая координаты каждого известного объекта в качестве параметров математического ожидания каждого неизвестного с определенной дисперсией $\delta_k^x = \delta_k^y$ и коэффициентом корреляции R_k.

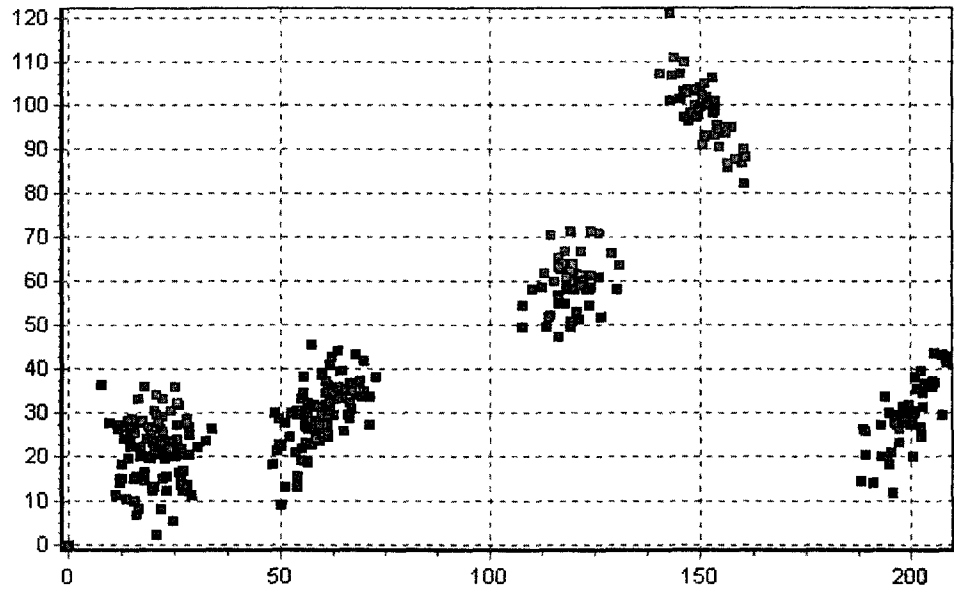

Рис. 6 Исходное множество объектов (M=350)

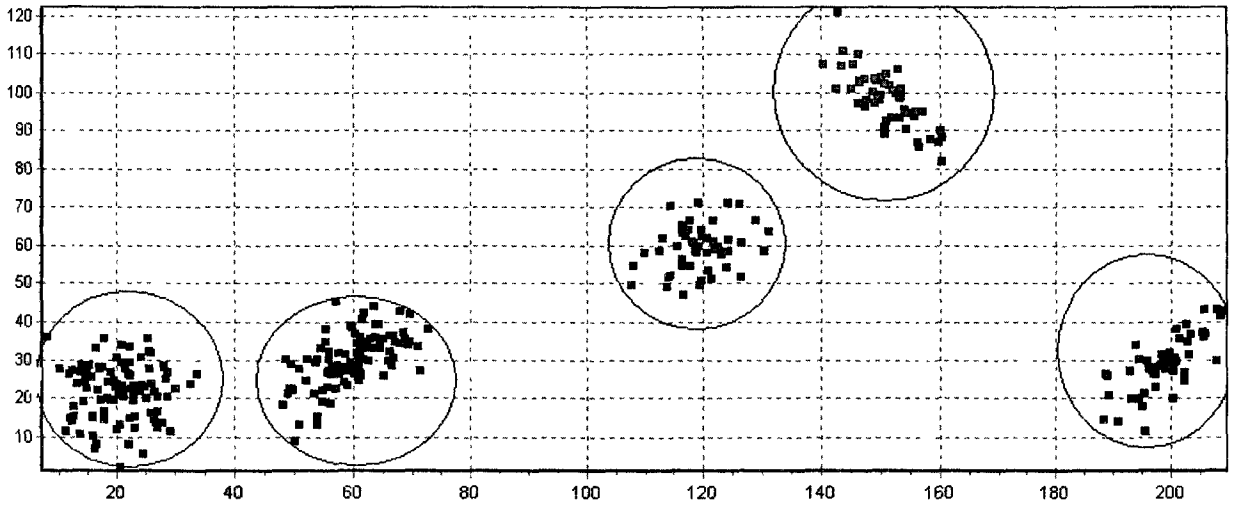

Рис. 7 Результат работы алгоритма классификации.

Полученные неизвестные объекты были распознаны, параллельно работающими алгоритмами вычисления оценок.

Работа алгоритмов вычисления оценок с использованием весов признаков и репрезентативностей классов, вычисленных по выражениям (4) и (16), и без их использования при коэффициенте корреляции R_k=0,7 и с разными значениями дисперсий $\delta_k^x = \delta_k^y$ показана в таблице и на рис.8 .

