Известия Гомельского государственного университета
имени Ф. Скорины, №4(37), 2006

*УДК 004.9*

# Об одном методе реализации процедуры обучения при построении системы распознавания образов

В. Г. Родченко

## Введение

Использование методов прикладной статистики и математической теории распознавания образов часто является единственно возможной альтернативой применению традиционного математического аппарата при проведении научных исследований и при решении большого числа прикладных задач, связанных с моделированием поведения многомерных объектов и сложных систем [1].

Современный подход к проблеме построения системы распознавания предусматривает решение двух взаимосвязанных задач – *обучения* и *принятия решения (контроля)* [2]. Только совместное решение этих задач решает проблему в целом и позволяет перейти к конкретному практическому использованию методов математической теории распознавания образов [3]. Реализация процедуры *обучения* непосредственно связана с анализом данных, которые априори представляют собой исходную *классифицированную обучающую выборку* (КОВ). Такая выборка образуется путем объединения данных об объектах, которые изначально классифицированы и описываются с помощью множества признаков, составляющих априорный словарь.

В реальных системах распознавания формирование *априорного словаря признаков* (АСП) начинается с построения *пространства наблюдений* (ПН) для соответствующих объектов. При этом необходимо в ПН включать такие характеристики объектов, которые обеспечивали бы разделение различных классов в этом пространстве. Фактически АСП получается в результате формализации пространства наблюдений и представляет собой выборку из *генерального словаря признаков*. Практический опыт построения АСП свидетельствует о том, что часть признаков, включаемых в априорный словарь, не несёт в себе разделяющую классы функцию, а потому эти признаки создают “шумы” при идентификации образов объектов и классов и, как следствие, приводят к искажению достоверности процедуры распознавания [4].

В данной статье предлагается метод реализации процедуры обучения на основе анализа данных из исходной классифицированной обучающей выборки, который предусматривает построение рабочего словаря, включающего в себя только информативные признаки с точки зрения разделения образов классов в соответствующем признаковом пространстве. Для выделения из априорного словаря указанных информативных признаков предлагается использовать статистические критерии однородности, а проверку достоверности процедуры принятия решения реализовать с помощью алгоритмов кластерного анализа типа FOREL-2, которые ориентированы на построение заранее заданного числа кластеров.

## Описание метода обучения

Процедура обучения при реализации системы распознавания строится на основе использования исходной классифицированной обучающей выборки. Процесс построения КОВ предполагает предварительное формирование *алфавита классов* и определение исходного *пространства наблюдений*. Отметим, что процедура формирования алфавита классов обычно особых затруднений не вызывает, тогда как определение пространства наблюдений является нетривиальной задачей, содержащей в себе ряд “тонких” моментов, которые могут в дальнейшем способствовать искажению достоверности результатов распознавания или даже приводить к серьезным ошибкам.

Процесс формирования пространства наблюдений часто трудно формализуется и плохо поддается автоматизации. При разработке систем распознавания для формирования ПН привлекаются квалифицированные эксперты, а это способствует повышению роли субъективного фактора. После определения пространства наблюдений проводится формализация, и в итоге строится априорный словарь признаков. Фактически любой априорный словарь представляет собой выборку из генерального словаря признаков, а следовательно, можно сформировать *приемлемый* для построения системы распознавания АСП, но нельзя сформировать оптимальный словарь. Кроме того, не исключается ситуация, когда в априорный словарь попадут такие признаки, которые не обеспечивают разделение классов в соответствующем признаковом пространстве, то есть в этом случае речь идет о *неприемлемом* для распознавания варианте априорного словаря.

Предположим, что имеется алфавит классов $A=\{A_1,A_2,...,A_k\}$ и сформирован априорный словарь признаков $P=\{P_1,P_2,...,P_n\}$. Каждый отдельный объект описывается n признаками из АСП и однозначно ассоциируется с одним из классов, а каждый класс $A_i$ (где $i=\overline{1,k}$) образуется путем объединения $m_i$ (где $i=\overline{1,k}$) многомерных объектов.

