OSTIS-2011
(«Open Semantic Technologies for Intelligent Systems»)

**УДК 004.942**

# МЕТОД АВТОМАТИЧЕСКОГО ВЫПОЛНЕНИЯ ПРОЦЕДУРЫ ОБУЧЕНИЯ ПРИ ПОСТРОЕНИИ СИСТЕМЫ РАСПОЗНАВАНИЯ

А.И. Жукевич (*san@grsu.by*)
Е.В. Олизарович (*e.olizarovich@grsu.by)*
В.Г. Родченко (*rovar@mail.ru)*
*Гродненский государственный университет имени Янки Купалы, г. Гродно, Республика Беларусь*

Предложен метод, которым предусматривается на основе использования априорного словаря признаков, алфавита классов и классифицированной обучающей выборки в автоматическом режиме реализовать процедуру обучения при построении системы распознавания. В результате формируется пространство решений, в котором формальные образы эталонов классов представляют собой компактные и разделенные кластеры.

*Ключевые слова:* алфавит классов, словарь признаков, классифицированная обучающая выборка, распознавание образов.

## Введение

При проведении научных исследований или при решении целого ряда прикладных задач в области естественнонаучных, социально-экономических и гуманитарных дисциплин применение аналитических методов оказывается весьма затруднительным в силу того, что исследуемые объекты априори характеризуются большим числом признаков, которые в свою очередь имеют различную природу и часто измеряются в разных шкалах. [Айвазян, 1998] В данном случае в качестве альтернативы более эффективным может оказаться использование аппарата математической теории распознавания образов.

В общем случае система распознавания должна предусматривать выполнение двух взаимосвязанных процедур – *обучения* и *непосредственно распознавания* (принятия решения). Причем в конечном итоге достоверность работы системы распознавания в определяющей степени зависит от качественного выполнения процедуры обучения. [Васильев, 1989]

Процесс обучения осуществляется путем анализа данных, которые предварительно формируются в матричном виде и представляют собой классифицированную обучающую выборку. Указанная обучающая выборка формируется в результате выполнения подготовительной относительно простой процедуры первичной обработки наблюдаемых данных на основе использования априорного словаря признаков.

Отметим, что вопросы, связанные с определением алфавита классов, набора наблюдаемых данных и построением априорного словаря признаков носят проблемно-ориентированный характер, и при этом являются отдельными и иногда нетривиальными задачами.

В априорный словарь необходимо включить признаки, которые описывают особые специфические характеристики объектов и соответствующих классов, обеспечивая разделение классов в многомерном признаковом пространстве решений. Однако на практике обычно оказывается, что часть из включенных в априорный словарь признаков не несут разделяющей классы функции. Они “размывают” формальные образы объектов и классов, что в конечном итоге приводит к серьезному искажению достоверного выполнения всей процедуры распознавания. [Родченко, 2007]

При построении реальных систем распознавания часто оказывается, что только

относительно небольшое число признаков, из первоначально включенных в априорный словарь, представляют интерес для качественного выполнения заключительной процедуры принятия решения. [Вакульчик, 2005]

Отметим, что неправильный выбор множества признаков при построении пространства решений, даже в случае относительно небольшого их числа, приводит в конечном итоге к содержательно ошибочной классификации, хотя при этом с формальной точки зрения она может выглядеть и достаточно обоснованной.

В данной работе предлагается метод автоматического выполнения процедуры обучения, которым предусматривается при реализации системы распознавания построение на основе анализа данных из классифицированной обучающей выборки такого словаря, который включает в себя только информативные признаки с точки зрения разделения формальных образов классов в соответствующем признаковом пространстве решений. В результате строится пространство решений, в котором в кластерном виде формируются компактные эталоны, представляющие собой формальные образы классов. Дальнейшее выполнение непосредственно процедуры распознавания принципиальных сложностей не вызывает. [Загоруйко, 1999]

## 1. Описание метода автоматического обучения

Пусть имеются алфавит классов $A=\{A_1,A_2,...,A_k\}$ и априорный словарь признаков $P=\{P_1,P_2,...,P_n\}$. Каждый класс $A_i$ (где $i=\overline{1,k}$) изначально определяется набором из $m_i$ (где $i=\overline{1,k}$) многомерных объектов. При этом каждый объект описывается $n$ признаками из априорного словаря и однозначно ассоциируется с одним из классов $A_i$ (где $i=\overline{1,k}$). Множество объектов одного класса образует формальное описание этого класса в априорном признаковом пространстве. Объединение всех объектов из всех классов $A_1,A_2,...,A_k$ образует исходную классифицированную обучающую выборку. Эта выборка представляет собой таблицу типа “объект–свойство” и формально представляется в виде матрицы размерности $n$ x $m$, где $m=m_1+m_2+\ldots+m_k$, и $m_i$ – количество объектов $i$-го класса.