Таблица

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

Знач. дисперсий $\delta_k^x = \delta_k^y$, 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

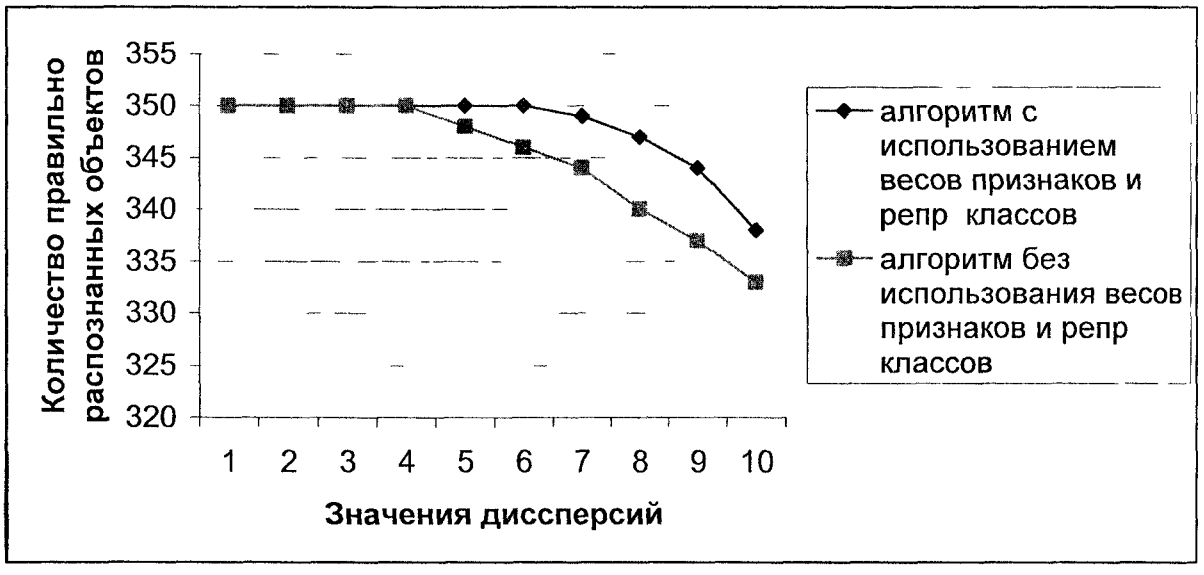

Рис 8 Результат работы алгоритмов вычисления оценок с использованием весов признаков и репрезентативностей классов и без их использования при коэффициенте корреляции R_k =0,7 и с разными значениями дисперсий $\delta_k^x = \delta_k^y$

4. ЗАКЛЮЧЕНИЕ

В заключении следует отметить то, что использование выражения (4) вполне возможно в задачах формирования признакового пространства при классификации обьектов и распознавании образов Об этом свидетельствует вычислительный эксперимент по сравнению значений весов признаков и оценки вероятности правильной классификации

Также применение выражений (4) и (16) в алгоритмах вычисления оценок не только придает им гибкость, но в некотором смысле увеличивает устойчивость работы АВО, что видно из таблицы и рис 8

Библиографический список

1 Горелик А Л , Скрипкин В А Методы распознавания образов Учебное пособие для вузов М Высшая школа, - 1977г

2. Фу К. Последовательные методы в распознавании образов и обучения машин. М.: Наука – 1971г.
3. Жиляков Е.Г., Маматов Е.М. Использование информационной меры в автоматической классификации объектов. \\ Международная научно-практическая конференция, посвященная 30-летию академии "Качество, безопасность, энерго- и ресурсосбережение в промышленности строительных материалов и строительстве на пороге XXI века". Белгород: БелГТАСМ – 2000г.
4. Журавлев Ю.И., Гуревич И.Б. Распознавание образов и распознавание изображений // Распознавание, классификация, прогноз. Математические методы и их применение.-Вып.2-М.: Наука, 1989 - с. 5-72.
5. Жиляков Е.Г., Маматов Е.М. Классификация месторождений сырья для производства строительных материалов.\\ Седьмые Академические чтения РААСН «Современные проблемы строительного материаловедения» Белгород БелГТАСМ 2001г.

DETERMINATION OF FEATURE PONDERABILITY IN PROBLEMS OF PATTERN RECOGNITION AND OBJECT CLASSIFICATION.

E.M. Mamatov

A method of determination of feature ponderability in problems of pattern recognition and object classification is offered in the article. This method makes it possible to reduce object dimention.

УДК 004.032.26

ПРОГНОЗИРОВАНИЕ СЛОЖНЫХ ТЕХНИЧЕСКИХ СИСТЕМ С ПРИМЕНЕНИЕМ ПЕРСЕПТРОНА.

Черных В.А.[1]

1 – Белгородский государственный технологический университет им. В.Г. Шухова

Рассматривается задача прогнозирования сложных технических систем с применением персептрона.

ВВЕДЕНИЕ

Для продления сроков эксплуатации сложных технических систем необходимо оценивать и прогнозировать их фактическое техническое состояние. Это необходимо, так как будущие явления или процессы, характеризующие изменения технического состояния, обладают большой значимостью для решения по управлению, принимаемых в данный момент.

Определить техническое состояние, т.е. состояние которое характеризуется в определенный момент времени, при определенных условиях внешней среды, значениями параметров, установленных технической документацией на объект, само по себе, представляет нетривиальную задачу. Применительно к технической сфере, в ходе применения уникальных и дорогостоящих систем не исключены ситуации, связанные с внезапными, непрогнозируемыми (традиционными методами) отказами элементов и подсистем объектов.

В таких системах становится выгодным и обоснованным применение диагностико-имитационных или математических моделей.[8]

На пути применения любых математических методов экстраполяции лежит ограничение, связанное с размерностью решаемой задачи прогноза. Наличие большого числа параметров некоторых объектов, порядка $10^3 \ldots 10^4$, а также эмерджентность прогнозируемых многомерных процессов делает затруднительным осуществление процедуры прогнозирования в реальном масштабе времени.