МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ

«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТРАНСПОРТА»

Кафедра микропроцессорной техники и информационно-управляющих систем

Н. В. РЯЗАНЦЕВА

КЛАССИФИКАЦИЯ ОБЪЕКТОВ С ИСПОЛЬЗОВАНИЕМ ТЕОРИИ РАСПОЗНАВАНИЯ ОБРАЗОВ

Лабораторный практикум

Ч а с т ь I

Гомель 2005

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТРАНСПОРТА»

Кафедра микропроцессорной техники
и информационно-управляющих систем

Н. В. РЯЗАНЦЕВА

КЛАССИФИКАЦИЯ ОБЪЕКТОВ С ИСПОЛЬЗОВАНИЕМ ТЕОРИИ РАСПОЗНАВАНИЯ ОБРАЗОВ

Лабораторный практикум

Ч а с т ь I

Одобрен методической комиссией электротехнического факультета Белорусского государственного университета транспорта

Гомель 2005

УДК Р 993 519.68

Р е ц е н з е н т – канд. техн. наук О.П. Гораев (БелГУТ).

Р 933 **Рязанцева Н. В.**
Классификация объектов с использованием теории распознавания образов: Лабораторный практикум Ч.I. – Гомель: БелГУТ, 2004. – 27 с.

Содержатся краткие сведения по основным понятиям и задачам теории распознавания. Описаны методы распознавания образов с использованием обучающей последовательности и условия их применимости.
Предназначен для студентов электротехнического факультета специализации «Микропроцессорные информационно-управляющие системы» для подготовки к выполнению лабораторных работ по дисциплине «Техническая кибернетика».

УДК 519.68

ВВЕДЕНИЕ

Проблема распознавания образов состоит из двух частей: обучения и распознавания. Обучение осуществляется путем показа отдельных объектов с указанием их принадлежности тому или другому образу. В результате обучения распознающая система должна приобрести способность реагировать одинаковыми реакциями на все объекты одного образа и различными - на все объекты различных образов. Очень важно, что процесс обучения должен завершиться только путем показов конечного числа объектов без каких-либо других подсказок. В качестве объектов обучения могут быть либо картинки, либо другие визуальные изображения (буквы), либо различные явления внешнего мира, например звуки, состояния организма при медицинском диагнозе, состояние технического объекта в системах управления и др. Важно, что в процессе обучения указываются только сами объекты и их принадлежность образу. За обучением следует процесс распознавания новых объектов, который характеризует действия уже обученной системы. Автоматизация этих процедур и составляет проблему обучения распознаванию образов.

Круг задач, которые могут решаться с помощью распознающих систем, чрезвычайно широк. Сюда относятся не только задачи распознавания зрительных и слуховых образов, но и задачи распознавания сложных процессов и явлений, возникающих, например, при выборе целесообразных действий руководителем предприятия или выборе оптимального управления технологическими, экономическими, транспортными или военными операциями. В каждой из таких задач анализируются некоторые явления, процессы, состояния внешнего мира, далее называемые объектами наблюдения. Прежде чем начать анализ какого-либо объекта, нужно получить о нем определенную, каким-либо способом упорядоченную информацию. Такая информация представляет собой характеристику объектов, их отображение на множестве воспринимающих органов распознающей системы.

Но каждый объект наблюдения может воздействовать по-разному, в зависимости от условий восприятия. Например, какая-либо буква, даже одинаково написанная, может в принципе как угодно смещаться относительно воспринимающих органов. Кроме того, объекты одного и того

же образа могут достаточно сильно отличаться друг от друга и, естественно, по-разному воздействовать на воспринимающие органы.
Каждое отображение какого-либо объекта на воспринимающие множества таких изображений, объединенные какими-либо общими свойствами, представляют собой образы.

При решении задач управления методами распознавания образов вместо термина "изображение" применяют термин "состояние". Состояние - это определенной формы отображение измеряемых текущих (или мгновенных) характеристик наблюдаемого объекта. Совокупность состояний определяет ситуацию. Понятие "ситуация" является аналогом понятия "образ". Но эта аналогия не полная, так как не всякий образ можно назвать ситуацией, хотя всякую ситуацию можно назвать образом.

Ситуацией принято называть некоторую совокупность состояний сложного объекта, каждая из которых характеризуется одними и теми же или схожими характеристиками объекта. Например, если в качестве объекта наблюдения рассматривается некоторый объект управления, то ситуация объединяет такие состояния этого объекта, в которых следует применять одни и те же управляющие воздействия. Если объектом наблюдения является военная игра, то ситуация объединяет все состояния игры.

Выбор исходного описания объектов является одной из центральных задач проблемы теории распознавания образов (ТРО). При удачном выборе исходного описания (пространства признаков) задача распознавания может оказаться тривиальной и, наоборот, неудачно выбранное исходное описание может привести либо к очень сложной дальнейшей переработке информации, либо вообще к отсутствию решения.

Лабораторная работа № 1

МЕТОДЫ АЛГОРИТМИЧЕСКОГО ЗАДАНИЯ РЕШАЮЩИХ ФУНКЦИЙ

Цель работы. Изучить формализованную постановку задачи распознавания, условия применения и суть методов ближайшего соседа и К-ближайших представителей, практическое освоение методов компьютерной реализации, методов алгоритмического задания решающих функций, применяемых для принятия решений в детерминированных системах распознавания.

1 Краткие сведения из теории

Распознавание образов можно представить как процесс принятия решения, устанавливающего принадлежность распознаваемого объекта к некоторому классу путем сравнения определенных его характеристик с характеристиками ранее изученных объектов.

