УДК 517

КЛАССИФИКАЦИЯ НА ОСНОВЕ КОМПОНЕНТНЫХ СТРУКТУР ДАННЫХ В ПРИЗНАКОВОМ ПРОСТРАНСТВЕ

Ильченко А. В.

Таврический национальный университет им. В. И. Вернадского,
факультет математики и информатики
пр-т Вернадского, 4, Симферополь, Крым, Украина, 95007

Abstract

A classifiers synthesis grounding, detection objectivism of clustering structure on basis of rejection sets in attribute space are considered.

Введение

Задачи классификации лежат в основе многих интеллектуализированных информационных систем и востребованы разработчиками программного обеспечения. Несмотря на обилие алгоритмов классификации, открытыми и актуальными остаются вопросы обоснования построения классификации, ее объективности.

Целью данной работы является исследование возможности тщательного анализа особенностей имеющихся эмпирических данных, которые могут быть использованы для построения классификаторов. При этом внимание уделяется оценке неслучайности появления таких особенностей в выборках. Для этой цели разработан новый подход к исследованию на основе введенного понятия компонентной структуры данных.

Пусть G — генеральная совокупность объектов, на которой определена вероятностная мера $P(G)$. G представимо как объединение конечного числа подмножеств:

$$G = \bigcup_{i=1}^{K} G_i \tag{1}$$

Эти подмножества называются классами. Семейство классов $\{G_i\}$ называется классовой структурой генеральной совокупности G. Рассматривается случай, когда классы не пересекаются, т.е. семейство $\{G_i\}$ образует разбиение множества G.

Для описания объектов множества G и указания о принадлежности объектов классам фиксируется некоторое множество признаков (переменных) $X = \{X_1, X_2, \ldots, X_n\}$ и целевая переменная Y.

$D_j = Dom\ X_j$ — множество возможных значений переменной $X_j,\ j = 1, 2, \ldots, n$.

Всюду далее считается, что для каждой переменной X_j множество ее значений D_j является вполне упорядоченным множеством.

Декартово произведение $D_X = \prod_{j=1}^{n} D_j$ задает n-мерное признаковое пространство.

$D_Y = Dom\, Y = 1, 2, \ldots, K$ — множество значений переменной Y. Переменная Y — номинальная. Ее значения обозначают номер класса.

Обозначим $D = D_X \times D_Y$ — декартово произведение признакового пространства D_X и множества D_Y.

Каждому объекту a из множества G соответствует точка множества D (проекция a на D). Эта точка определяется набором значений $X(a) = (X_1(a), X_2(a), \ldots, X_n(a)) = (x_1, x_2, \ldots, x_n)$ и $Y(a) = y$, где

$x_j = X_j(a)$ — значение переменной X_j для объекта a;

$y = Y(a)$ — значение переменной Y для объекта a (номер класса для a).

Такая проекция множества G на множество D, с учетом существования вероятностной меры $P(G)$, порождает в множестве D вероятностную меру $P(D)$ [3].

Всюду далее будем считать, что на множестве D определено совместное распределение вероятностей $P(y, x)$, где $x = (x_1, x_2, \ldots, x_n)$. Это распределение может быть записано в виде

$$P(x, y) = P(y)P(x/y), \text{ где}$$

$P(y)$ — априорная вероятность появления объектов класса y,

$P(x/y)$ — условное распределение величины x по классу y в пространстве D_X.

Отметим [3], что если $X_1, X_2, \ldots, X_n$ — дискретные переменные, то $P(y/x)$ — условное распределение вероятности в пространстве D_X. Если $X_1, X_2, \ldots, X_n$ — непрерывные переменные, то $P(y/x)$ — условное распределение плотности вероятности в пространстве D_X.

Всякое подмножество признакового пространства характеризуется величиной

$$\mu(H) = \frac{|H|}{|D_X|}, \tag{2}$$

где μ — относительная мера подмножества H.

Для дискретных переменных $|H|,\ |D_X|$ — мощность этих подмножеств.

Для непрерывных переменных $|H|,\ |D_X|$ — мера Лебега этих подмножеств.

