ФОРМИРОВАНИЕ ОПТИМАЛЬНОГО ПОДМНОЖЕСТВА ВХОДНЫХ ПРИЗНАКОВ НЕЧЕТКОГО КЛАССИФИКАТОРА МЕТОДОМ ЧИУ[1]

К.С. Сарин

Научный руководитель: профессор, д.т.н. И.А. Ходашинский

Томский государственный университет систем управления и радиоэлектроники,

Россия, г. Томск, пр. Ленина, 40, 634050

E-mail: sks@security.tomsk.ru

FORMATION OF THE OPTIMUM INPUT VARIABLE SET FOR A FUZZY CLASSIFIER BY THE METHOD OF CHIU

K.S. Sarin

Scientific Supervisor: Prof., Dr. I.A. Hodashinsky

Tomsk State University of Control Systems and Radioelectronics, Russia, Tomsk, Lenin str., 40, 634050

E-mail: sks@security.tomsk.ru

Abstract. *In the present study, a method for selecting important variables for constructing fuzzy classifiers is proposed. An example of selecting important variables and identification classifiers on real data is considered.*

Введение. Целью настоящей работы является выявления информативных переменных при построении нечетких классификаторов на основе анализа данных. За основу взят метод Чиу, предложенный в статье [1], в которой выбор важных переменных осуществлялось для построения нечетких аппроксиматоров.

Постановка задачи. Нечеткий классификатор состоит из базы нечетких правил следующего вида:

$$R_i\text{: IF } x_1 = A_{i,1} \text{ AND } \ldots \text{AND } x_n = A_{i,n} \text{ THEN } y = L_i,$$

где n – размерность пространства входных данных; $\{A_{i1},\ldots, A_{i,n}\}$ – множество нечетких термов, оценивающих советующую переменную; L_i – выход нечеткого правила, соответствующий элементу множества меток класса $\{1,\ldots, K\}$, K – число классов. Классификация нового входного вектора **x** проводится с помощью нахождения метки класса среди $\{1,\ldots, K\}$ с наибольшей суммой влияний правил, имеющих соответствующую классу выходную метку:

$$y = \arg \max_{k=1,\ldots,K} \sum_{\substack{i=1,\ldots,r;\\ L_i=k}} \prod_{j=1}^{n} \mu_{i,j}\left(x_j\right),$$

где П – t-норма операции конъюнкции, в данной работе используется произведение; $\mu_{i,j}$ – функция принадлежности для нечетких множеств i-го правила j-й переменной, которая определяет нечеткий терм $A_{i,j}$, в работе используются функции гауссового типа.

Основой для построения нечеткого классификатора с возможностью отбора информативных признаков является продукционное правило следующего вида:

$$R_i\text{: IF } s_1 \vee x_1 = A_{i,1} \text{ AND } \ldots \text{AND } s_n \vee x_n = A_{i,n} \text{ THEN } y = L_i,$$

где запись $s_i \vee x_i$ указывает на наличие ($s_i = 0$) или отсутствие ($s_i = 1$) признака в классификаторе.

[1]Исследование выполнено в рамках базовой части государственного задания министерства образования и науки Российской Федерации на 2017-2019 гг. Номер 8.9628.2017/БЧ.

Идентификация нечеткого классификатора. При построении классификатора с помощью анализа данных обычно выделяют три этапа: 1) генерация структуры; 2) оптимизация параметров; 3) оценка точности классификатора. Экспериментальные данные представляют собой таблицу наблюдений с входными и выходными значениями $\{(\mathbf{x}_p, c_p), p = 1,..., m\}$, причем обычно она разбивается на две части; одна часть, называемая обучающими данными, используется для построения классификатора, а другая, называемая тестовыми данными, для оценки точности его работы.

Построение нечеткого классификатора осуществляется таким образом:

$$E = \frac{\sum_{p=1}^{z} \begin{cases} 0, \text{если } c_p = f(\mathbf{x}_p) \\ 1, \text{иначе} \end{cases}}{z} \to \min,$$

где z – количество обучающих данных, $f(\mathbf{x}_p)$ – выход классификатора для входа $\mathbf{x}_p$.

