Известия Гомельского государственного университета
имени Ф. Скорины, №5(62), 2010

УДК 004.02:004.9

Об одном методе построения формальных образов классов при реализации систем распознавания

А. И. Жукевич, В. Г. Родченко

Введение. Аппарат математической теории распознавания образов часто оказывается наиболее приемлемым, а нередко единственным инструментом, который предоставляет исследователю реальную возможность изучать явления и объекты сложной природы, характеризующиеся большим числом разнообразных по своей природе признаков.

При построении систем распознавания образов предполагается качественное выполнение двух основных процедур, к которым относится процедура обучения и процедура принятия решения или контроля. Если процедуру обучения удается реализовать эффективно, то выполнение второй процедуры носит технический характер и затруднений не вызывает. В практических задачах именно процесс обучения является наиболее трудоемким с точки зрения реализации, поскольку в реальных системах исследуются объекты, которые характеризуются большим количеством разнообразных признаков, имеющих сложную природу и распределенными по разным законам [1].

Процедура обучения проводится на основе исходных данных, представляющих собой классифицированную обучающую выборку. В результате выполнения этой процедуры должна быть установлена закономерная связь между значениями признаков и соответствующими классами объектов. Получаемая закономерность выражается в виде решающего правила, на основе которого выполняется процедура принятия решения, и исследуемый объект либо относится к одному из исходных классов, либо выделяется в отдельный самостоятельный класс, так называемый *джокер-класс*.

В идеальном случае построение классифицированной обучающей выборки осуществляется на основе использования такого априорного словаря признаков (АСП), в котором содержаться только разделяющие классы признаки. В свою очередь, априорный словарь представляет собой выборку из соответствующего генерального словаря. В реальных задачах априори сформированная выборка не гарантирует полной объективности, а часто строится с учетом ограничений, присущих имеющимся ресурсам. Опыт показывает, что редко удается априори определить соответствующий набор признаков, а потому в АСП включаются и признаки, не несущие разделяющей классы функции, а значит создающие “помехи” при распознавании и существенно ухудшающие ее качество. С целью обеспечения достоверного выполнения процедуры распознавания в системах предусматривается реализация дополнительного этапа исследований, который связан, во-первых, с анализом данных из классифицированной обучающей выборки, и, во-вторых, с формированием пространства решений, обеспечивающим разделение эталонов классов в нем [2]. Построение пространства решений осуществляется на основе получаемого из АСП уточненного рабочего словаря признаков (РСП).

После завершения процедуры построения рабочего словаря признаков предусматривается этап аттестации достоверности процедуры распознавания в сформированном именно на основе признаков из РСП пространстве решений. В данной статье предлагается метод построения формальных образов классов при реализации систем распознавания, который базируется на использовании оригинального алгоритма кластеризации.

Постановка задачи. Существует целый ряд методов реализации систем распознавания образов, которые базируются на построении разделяющих классы поверхностей в многомерном признаковом пространстве. Так или иначе, в этом случае предполагается, что соответствующие поверхности могут быть построены. В первую очередь указанные предположе-

ния обосновываются тем, что словарь признаков, на основе которого производится описание исследуемых объектов, включает именно такие признаки, которые обеспечивают разделение образов классов в соответствующем признаковом пространстве. Практическое же использование математического аппарата теории распознавания образов демонстрирует совершенно другую тенденцию, которая связана с тем, что в исходный вариант априорного словаря реально попадают признаки, во-первых, не несущие разделяющей классы функции, и, во-вторых, создающие шумы при выполнении непосредственно процедуры распознавания [3, 4]. Множество признаков такого словаря не обеспечивает выполнение гипотезы компактности, а потому использование этих признаков для проведения процедуры распознавания в большинстве случаев будет приводить к серьезным искажениям и ошибочным результатам.

Для обеспечения качественного распознавания необходимо предусмотреть выполнение этапа исследования, связанного с анализом информативности каждого из признаков, включенного в априорный словарь, с целью формирования такого уточненного словаря (рабочего словаря признаков), который содержал бы только признаки, обеспечивающие разделение всех образов классов попарно между собой в многомерном пространстве решений. Фактически в этом случае образы классов в признаковом пространстве будут представлять собой отдельно размещенные непересекающиеся кластеры, и при этом будет обеспечиваться условие, связанное с выполнением гипотезы компактности.