Максимальные связные подмножества всякого покрытия признакового пространства D_X называются компонентами этого покрытия.

Всякое покрытие признакового пространства, составляющим подмножествам которого приписаны номера классов, называется классовой структурой признакового пространства.

Всякое покрытие признакового пространства, составляющим компонентам которого приписаны номера классов, называется компонентной структурой признакового пространства.

Каждому классу G_i соответствует подмножество точек признакового пространства D_X. Обозначим это подмножество H. Подмножество H не обязательно односвязное. Именно условная плотность (вероятность) $P(x/y)$ порождает в признаковом

пространстве D_X подмножества H_i:

$$H_i = \{x \in D_X | P(x/y = t) > 0\}\,,\ i = 1, 2, \ldots, K \tag{3}$$

Отметим, что подмножества H_i определяются однозначно.

Определение 1. *Компонентой класса* называется максимальное односвязное подмножество признакового пространства, состоящее из точек, соответствующих объектам этого класса.

Любое подмножество H_i может быть единственным способом представлено в виде объединения компонент класса G_i:

$$H_i = \bigcup_{j=1}^{K_i} H_{ij}, \tag{4}$$

где H_{ij} — компоненты класса G_i, K_i — число компонент класса G_i.

Таким образом, каждый класс объектов G_i порождает в признаковом пространстве D соответствующее этому классу семейство компонент $\{H_{ij}\}$. Это семейство компонент называется компонентной структурой класса G_i (компонентной структурой подмножества H_i), порождаемой условной плотностью (вероятностью) $P(x/y = i)$.

Обозначим H_0 те точки признакового пространства, которые не соответствуют никаким объектам из множества G.

Определение 2. *Компонентой подмножества* H_0 называется максимальное односвязное подмножество признакового пространства, не содержащее точек, соответствующих каким-либо точкам множества G.

H_0 также единственным способом можно представить в виде объединения его компонент

$$H_0 = \bigcup_{j=1}^{K_0} H_{0j}, \tag{5}$$

Это семейство компонент называется компонентной структурой подмножества H_0.

Иногда, компоненты H_{0j} удобно называть «пустыми» компонентами, а подмножество H_0 — областью запрета [2].

Если компоненты, порождаемые классами множества G и компоненты, представляющие подмножество H_0, известны (определены), то признаковое пространство D_X представимо в виде

$$D_X = \bigcup_{i=0}^{K} H_i = \bigcup_{i=0}^{K} \left(\bigcup_{i=1}^{K_i} \right), \tag{6}$$

т.е. D_X представимо как объединение компонент, порождаемых классами семейства $\{G_i\}$ и компонент, представляющих подмножество H_0.

Семейство компонент H_{ij} из (6) называется *компонентной структурой признакового пространства D_X, порождаемой условной плотностью (вероятностью) $P(x/y)$.*

Когда компонентная структура (6) не содержит пересекающихся компонент, соответствующих различным классам, говорят, что разделяющей способности признакового пространства достаточно для отделения компонент различных классов. В противном случае, когда существует хотя бы одна пара пересекающихся компонент из различных классов, разделяющей способности признакового пространства не достаточно для отделения компонент различных классов. Для этого случая, используя байесовскую стратегию распознавания [1, 3], можно получить более тонкую компонентную структуру — байесовские компоненты (B-компоненты).

Всюду далее рассматривается случай, когда компонентная структура (6) не содержит пересекающихся компонент.

1. Использование компонентной структуры

Компонентная структура признакового пространства (6) может быть использована для:

1. Описания областей признакового пространства, соответствующих классам генеральной совокупности G;
2. Формирования решающего правила о принадлежности классу предъявленного объекта из множества G.

Обозначения:

1. «Точка» — точка признакового пространства, соответствующая предъявленному объекту.
2. РП — решающее правило.
3. P_{error} — вероятность ошибочной классификации (доля неправильно классифицируемых объектов генеральной совокупности).

РП — «Точка» принадлежит некоторой компоненте. Объект относится к тому классу, компонентой которого является компонента, содержащая «точку».