Основными понятиями теории распознавания являются: объект, признак, класс, решающее правило, решающая функция. Каждому объекту соответствует n-мерный вектор, то есть объект можно рассматривать как точку в n-мерном признаковом пространстве (рисунок 1). Каждому классу V_i соответствует некоторая область признакового пространства. Задача распознавания заключается в построении на основе имеющейся информации решающего правила $R(X)$, позволяющего по заданным значениям признаков определить принадлежность объекта к какому-либо классу. Допустимы различные способы описания $R(X)$, но чаще всего $R(X)$ конструируется с использованием решающих функций $G(X)$, которые задают описание классов на языке признаков.

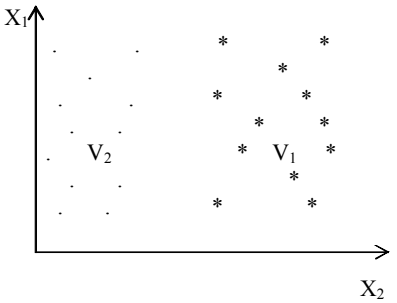

Рисунок 1

Далеко не всегда $G(X)$ можно записать в аналитическом виде, тогда обращаются к алгоритмическому представлению, которое является правилом (алгоритмом) определения значения решающей функции по ее аргументу X.

Рассмотрим простой пример алгоритмического задания $G(X)$, кото-

рый получил название алгоритма Фикса–Ходжеса или метода ближайшего соседа.

Правило ближайшего соседа является наиболее простым методом классификации объектов.

Суть метода заключается в следующем: ищется ближайшая к распознаваемому объекту точка из обучающей последовательности; объект Z_i относят к классу, к которому принадлежит эта точка.

Таким образом, в основе решения этой задачи лежит понятие расстояния между объектами – точками признакового пространства. В зависимости от свойств объектов и исследуемого процесса могут быть использованы различные формализованные определения расстояния между точками

$$Z_1 (X_1^1, X_2^1, X_3^1, ..., X_n^1) \text{ и } Z_2 (X_1^2, X_2^2, X_3^2, ..., X_n^2).$$

Наиболее распространенными являются следующие определения:

ρ_1 – евклидово расстояние

$$\rho_1 = \sqrt{\sum_{i=1}^{n} (X_i^2 - X_i^1)^2} \ ; \qquad (1)$$

ρ_2 – расстояние по Манхэттену

$$\rho_2 = \max\left\{ \left| X_i^2 - X_i^1 \right| \right\} ; \qquad (2)$$

ρ_3 – Чебышевское расстояние

$$\rho_3 = \sum_{i=1}^{n} \left| X_i^2 - X_i^1 \right| \ . \qquad (3)$$

Здесь через X_i^j обозначена i-я составляющая j-го вектора.

$$\rho_4 = arccos \frac{\overline{z_i}\,\overline{z_j}}{\left\| \overline{z_i} \right\| \left\| \overline{z_i} \right\|}, \qquad (4)$$

где $\left\| z_i \right\|$ и $\left\| z_j \right\|$ – нормы соответствующих векторов.

В практических задачах адекватный вид ρ выбирается на основании исследования физической сущности задачи. Конкретные условия могут потребовать составления иных, отличных от (1) – (3) соотношений.

Для решения задачи распознавания любого объекта, по правилу ближайшего соседа необходимо хранить в памяти всю обучающую после-

довательность данных. По известному виду зависимости ρ рассчитываются расстояния от исследуемой точки до точек обучающей последовательности. Определяется минимальное расстояние и считается, что распознаваемая точка принадлежит к тому же классу, что и ближайшая к ней.

Например, в таблице 1 дана обучающая последовательность точек. Требуется определить по правилу ближайшего соседа принадлежность к классам точки Z (3;3) с помощью формул (2) и (3).

Таблица 1

Объекты	X_1	X_2	V
Z_1	1	5	1
Z_2	-2	4	1
Z_3	-1	3	1
Z_4	7	5	2
Z_5	5,5	6	2
Z_6	7	2	2

Сравнение полученных результатов показывает, что ближайшей к Z является точка Z_1. Если пользоваться (3), то ближайшей точкой окажется Z_4. В данном примере обе эти точки принадлежат к первому классу, поэтому принадлежность распознаваемой точки не изменилась. Очевидно, однако, что при использовании различных видов ρ принадлежность точки классам может измениться. Применение правила ближайшего соседа значительно упрощает процесс решения задачи распознавания, но предъявляет к качеству исходных данных повышенные требования: координаты точек обучающей последовательности и их классификация должны быть безошибочными, набор – полным, то есть отражать все возможные состояния исследуемого процесса. Необходимым условием применения метода, является хорошая разделимость классов.

Метод К – ближайших представителей

Обобщением метода ближайшего соседа является метод К - ближайших представителей, согласно которому объект относят к тому классу, к которому принадлежит большинство из К ближайших к нему объектов обучающей последовательности. Для того, чтобы избежать неопределенности, удобнее К выбирать нечетным. Итак, К - произвольное не-

четное число, величина которого выбирается исходя из требований точности классификации с одной стороны и уменьшения машинных расчетов с другой. При больших К вероятность ошибки классификации мала, но возрастают затраты машинного времени.

К недостаткам метода следует отнести необходимость хранения в памяти всей обучающей последовательности, вычисления при классификации расстояний между всеми точками обучающей последовательности и классифицируемым объектом. Этот метод желательно применять при небольших объемах обучающих и распознаваемых выборок.

Рассмотрим пример классификации объекта, заданного вектором Z (2;3). Обучающая последовательность представлена в таблице 2.

