УДК 519.6

В. И. Бритик, Е. А. Байдикова

# АЛГОРИТМ ОБУЧЕНИЯ ИДЕНТИФИКАЦИИ МНОГОМЕРНЫХ ОБЪЕКТОВ

## 1. Введение

Распознавание образов является одной из актуальных задач в различных сферах жизнедеятельности общества: в биологии, геологии, социологии, астрономии, медицине и прочих областях [1]. В зависимости от цели, задачу распознавания можно представить как задачу:

1) классификации объектов;

2) поиска и идентификации объекта.

Задача классификации предусматривает наличие набора статистически достоверных признаков, характеризующих исследуемое множество объектов. Решение данной задачи заключается в поиске границ областей, образуемых подмножествами множества исследуемых объектов, которые обладают свойством общности (подобности) признаков. Случай, когда результатом классификации является множество непересекающихся областей (классов) по каждому характерному признаку, является идеальным. Однако при решении реальных задач классификации эта ситуация далеко не всегда имеет место.

Наличие пересекающихся классов, полученных в процессе классификации, ставит перед исследователями дополнительные вопросы относительно поиска отличительных признаков объектов из области пересечения, которые являются причиной неопределенности результатов классификации. Поиск вышеуказанных признаков, т. е. поиск отличий объектов, представляет собой задачу идентификации.

## 2. Постановка проблемы

При наличии достаточного количества признаков задача идентификации решается методом полного перебора всех признаков и их сочетаний. Однако этот механизм практически не реализуем из-за большой вычислительной сложности, связанной с размерностью реальных объектов.

Алгоритм поиска идентификационных признаков обычно предполагает рекурсивную процедуру, условием остановки которой является наличие признака или совокупности признаков, позволяющих решить поставленную задачу идентификации. Однако возможно, что такие признаки могут быть и не найдены при заданном наборе признаков. Например, при наличии одинаковых значений признаков на ограниченном наборе признаков рекурсия не будет остановлена.

Актуальность данного исследования состоит в разработке механизма определения отличительных идентификационных признаков объектов.

Целью данного исследования является разработка алгоритма автоматического обучения идентификации образов, основанного на вычислении оценок, характеризующих отличительные признаки исследуемых объектов или объекта. Достижение поставленной цели предполагает решение следующих основных подзадач:

1) формирование набора параметров, характеризующих множество исследуемых объектов, и предварительная обработка этих исходных данных;

2) задание числа предполагаемых классов и грубое разбиение исходного множества объектов на заданное число классов;

3) разработка механизма, позволяющего оценить отличительные характеристики многомерных объектов;

4) построение решающих правил на основе полученных оценок отличительных характеристик признаков объекта.

Подзадачи 1)–2) представляют собой первый этап алгоритма автоматического обучения. Данные подзадачи в той или иной мере решались разными авторами [2].

Подзадачи 3)–4) являются завершающими в решении задачи идентификации. Их решение предусматривает определение и анализ отличительных признаков идентифицируемых объектов и характерных параметров (признаков) объектов каждого грубо сформированного класса, а также построение решающих правил.

## 3. Исходные данные

Любой объект или образ, подлежащий классификации или идентификации, может быть описан с помощью бесконечного количества каких-либо характеристик, и только некоторые из них могут быть использованы в качестве признаков для решения конкретной задачи распознавания, поставленной перед исследователем.

Решение задачи выбора признаков является необходимым в процессе синтеза любой системы распознавания. Однако универсального формализованного подхода для формирования набора признаков не существует. Конечное множество признаков зависит от предметной области и предполагает анализ значительного количества параметров исследуемых объектов, т. е. многоразовое экспериментирование с учетом знаний экспертов данной области.