Другая формулировка РП:

РП — Компонента определяет класс.

$$P_{error} = 0$$

2. Элементарные конструкции и их сложность

Определение 3. Подмножество называется *элементарным*, если оно представимо в виде объединения конечного числа интервалов признакового пространства D_X [4].

Определение 4. Компонента называется *элементарной*, если она является элементарным подмножеством.

Определение 5. *Компонентная структура класса* G_i (подмножество H_i) называется *элементарной*, если она состоит из элементарных компонент.

Определение 6. *Компонентная структура* $\{H_{ij}\}$ *признакового пространства* называется *элементарной компонентной структурой*, если все ее составляющие компоненты — элементарные.

Для элементарных компонент разными способами вводится понятие сложности.

Определение 7. *Сложностью элементарной компоненты* называется минимальное число попарно не пересекающихся интервалов, представляющих эту компоненту.

Определение 8. *Сложностью элементарной компоненты* называется минимальное число максимальных интервалов, представляющих эту компоненту.

На основании сложности компоненты определяются понятия сложности для элементарных подмножеств H_i и соответствующей элементарной компонентной структуры.

Определение 9. *Сложность элементарного подмножества* H_i — это сумма сложностей составляющих это подмножество элементарных компонент.

Определение 10. *Сложность элементарной компонентной структуры* — это сумма сложностей элементарных подмножеств H_i, составляющих эту компонентную структуру.

Понятие сложности порождает отношение порядка на подклассах элементарных компонент (соответственно на подклассах элементарных подмножеств H_i, на подклассах элементарных компонентных структур).

Обозначим C_k — подкласс всех элементарных компонент, сложности не больше k. Множество всех таких подклассов $\{C_k\}$, представляет собой неубывающую последовательность $C_1 \subset C_2 \subset \ldots \subset C_j \subset \ldots$. Отношение вложенности на подклассах C_k и есть отношение порядка на подклассах C_k элементарных компонент.

Аналогично определяются отношения порядка на подклассах элементарных подмножеств H_i на подклассах элементарных компонентных структур.

3. P-мера различия подмножеств признакового пространства

Подмножества признакового пространства D_X отличаются друг от друга по составу входящих в них элементов. Симметрическая разность двух подмножеств определяет подмножество, состоящее из элементов, которыми различаются между собой эти два подмножества. Если A и B — некоторые подмножества D_X, то вероятностная мера их симметрической разности $P(A \triangle B)$ может рассматриваться как числовое выражение различия подмножеств A и B по составу элементов (P-различие подмножеств A и B).

Если A — некоторая компонента компонентной структуры пространства D_X, B — элементарное подмножество из пространства D_X, то величина $P(A \triangle B)$ показывает точность P-аппроксимации компоненты A элементарным подмножеством B.

В силу полноты множества элементарных компонент [4] относительно P меры, любая компонента с любой степенью точности может быть аппроксимирована элементарной компонентой.

4. P-мера различия компонентных структур признакового пространства

Пусть H и $\tilde{H}$ — компонентные структуры признакового пространства D_X:

$$H = (H_0, H_1, \ldots, H_K) = (H_{01}, \ldots, H_{0K_0}, \ldots, H_{K1}, \ldots, H_{KK_K})$$

$$\tilde{H} = \left(\tilde{H}_0, \tilde{H}_1, \ldots, \tilde{H}_K\right) = \left(\tilde{H}_{01}, \ldots, \tilde{H}_{0K_0}, \ldots, \tilde{H}_{K1}, \ldots, \tilde{H}_{KK_K}\right)$$

Симметрическая разность H и $\tilde{H}$:

$$H \triangle \tilde{H} = \left(H_0 \triangle \tilde{H}_0, H_1 \triangle \tilde{H}_1, \ldots, H_K \triangle \tilde{H}_K\right) = \left(H_{01} \triangle \tilde{H}_{01}, \ldots, H_{KK_K} \triangle \tilde{H}_{KK_K}\right)$$

P-мера подмножества $H \triangle \tilde{H}$ отражает P-различие H и $\tilde{H}$ по составу элементов:

$$P(H \triangle \tilde{H}) = \frac{1}{2}\sum_{i=0}^{K} P(H_i \triangle \tilde{H}_i) = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K_i} P(H_{ij} \triangle \tilde{H}_{ij}) \tag{7}$$

С учетом полноты множества элементарных компонент относительно P-меры и учетом того, что число компонент компонентной структуры конечно, можно сделать вывод о полноте элементарных компонентных структур относительно P-меры, определяемой выражением (7).

