



МАТЕМАТИКА И ИНФОРМАТИКА

УДК 519.854.33

Р.И. Кузьмич, А.И. Виноградова

ПРИМЕНЕНИЕ ПРОЦЕДУРЫ КЛАСТЕРИЗАЦИИ ДЛЯ ГЕНЕРИРОВАНИЯ ОБЪЕКТОВ С ЦЕЛЬЮ СОКРАЩЕНИЯ ЧИСЛА ПАТТЕРНОВ В МОДЕЛИ КЛАССИФИКАЦИИ

В статье рассматривается способ сокращения числа паттернов в модели классификации, основанный на применении процедуры кластеризации к объектам обучающей выборки.

Ключевые слова: алгоритм k -средних, центр оид, паттерн.

R.I. Kuzmich, A.I. Vinogradova

CLUSTERING PROCEDURE APPLICATION FOR OBJECT GENERATION WITH THE PURPOSE OF PATTERN NUMBER REDUCTION IN CLASSIFICATION MODEL

The method of the pattern number reduction in the classification model based on the clustering procedure application to training sample objects is considered in the article.

Key words: k -average algorithm, centroid, pattern.

Введение. Большое количество задач распознавания, привлекающих внимание исследователей как в медицине, так и во множестве других областей, может быть сформулировано следующим образом. Имеется выборка данных, которая состоит из двух непересекающихся множеств Ω^+ и Ω^- n -мерных векторов. Каждый вектор соответствует некоторому пациенту, векторы множества Ω^+ соответствуют пациентам, находящимся в некотором медицинском состоянии (например, болен или имеет осложнение заболевания), а векторы Ω^- не соответствуют этому состоянию. Компоненты векторов (признаки) представляют собой результаты определенных измерений, тестов или просто показывают присутствие или отсутствие определенных симптомов. Эти компоненты могут быть численными, номинальными или бинарными.

Задачи и методы исследований. Задача состоит в том, чтобы на основании имеющейся выборки данных (классифицированных ранее наблюдений) извлечь информацию о состоянии «нового» пациента, наблюдение которого не содержится в выборке [1].

Для решения этой задачи исследуется метод анализа данных, в основе которого лежит принцип вывода логических закономерностей или правил. Каждое правило должно покрывать достаточно много объектов одного класса и практически не покрывать объекты другого класса. Взяв вместе некоторое количество правил, можно получить алгоритм (модель, решающее правило), который будет решать поставленную задачу классификации.

В основе рассматриваемого подхода лежит понятие паттерна. Для исключения избыточных переменных в исходной выборке данных в множестве переменных определяется некоторое подмножество S , используя которое можно отличать положительные наблюдения от отрицательных. Далее для работы метода используются проекции Ω_s^+ и Ω_s^- множеств Ω^+ и Ω^- на S . Положительным паттерном называется подкуб пространства булевых переменных B_2^t , который пересекается с множеством Ω_s^+ и не имеет общих элементов с множеством Ω_s^- . Отрицательный паттерн задается аналогично.

Положительный ω -паттерн для $\omega \in \{0,1\}^t$ – это паттерн, содержащий в себе точку ω . Для каждой точки $\omega \in \Omega_s^+$ найдем максимальный ω -паттерн, то есть покрывающий наибольшее число точек Ω_s^+ .

Соответствующий подкуб зададим с помощью переменных y_j :

$$y_j = \begin{cases} 1, & \text{если } x_i \text{ зафиксирована в подкубе,} \\ 0, & \text{в противном случае,} \end{cases}$$

где x_i – i -й признак, $i=1, \dots, t$.

То есть путем фиксирования l переменных исходного куба размерностью t получаем подкуб размерностью $(t-l)$ и с числом точек 2^{t-l} .

Условие, говорящее о том, что положительный паттерн не должен содержать ни одной точки Ω_s^- , требует, чтобы для каждого наблюдения $\rho \in \Omega_s^-$ переменная y_j принимала значение 1 по меньшей мере для одного j , для которых $\rho_j \neq \omega_j$:

$$\sum_{\substack{j=1 \\ \rho_j \neq \omega_j}}^t y_j \geq 1 \text{ для любого } \rho \in \Omega_s^-.$$

Усиление ограничения для повышения устойчивости к ошибкам производится путем замены числа 1 в правой части неравенства на целое положительное число d .

С другой стороны, позитивное наблюдение $\sigma \in \Omega_s^+$ будет тогда входить в рассматриваемый подкуб, когда переменная y_j принимает значение 0 для всех индексов j , для которых $\sigma_j \neq \omega_j$. Таким образом, число положительных наблюдений, покрываемых ω -паттерном, может быть вычислено как

$$\sum_{\sigma \in \Omega_s^+} \prod_{\substack{j=1 \\ \sigma_j \neq \omega_j}}^t (1 - y_j).$$

Таким образом, имеем задачу условной псевдодвулевой оптимизации с алгоритмически заданными функциями:

$$\sum_{\sigma \in \Omega_s^+} \prod_{\substack{j=1 \\ \sigma_j \neq \omega_j}}^t (1 - y_j) \rightarrow \max, \quad (1)$$

$$\sum_{\substack{j=1 \\ \rho_j \neq \omega_j}}^t y_j \geq d \text{ для любого } \rho \in \Omega_s^-, y \in \{0,1\}^t. \quad (2)$$

Аналогично формулируется задача нахождения максимальных отрицательных паттернов.