Таблица 2

Объекты	X_1	X_2	V
Z_1	0	0	1
Z_2	1	2	1
Z_3	4	4	2
Z_4	2	1	1
Z_5	3	5	2
Z_6	5	4	2
Z_7	4	5	2

Принимать решение будем по К = 5. Вычислим расстояния между точками обучающей последовательности и точкой (2; 3), сравнив их между собой, видим, что ближайшими к Z (2; 3) являются точки Z_2, Z_3, Z_4, Z_5, Z_7. Так как три из пяти точек принадлежат ко второму классу, то и точку Z(2; 3) следует отнести ко второму классу. Совершенно очевидно, что при изменении К может измениться и решение о принадлежности распознаваемой точки. То же самое может произойти, если мы изменим способ определения расстояний.

2 Порядок выполнения работы

1 Смоделировать распознающую систему с использованием метода ближайшего соседа и К - ближайших представителей, для чего написать и отладить программу на языке программирования высокого уровня для классификации объектов с произвольными значениями признаков. Программа должна запрашивать у пользователя значения обучающей выборки (или читать их из файла), метод распознавания, число К, выдавать на экран или принтер исходные данные и результат

классификации.

2 Для проверки работоспособности программы по своему варианту (номер в журнале) ввести исходные данные, распечатать листинг программы, таблицу исходных данных и результатов. При выполнении задания исходные данные берутся из таблицы 3, при этом номер варианта прибавляется к значениям X_1, если он четный и к значениям X_2, если он нечетный.

Таблица 3

Объекты	X_1	X_2	V
Z_1	5	7	1
Z_2	3	2	1
Z_3	1	4	2
Z_4	10	5	2
Z_5	5	0	1
Z_6	4	1	2
Z_7	2	8	1

3 Сделать выводы о применимости метода.

4 Оформить отчет, в котором привести листинг программы, копии экранов и результаты работы программы.

Контрольные вопросы

1 Сформулируйте преимущества и недостатки применения метода ближайшего соседа и метода К - ближайших представителей, дайте их сравнительный анализ.

2 Дайте сравнительный анализ различных мер близости между объектами. Укажите условия применимости каждой меры.

3 Обучающая последовательность задана таблицей 4:

Таблица 4

Объекты	X_1	X_2	Vi
Z1	-1	4	V_1
Z2	0	2	V_1
Z3	6	5	V_2
Z4	7	3	V_2
Z5	1	6	V_1

Определите принадлежность к классам точки Z (3;4). Как изменится результат, если:

- допущена ошибка в задании координат точки Z_5 и истинные значения $X_1 = 1$, $X_2 = 5$.
- неверно классифицирована точка Z_5 , то есть $V(Z_5) = 2$.

Лабораторная работа № 2

МЕТОД ЭТАЛОНА

Цель работы. Изучить условия применения и механизм реализации метода эталона в двух постановках: с построением разделяющей границы и без.

1 Краткие сведения из теории

Метод К - ближайших представителей (в том числе и правило ближайшего соседа) обладает существенными недостатками, затрудняющими его применение в практических задачах:

- требуются значительные объемы машинной памяти для хранения всех объектов обучающей последовательности;
- вычисление мер близости накладывает жесткие ограничения на выбор программного и технического обеспечения синтезируемых систем управления;
- большие затраты машинного времени на расчеты ограничивают круг решаемых задач, требующих принятия решения в реальном масштабе времени.

В связи с этим получил большое распространение метод эталона, свободный от этих недостатков.

Метод эталона в своей простейшей постановке может рассматриваться как частный случай метода К - ближайших представителей, если в качестве К - ближайших точек рассматриваются все точки обучающей последовательности. Возникает задача определения расстояния между точкой и классом. Расстоянием между точкой и классом будем считать расстояние между этой точкой и эталоном класса - его центром тяжести.

В качестве меры близости точки и класса можно использовать следующее представление:

$$\rho(Z;V_i) = \rho(Z;Z_э),$$

где Z – объект, подлежащий классификации; $Z_э$ – эталонный объект (эталон) соответствующего класса.

Координаты эталона рассчитываются по формуле

$$X_i = \frac{1}{K_j} \sum_{j=1}^{K_j} X_i^j \qquad (1)$$

где X_i – i-я координата эталона; K_j – число объектов обучающей последовательности j -го класса; X_i^j – i-я координата *l-той* точки j -го класса.

Метод эталона позволяет резко сократить объемы хранимой информации, объемы вычислительной работы при распознавании объектов за счет построения разделяющей границы между классами и формировании на основе аналитического выражения для решающего правила.

Алгоритм построения разделяющей границы между классами и методом эталона заключается в следующем:

1 Для каждого класса по формуле (1) рассчитывается эталон – наиболее типичный представитель класса.

2 Все признаковое пространство делится на две части гиперповерхностью, все точки которой одинаково отстоят от эталонов.

Распознавание объектов можно производить без построения разделяющей границы на основании сравнения расстояний от исследуемой точки до эталона. В этом случае в памяти необходимо хранить только эталоны и вычислить всего два расстояния.

Линейная разделяющая граница (гиперплоскость) между классами:

$$g(X) = C_0 + C_1 X_1 + C_2 X_2 + \ldots + C_n X_n = 0,$$

где $C_1, C_2, \ldots, C_n$ – весовые коэффициенты; $X_1, X_2, \ldots, X_n$ – координаты текущей точки гиперплоскости.

Для удобства записи и вычислений точку доопределяют нулевой координатой X_0, тождественно равной 1. Тогда $g(X)$ можно представить в виде :

$$g(X) = \sum_{i=0}^{n} C_i X_i .$$

На основании g(X) легко формируется решающее правило:

$$R(X) = \begin{cases} 0, \text{ если } g(X) > 0, \\ \\ 1, \text{ если } g(X) \le 0. \end{cases} \qquad (2)$$

$Z \in V_1$, если $R(X) = 0$ и $Z \in V_2$, если $R(X) = 1$.

