УДК 004.8

КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ КЛАСТЕРНОГО ГЕНЕТИЧЕСКОГО АЛГОРИТМА И НЕЧЕТКОЙ ЛОГИКИ

© Новоселова Н.А., Том И.Э.

Объединенный институт проблем информатики НАН Беларуси
ул. Сурганова 6, г.Минск, Беларусь, 220012

e-mail: *tom@newman.bas-net.by*

Abstract. The paper describes the approach to the accuracy increasing of classification rules, obtained by genetic clustering algorithm. Proposed approach uses the theory of fuzzy sets, allowing to lower the uncertainty during classification process. The approach permits to take decisions, considering the whole set of rules, activated by the experimental observation.

Введение

Нейросетевые модели для классификации [1] имеют как положительный, так и негативный опыт применения. С одной стороны нейронные сети позволяют строить модели различной степени сложности, эффективны для обнаружения скрытых закономерностей в многомерных неоднородных данных и, соответственно, способны эффективно решать задачи классификации многомерных объектов. С другой стороны нейросетевые модели трудны для понимания непрофессионалами, так как представляют собой «черный ящик» с множеством параметров и потому плохо воспринимаются конечными пользователями. Это определило необходимость *постановки следующей проблемы*: для практического применения, особенно для медицинских задач, построить такие классификационные модели, которые можно легко интерпретировать.

Согласно проведенному *анализу последних достижений и публикаций*, в последнее время большое внимание уделяется разработке алгоритмов, позволяющих явным образом получить знания, скрытые в нейросетевой модели классификации, построенной на основе имеющихся в распоряжении данных. Параметры такой нейросетевой модели определяются на основе имеющихся данных с использованием специальных алгоритмов [2]. Если нейросеть обучена до желаемого уровня точности, то можно говорить, что знания, представленные этой нейросетью, закодированы в весах связей [3]. Веса связей определяют значения активирующих функций, поэтому нейроны скрытого слоя могут быть названы детекторами переменных, т.к. они группируют входные переменные, выявляя наиболее значимые их комбинации. Получение знаний из нейросетей заключается в применении алгоритмов, использующих или веса связей, или активирующие значения нейронов скрытого слоя.

При рассмотрении одного из таких алгоритмов [4], а именно кластерного генетического алгоритма для получения классифицирующих правил, основанного на активирующих значениях нейронов скрытого слоя, можно заключить, что *нерешенным остается вопрос сохранения достигнутого уровня точности классификации при переходе от нейросетевой модели к набору четких правил.*

Целью настоящей работы является разработка подхода к повышению точности классификации объектов данных с помощью правил, полученных с использованием кластерного генетического алгоритма и проведение сравнительного анализа полученных результатов.

1. Классификация с использованием кластерного генетического алгоритма

Предложенный в [4] метод классификации с помощью генетического кластерного алгоритма работает вполне эффективно, однако безошибочность классификации, к сожалению, не превышает 95-96%, что послужило мотивацией предложить некоторые усовершенствования метода, улучшающие его свойства.

Рассмотрим последовательность работы модифицируемого метода. Сначала на исходном наборе данных $\mathrm{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, где $\mathbf{x}_t = \{x_t^1, x_t^2, \ldots, x_t^n\}$ – n-мерный вектор входных переменных, характеризующий элемент данных, обучается нейросетевая модель Активирующие значения $a_i (i = 1, \ldots, q$ – количество нейронов скрытого слоя) скрытого слоя нейронной сети выступают в качестве входных переменных кластерного генетического алгоритма. На выходе генетического алгоритма формируется набор кластеров в которые группируются объекты (наблюдения), принадлежащие только одному классу $j, j = 1, \ldots, M$. Из полученных кластеров отбираются наиболее эффективные кластеры С согласно следующей схеме:

1. Для каждого l-го кластера вычисляется процент безошибочной классификации $p_l = \frac{n_l + m_l}{N}$, где $l = 1, \ldots, k_j$ – количество кластеров для j-го класса, n_l – количество элементов l-го кластера, принадлежащих соответствующему (кластеризуемому) классу, m_l – количество элементов, не принадлежащих l-у кластеру и не принадлежащих соответствующему классу, N – общее количество наблюдений.
2. Кластеры ранжируются по параметру p_l. Далее проверяются все возможные комбинации $\bigcup_{j=1}^{M} A_j$ кластеров для каждого класса такие, что $A_j \subset \{C_1^j, C_2^j, \ldots, C_{k_j}^j\}$, где j – номер класса, M – количество классов, k_j – количество кластеров для j-го класса. Отбирается комбинация кластеров, обеспечивающая наилучшую классификацию наблюдений и содержащая не менее одного кластера для каждого класса.
3. Для отобранных кластеров в соответствии с входными значениями каждого нейрона скрытого слоя формируются соответствующие классифицирующие правила в следующей форме: *Если* $((\nu_{min}^1 \leq a_1 \leq \nu_{max}^1)$ $u \ldots u$ $(\nu_{min}^q \leq a_q \leq \nu_{max}^q)$, *то класс* $= j$, где a_i – активирующее значение нейрона скрытого слоя нейросетевой модели (далее по тексту интегральная переменная), q – количество интегральных переменных (нейронов скрытого слоя), $\nu_{min(max)}$ – минимальные и максимальные значения интегральных переменных.

Для случая двух классов лучший результат дает набор из нескольких лучших кластеров для одного класса, с принятием решения по умолчанию. Если ни одно из

условий правил для этого класса не выполнено, то принимается гипотеза о принадлежности наблюдения к классу по умолчанию. Однако такой метод не дает представления о характеристиках остальных классов, поэтому рекомендуется использовать набор правил, охватывающий все классы.

Для оценки сложности набора правил применяется мера, предложенная Гейнсом [5]:

$$Sm = 0.6 * Rn + 0.4 * Inp\,,$$

где Sm – мера сложности набора правил, Rn – количество классифицирующих правил, Inp – количество входных переменных. Чем меньше значение Sm, тем лучше, с точки зрения компактности, является набор правил.

Таким образом, к достоинствам рассматриваемого метода Hruschka E., Ebecken N. можно отнести:

- получение достаточно компактного набора классифицирующих правил;
- использование эволюционного подхода для определения числа кластеров.

Недостатками являются:

- при применении генетического алгоритма к активирующим значениям скрытого слоя нейронной сети не достигается желаемая интерпретируемость результатов классификации, так как в качестве предпосылок правил выступают не сами входные переменные, а их линейные комбинации;
- невозможность классифицировать наблюдение, активирующее несколько правил, относящих наблюдение к различным классам.

2. Улучшение классификации с помощью аппарата нечеткой логики

Недостаточная точность классификации многомерных данных с помощью алгоритма [4] объясняется двумя причинами. Во-первых, из-за того, что часть не эффективных кластеров отбрасывается, то некоторые области пространства значений входных переменных выпадают из рассмотрения. Во-вторых, представление кластеров в виде гиперпаралеллепипеда создает ряд перекрывающихся областей. Наблюдения, попадающие в такую область пересечения нескольких кластеров, активируют одновременно несколько правил и в условиях такой неопределенности решение по методу [4] невозможно принять, а потому наблюдения относятся к не классифицированным.

Предлагаемое нами развитие метода [4], обеспечивающее повышение точности классификации многомерных данных за счет снижения числа объектов, которые не удается классифицировать, заключается в следующем:

1. Снижение неопределенности при классификации наблюдений, лежащих в перекрывающихся областях кластеров, за счет проецирования их границ на оси a_i, и задания на соответствующих интервалах-проекциях нечетких множеств с функциями принадлежности трапециевидной или гауссовой формы.
2. Принятие решений о классификации наблюдений с помощью специальной процедуры на основе аппарата нечеткой логики, учитывающей все множество правил, активированных наблюдением.