Для повышения устойчивости метода к выбросам следует ослабить ограничение (2) – сделать возможным, чтобы паттерн захватывал некоторое малое число объектов другого класса. Тогда степень вычисляемых паттернов уменьшится, а покрытие увеличится.

Ограничение оптимизационной модели будет выглядеть следующим образом.

$$\sum_{\rho \in \Omega_s^-} z_\rho \leq D, \text{ где } z_\rho = \begin{cases} 0, & \text{если } \sum_{\substack{j=1 \\ \rho_j \neq \omega_j}}^t y_j \geq d, \\ 1, & \text{в противном случае,} \end{cases}$$

где D – число объектов другого класса, которые допускаются быть покрытыми паттерном (целое неотрицательное число) [1].

В итоге получаем семейство максимальных паттернов, число которых ограничено мощностью выборки данных $|\Omega^+ \cup \Omega^-|$. Обозначим $M_1^+, \dots, M_p^+$ и $M_1^-, \dots, M_q^-$ соответственно множества положительных и отрицательных паттернов.

Чтобы классифицировать новое наблюдение, воспользуемся следующим решающим правилом:

- 1) если наблюдение удовлетворяет условиям одного или нескольких положительных паттернов и не удовлетворяет условиям ни одного из отрицательных, то оно классифицируется как положительное;
- 2) если наблюдение удовлетворяет условиям одного или нескольких отрицательных паттернов и не удовлетворяет условиям ни одного из положительных, то оно классифицируется как отрицательное;
- 3) если наблюдение удовлетворяет условиям p' из p положительных паттернов и q' из q отрицательных, то «знак» наблюдения определяется как $p'/p - q'/q$;
- 4) в случае если наблюдение не удовлетворяет условиям ни одного паттерна, положительного или отрицательного, то оно остается неклассифицированным.

Паттерн строится относительно каждого объекта (n -мерной точки) в обучающей выборке. Как известно, на практике объем обучающей выборки данных может быть значителен, что приводит к увеличению временных ресурсов при построении модели, а также снижению наглядности построенной модели. Поэтому встает вопрос о возможности сокращения числа паттернов в полученной модели при условии сохранения способности модели классифицировать те же наблюдения, которые можно классифицировать с помощью полной системы паттернов.

Предлагается такой способ сокращения числа паттернов в модели классификации, основанный на применении процедуры кластеризации к объектам классификации в обучающей выборке. Для этого необходимо объекты каждого класса разбить на небольшое количество кластеров, используя алгоритм k -средних [2]. Алгоритм k -средних относит каждое обучающее наблюдение к одному из k кластеров (k задается исследователем) таким образом, чтобы каждый кластер был представлен центроидом соответствующих наблюдений, а каждое наблюдение отстояло бы от центроида своего кластера меньше, чем от центроидов всех других кластеров. Цель здесь состоит в том, чтобы найти набор центроидов, наилучшим образом представляющий распределение обучающих наблюдений. В [2] приведены шаги, из которых состоит алгоритм k -средних.

В результате реализации алгоритма k -средних получаем набор центроидов для каждого класса. Данные центроиды являются объектами, на базе которых строятся паттерны для новой модели классификации.

Таким образом, реализуя описанную выше эвристическую процедуру, получаем новую модель классификации, состоящую из меньшего числа паттернов. Ясно, что точность классификации, получаемая в результате использования новой модели, зависит от количества центроидов, полученных в результате реализации алгоритма k -средних, поэтому необходимо провести несколько экспериментов с разным количеством кластеров, чтобы определить зависимость точности классификации от количества кластеров, на которые делятся исходные классы. Апробацию данного способа проведем на задаче прогнозирования осложнений инфаркта миокарда: фибрилляция желудочков, фибрилляция предсердий [3].

Результаты исследований и их обсуждение. Для задачи фибрилляция желудочков построим по 10 центроидов для каждого класса, используя метод k -средних в программе WEKA [4]. Добавим вновь построенные объекты (центроиды) в выборку и построим на их базе паттерны. В данной задаче для тестирования использовалось 15 % выборки, состоящей из 80 положительных и 80 отрицательных объектов; бинарных признаков 200 из 112 исходных.

Для задачи фибрилляция предсердий построим по 15 центроидов для каждого класса, используя метод k -средних в программе WEKA. Добавим вновь построенные объекты (центроиды) в выборку и построим на их базе паттерны. В данной задаче для тестирования использовалось 10 % выборки, состоящей из 184 положительных и 184 отрицательных объектов. Бинарных признаков 215 из 112 исходных. В процессе проведения многочисленных экспериментов подобраны веса для классов и количество объектов другого класса, которое может захватывать паттерн, в решаемых задачах. Результаты классификации приведены в табл. 1–4.