Оптимальность набора признаков обосновать практически невозможно, можно говорить только об удовлетворяющем качестве сформированного набора признаков. Как правило, формирование набора признаков осуществляется в связи с задаваемым качеством обработки: эффективность выбранных признаков непосредственно связана с качеством распознающей системы. Обычно эта связь выражается

в терминах вероятности правильного распознавания (классификации или идентификации) [3].

Тем не менее, очевидно, что количество выбранных признаков, необходимое для успешного решения задачи распознавания, зависит от информативности этих признаков. Поэтому в процессе их выделения исследователи сталкиваются с проблемными ситуациями, решение которых заключается в поиске компромисса между качеством системы распознавания и набором признаков.

Например, результаты измерения признаков могут содержать дублированную информацию (т. е. имеет место корреляция признаков), что приводит к усложнению схемы соответствующей классификации или идентификации. При возникновении такой ситуации определяется информативность признаков и выбираются более существенные из них, чтобы создать более эффективную и точную систему. Эту процедуру часто называют предварительной обработкой с целью выделения признаков. С другой стороны, часто наиболее информативные признаки не всегда легко измерить или же возможности измерения ограничиваются экономическими факторами, что в свою очередь усложняет задачу выбора признаков и влияет на качество системы распознавания. Это приводит к ситуации, когда, в частности, приходится использовать менее информативные признаки, получение которых обходится дешевле.

Следует отметить, что при решении задачи выделения (измерения) признаков исследователи также сталкиваются с проблемой определения значений качественных признаков. Как правило, значения качественных признаков подвергаются преобразованию с целью их дальнейшего описания в более строгих шкалах измерений (двоичная шкала, шкала отношений и пр.).

При формировании набора признаков исследователи нередко обращаются к математическим признакам, например, статистическим средним, коэффициентам корреляции, характеристическим числам, собственным векторам ковариационных матриц и пр. Это позволяет получить дополнительную информацию, которую можно использовать при решении задачи распознавания в качестве новых признаков.

Таким образом, формирование набора признаков, который учитывает трудности, связанные с реализацией процессов выделения или выбора признаков, и в то же время обеспечивает необходимое качество классификации или идентификации, представляет собой одну из наиболее трудных задач при построении систем распознавания.

Однако большинство компьютерных экспертных систем интеллектуальной поддержки принятия решений, применяемых сегодня на практике, основаны на использовании формализованных программистами знаний специалистов какой-либо прикладной области без привлечения математиков, занимающихся проблемой обработки многомерных сигналов. Такие экспертные системы лишь отражают субъективное мнение их создателей и не выявляют необходимые для анализа взаимосвязи признаков объектов. Например, экспертная система диагностики ишемической болезни сердца [4]. В данной системе набор признаков классификации и число классов заданы в соответствии со сложившейся клинической практикой. Эти данные определяются один раз и не подвергаются сомнению.

При формировании исходных данных для нашего алгоритма предполагаем, что данные представляются в виде матрицы «объект-свойство», где каждая строка матрицы соответствует значениям выбранных параметров, полученных в результате измерения и характеризующих данный объект.

С целью упрощения анализа этих данных исходную матрицу преобразуем в матрицу нового вида – стандартизированную матрицу [2].

Введем следующие обозначения. Пусть исследователем выбраны для анализа $m$ объектов $A$, описываемые $n$ параметрами.

Если исходную матрицу записать в виде $Z=\{z_{ij}\}$, $i=\overline{1,m}$, $j=\overline{1,n}$, то переход к стандартизованной матрице $X=\{x_{ij}\}$, $i=\overline{1,m}$, $j=\overline{1,n}$, может быть осуществлен по следующим формулам:

$$z^{-j}=\frac{1}{m}\sum_{i=1}^{m} z_{ij},\ j=\overline{1,n}, \qquad (1)$$

$$\sigma_j^{2}=\frac{1}{m}\sum_{i=1}^{m}(z_{ij}-z^{-j})^{2},\ j=\overline{1,n}, \qquad (2)$$

$$x_{ij}=\frac{z_{ij}-z^{-j}}{\overline{\sigma}_j},\ i=\overline{1,m},\ j=\overline{1,n}, \qquad (3)$$

где $z_{ij}$ — значение $j$-го параметра для $i$-го объекта исходной матрицы; $x_{ij}$ — значение $j$-го параметра для $i$-го объекта стандартизированной матрицы.

