OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.02 : 004.9

ПРИМЕНЕНИЕ КЛАСТЕРНЫХ СТРУКТУР ДЛЯ ПРЕДСТАВЛЕНИЯ КЛАССОВ ПРИ ПОСТРОЕНИИ СИСТЕМ РАСПОЗНАВАНИЯ

Анищенко И.С., Жукевич А.И., Родченко В.Г.

Гродненский государственный университет имени Янки Купалы, г. Гродно, Республика Беларусь

ivan.ani@mail.ru

san@grsu.by

rovar@mail.ru

Реализация системы распознавания начинается с определения алфавита классов, формирования априорного словаря признаков и построения обучающей выборки. Внутри этой выборки каждый объект представляется в виде многомерного вектора в пространстве, сформированном на основе априорного словаря. Применение кластерных структур позволяет формализовать процедуру представления образов классов и производить оценку взаимного размещения образов и эталонов классов при построении систем распознавания.

Ключевые слова: интеллектуальная система, распознавание образов, алфавит классов, словарь признаков, классифицированная обучающая выборка, кластерная структура.

ВВЕДЕНИЕ

Процесс построения системы распознавания на основе анализа наблюдаемых данных предполагает выполнение трех основных этапов. Первый из них является *подготовительным* и связан с определением алфавита классов, формированием априорного словаря признаков и построением исходной классифицированной обучающей выборки. Второй этап является центральным и направлен на реализацию процедуры *обучения* [Васильев, 1989]. В результате выполнения из априорного словаря признаков (АСП) исключаются все малоинформативные признаки, которые не обеспечивают разделение формальных образов эталонов классов в признаковом пространстве принятия решений. Заключительный третий этап связан с выполнением процедуры *принятия решения*, когда исследуемый образ либо классифицируется к одному из предусмотренных классов, либо выделяется в отдельный джокер-класс [Родченко, 2006].

Формально весь процесс распознавания может быть реализован в результате выполнения следующей последовательности преобразований:

$$S \xrightarrow{F_1} C \xrightarrow{F_2} A \xrightarrow{F_3} T \xrightarrow{F_4} A^* \xrightarrow{F_5} E \xrightarrow{F_6} R$$

где S – алфавит классов; C – словарь наблюдаемых (измеряемых) характеристик; A – априорный словарь признаков; T – классифицированная обучающая выборка; A^* – уточненный словарь признаков для построения пространства решений; E – множество эталонов классов; R – множество решений; F_1 – алгоритм получения наблюдаемых характеристик; F_2 – алгоритм построения априорного словаря признаков; F_3 – алгоритм формирования классифицированной обучающей выборки; F_4 – алгоритм сепарирования признаков из априорного словаря по степени их информативности для построения пространства решений; F_5 – алгоритм построения образов эталонов классов в пространстве решений; F_6 – алгоритм процедуры принятия решения.

Для формального представления образов классов или эталонов классов в многомерном признаковом пространстве предлагается на основе использования всех экземпляров классов строить соответствующие кластерные структуры.

Построение кластерных структур

Предположим, что определены алфавит классов $S=\{S_1,S_2,...,S_k\}$ и априорный словарь признаков $A=\{A_1,A_2,...,A_n\}$.

Пусть каждый отдельный объект класса описывается n признаками из АСП в виде вектор-столбца $x^T=(x_1,x_2,\ldots,x_n)$, где x_i – значение i-го признака, и однозначно ассоциируется с одним из классов.

Набор объектов отдельного класса образует исходное описание этого класса в априорном признаковом пространстве. Объединение всех объектов из всех классов образует классифицированную обучающую выборку, которая описывается в виде таблицы типа *“объект-свойство”* и формально представляется в виде матрицы $X_{n \times m}$, где $m=m_1+m_2+...+m_k$, а m_i - количество объектов i-го класса.

Для формального представления образов классов предлагается на основе всех экземпляров класса строить соответствующую кластерную структуру. Поскольку каждый отдельный 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, то объединение всех векторов одного класса в кластерную структуру и будет представлять собой формальное описание класса.