5. Задача аппроксимации компонентной структуры элементарной компонентной структурой

Пусть H — фиксированная компонентная структура пространства D_X.

E^l — семейство всех элементарных компонентных структур признакового пространства D_X сложности не выше l.

Требуется среди элементов семейства E^l найти компонентную структуру, которая минимально отличается от H в смысле P-меры, и для которой максимальна μ-мера области запрета E_0:

$$E^* = \arg\inf_{E \in E^l} P(H \triangle E) \tag{8}$$

$$\mu(E_0^*) \to \max \tag{9}$$

Если решение задачи (8)-(9) существует и найдено, то элементарная компонентная структура E^* используется в качестве приближенного описания компонентной структуры H и приближенного решающего правила, для которого $P_{error} = P(H \triangle E^*)$.

Далее рассматривается случай, когда компоненты различных классов образуют разбиение признакового пространства и не имеют общих участков границ («компактные» компоненты).

Отметим, что для таких разбиений, повышая сложность семейства E^l, можно получить аппроксимацию, для которой $P_{error} = 0$.

6. Компонентная структура признакового пространства в задаче обучения распознаванию образов

В задаче обучения распознаванию образов совместное распределение вероятностей $P(y, x)$ неизвестно. Задана обучающая выборка в виде таблицы данных $T_{m,n+1}$, состоящей из m строк (объем выборки) и $n + 1$ столбцов с номерами $0, 1, \ldots, n$. Нулевой столбец содержит номер класса для каждого объекта. Остальные столбцы — значения соответствующих признаков на каждом объекте обучающей выборки.

Если, для каждой строки таблицы, кроме номера класса, указан и номер компоненты, то, каждую компоненту можно рассматривать как отдельный класс. Для этого случая, используя алгоритмы теории распознавания (например, алгоритмы [3, 5]), можно получить аппроксимацию компонентной структуры признакового пространства элементарной компонентной структурой.

Если же, в таблице данных номера компонент не указаны, то для аппроксимации компонентной структуры необходимо:

1. либо использовать алгоритмы, которые автоматически восстанавливают компонентную структуру;
2. либо использовать алгоритмы, которые определяют, существует ли разбиение классов на компоненты и, если существует, то распределяют элементы обучающей выборки по компонентам классов.

Для рассматриваемого случая разбиения признакового пространства («компактные» компоненты) вопрос о существовании компонент классов может быть, в какой-то мере, исследован с использованием области запрета [2].

В существующей точной компонентной структуре признакового пространства компоненты классов отделены друг от друга «пустыми» компонентами подмножества H_0. Если подмножество H_0 построено, то дополнение H_0 до D_X и порождает компонентную структуру признакового пространства.

Это же отделение компонент сохраняется и на элементах обучающей выборки (на элементах выборки отделение может возрасти). Если, используя обучающую выборку, удается построить достаточно «хорошую» аппроксимацию E_0 подмножества H_0, то дополнение E_0 до D_X порождает искомую аппроксимацию компонентной структуры.

Напомним, что компоненты подмножества E_0 представляют собой объединение «пустых» интервалов. Именно эти интервалы необходимо построить.

Способ оценки закономерности как не случайности, предложенный А.Д. Закревским [2], позволяет, используя обучающую выборку, оценить размеры интервалов, которые можно считать «пустыми». Такая оценка выражается через вероятность того, что интервал оказался пустым случайно. Если эта вероятность мала, тогда велика вероятность того, что такой интервал пуст не случайно (закономерно).