Процедуру расчета разделяющей границы рассмотрим на конкретных примерах.

Пример 1. Обучающая последовательность точек задана таблицей 5. Требуется построить разделяющую границу между классами и определить принадлежность точки Z(3; 4) к классам без применения “разделяющей границы” и с ее помощью.

Таблица 5

Объекты	X_1	X_2	V_i
Z_1	1	3	V_1
Z_2	4	4	V_2
Z_3	5	5	V_2
Z_4	5	3	V_2
Z_5	0	1	V_1
Z_6	6	4	V_2
Z_7	2	2	V_1
Z_8	1	1	V_1
Z_9	1	2	V_1
Z_{10}	6	5	V_2

Найдем координаты эталона первого класса:

$$X_{1э1} = \frac{1 + 0 + 2}{3} = 1; X_{2э1} = \frac{3 + 1 + 2}{3} = 2.$$

Тогда $Z_{Э1}$(1; 2). Аналогично определим второй эталон (рисунок 2).

$Z_{э2}$ = (5; 4)

$$X_{1э2} = \frac{4 + 5 + 5 + 6}{4} = 5; X_{2э2} = \frac{4 + 5 + 3 + 4}{4} = 4.$$

Найдем координат ы середины отрезка $Э_1$ $Э_2$ (точки М):

$$X_{1M} = \frac{1+5}{2} = 3; X_{2M} = \frac{2+4}{2} = 3$$

$$X_{1M} = \frac{1+5}{2} = 3; X_{2M} = \frac{2+4}{2} = 3$$

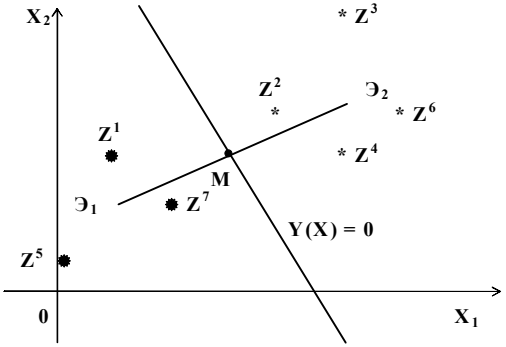

Рисунок 2

Разделяющая граница между классами V_1 и V_2 проходит через точку М перпендикулярно прямой $Э_1Э_2$ и имеет вид:

$$g(X) = C_0 X_0 + C_1 X_1 + C_2 X_2 = 0 \qquad (3)$$

Вектор $Э_2Э_1$ определяет коэффициенты C_1 и C_2 (из условия ортогональности прямых). Координаты вектора $Э_2Э_1$:

$C_1 = 1 - 5 = 4$ и $C_2 = 2 - 4 = -2$.

Уравнение (3) принимает вид:

$$C_0 X_0 - 4 X_1 - 2 X_2 = 0. \qquad (4)$$

Учитывая, что $X_0 \equiv 1$ и, что координаты точки М удовлетворяют соотношению (3), получим $C_0 - 4 \cdot 3 - 2 \cdot 3 = 0$, откуда $C_0 = 18$.

Следовательно,

$$g(X) = 18 - 4 \cdot X_1 - 2 \cdot X_2 = 0.$$

Выясним, к какому классу относится точка X (3; 4).

Первый способ.

Найдем ρ_1 (Z, $Э_1$) = $\sqrt{(3-1)2+(4-2)2} = 2\sqrt{2}$

и ρ_1 (Z, $Э_2$) = $\sqrt{(3-5)2+(4-4)2} = 2$.

ρ_1 (Z, $Э_1$) > ρ_1 (Z, $Э_2$), следовательно Z ∈ V .

Второй способ.

Найдем Y(X) = 18 – 4 · 3 – 2 · 4 = –2 < 0 . Это значит, что Z ∈ V_2 согласно (2). Во втором случае меньше и проще вычисления. Но требуется предварительный расчет функции Y(X).

Пример 2. В трехмерном признаковом пространстве заданы точки

Z_1 = (0;2;6); Z_2 = (1;3;5); Z_3 = (2;1;7); Z_4 = (5;4;0);

Z_5 = (6,5;3;1); Z_6 = (6,5;2;–1).

Известно, что X_1, Z_2, Z_3 ∈ V_1 , а Z_4, Z_5 , Z_6 ∈ V_2 . Найти разделяющую границу между V_1 и V_2 . К какому классу принадлежат точки Z_7 (0;0;0), Z_8 (0;8;1). Найдем эталоны классов:

$$z^{Э1} = (\frac{0+1+2}{3}, \frac{2+3+1}{3}, \frac{6+5+7}{3}) = (1;2;6);$$

$$z^{Э2} = (6;3;0).$$

Вектор $Э_2Э_1$, ортогональный искомой прямой

$$Y(X) = C_0X_0 + C_1X_1 + C_2X_2 + C_3X_3 \quad ,$$

имеет вид $Э_2Э_1$ = (1 – 6 ; 2 – 3 ; 6 – 0) = (–5 ; –1 ; + 6) , то есть C_1= – 5,

$$C_2 = -1, \; C_3 = +6.$$

Середина отрезка $Э_1Э_2$ – точка Q характеризуется координатами:

$$X_{Q1} = \frac{1+6}{2} = 3{,}5; \; X_{Q2} = \frac{2+3}{2} = 2{,}5; \; X_{Q3} = \frac{6+0}{2} = 3.$$

Из условия принадлежности точки Q прямой Y(X) = 0 определим C_0 :

$$-5X_1 - X_2 + 6X_3 + C_0 = 0 ;$$

$$-5 \cdot 3{,}5 - 2{,}5 + 6 \cdot 3 + C_0 = 0 ; \qquad C_0 = 2 .$$

Откуда следует

$$g(X) = -5X_1 - X2 + 6X_3 + 2 = 0 .$$

Исследуем положение точек Z_7 = (0;0;0) и Z_8 = (0;8;1)

$g(Z_7) = 2 > 0$, следовательно, $Z_7 \in V_1$.