Преобразования (1)–(3) можно рассматривать как приведение всех параметров к некоторой стандартной единой шкале. Это позволяет:

1) проводить сравнение параметров, имеющих различный физический смысл;

2) получить матрицу выборочных коэффициентов корреляции;

3) саму матрицу $X=\{x_{ij}\}$, $i=\overline{1,m}$, $j=\overline{1,n}$, рассматривать как $n$-мерное пространство, оси которого соответствуют отдельным измерениям объектов. Кроме этого, в результате данного преобразования мы получаем дополнительные признаки — статистические средние и выборочные дисперсии каждого признака.

### 4. Классификация многомерных объектов

Следует также отметить, что исследователь в процессе анализа исходного набора данных имеет дело не с самим объектом, а с его отображением — математическим объектом, который формируется путем

выбора и использования процедур измерения, отобранных исследователем некоторых характеристик анализируемого объекта. Произвольный выбор исследователем измерительных процедур характеристик объекта приводит к порождению различных математических объектов и отражается на результатах классификации. Это может привести к изменению числа классов объектов, которые согласно шкале наименований идентифицируются однообразно. Кроме того, различные методы классификации могут дать отличающиеся результаты [5].

В данном алгоритме число предполагаемых классов ($k$) анализируемых объектов задается на основании анализа некоторого числа выбранных параметров. В нашем случае — двух столбцов матрицы «объект-свойство», с минимальной и максимальной энтропией (количество измерений рассматривается как конечное множество событий). Анализируя каждый их этих параметров (признаков классификации) отдельно, можно автоматически определить возможное число разбиения объектов на классы при использовании соответственно только этих признаков.

Следует отметить, что описанный выше подход к заданию числа предполагаемых классов не единственный. В общем случае эту подзадачу можно сформулировать следующим образом: нужно определить число параметров из заданного множества (набора), которое необходимо проанализировать, чтобы решить вопрос о количестве классов исследуемых объектов. Результат решения данной подзадачи зависит от рассматриваемой предметной области, тенденций поведения значений параметров, знаний специалистов данной предметной области и пр.

Таким образом, число классов анализируемых объектов зависит от числа признаков, используемых при классификации, шкал измерения этих признаков и в значительной степени носит субъективный характер. Именно поэтому при решении задач классификации одним из главных является вопрос о числе предполагаемых классов анализируемых объектов.

Число предполагаемых классов определяется при грубом разбиении распознаваемых объектов на $k$ классов. Использование алгоритма грубого разбиения существенно влияет на результаты классификации, но в рамках данной работы мы его не рассматриваем.

### 5. Методика определения значений функций принадлежности

Механизм оценивания отличительных характеристик многомерных объектов в нашем алгоритме идентификации основывается на вычислении и анализе значений функций принадлежности исследуемых объектов к классам, полученным в результате грубой классификации по каждому признаку из исходного набора. Наличие пересекающихся областей, полученных в результате решения задачи грубой классификации, позволяет утверждать о размытости классов, поскольку в данном случае для некоторого объекта $x$, находящегося в области пересечения некоторых классов $K_i$ или $K_j$, стоит вопрос в том, до какой степени $x$ принадлежит классу $K_i$ и $K_j$.