Таблица 1

Результаты классификации для задачи прогнозирования осложнения фибрилляции желудочков с использованием вновь сгенерированных объектов для построения правил

Номер опыта	Множество паттернов	Покрытие отрицательных объектов	Покрытие положительных объектов	Степень паттерна	Веса для классов	Точность классификации для новой модели, %	Средняя точность классификации для новой модели, %
1	Отрицательные	28	15	5	0,5	90	88
	Положительные	15	40	2	0,5	93	84,6
2	Отрицательные	30	15	6	0,5	90	-
	Положительные	15	42	3	0,5	86	-
3	Отрицательные	29	15	5	0,5	90	-
	Положительные	15	35	2	0,5	79	-
4	Отрицательные	31	15	5	0,5	80	-
	Положительные	15	37	2	0,5	86	-
5	Отрицательные	30	15	6	0,5	90	-
	Положительные	15	39	3	0,5	79	-

Таблица 2

Результаты классификации для задачи прогнозирования осложнения фибрилляции желудочков с использованием всех объектов обучающей выборки для построения правил

Номер опыта	Множество паттернов	Покрытие отрицательных объектов	Покрытие положительных объектов	Степень паттерна	Веса для классов	Средняя точность классификации для исходной модели, %	Точность классификации для исходной модели, %
1	Отрицательные	40	15	5	0,5	86	90
	Положительные	15	33	2	0,5	75,8	79
2	Отрицательные	40	15	6	0,5	-	80
	Положительные	15	34	2	0,5	-	86
3	Отрицательные	40	15	5	0,5	-	90
	Положительные	15	33	2	0,5	-	64
4	Отрицательные	40	15	5	0,5	-	80
	Положительные	15	34	2	0,5	-	79
5	Отрицательные	41	10	5	0,5	-	90
	Положительные	15	33	2	0,5	-	71

Таблица 3

Результаты классификации для задачи прогнозирования осложнения фибрилляции предсердий с использованием вновь сгенерированных объектов для построения правил

Номер опыта	Множество паттернов	Покрытие отрицательных объектов	Покрытие положительных объектов	Степень паттерна	Веса для классов	Средняя точность классификации для новой модели, %	Точность классификации для новой модели, %
1	2	3	4	5	6	7	8
1	Отрицательные	77	30	8	0,66	56	69
	Положительные	30	78	6	0,34	89	73,6
2	Отрицательные	70	30	8	0,66	72	-
	Положительные	30	71	7	0,34	63	-
3	Отрицательные	71	30	8	0,66	83	-
	Положительные	30	75	5	0,34	68	-

Окончание табл. 3

1	2	3	4	5	6	7	8
4	Отрицательные	70	30	7	0,66	67	-
	Положительные	30	72	5	0,34	74	-
5	Отрицательные	72	30	7	0,66	67	-
	Положительные	30	79	4	0,34	74	-

Таблица 4

Результаты классификации для задачи прогнозирования осложнения фибрилляции предсердий с использованием всех объектов обучающей выборки для построения правил

Номер опыта	Множество паттернов	Покрывие отрицательных объектов	Покрывие положительных объектов	Степень паттерна	Веса для классов	Средняя точность классификации для исходной модели, %	Точность классификации для исходной модели, %
1	Отрицательные	70	30	8	0,66	64,4	61
	Положительные	30	79	4	0,34	79	74
2	Отрицательные	69	30	8	0,66	-	72
	Положительные	30	80	4	0,34	-	84
3	Отрицательные	69	30	8	0,66	-	67
	Положительные	30	78	5	0,34	-	79
4	Отрицательные	71	30	7	0,66	-	61
	Положительные	30	82	4	0,34	-	79
5	Отрицательные	71	30	8	0,66	-	61
	Положительные	30	81	4	0,34	-	79

Выводы

Предложена процедура генерирования объектов для построения паттернов с целью упрощения модели классификации. В результате использования способа сокращения паттернов в модели получили приемлемое изменение точности классификации и сокращение паттернов в модели классификации для задачи фибрилляции предсердий в 10 раз, для задачи фибрилляции желудочков – в 6 раз. Таким образом, процедура кластеризации для генерирования объектов является эффективной с точки зрения применимости для построения паттернов, образующих новую упрощенную модель классификации.

Литература

1. Модель логического анализа для решения задачи прогнозирования инфаркта миокарда / С.Е. Головенкин, Т.К. Гулакова, Р.И. Кузьмич [и др.] // Вестн. СибГАУ. – 2010. – Вып. 4. – С. 68–73.
2. Прикладная статистика: классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков [и др.]. – М.: Финансы и статистика, 1989.
3. Осложнения инфаркта миокарда: база данных для апробации систем распознавания и прогноза / С.Е. Головенкин, А.Н. Горбань, В.А. Шульман [и др.]. – Красноярск, 1997.
4. Weka the University of Waikato [Электронный ресурс]. – Режим доступа: <http://www.cs.waikato.ac.nz/~ml/weka/index.html>, свободный. – Загл. с экрана.