Для генерации структуры классификаторов в данной работе использовался алгоритм на основе горной кластеризации [2]. Оптимизации параметров классификатора осуществлялась метаэвристическим алгоритмом «кукушкин поиск» [3].

Метод отбора информативных признаков. Процедура отбора информативных входных переменных представлена следующими шагами:

Шаг 1. Построить классификатор на всех входных переменных только алгоритмом генерации структуры и оценить его точность.

Шаг 2. Оценить точность модели, временно отключив поочередно каждую входную переменную.

Шаг 3. Удалить переменную, отключение которой показало лучший результат точности на шаге 2. Запомнить эту переменную и записать значение точности классификатора.

Шаг 4. Если есть не удаленные переменные, то перейти на шаг 2, иначе на шаг 5.

Шаг 5. Выбрать лучшие переменные из запомненных на шаге 3.

Эксперимент. Для эксперимента по выявлению информативных признаков был рассмотрен набор данных WINE с 13 признаками из репозитория KEEL [4].

Метод отбора информативных признаков выявил следующий по важности порядок переменных: 7, 13, 10, 4, 11, 2, 1, 12, 9, 8, 3, 5, 6. На рисунке 1 показана зависимость ошибки классификации от количества переменных на шаге 3 метода отбора, так же на графике указаны удаляемые переменные.

Рис. 1. Зависимость ошибки классификации от количества переменных на Шаге 3 метода отбора признаков

На отобранных по важности наборах переменных были сформированы нечеткие классификаторы, оценка ошибки классификации и количества правил которых проводилась методом пятикратной кросс-валидации. В таблице 1 указаны ошибки на обучающих (*Etra*) и тестовых (*Etst*) данных, а так же количество нечетких правил (*R*).

Таблица 1

Характеристики классификаторов на разных наборах входных переменных

Наборы переменных	*Etra*	*Etst*	*R*	*Кол-во признаков*
7	0.245	0.291	3.2	1
7, 13	0.065	0.152	6.8	2
7, 13, 10	0.013	0.057	7	3
7, 13, 10, 4	0.017	0.079	4.4	4
7, 13, 10, 4, 11	0.024	0.068	4.8	5
7, 13, 10, 4, 11,2	0.021	0.068	5.6	6
7, 13, 10, 4, 11, 2, 1	0.011	**0.045**	6	7
7, 13, 10, 4, 11, 2, 1, 12	0.015	0.051	6.6	8
7, 13, 10, 4, 11, 2, 1, 12, 9	0.021	0.079	5	9
7, 13, 10, 4, 11, 2, 1, 12, 9, 8	0.022	0.084	5.8	10
7, 13, 10, 4, 11, 2, 1, 12, 9, 8, 3	0.025	0.067	5	11
7, 13, 10, 4, 11, 2, 1, 12, 9, 8, 3, 5	0.037	0.079	5	12
Все	0.042	0.056	5.4	13

Из таблицы видно, что точность классификатора выше на наборе переменных 7, 13, 10, 4, 11, 2, 1. Данные переменные и будут составлять оптимальное множество входных признаков.

Заключение. В статье предложен метод Чиу для отбора признаков при построении нечетких классификаторов. Преимуществом данного метода являются то, что при нахождении важных переменных не требуется всякий раз перестраивать классификаторы, а достаточно использовать один, построенный на всех переменных. Эксперименты на реальных данных показали, что данный метод находит подмножество входных переменных, на которых построенный классификатор имеет высокую точность.

СПИСОК ЛИТЕРАТУРЫ

1. Chiu S.L. Selecting Input Variables for Fuzzy Models // Journal of Intellegent and Fuzzy Systems. – 1996. – Vol. 4, №. 4. – P. 243-256.
2. Chiu S.L. Fuzzy model identification based on cluster estimation // Journal of Intelligent and Fuzzy System. – 1994. – Vol. 2, № 3. – P. 267-278.
3. Ходашинский И.А., Минина Д.Ю., Сарин К.С. Идентификация параметров нечетких аппроксиматоров и классификаторов на основе алгоритма «кукушкин поиск» // Автометрия. – 2015. – Том 51, №3. – С.27-34.
4. KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on) [Электронный ресурс]. – Режим доступа: http://www.keel.es/. – 23.02.17.