$g(Z_8) = 0$, точка Z_8 лежит на границе между классами.

По определению (2) $Z_8 \in V_2$.

2 Порядок выполнения работы

1 Смоделировать распознающую систему с использованием эталона, написать и отладить программу на языке программирования высокого уровня для классификации объектов с произвольными значениями признаков. Программа должна запрашивать у пользователя значения обучающей выборки (или читать их из файла), выдавать на экран или принтер исходные данные и результат классификации.

2 Для проверки работоспособности программы по своему варианту (номер в журнале) ввести исходные данные, распечатать листинг программы, таблицу исходных данных и результатов. При выполнении задания исходные данные берутся из таблицы 5, при этом номер варианта прибавляется к значениям X_1, если он четный и к значениям X_2, если он нечетный.

3 Сделать выводы о применимости метода.

4 Оформить отчет, в котором привести листинг программы, копии экранов и результаты работы программы.

Контрольные вопросы

1 Сформулировать условия применимости метода эталона. Оценить чувствительность метода к ошибкам в исходных данных.

2 Решить задачу распознавания методом эталона для точек Z1 и Z2. Выполнить чертеж и охарактеризовать расположение классов и разделяющей границы (Z1 = (3; 4), Z2 = (4; 3)). Обучающая последовательность задана таблицами:

а)

Объекты	*X1*	*X2*	*Vi*
Z_3	0	3	V1
Z_4	1	4	V1
Z_5	2	2	V1
Z_6	2	6	V1
Z_7	3	3	V2

Объекты	*X1*	*X2*	*Vi*
Z9	4	1	V1
Z10	4	2	V2
Z11	4	4	V1
Z12	5	7	V2
Z13	6	2	V2

Z_8	3	5	V2

Z14	6	5	V2

3 $Z_1 = (3;3;3;3)$; $Z_2 = (0;1;1;0)$.

б)

Объекты	*X1*	*X2*	*X3*	*X4*	*Vi*
Z_3	0	4	6	1	V1
Z_4	1	6	5	1	V1
Z_5	4	1	3	7	V2
Z_6	5	2	0	5	V2
Z_7	2	5	7	1	V1

Л а б о р а т о р н а я р а б о т а № 3

АДАПТИВНЫЕ МЕТОДЫ РАСПОЗНАВАНИЯ

Цель работы: Изучить условия применения и механизм реализации адаптивных методов распознавания.

1 Краткие сведения из теории

Адаптивные методы построения разделяющих границ в ряде практических задач являются наиболее предпочтительными или даже единственно возможными. Они позволяют организовать последовательный алгоритм уточнения коэффициентов по мере поступления информации и незаменимы в условиях нестационарности исследуемого процесса. В основу построения адаптивных алгоритмов расчета решающих функций g (C,X) = C X может быть положен метод градиента, определяющий (k + 1) – приближение коэффициентов C по формуле:

$$C(k+1) = C(k) - \alpha \{ \frac{\partial J(c, x)}{\partial c} \},$$

$$C = C(k),$$

где α – шаг коррекции; J (C,X) – функционал, отражающий величину ошибки классификации объектов в зависимости от вектора коэффициентов C . При достижении минимума J (C,X) , то есть при

$$\frac{\partial J}{\partial c} = 0$$

значение вектора C не корректируется.

Производная $\partial J/\partial C$ определяет направление и величину возрастания J .

При малых α достигается хорошая точность коэффициентов, но требуется много шагов для нахождения min J. Оптимальной является стратегия, когда начальные α (определяющие первые шаги движения к min J) достаточно большие, затем уменьшаются до 0. Часто в качестве α принимают выражение 1/k, где k – номер итерации. Вид функционала определяется конкретным физическим содержанием задачи. Ниже в качестве J будем рассматривать величину |C X| , определяющую расстояние точки от гиперплоскости C X = 0. Процедура (1) в этом случае примет вид

$$C(k+1) = C(k) + \alpha X^k, \qquad (2)$$

если $C(k)\, X^k \le 0$, а $X^k \in V_1$, и

$$C(k+1) = C(k) - \alpha X^k \qquad (3)$$

в противном случае.

Рассмотрим конкретный пример, иллюстрирующий адаптивный алгоритм построения разделяющей границы. Класс V_1 представлен точками $Z_1(0;\ 1)$, $Z_2(1;\ 2)$, $Z_3(1;1)$. Точки $Z_4(4;\ 2)$, $Z_5(4;\ 4)$ и $Z_6(3;2)$ принадлежат классу V_2. Каждую точку доопределим для удобства математических выкладок нулевой координатой $x_0 \equiv 1$ и будем искать разделяющую границу в виде

$$g(x) = c_0x_0 + c_1x_1 + c_2x_2 = 0.$$

Необходимо найти вектор $(c_0;\ c_1;\ c_2)$ такой, что для всех точек класса V_1 значения $g(x) \ge 0$, для точек класса V_2 – $g(x) < 0$. Пусть $\alpha = 1$ и $C(1) = (1;\ 1;\ 1)$. Этому вектору соответствует уравнение

$$1 + x_1 + x_2 = 0.$$

Легко проверить, что все точки класса V_1 классифицируются верно. Действительно $C(1)\, Z_1 = (1;1;1;)(1;0;1)^T = 2 > 0$ и т.д. Вместе с тем для точки Z_4 имеем:

$$C(1)\ Z_4 = (1;1;1;)(1;4;2)^T = 7 > 0,$$

поэтому согласно (3) при $\alpha = 1$ второе приближение вектора будет равно

$$C(2) = C(1) - Z_4 = (1;1;1;) - (1;4;2) = (0;-3;-1).$$

Этому вектору соответствует прямая

$$g(x) = -3x_1 - x_2 = 0.$$

Проверим правильность распознавания точек обучающей последовательности: подставим координаты точек обучающей последовательности в g(x) и сравним с нулем. Здесь верно классифицируются Z_4 , Z_5 и Z_6 , зато ни одна из точек первого класса не классифицируется верно. Для получения C(3) используем Z_1:

$$C(3) = C(2) - Z_1 = (0;-3;-1) - (1;0;1) = (1;-3;0).$$

Этому вектору соответствует

$$g(x) = -3x_1 + 1 = 0.$$

Опять проверяем правильность классификации и находим, что Z_2 классифицируется неверно, тогда

$C(4) = C(3) - Z_2 = (1;-3;0) + (1;1;1) = (2;-2;1)$, что соответствует

$$g(x) = 2 - 2x_1 + x_2 = 0.$$

Не трудно убедиться, что это уравнение обеспечивает правильное распознавание всех точек обучающей последовательности.

2 Порядок выполнения работы

1 Смоделировать распознающую систему, написать и отладить программу на языке программирования высокого уровня для классификации объектов с произвольными значениями признаков. Программа должна запрашивать у пользователя значения обучающей выборки (или читать их из файла), выдавать на экран или принтер исходные данные и результат классификации.

2 Для проверки работоспособности программы по своему варианту (номер в журнале) ввести исходные данные, распечатать листинг программы, таблицу исходных данных и результатов. При выполнении задания исходные данные берутся из таблицы 5, при этом номер варианта прибавляется к значениям X_2, если он четный и к значениям X_1, если он нечетный.

3 Оформить отчет, в котором привести листинг программы, копии экранов и результаты работы программы. Сделать выводы о применимости метода.

Контрольные вопросы

1 Кратко изложить суть адаптивных методов распознавания. Дать геометрическую интерпретацию.

2 Определить условия применения, преимущества и недостатки метода.

3 Как изменяются свойства адаптивного алгоритма построения решающего правила в зависимости от значений α – шага коррекции?

4 Построить разделяющую границу между классами адаптивным методом, если известно, что $Z_1(0;5)$; $Z_2(1;4)$; $Z_3(1;6) \in V_1$, а $Z_4(4;0)$; $Z_5(5;1)$; $Z_6(7;2) \in V_2$. Определить, к какому классу принадлежат объекты $Z_7(1;4)$ и $Z_8(0;0)$.

Лабораторная работа № 4

МЕТОДЫ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ. ПРОСТЕЙШИЙ АЛГОРИТМ ВЫЯВЛЕНИЯ КЛАСТЕРОВ

Цель работы. Изучить условия применения и механизм реализации метода автоматической классификации и простейшего алгоритма выявления кластеров

1 Краткие сведения из теории

Задача автоматической классификации заключается в следующем. Имеется выборка объектов $Z^1, Z^2, \ldots, Z^j$, каждый из которых характеризуется набором признаков $X_1, X_2, \ldots, X_n$. Требуется разбить это множество на m классов таким образом, чтобы объекты, объединяющиеся в один класс, были похожи друг на друга по достаточно большому числу признаков. При этом используют гипотезу " компактности " образов, то есть считают, что степень сходства между собой у объектов, принадлежащих одному классу, должна быть больше, чем у объектов, относящихся к разным классам. Геометрически это означает, что классам объектов в признаковом пространстве соответствуют кластеры – однородные, обособленные скопления точек. Число классов m часто является неизвестным и определяется в ходе решения задачи.

В качестве меры сходства объектов Z^p и Z^q используется расстояние между точками признакового пространства, которые соответствуют этим объектам:

$$\rho(Z^p, Z^q) = \sqrt{\sum_{i=1}^{n} \left(X_i^p - X_i^q\right)^2} \ ,$$

при этом, чем меньше расстояние между ними, тем больше сходство.

Алгоритмы выделения кластеров (классов) строятся либо на основе некоторых эвристических соображений (например, простейший алгоритм выявления кластеров), либо основываются на минимизации (или максимизации) какого-нибудь показателя качества получаемого разбиения объектов на классы (например, алгоритм *m* внутригрупповых средних).

При эвристическом подходе задается набор правил, которые обеспечивают использование меры сходства для отнесения точек к одному из кластеров. Для установления приемлемой степени сходства двух объектов, объединяемых в один класс, часто вводится некоторый порог *T*.

Подход, предусматривающий использование показателя качества, связан с разработкой процедур, обеспечивающих минимизацию выбранного показателя качества. Наиболее часто используемым показателем является сумма квадратов отклонений от средних значений

$$Q = \sum_{i=1}^{m} \sum_{x \in V_i} (X - Z^i)^2,$$

где *m* – число кластеров;

$$Z^i = \frac{1}{L_i} \sum_{x \in V_i} x,$$

i – центр кластера V_i;

L_i – количество объектов, попавших в i - й кластер.

Очевидно, что чем меньше *Q*, тем лучше классификация.

Отсутствие априорной информации о значении пороговой величины *T* меры сходства объектов и числе выделяемых классов *m* требует проведения многочисленных экспериментов с различными значениями этих параметров для получения приемлемых результатов. Лучшей из

полученных при этом классификаций является та, которой соответствует минимальное значение критерия качества Q.

Кроме того, для оценки результатов работы алгоритмов выявления кластеров используется анализ таблиц расстояний между центрами кластеров, количества объектов в классах и дисперсий по каждому признаку для каждого кластера. Эта информация позволяет изменять параметры алгоритмов в направлении улучшения качества получаемого разбиения объектов на классы.