С точки зрения представления разделяющих между собой классы характеристик, попадающие в априорный словарь признаки следует классифицировать на три вида. [Родченко, 2004]

К первому виду следует отнести те признаки из априорного словаря, значения которых фактически подчиняются одному и тому же закону распределения для всех классов $A=\{A_1,A_2,...,A_k\}$. Сущность этих признаков такова, что они не несут разделяющей разные классы функции, а потому будут создавать “помехи” как на этапе обучения системы распознавания при построении эталонов классов, так и в процессе выполнения непосредственно процедуры распознавания.

Следует признак из априорного словаря отнести ко второму виду, если в результате сопоставления всех пар выборок значений этого признака из разных классов оказалось, что не выполняются соответствующие критерии однородности. Именно признаки этого вида по своей природе обеспечивают разделение формальных образов классов в многомерном признаковом пространстве. Они и образуют словарь информативных признаков, на основе которого будут выполняться процедура построения компактных и разделенных в многомерном признаковом пространстве эталонов классов и процедура распознавания исследуемого образа.

Если же в результате проводимого анализа признак не относится к первому или второму виду, то его будем относить к третьему виду. Сущность признаков третьего вида такова, что они не отражают какие-либо четко выраженные межклассовые различия, а потому будут создавать “помехи” как на этапе обучения системы, так и при выполнении заключительного этапа процедуры распознавания.

Исходная классифицированная обучающая выборка подвергается перестройке путем исключения из нее всех строк, содержащих значения признаков первого и третьего видов, и в результате получается промежуточная выборка. На основе содержимого этой выборки проводится нормировка значений к единичному интервалу, и в итоге формируется классифицированная выборка эталонных значений в признаковом пространстве решений, сформированном с использованием признаков второго вида. Затем строятся эталоны-кластеры классов, и процедура обучения системы распознавания завершается.

## 2. Алгоритм реализации метода

Для выполнения алгоритма автоматического обучения при построении системы распознавания необходимо предварительно сформировать алфавит классов, априорный словарь признаков и исходную классифицированную обучающую выборку. Соответствующий алгоритм подготовительной процедуры предполагает выполнение следующих трех шагов:

1. На основе проведения предварительного анализа исследуемых объектов формируются алфавит классов $A=\{A_1,A_2,...,A_k\}$ и априорный словарь признаков $P=\{P_1,P_2,...,P_n\}$.

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

3. Объединение всех соответствующих вектор-столбцов из всех классов образуют классифицированную обучающую выборку. Выборка представляет собой прямоугольную таблицу типа "объект-свойство", которая состоит 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):

$$X^i_{n\ x\ m_i} = \begin{pmatrix} x^i_{11} & x^i_{12} & \ldots & x^i_{1m_i} \\ x^i_{21} & x^i_{22} & \ldots & x^i_{2m_i} \\ \ldots & \ldots & \ldots & \ldots \\ x^i_{n1} & x^i_{n2} & \ldots & x^i_{nm_i} \end{pmatrix}, \text{ где } i=\overline{1,k} \qquad (1)$$

Результатом выполнения этого шага алгоритма будет являться классифицированная обучающая выборка $X_{n\ x\ m} = \Gamma_{i=1}^{k} X^i_{n\ x\ m_i}$ (где $n$ – количество признаков априорного словаря, а $m=m_1+m_2+\ldots+m_k$), которая получается при объединении всех матриц $X^i_{n\ x\ m_i}$:

$$X_{n\ x\ m} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix} \qquad (2)$$

Непосредственно алгоритм обучения предусматривает выполнение следующих четырех шагов в автоматическом режиме:

Шаг 1. Последовательно анализируются все признаки из априорного словаря $P=\{P_1,P_2,...,P_n\}$, и в результате они разбиваются на три вида: $P^{(1)}=\{P_1^{(1)},P_2^{(1)},...,P_f^{(1)}\}$, $P^{(2)}=\{P_1^{(2)},P_2^{(2)},...,P_q^{(2)}\}$, $P^{(3)}=\{P_1^{(3)}, P_2^{(3)},..., P_t^{(3)}\}$, где $P= P^{(1)} \cup P^{(2)} \cup P^{(3)}$ и $f + q + t = n$.