В соответствии с изложенным ранее и с учетом фактора размытости классов обозначим через $X=\{X_1, X_2, ..., X_j, ..., X_n\}$ и $X_j=\{x_{1j}, x_{2j}, ..., x_{mj}\}^T$, $j=\overline{1, n}$, пространство и подпространство параметров соответственно. Подпространство параметров $X_j$, $j=\overline{1, n}$, в свою очередь может быть разбито на $k$ размытых множеств $\{A_{j1}, A_{j2}, ..., A_{jk}\}$, $j=\overline{1, n}$, где $k$ — число предполагаемых классов. Размытое множество $A_{js}$, $s=\overline{1, k}$, в $X_j$ есть совокупность упорядоченных пар $A_{js}=\{x_{ij}, \mu_{A_{js}}(x_{ij})\}$, $s=\overline{1, k}$, $i=\overline{1, m}$, $j=\overline{1, n}$, а $\mu_{A_{js}}(x_{ij})$ — функция принадлежности объекта $x_{ij}$ к $A_{js}$ по параметру $j$, характеризующая соответственно степень принадлежности $A_{js}$ к $X_j$.

Методика определения функции принадлежности значения исследуемого $j$-го признака к $s$-му классу основана на определении уровневых подмножеств заданного нечеткого множества и анализе частоты встречаемости значений $j$-го признака, относящегося к $s$-му классу [6].

Пусть $A_{js}$ — нечеткое множество конечного множества $X_j=\{x_{1j}, x_{2j}, ..., x_{mj}\}^T$, где $x_{ij}$, $i=\overline{1, m}$ — дискретное значение исследуемого признака $j$. Тогда нечеткому подмножеству $\alpha$-уровня нечеткого множества $A_{js}$ можно поставить в соответствие четкое подмножество $A_{js}^{\alpha}$ множества $X_j$, содержащее все элементы, степень принадлежности $\mu_{A_{js}}(x_{ij})$ которых не менее $\alpha$. То есть

$$A_{js}^{\alpha}=\{x_{ij} \mid \mu_{A_{js}}(x_{ij}) \geq \alpha,\ x_{ij} \in X_j\},\ \alpha \in [0, 1],\ i=\overline{1, m}.$$

Таким образом, нечеткое множество $A_{js}$, несмотря на нечеткость его границ, может быть точно определено путем сопоставления каждому значению числа $x_{ij}$, лежащего между нулем и единицей, которое представляет собой степень его принадлежности к $A_{js}$.

Методика определения степени принадлежности значений признака состоит в следующем:

1) Строится гистограмма, где по оси ординат откладываются количества встречаемости $j$-го признака $s$-го класса, полученного при грубом разбиении. По оси абсцисс откладываются дискретные значения этого признака.

2) Диапазон изменения количества встречаемости значений $j$-го признака, измеряемый от 0 до некоторого максимального значения, ассоциируется с единичным интервалом изменения степени принадлежности. Этот диапазон разбивается на $M$ частей равной длины. Число $M$ задается априорно, например, $M$ равно удвоенному числу классов, полученных при грубом разбиении.

3) В зависимости от принадлежности значений признака соответствующим поддиапазонам, полученным на предыдущем шаге, производится их

группирование. Вычисляются суммарные количества встречаемости значений признака $N_0, N_1, \ldots N_{M-1}$ для каждого поддиапазона.

4) Вычисляются степени принадлежности значений $j$-го признака, являющиеся элементами четкого подмножества $A_{js}^{\alpha_0}$ $(1 \geq \alpha_0 > 1-\frac{1}{M})$, т. е. тех элементов, значения встречаемости которых соответствуют поддиапазону от максимального значения (1) до $\frac{1}{M}$. Формула расчета принадлежности некоторого значения $x_{ij}$ нечеткому множеству $A_{js}$ получается из пропорции, составленной на том основании, что неизвестной степени принадлежности соответствует известная частота, вычисляемая как отношение количества встречаемости $q$-го значения $j$-го признака $\alpha_0$-го уровня $N_0(x_q)$ к общему количеству встречаемости значений признака $\alpha_0$-го уровня ($N_0$). А частоте встречаемости значений признака $\alpha_0$-го уровня четкого подмножества, при условии равенства количества встречаемости каждого значения признака, соответствует степень принадлежности, равная $\sup \alpha_0 = 1$. Тогда значение степени принадлежности $q$-го значения $j$-го признака классу $s$ равно

$$\mu_{A_{js}}(x_q) = \frac{N_0(x_q)N}{N_0 N_{\text{общ}}},$$

где $N = \sum_{m=0}^{M-1} N_m$ — суммарное количество встречаемости значений признака, степень принадлежности которых еще не определена; $N_{\text{общ}}$ — общее количество значений встречаемости исследуемого признака.