Построение кластерных структур можно осуществить на основании универсального алгоритма, предусматривающего выполнение 4 шагов:

1. Начальная инициализация центров кластеров (например, можно задать начальные значения случайно выбранными точками из пространства, в котором определены данные; можно выбрать случайные точки из входного массива данных и т.д.).
2. E-шаг (expectation): происходит ассоциация между элементами данных и кластерами, которые представлены в виде своих центров (центроидами).
3. M-шаг (maximization): пересчитываются центры кластеров, как средние значения от данных, которые были включены с соответствующий кластер (другими словами, происходит модификация параметров модели таким образом, чтобы максимизировать вероятность попадания элемента в выбранный кластер). В случае, если кластер после шага 2 оказался пустым, то происходит инициализация каким-либо другим способом.
4. Шаги 2-3 повторяются до сходимости, либо пока не выполнится другой критерий остановки алгоритма (например, превышение некоторого числа итераций).

Однако использование данного подхода имеет и ряд недостатков, в частности:

- Алгоритм не гарантирует определения лучшего из возможных расположений центров гиперсфер (достижение глобального минимума суммарного квадратичного отклонения), но гарантирует сходимость к какому-либо решению, т.е. итерации не зациклятся.
- Результат построения кластерной структуры зависит от выбора исходных центров гиперсфер.

Об одной реализации общего подхода

Процесс построения кластерной структуры начинается с поиска наиболее удаленного от других вектора. Затем определяется ближайший к найденному вектору представитель класса, и он включается в состав кластера. Затем вычисляются и запоминаются значения координат вспомогательного вектора, который указывает на середину отрезка, соединяющего два очередных экземпляра класса. В результате формируется “скелет” кластерной структуры для построения формального 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” выступает в качестве центра гиперсферы при построении кластерной структуры, представляющей собой объединение областей, образованных пересекающимися гиперсферами.

1. Возьмем наиболее удаленный от всех экземпляр класса $X^{(1)}$ и найдем для него ближайший экземпляр $X^{(2)}$, и расстояние между ними обозначим $l^{(1)}$.

2. 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 $O^{(1)}$.

3. 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.

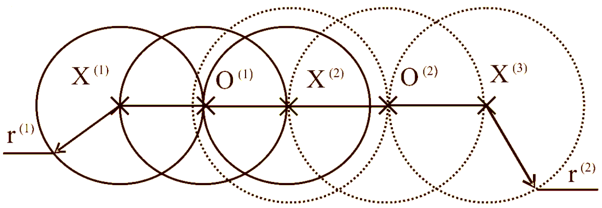

Рисунок 1 - Сферы с центрами $X^{(1)}$, $O^{(1)}$, $X^{(2)}$, $O^{(2)}$, $X^{(3)}$

Аналогично продолжим построение сфер для остальных экземпляров класса. Опишем результат такого построения с помощью таблицы, в которой первый столбец содержит порядковые номера сфер, второй столбец содержит координаты центра сферы с радиусом, значение которого в третьем столбце. Четвертый столбец 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 используются гиперсферы, 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 [Розенфельд, 1966].

Объем построенного кластера можно вычислить по формуле $V=\sum_{j=1}^{2m_i-1}V^{(j)}-U$, где U представляет собой объем пересечения сфер, образующих кластер. Для вычисления значения U можно воспользоваться методом Монте-Карло.

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.

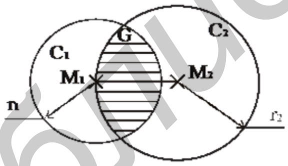

Рис.2. Пересечение сфер C_1 и C_2

Для вычисления объема V_G. 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 $|M_2-M_G|\le r_2$.

Очевидно, что при большом значении N, $\frac{N'}{N}\approx\frac{V_G}{V_1}$, а значит $V_G\approx V_1\frac{N'}{N}$. Тогда объем области объединения двух сфер C_1 и C_2 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 кластерной структуры. 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 $U=\sum_{j=1}^{2m_i-1}U^{(j)}$.

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

Плотность кластера ρ можно вычислить по формуле $\rho=V/k$, где k количество экземпляров класса.

Заключение

При построении систем распознавания первоначально каждый отдельный экземпляр класса формально может быть представлен в виде вектора в многомерном признаковом пространстве. Формальное же описание образа класса предложено реализовать в виде кластерных структур, получаемых путем объединения соответствующих гиперсфер.

В статье описан метод, который позволяет представить образ класса в виде кластерной структуры, получаемой путем объединения гиперсфер в многомерном признаковом пространстве.

Реализован оригинальный алгоритм построения кластерных структур, которым предусматривается возможность вычисления объема и плотности кластерной структуры, что позволяет производить оценку ее компактности.