Вектор-столбец $\begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \\ i \end{pmatrix}$ (где $j=\overline{1,m_i}$) задает формальное описание j-го объекта i-го класса. Объединение всех таких векторов, описывающих объекты одного и того же класса, образует матрицу, которая представляет собой формальное описание соответствующего i-го класса в многомерном априорном признаковом пространстве и имеет вид:

$$X^i_{n\,x\,m_i} = \begin{pmatrix} x^i_{11} & x^i_{12} & \dots & x^i_{1m_i} \\ x^i_{21} & x^i_{22} & \dots & x^i_{2m_i} \\ \dots & \dots & \dots & \dots \\ x^i_{n1} & x^i_{n2} & \dots & x^i_{nm_i} \end{pmatrix}, \text{ где } i=\overline{1,k};\ j=\overline{1,m_i}.$$

В итоге классифицированная обучающая выборка $X_{n\,x\,m}$ (где n – количество признаков априорного словаря, а $m=m_1+m_2+\dots+m_k$, где k – количество классов) будет получаться путем объединения всех соответствующих матриц вида $X^i_{n\,x\,m_i}$, т.е. $X_{n\,x\,m}=\bigcup_{i=1}^{k} X^i_{n\,x\,m_i}$.

После завершения формирования классифицированной обучающей выборки сразу же можно воспользоваться процедурой кластеризации с целью исследования возможности использования признаков из АСП для построения эталонов классов. В данном случае, во-первых, следует воспользоваться алгоритмами кластеризации, которые ориентированы на построение точно заданного количества кластеров, во-вторых, число кластеров должно равняться количеству классов k. Если в итоге выполнения процедуры кластеризации оказалось, что в состав соответствующих кластеров вошли в подавляющем большинстве представители одного класса и не превышен допустимый пороговый уровень ошибочных включений объектов в другие классы, то на основе признаков из АСП можно строить эталоны классов и проводить процедуру принятия решения. Иначе же необходимо провести дополнительный разведочный анализ информативности признаков.

По степени информативности (с точки зрения отражения характерных для каждого класса индивидуальных свойств) признаки, попадающие в априорный словарь, следует сепарировать на три вида [5].

К первому виду относятся те признаки из АСП, значения которых фактически подчиняются одному и тому же закону распределения для всех классов $A=\{A_1,A_2,...,A_k\}$. Природа этих признаков такова, что они не несут разделяющей разные классы функции, а потому со-

здают “шумы” и на этапе обучения системы (размывая эталоны классов), и в процессе выполнения процедуры принятия решения (искажая распознаваемый образ).

Признаки будут отнесены ко второму виду, если в результате сопоставления всех пар выборок значений этих признаков из разных классов оказалось, что не выполняются соответствующие статистические критерии однородности. Свойства признаков этого вида таковы, что обеспечивается разделение образов классов в многомерном признаковом пространстве. Именно эти признаки будут включаться в словарь информативных признаков, на основе которого будет выполняться процедура построения компактных и разделенных в многомерном признаковом пространстве эталонов классов, а затем будет выполняться и процедура принятия решения (распознавания исследуемого объекта).

Если в процессе проводимого анализа очередной признак не является представителем ни первого, ни второго вида, то его относят к третьему виду. Такие признаки не отражают какие-либо четко выраженные внутриклассовые и межклассовые особенности, а потому размывают образы эталонов классов на этапе обучения, тем самым провоцируя серьезные помехи для качественного выполнения процедуры принятия решения.

**Описание алгоритма реализации метода**

Алгоритм реализации процедуры обучения при построении системы распознавания образов предполагает выполнение следующих семи шагов:

Шаг 1. Формируются алфавит классов $A=\{A_1,A_2,...,A_k\}$ и определяется исходное пространство наблюдений, на основе которого в итоге строится априорный словарь признаков $P=\{P_1,P_2,...,P_n\}$.

Шаг 2. Каждый класс $A_i$ (где $i=\overline{1,n}$) изначально определяется совокупностью объектов. В свою очередь, каждый отдельный объект (на основе признаков из априорного словаря) описывается в многомерном признаковом пространстве в виде вектора-столбца $x^T=(x_1,x_2,...,x_n)$, где $x_i$ – значение i-го признака.

Шаг 3. Формируется классифицированная обучающая выборка путем объединения всех соответствующих векторов из всех классов. Эта выборка представляет собой прямоугольную матрицу, состоящую из n строк и m столбцов (где $m=m_1+ m_2+ \ldots+m_k$, а $m_i$ – количество объектов i-го класса). При этом для $\forall\ A_i \subset A$ ($i=\overline{1,k}$) формируется матрица $X_i$ размерности n x $m_i$, где $m_i$ – число объектов i-го класса.