Отметим, что для $\alpha_0$-го уровня $N = N_{\text{общ}}$.

5) Вычисляются степени принадлежности значений признака, являющихся элементами четкого подмножества $A_{js}^{\alpha_1}$, где $(1-\frac{1}{M} \geq \alpha_1 > 1-\frac{2}{M})$. Формулы расчета степеней принадлежности аналогичны предыдущему шагу, за исключением того, что при их получении не учитываются элементы, степень принадлежности которых уже определена ранее.

Таким образом, рекуррентная формула вычисления степени принадлежности $q$-го значения $j$-го признака классу $s$ может быть представлена в следующем виде:

$$\mu_{A_{js}}(x_q) = \frac{N_k(x_q)(N - \sum_{k=0}^{K} N_k)M}{N_k N_{\text{общ}}(M-K)},$$

где $N_k(x_q)$ — количество встречаемости $q$-го значения $j$-го признака, соответствующего $K$-му уровню, $K = 0, 1, \ldots, M-1$; $N = \sum_{m=0}^{M-1} N_m$ — суммарное количество встречаемости неучтенных значений признака, соответствующих вычислительной процедуре данного уровня; $\sum_{k=0}^{K} N_k$ — количество встречаемости значений признака, учтенных на момент анализа $K$-го уровня; $M$ — число, определяющее единицу изменения (шкалу) анализируемых значений исследуемого $j$-го признака ($M > K$); $N_k$ — суммарное количество встречаемости значений признака $k$-го уровня; $N_{\text{общ}}$ — суммарное количество значений $j$-го признака, соответствующих анализируемому размытому множеству.

## 6. Заключение

Научная новизна данного исследования заключается в разработке алгоритма обучения идентификации, который основан на вычислении оценок отличительных характеристик многомерных объектов. Механизм оценивания заключается в определении значений функций принадлежности исследуемых объектов к классам, полученным в результате грубой классификации, по каждому признаку из исходного набора. Данный подход позволяет «уйти» от полного перебора всех признаков и их сочетаний, что приводит к упрощению вычислительной схемы алгоритма идентификации.

Практическая значимость данного исследования — возможность создания интеллектуальных систем распознавания, которые могут быть использованы в различных прикладных областях, характеризующихся большой размерностью исследуемых объектов.

**Список литературы:** 1. *Загоруйко Н. Г.* Методы распознавания и их применение. — М.: Советское радио, 1972. 2. *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирических данных. — М.: Наука. Главная редакция физико-математической литературы, 1983. 3. *Ту Дж., Гонсалес Р.* Принципы распознавания образов / Пер. с англ. Н. Б. Гуревича под ред. Ю. И. Журавлева. — М.: Мир, 1978. 4. *Рощейн А. П.* Интеллектуальные технологии идентификации: нечеткие множества, генетические алгоритмы, нейронные сети. — Винница: УНИВЕРСУМ-Винница, 1999. 5. *Шапиро Л., Стокман Дж.* Компьютерное зрение / Пер. с англ. — М.: БИНОМ. Лаборатория знаний, 2006. — 752 с. 6. *Бритик В. И.* Локально-адаптивные фильтры в задачах обучения распознаванию образов // Автоматизированные системы управления и приборы автоматики. — Харьков, 1988.

*Поступила в редколлегию 04.10.2006*