Предложенный подход базируется на применении аппарата кластерного анализа, а для вычисления объема пересечения гиперсфер в многомерном признаковом пространстве используется метод Монте-Карло.

Библиографический список

[Васильев, 1989] Васильев, В.И. Проблема обучения распознаванию образов / В.И. Васильев – К: Выща шк. Головное изд-во, 1989.

[Родченко, 2006] Родченко, В.Г. Об одном методе реализации процедуры обучения при построении системы распознавания образов / В.Г. Родченко. // Известия Гомельского государственного университета имени Ф.Скорины. - 2006. - №4. - С.73-76

[Розенфельд, 1966] Розенфельд, Б.А. Многомерные пространства / Б.А. Розенфельд – М: Наука, 1966.

THE USE OF CLUSTER STRUCTURES TO REPRESENT OF CLASSES FOR CONSTRUCTION OF RECOGNITION SYSTEMS

Anishchenko I.S., Zhukevich A.I.,
Rodchenko V.G.

Yanka Kupala Grodno State University, Grodno, Republic of Belarus

ivan.ani@mail.ru

san@grsu.by

rovar@mail.ru

The construction of the recognition system begins with the definition of the alphabet classes, the formation of a priori dictionary features and builds a learning sample. Within this sample of each object is represented in the form of multidimensional vector in the space, formed on the basis of a priori dictionary. The use of cluster structures allows to formalize the procedure of submission of patterns classes and to assess the mutual placement of patterns and etalons classes when building of the recognition systems.

INTRODUCTION

At building of real systems on the basis of use of pattern recognition theory apparatus there is a necessity of formal representation of classes in corresponding multidimensional feature space. Such space is built on the basis of a priory feature vocabulary, or on the basis of the specified feature vocabulary.

For formal representation of pattern class in multidimensional feature space building of corresponding cluster structure on the basis of all class standards is suggested. Each class standard represents a vector in space R^n with top coordinates (x_1, x_2, …, x_n), where x_i – i feature value. Association of all vectors of one class in cluster structure also will represent the formal description of a class in multidimensional feature space.

MAIN PART

Cluster structure building process begins with search of such example class which is the most removed from others. Then the nearest to a found example representative of a class is defined, and it joins the cluster structure as a following element. Besides, are calculated and if necessary are remembered coordinates values of an auxiliary vector which indicates the midpoint, connecting two next class examples. As a result cluster "skeleton" for building of pattern class is formed.

Let to a priori feature vocabulary be formed $P=\{P_1,P_2,...,P_n\}$, and every class example is described on the basis of n features from this vocabulary. Formally each such example is represented in the form of vector-column $X^T = (x_1, x_2,…,x_n)$, where x_i – i feature value.

We will begin cluster structure building with search of the most remote from all class examples. We will designate it $X^{(1)}$. The given example will be the first element of formed cluster structure.

Then we will find for $X^{(1)}$ the nearest example $X^{(2)}$, the distance between them will be designated as $l^{(1)}$. We will build hyperspheres (further spheres) of radius $r^{(1)}$, where $r^{(1)} = l^{(1)}/2$ with centers in $X^{(1)}$, $X^{(2)}$. We will designate the contact point of two spheres – $O^{(1)}$ with coordinates $(o_1^{(1)},...,o_n^{(1)})$ and we will build radius sphere $r^{(1)}$ with center in $O^{(1)}$. Further for the example $X^{(2)}$ we will find the nearest example $X^{(3)}$ and distance between them will be designated as $l^{(2)}$, and is excluded $X^{(1)}$. We will build radius spheres $r^{(2)}$, where $r^{(2)} = l^{(2)}/2$ with centers in $X^{(2)}$, $X^{(3)}$ and we will get the sphere contact point $O^{(2)}=(o_1^{(2)},...,o_n^{(2)})$. We will build radius sphere $r^{(2)}$ with center $O^{(2)}$. The example $X^{(2)}$ is a center of two radius spheres $r^{(1)}$ and $r^{(2)}$, the maximum radius is chosen for the cluster structure description.

As a result we received that the area uniting all spheres will represent cluster structure which is subject of a formal pattern class in multidimensional feature space.

For cluster structure characteristics estimation the possibility of cluster structure volume and density calculation is provided.

CONCLUSION

In the article the method which allows presenting a pattern class in cluster structure, received by association of hyperspheres in multidimensional feature space, is described.

This method can be used at designing and building of recognition systems for procedure performance of the analysis of mutual placing of class patterns and standards in multidimensional feature space.