Полное решение задачи автоматической классификации, как правило, предполагает также построение описания полученных кластеров и решающего правила, различающего выделенные классы. В этом случае основное отличие задачи автоматической классификации от задачи распознавания заключается в том, что классы, на которые надо разбить объекты, не заданы, т. е. отсутствует обучающая последовательность. Так как обучающая последовательность первоначально не задана, а формируется в процессе решения задачи, то построение решающего правила называют самообучением распознаванию образов.

Простейший алгоритм выявления кластеров заключается в следующем. В качестве центра первого кластера $Z_{э_1}$ выбирается любой из заданных векторов объектов из выборки, подлежащей классификации. Далее задается произвольная неотрицательная пороговая величина T; для удобства можно считать, что $Z_{э_1} = Z^1$. После этого вычисляется расстояние $\rho(Z^2;Z^1)$ между вектором Z^2 и центром кластера Z^1. Если $\rho(Z^2;Z^1) > T$, то $Z^2 = Z_{э_2}$ – центр нового кластера V_2, иначе Z^2 зачисляется в первый кластер V_1. Подобным образом расстояния от каждого нового объекта до каждого из полученных центров кластеров вычисляются и сравниваются с пороговой величиной T. Если все эти расстояния превосходят значение порога T, то учреждается новый центр кластера, в противном случае вектор зачисляется в кластер с самым близким к нему центром.

Результаты описанного метода зависят от геометрических характеристик данных, а также от значения пороговой величины T и порядка просмотра объектов.

Поэтому получение хороших результатов с помощью этого метода требует многочисленных экспериментов с различными значениями порога и порядком просмотра объектов.

При этом следует придерживаться следующих рекомендаций:

1 Поскольку обычно изучаемые объекты имеют высокую размерность и визуальная интерпретация результатов исключается, то необходимая информация может быть получена при помощи составления после каждого цикла просмотра данных таблиц расстояний, разделяющих центры кластеров, и таблиц количества объектов, вошедших в различные кластеры.

2 При выполнении очередного просмотра данных в первую очередь просматриваются точки наиболее близкие к центрам кластеров, а затем все остальные.

3 Если число полученных кластеров велико, следует увеличивать T при очередном применении алгоритма.

2 Порядок выполнения работы

1 Смоделировать распознающую систему, написать и отладить программу для классификации объектов с произвольными значениями признаков. Программа должна запрашивать у пользователя значения обучающей выборки (или читать их из файла), выдавать на экран или принтер исходные данные и результат классификации.

2 Для проверки работоспособности программы по своему варианту (номер в журнале) ввести исходные данные, распечатать листинг программы, таблицу исходных данных и результатов. При выполнении задания исходные данные берутся из таблицы 5, при этом номер варианта прибавляется к значениям X_2, если он четный и к значениям X_1, если он нечетный.

3 Сделать выводы о применимости метода.

4 Оформить отчет, в котором привести листинг программы, копии экранов и результаты работы программы.

Контрольные вопросы

1 Сформулируйте, какие достоинства и недостатки имеет простейший алгоритм выявления кластеров.

2 Примените простейший алгоритм отыскания кластеров к следующему множеству данных $\{Z_1 = (0;0), Z_2 = (0;1), Z_3 = (5;4),$
$Z_4 = (5;5), Z_5 = (4;5), Z_6 = (1;0)\}$.

3 Вычислите значения критерия Q для полученного в п. 2 разбиения объектов на классы.

4 Составьте и проанализируйте таблицы расстояний между центрами кластеров, количества объектов в классах и дисперсий значений признаков в каждом классе.

Лабораторная работа № 5

МЕТОД ОБЪЕДИНЕНИЯ

Цель работы. Изучить условия применения и механизм реализации метода объединения.

1 Краткие сведения из теории

Суть метода объединения заключается в объединении ближайших точек в одну с учетом их весов. Для этого задается число классов m. Сначала все точки классифицируемой выборки имеют единичный вес. Среди них отыскиваются две ближайшие точки Z^p и Z^q, объединяющиеся в одну, расположенную в их центре тяжести. Их координаты вычисляются по формуле

$$X^{p,q} = \frac{1}{2}(X^p + X^q).$$

Новая точка включается в выборку с весом 2 вместо точек Z^p и Z^q и т. д. Если в процессе такого объединения встречаются точки Z^i и Z^j с весами W_i и W_j, то они объединяются в одну, расположенную в их центре тяжести

$$X^{i,j} = \frac{1}{W_i + W_j}(X^i W_i + X^j W_j),$$

которая записывается в выборку с весом $W_i + W_j$.

Рассмотренная процедура выполняется до тех пор, пока останется m точек, которые и следует считать эталонами (центрами) полученных кластеров.

При неизвестном числе классов момент прекращения процедуры объединения можно определять следующим образом.

На каждом *к*-м шаге объединения определяется суммарное расстояние между всеми точками

$$S_k = \sum_{i,j} W_i W_j \rho(X^i, X^j).$$

Это расстояние будет уменьшаться на каждом шаге на величину

$$\Delta S_k = S_k - S_{k-1}.$$

Резкое увеличение (скачок) величины ΔS_k на очередном шаге связано с объединением удаленных скоплений точек, т. е. потерей информации о различии классов. При появлении такого скачка процедура объединения прекращается, и в качестве ее результатов берется классификация, полученная на предыдущем шаге.

2 Порядок выполнения работы

1 Смоделировать распознающую систему, написать и отладить программу для классификации объектов с произвольными значениями признаков. Программа должна запрашивать у пользователя значения обучающей выборки (или читать их из файла), выдавать на экран или принтер исходные данные и результат классификации.