Оценка зависит от характеристик рассматриваемого признакового пространства и от объема обучающей выборки.

Когда признаковое пространство $D_X = B^n$ является булевым пространством, размерности n, оценка вероятности $P(m, n, r)$ того, что существует интервал ранга r, не пересекающийся с обучающейся выборкой объема m, приведена в [2]:

$$P(m, n, r) \leqslant C_n^r 2^r (1 - 2^{-r})^m \tag{10}$$

Когда признаковое пространство D_X дискретное размерности n и числом значений по каждому признаку $N_1, N_2, \ldots, N_n$, оценка вероятности $P(m, n, r_1, \ldots, r_n)$ того, что существует интервал размера $r_1 \times r_2 \times \ldots \times r_n$, не пересекающийся с обучающейся выборкой объема m,

$$P(m, n, r_1, \ldots, r_n) \leqslant W(m, n, r_1, \ldots, r_n) = \\ = \left(\prod_{i=1}^{n} (N_i - r_i + 1) \right) \left(1 - \frac{r_1 \times r_2 \times \ldots \times r_n}{N_1 \times N_2 \times \ldots \times N_n} \right)^m \tag{11}$$

В конце статьи приведена таблица для случая двух дискретных количественных признаков, в которой приводится оценка вероятности появления пустого интервала $\frac{r_1 \times r_2}{N_1 \times N_2}$, когда $m = 200$, $N_1 = N_2 = 10$.

Приведенная таблица показывает, что в данном случае, начиная с интервалов относительной μ-меры равной $0,03$, может быть введен порог, начиная с которого интервалы можно считать почти достоверно пустыми.

7. Компонентная структура признакового пространства в задаче кластеризации

В задаче кластеризации компоненты, аппроксимирующие отдельные кластеры, представляют собой дополнение области запрета до признакового пространства. Если область запрета определена и ее дополнение распадается на отдельные компоненты, то каждую такую компоненту и предлагается рассматривать как область, соответствующую отдельному кластеру. Иначе, в рамках принятой модели, выборка рассматривается как однородная.

Понятие «пустой» интервал — относительно. Оно зависит от μ-меры признакового пространства. Это иногда позволяет определять кластерную структуру иерархически.

r_1	r_2	$W(200, 2, r_1, r_2)$	Относительная μ-мера интервала
1	1	13,39796749	0,01
2	1	1,582915195	0,02
1	2	1,582915195	0,02
3	1	0,180899281	0,03
1	3	0,180899281	0,03
4	1	0,019922537	0,04
2	2	0,023053222	0,04
1	4	0,019922537	0,04
5	1	0,00210316	0,05
6	1	0,000211126	0,06
3	2	0,000304021	0,06
2	3	0,000304021	0,06
1	6	0,000211126	0,06
7	1	1,98907E-05	0,07
1	7	1,98907E-05	0,07
8	1	1,71667E-06	0,08
4	2	3,60501E-06	0,08
2	4	3,60501E-06	0,08
1	8	1,71667E-06	0,08
9	1	1,2862E-07	0,09
3	3	4,11584E-07	0,09
1	9	1,2862E-07	0,09
10	1	7,05508E-09	0,1

Список литературы

1. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. — 607с.
2. *Закревский А.Д.* Логика распознавания. – Мн.: Наука и техника, 1988. — 118с.
3. *Лбов Г.С., Старцева Н.Г.* Логические решающие функции и вопросы статистической устойчивости решений. – Новосибирск: Изд-во Ин-та математики, 1999. – 212с.
4. *Колмогоров А.Н., Фомин С.В.* Элементы теории функций и функционального анализа. Учебник для вузов. 6-е изд., испр., – М.: Наука. Гл. ред. физ.-мат. лит. 1989. – 624с.
5. *Agrawal R., Gehrke J., Gunopulos D.P.* Ragkavan Automatic Subspace Clustering of High Dimensional Data for Data Miining Applications. // In Proc. of the 1998 ACM-SIGMOD Conf. On the Management of Data. – 1998. – Seattle, Washington. – P.94-105.