Наличие исходной классифицированной обучающей выборки при построении системы распознавания образов предоставляет возможность реализовать процедуру обучения путем проведения анализа данных из этой выборки. Сепарирование признаков по степени их информативности с точки зрения разделения образов классов в многомерном признаковом пространстве осуществляется на основе компаративного анализа данных, первоначально размещаемых в классифицированной обучающей выборке. Такая выборка образуется путем объединения всех векторов, формально описывающих образы экземпляров всех классов.

Пусть имеется алфавит классов $A=\{A_1,A_2,...,A_k\}$ и сформирован априорный словарь признаков $P=\{P_1,P_2,...,P_n\}$. Каждый объект описывается n признаками из априорного словаря признаков в виде вектор-столбец $x^T=(x_1,x_2,\ldots,x_n)$, где x_i – значение i-го признака, однозначно идентифицируется с одним из классов, а все выборки значений признаков каждого класса имеют непрерывные функции распределения. Множество объектов отдельного класса образует исходное описание этого класса в априорном признаковом пространстве, а объединение всех объектов из всех классов представлять собой классифицированную обучающую выборку. Эта выборка описывается в виде таблицы типа *“объект-свойство”* и формально представляется в виде матрицы $X_{n\,x\,m}$, где $m=m_1+m_2+...+m_k$, и m_i – количество объектов i-го класса.

При решении реальных задач на основе использования методов математической теории распознавания образов, исследователи сталкивается с необходимостью формального представления образов классов в соответствующем многомерном признаковом пространстве, построенном либо на основе априорного словаря, либо на основе уточненного рабочего словаря признаков. Вариант размещения образов классов представлен на следующем рисунке:

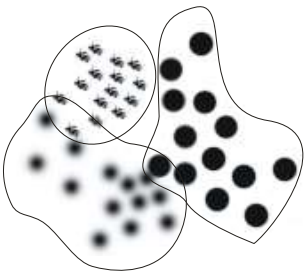

Рисунок 1 – Взаимное размещение образов трех классов в пространстве R^2

С целью формального представления образа класса в многомерном признаковом пространстве предлагается на основе всех экземпляров класса построить соответствующий кла-

стер. Каждый экземпляр класса представляет собой вектор в пространстве R^n с координатами вершины $(x_1,x_2,\ldots,x_n)$, где x_i – значение i-го признака. Объединение всех векторов одного класса в кластер и будет представлять собой формальное описание класса в соответствующем многомерном признаковом пространстве. Процесс построения кластера начинается с поиска наиболее удаленного от других экземпляра класса. Затем определяется ближайший к найденному экземпляру представитель класса, и он включается в состав кластера. Кроме того, вычисляются и запоминаются значения координат вспомогательного вектора, который указывает на середину отрезка, соединяющего два очередных экземпляра класса. В итоге формируется "скелет" кластера для построения образа класса, который содержит $2*m_i-1$ векторов, из которых m_i векторов-экземпляров i-го класса (где m_i – количество объектов i-го класса) и . m_i-1 вспомогательных векторов. В дальнейшем каждый экземпляр "скелета" выступают в качестве центра гиперсферы при построении кластера, представляющего собой объединение областей, образованных пересекающимися гиперсферами.

Описание алгоритма реализации метода. Пусть имеется словарь признаков $P=\{P_1,P_2,...,P_n\}$, и пусть каждый экземпляр класса описывается на основе n признаков из этого словаря, т.е. каждый экземпляр формально представляется в виде вектора-столбца $x^T=(x_1,x_2,\ldots,x_n)$, где x_i – значение i-го признака. Объединение всех соответствующих векторов-столбцов будет представлять собой матрицу размерности n x k, (где k – количество экземпляров класса).

Возьмем наиболее удаленный от всех экземпляр класса $X^{(1)}$ и найдем для него ближайший экземпляр $X^{(2)}$, и расстояние между ними обозначим $l^{(1)}$. Построим гиперсферы (далее сферы) радиуса $r^{(1)}=\frac{l^{(1)}}{2}$ с центрами в $X^{(1)}$, $X^{(2)}$. Обозначим точку касания двух сфер – $O^{(1)}$ с координатами $(o_1^{(1)},...,o_n^{(1)})$ и построим сферу радиуса $r^{(1)}$ с центром в $O^{(1)}$. Для экземпляра $X^{(2)}$ найдем ближайший экземпляр $X^{(3)}$, и расстояние между ними обозначим $l^{(2)}$, причем из поиска исключаем $X^{(1)}$. Построим сферы радиуса $r^{(2)}=\frac{l^{(2)}}{2}$ с центрами в $X^{(2)}$, $X^{(3)}$ и получим точку касания сфер $O^{(2)}=(o_1^{(2)},...,o_n^{(2)})$. Построим сферу радиуса $r^{(2)}$ с центром $O^{(2)}$. Поскольку $X^{(2)}$ является центром двух сфер радиуса $r^{(1)}$ и $r^{(2)}$, то для данного экземпляра выбираем сферу с максимальным радиусом.