2 Для проверки работоспособности программы по своему варианту (номер в журнале) ввести исходные данные, распечатать листинг программы, таблицу исходных данных и результатов. При выполнении задания исходные данные берутся из таблицы 3, при этом номер варианта прибавляется к значениям X_2, если он четный и к значениям X_1, если он нечетный.

3 Сделать выводы о применимости метода.

4 Оформить отчет, в котором привести листинг программы, копии экранов и результаты работы программы.

Контрольные вопросы

1 Какой из рассмотренных методов более трудоемкий – метод объединения или простейший алгоритм выявления кластеров?

2 Примените метод объединения к множеству данных, приведенному в п. 2 предыдущей работы.

3 Для задачи 2 вычислите ΔS суммарного расстояния между точками на каждом шаге объединения. Убедитесь, что оптимальное число классов равно двум, так как объединение всех точек в один класс вызывает резкое возрастание величины ΔS по сравнению с предыдущими шагами.

Л а б о р а т о р н а я р а б о т а № 6
АЛГОРИТМ М-ВНУТРИГРУППОВЫХ СРЕДНИХ.

Цель работы. Изучить условия применения и механизм реализации алгоритма *m*-внутригрупповых средних.

1 Краткие сведения из теории.

Алгоритм *m*-внутригрупповых средних основан на минимизации критерия качества Q и заключается в следующем.

1 Выбираются произвольно *m* исходных центров кластеров

$$Z^1(1), Z^2(1), Z^m(1),$$

где 1 – номер итерации, в данном случае первый.

2 На *к*-м шаге алгоритма все объекты распределяются по кластерам по правилу

$$Z \in V_i(k), \text{ если}$$

$$\rho(Z, Z^i(k)) < \rho(Z, Z^j(k)),\ j = 1,2,\ldots,m;\ i \neq j,$$

где $V_i(k)$ – кластер, центром которого является $Z^j(k)$.

3 Выбираются новые центры кластеров $Z^j(k+1)$, чтобы минимизировать сумму квадратов расстояний между объектами, принадлежащими $V_j(k)$, и новым центром кластером, т. е. в качестве $Z^j(k+1)$ берется выборочное среднее, определенное по множеству

точек, входящих в $V_j(k)$:

$$X^j(k+1) = \frac{1}{l_j} \sum_{X \in V_j(k)} X, \quad j = 1, 2, \ldots, m.$$

4 Если $Z^j(k+1) = Z(k)$, то алгоритм заканчивает работу, иначе он повторяется с шага 2.

Качество работы рассмотренного алгоритма зависит от выбора исходных центров кластеров и их числа, а также от геометрических особенностей данных. Поэтому в большинстве случаев практическое применение алгоритма *m*-внутригрупповых средних требует проведения

экспериментов, связанных с выбором расчетных значений параметра *m* и исходного расположения центров кластеров.

2 Порядок выполнения работы

1 Смоделировать распознающую систему, написать и отладить программу на языке программирования высокого уровня для классификации объектов с произвольными значениями признаков. Программа должна запрашивать у пользователя значения обучающей выборки (или читать их из файла), выдавать на экран или принтер исходные данные и результат классификации.

2 Для проверки работоспособности программы по своему варианту (номер в журнале) ввести исходные данные, распечатать листинг программы, таблицу исходных данных и результатов. При выполнении задания исходные данные берутся из таблицы 5, при этом номер варианта прибавляется к значениям X_2, если он четный и к значениям X_1, если он нечетный.

3 Сделать выводы о применимости метода.

4 Оформить отчет, в котором привести листинг программы, копии экранов и результаты работы программы.

Контрольные вопросы

1 Почему рассмотренный алгоритм называется алгоритмом *m*-внутригрупповых средних?

2 Докажите, что для множества точек $V=(Z_1, Z_2, \ldots, Z_l)$ центр кластера Z, обеспечивающий минимизацию суммы квадратов расстояний до каждой точки из V, представляет собой выборочное среднее,

$$X = \frac{1}{l}\sum_{i=1}^{l} X^i .$$

3 Примените алгоритм *m* внутригрупповых средних к множеству данных, приведенному в задаче 2 из лабораторной работы “Простейший алгоритм выявления кластеров”.

Список литературы

1. *Лябах Н.Н.* Математические основы разработки и использования машинного интеллекта. –Ростов н/Д , 1991. –157 с.

2. *Коршунов Ю.М.* Математические основы кибернетики. – М., 1987 – 286 с.

3. *Змитрович А.И.* Интеллектуальные информационные системы. – Минск : Тэтрасистэм, 1997. – 354 с.

4. *Фор А.* Восприятие и распознавание образов. – М. Машиностроение, 1980. – 272 с.

СОДЕРЖАНИЕ

Учебное издание
РЯЗАНЦЕВА Наталья Васильевна

КЛАССИФИКАЦИЯ ОБЪЕКТОВ С ИСПОЛЬЗОВАНИЕМ ТЕОРИИ РАСПОЗНАВАНИЯ ОБРАЗОВ

Лабораторный практикум
Часть I
Редактор Т.М. Ризевская
Технический редактор В. Н. Кучерова
Корректор
Подписано в печать . Формат 60x84 $^{1}/_{16}$. Бумага газетная
Гарнитура *Times New Roman.* Печать на ризографе.
Усл. печ. л . Уч.-изд. л . Тираж 100 экз.
Зак. № . Изд. №

Редакционно-издательский отдел УО «БелГУТ», 246653, г. Гомель,
ул. Кирова, 34.
Лиц. № 02330/0133394 от 19.07.2004 г.
Типография УО «БелГУТ», 246022, г. Гомель, ул. Кирова, 34.

Лиц. № 02330/0148780 от 30.04.2004 г.