Шаг 4. Используя алгоритм кластеризации типа Forel-2, строится k кластеров, и проводится анализ их содержимого. Если каждый кластер содержит в основном объекты одного класса и не превышается допустимое пороговое значение ошибочных включений объектов в кластеры, то все признаки из априорного словаря переносятся в словарь информативных признаков. Далее осуществляется переход к шагу 6 этого алгоритма, а иначе выполняется следующий шаг.

Шаг 5. Последовательно анализируются все признаки из априорного словаря $P=\{P_1,P_2,...,P_n\}$, и в результате они сепарируются на три вида: $P^{(1)}=\{P_1^{(1)},P_2^{(1)},...,P_{n1}^{(1)}\}$, $P^{(2)}=\{P_1^{(2)},P_2^{(2)},...,P_{n2}^{(2)}\}$, $P^{(3)}=\{P_1^{(3)}, P_2^{(3)},..., P_{n3}^{(3)}\}$, где $P= P^{(1)} \cup P^{(2)} \cup P^{(3)}$ и $n_1+n_2+n_3=n$.

Очередной признак $P_i$ (где $i=\overline{1,n}$) классифицируется к одному из указанных выше видов на основе правила:

– если для всех пар классов соответствующий критерий однородности не показал существенного различия между выборками значений этого признака для двух сравниваемых классов, то $P_i$ относится к первому виду;

– если для всех пар классов соответствующий критерий однородности показал существенное различие между выборками значений этого признака для двух сравниваемых классов, то $P_i$ является признаком второго вида;

– если для признака $P_i$ не выполнилось ни одно из двух предыдущих условий, то он относится к третьему виду.

В результате только признаки второго вида $P^{(2)}=\{P_1^{(2)},P_2^{(2)},...,P_{n2}^{(2)}\}$ включаются в словарь информативных признаков.

Отметим, что не исключена ситуация, когда этот словарь оказывается пустым. В этом случае необходимо вернуться к формированию нового варианта априорного словаря, и затем сначала повторить процедуру анализа информативности признаков.

Шаг 6. Вновь, как и на шаге 4, с помощью алгоритма кластеризации типа Forel-2 строится k кластеров и проводится анализ их содержимого. Если каждый кластер содержит в основном объекты одного класса и не превышается допустимое пороговое значение ошибочных включений объектов в кластеры, то все признаки из словаря $P^{(2)}=\{P_1^{(2)},P_2^{(2)},...,P_{n2}^{(2)}\}$ переносятся в словарь информативных признаков.

Шаг 7. На основе классифицированной обучающей выборки и полученного словаря информативных признаков формируются эталоны классов. На этом процедура обучения заканчивается.

## Заключение

При построении системы распознавания фактически реализуются две основные процедуры, первая из которых связана с обучением системы, а вторая – с принятием решения (непосредственно распознаванием). Достоверность второй процедуры напрямую зависит от качественной реализации процедуры обучения. Если на этапе обучения удается сформировать такое рабочее признаковое пространство, в котором эталоны классов разделены и компактны, то выполнение процедуры принятия решения принципиальных затруднений не вызывает и носит чисто технический характер. В статье представлен метод реализации процедуры обучения на основе исследования информативности признаков с использованием алгоритмов кластерного анализа. Сепарирование признаков по степени их информативности на три вида, прежде всего, ориентировано на исключение признаков, искажающих образы эталонов классов. В конечном итоге происходит не только “фокусировка” образов, но и сжатие размерности пространства, что требует меньших ресурсов для выполнения процедур обучения и распознавания.

**Abstract.** The paper presents a method of the realization of the instruction procedure (instruction of the system) on the basis of studying self-descriptiveness of the signs with the use of cluster analysis algorithms.

## Литература

1. Распознавание образов: состояние и перспективы: Пер. с англ./ К.Верхаген, Р.Дёйн, Ф.Грун и др. – М.: Радио и связь, 1985. – 104 с.

2. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Издательство Института математики, 1999. – 266 с.

3. Васильев В.И. Проблема обучения распознаванию образов. – К.: Выща шк. Головное изд-во, 1989. – 64 с.

4. Родченко В.Г. Технология атрибуционных исследований на основе методов теории распознавания образов // Веснік ГрДУ. Сер. 2. – 2001. – №2(28). – С. 89 – 94.

5. Родченко В.Г. Об одном методе построения компактных эталонов классов при проектировании систем распознавания образов // Известия Гомельского государственного университета имени Ф.Скорины. – 2004. – №4(25). – С.114–117.

Гродненский государственный
университет имени Янки Купалы

Поступило 10.05.06