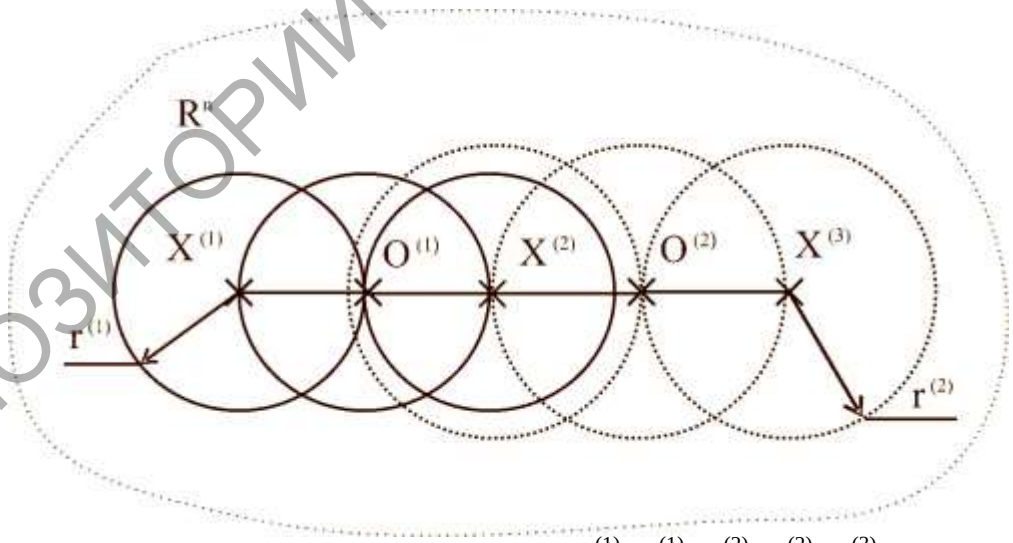

Рисунок 2 – Сферы с центрами $X^{(1)}$, $O^{(1)}$, $X^{(2)}$, $O^{(2)}$, $X^{(3)}$

Аналогично продолжим построение сфер для остальных экземпляров класса. Опишем результат такого построения с помощью таблицы, в которой первый столбец *"№"* содержит порядковые номера сфер, столбец *"Центр"* – координаты центра сферы с радиусом в столбце *"Радиус"*, столбец *"Признак"* принимает значение 1 – если центром сферы является x_i и 0 – если центром сферы оказывается точка пересечения o_i.

№	Центр	Радиус	Признак
1	$X^{(1)}$	$r^{(1)}$	1
2	$O^{(1)}$	$r^{(1)}$	0
3	$X^{(2)}$	$r^{(2)}$	1
….	….	….	…
2m-2	$O^{(m-1)}$	$r^{(m-1)}$	0
2m-1	$X^{(m)}$	$r^{(m-1)}$	1

Экземпляры $X^{(1)}$ и $X^{(m)}$ являются “крайними”, а потому для них в таблицу записываем радиусы $r^{(1)}$ и $r^{(m-1)}$. Для каждого экземпляра класса $X^{(2)},...,X^{(m-1)}$ в таблицу записываем максимальное значение радиуса, связанных с построением соответствующих сфер.

В результате получаем, что область, объединяющая все сферы, и будет представлять собой кластер класса.

При построении кластеров мы использовали сферы, для которых объем V радиуса r в пространстве R^n равен для n четных $V = \frac{2^{\frac{n}{2}}\pi^{\frac{n}{2}}}{n!!} r^n$ и нечетных $V = \frac{2^{\frac{n+1}{2}}\pi^{\frac{n-1}{2}}}{n!!} r^n$ соответственно [5].

Объем построенного кластера можно вычислить по формуле:

$$V = \sum_{j=1}^{2m_i-1} V^{(j)} - U, \qquad (1)$$

где U представляет собой объем пересечения сфер, образующих кластер. Для вычисления значения U воспользуемся методом Монте-Карло.

Для начала рассмотрим область пересечение G двух сфер C_1 с центром в точке M_1 и радиусом r_1, и C_2 с центром в точке M_2 и радиусом r_2 в пространстве R^n.