Отнесение очередного анализируемого признака $P_i$ (где $i=\overline{1,n}$) к одному из трех видов производится по следующему правилу:

- если для всех пар классов соответствующий критерий однородности не показал существенного различия между выборками значений этого признака для двух сравниваемых классов, то соответствующий признак $P_i$ 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 признак $P_i$ является признаком второго вида;
- если для признака $P_i$ не выполнилось ни одно из двух предыдущих условий, то он относится к третьему виду.

Шаг 2. На основе полученного словаря из признаков второго вида $P^{(2)}=\{P_1^{(2)},P_2^{(2)},...,P_q^{(2)}\}$ формируется пространство решений. Из классифицированной обучающей выборки исключаются

все строки, содержащие значения признаков, не попавших в словарь $P^{(2)}$. Получается промежуточная выборка $X_{q\ x\ m}$.

Шаг 3. На основе содержимого промежуточной выборки проводится нормировка значений к единичному интервалу по следующей формуле:

$$y_{ij} = \frac{(x_{ij} - x_{i\min})}{(x_{i\max} - x_{i\min})}, \text{ где } i=\overline{1,q}, \ j=\overline{1,m}. \quad (3)$$

В результате формируется классифицированная выборка эталонных значений, которая представляет собой матрицу $Y_{q\ x\ m}$:

$$Y_{q\ x\ m} = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1m} \\ y_{21} & y_{22} & \dots & y_{2m} \\ \dots & \dots & \dots & \dots \\ y_{q1} & y_{q2} & \dots & y_{qm} \end{pmatrix}, \text{ где } y_{ij} \in [0,1] \quad (4)$$

Шаг 4. На основе полученной классифицированной выборки эталонных значений строятся эталоны-кластеры классов в полученном многомерном пространстве решений, и завершается процедура обучения.

Отметим, что может возникать ситуация, когда в результате выполнения первого шага алгоритма словарь $P^{(2)}=\{P_1^{(2)},P_2^{(2)},...,P_q^{(2)}\}$ оказывается пустым, и тогда необходимо возвращаться и формировать новый вариант априорного словаря признаков, а затем повторно выполнять процедуру обучения системы.

## 3. Заключение

Разработаны метод и соответствующий алгоритм для автоматического выполнения процедуры обучения на основе использования классифицированной обучающей выборки. Предусматривается сепарирование признаков по степени их информативности на три вида, что обеспечивает исключение признаков, “размывающих” образы эталонов-кластеров классов (признаки первого и третьего видов) и выделение наиболее информативных с точки разделения образов классов (признаки второго вида).

В результате обучения системы осуществляется анализ информативности всех признаков из априорного словаря. При этом обеспечивается не только “фокусировка” формальных образов классов, но и сжатие размерности пространства решений, что в конечном итоге потребует меньших вычислительных ресурсов для распознавания.

## Библиографический список


**[Айвазян, 1998]** Айвазян, С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян – М., 1998.

**[Вакульчик и др., 2005]** Вакульчик, В.Г. Об одном методе построения математической модели исследования патологических процессов: диагностика острого аппендицита у детей / В. Г. Вакульчик [и др.] // Известия Гомельского государственного университета имени Франциска Скорины - 2005. - №5(35). - С.16-19.

**[Васильев, 1989]** Васильев, В.И. Проблема обучения распознаванию образов / В.И. Васильев – К., 1989.

**[Загоруйко, 1999]** Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко – Новосибирск, 1999.

**[Родченко, 2004]** Родченко, В.Г. Об одном методе построения компактных эталонов классов при проектировании систем распознавания образов / В. Г. Родченко // Известия Гомельского государственного университета имени Франциска Скорины - 2004. - № 4(25). - С.114-117.

**[Родченко, 2007]** Родченко, В.Г. Метод реализации стилеметрических исследований на основе применения аппарата математической теории распознавания образов / В. Г. Родченко // Известия Гомельского государственного университета имени Франциска Скорины - 2007. - № 5(44). - С.58-62.