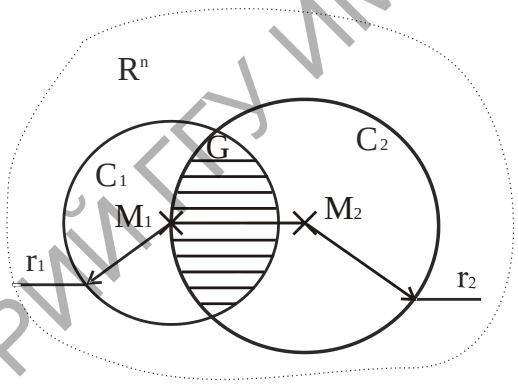

Рисунок 3 – Пересечение сфер C_1 и C_2.

Оценим объем V_G. Для этого возьмем сферу C_1, объем которой V_1. Выберем N случайных точек, равномерно распределенных в C_1, и обозначим через N' количество точек, попавших в G. Точка $M_G \in G$ тогда и только тогда, когда $|M_1 - M_G| \le r_1$ и $|M_2 - M_G| \le r_2$.

При большом значении N, очевидно $\frac{N'}{N} \approx \frac{V_G}{V_1}$, откуда $V_G \approx V_1 \frac{N'}{N}$ [6].

Тогда объем области объединения двух сфер C_1 и C_2 в пространстве R_n рассчитывается по формуле $V=V_1+V_2-V_G$ или $V \approx V_2 + V_1\left(1 - \frac{N'}{N}\right)$.

Теперь перейдем к вычислению объема U пересечения сфер кластера. Найдем для j-ой сферы, где $j=\overline{1,2m-1}$, все сферы с которыми она пересекается, причем к рассмотрению будем брать сферы, у которых порядковый номер больше j. После чего оценим объем по формуле $U^{(j)} \approx V^{(j)} \frac{N_j'}{N_j}$. Для последней сферы, порядковый номер которой равен 2m-1, имеем $U^{(2m-1)}=0$.

Тогда суммарный объем пересечения сфер кластера вычисляется по формуле:

$$U = \sum_{j=1}^{2m_i-1} U^{(j)} \quad (2)$$

Подставляя полученное значение U в формулу (1) получим искомое значение объема $V \approx \sum_{j=1}^{2m_i-1} V^{(j)} - \sum_{j=1}^{2m_i-1} U^{(j)}$.

Далее вычисляем плотность кластера ρ по формуле $\rho=V/k$, где k количество экземпляров класса.

Заключение. При построении систем распознавания первоначально формальное определение каждого отдельного экземпляра класса представляет собой вектор в многомерном признаковом пространстве. В свою очередь, первоначальное формальное описание класса можно представить в виде матрицы, получаемой путем объединения соответствующих векторов.

В статье описан метод, который позволяет представить образ класса в виде кластера, получаемого путем объединения гиперсфер в многомерном признаковом пространстве. Предложен оригинальный алгоритм построения кластера, который предусматривает возможность вычисления объема и плотности кластера, что позволяет производить оценку компактности кластера.

Разработанный алгоритм базируется на использовании аппарата кластерного анализа, а для вычисления объема пересечения гиперсфер в соответствующем многомерном признаковом пространстве используется метод Монте-Карло.

Резюме. При построении систем распознавания возникает задача формального представления либо образа класса, либо образа эталона класса в многомерном признаковом пространстве. В статье предлагается такой образ представлять в виде кластера и описывается оригинальный алгоритм построения соответствующего кластера. Предусматривается вычисление объема и плотности кластера для оценки его компактности.

Abstract. During the construction of recognition systems there occurs a problem of formal representation of a class image, or a class standard image in multidimensional sign space. In the article it is offered to represent this image in a kind of cluster. The original algorithm of corresponding cluster construction is also described. Volume and density cluster calculation for an estimation of its compactness is provided.

Литература

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

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

3. Марусенко, М.А. Атрибуция анонимных и псевдоанонимных литературных произведений методами распознавания образов / М.А.Марусенко; Л.: Издательство Ленинградского университета, 1990. – 168 с.

4. Гуща, Ю.В. Об использовании одного алгоритма кластерного анализа при построении системы диагностики острого аппендицита у детей / Ю.В.Гуща // Известия Гомельского государственного университета имени Ф.Скорины. Гомель, 2007. – №5(44). – С.21-26.

5. Розенфельд, Б.А. Многомерные пространства. / Б.А. Розенфельд; М.: Наука, 1966. – 647 с.

6. Соболь, И.М. Численные методы Монте-Карло / И.М.Соболь; М.: Наука, 1973. – 311 с.

Гродненский государственный университет им. Янки Купалы

Поступило 27.04.10