Таблица 1. Результаты классификации без применения нечеткой логики

№	Правильно классифицировано	Ошибка	Невозможно классифицировать
1	144	1	5
2	143	2	5
3	144	1	5
4	143	2	5
5	144	1	5
Среднее, %	**95.73**	0.93	**3.33**

Принятие решений происходит следующим образом: новое наблюдение проверяется на соответствие каждому из классифицирующих правил. Если не активировано ни одно из правил, то наблюдение не может быть классифицировано. Если активировано только одно правило, классификация закончена. Если активируются более одного правила, то анализируются следствия этих правил, т.е. каким классам они соответствуют. Зачастую эти правила относят наблюдение к одному и тому же классу, что позволяет сразу же принять однозначное решение о принадлежности наблюдения к соответствующему классу. И самый сложный случай – когда правила относят наблюдение к разным классам. В таком случае используется специальная процедура принятия решения.

Для наблюдения x_t, активирующего одновременно два правила, относящих его разным классам, необходимо вычислить степень его принадлежности μ^k каждому из правил. На интервале значений каждой интегральной переменной a_i , входящей в предпосылку правила, определяются функции принадлежности $\mu_i^k, i = 1, \ldots, q, k$ – номер правила. Функции принадлежности имеют трапециевидную форму и формируются специальным образом. Степень принадлежности наблюдения правилу определяется произведением функций принадлежностей по каждой из осей a_i: $\mu^k = \prod_{i=1}^{q} \mu_i^k(a_i)$, где k – номер правила, q – количество интегральных переменных, a_i – значение i-ой интегральной переменной для наблюдения x_t. В качестве классифицирующего принимается правило, степень принадлежности наблюдения которому максимальна. Функции принадлежности имеют трапециевидную форму и для случая двумерных кластеров формируются как показано на рисунке 1. Очевидно, что область неопределенности сужается до объектов, лежащих на пересечении функций принадлежности, построенных на соответствующих осях (интегральных переменных – активирующих значениях нейронов скрытого слоя нейросетевой модели).

3. Результаты сравнения

Тестирование проводилось на наборе данных *Iris.data* (*http://joc.pubs.informs.org/Supplements/Lee/iris.data*), содержащем 150 наблюдений, 4 переменных. Наблюдения делятся на 3 класса, первый из которых отделен, второй и третий пересекаются.

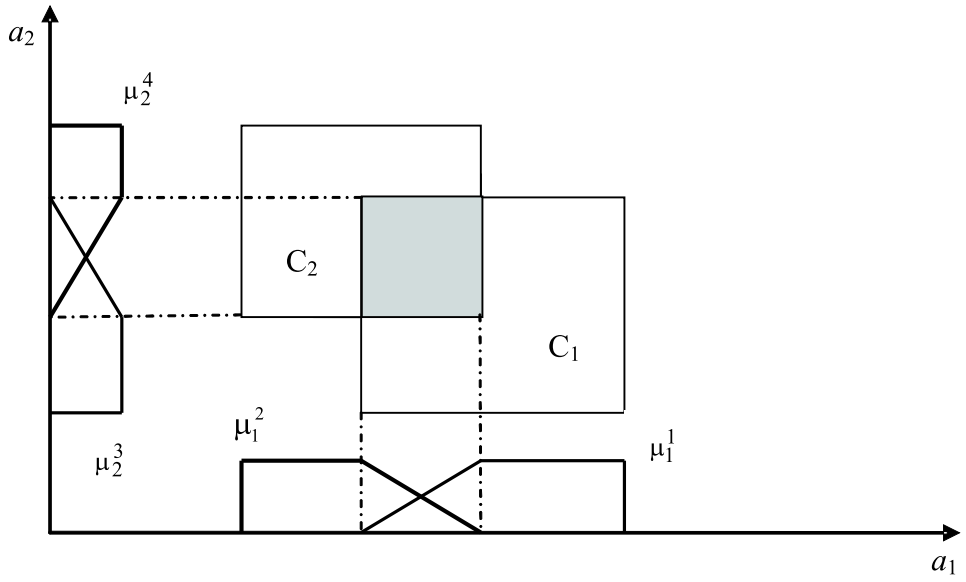

Рис. 1. Пример формирования функций принадлежности для пересекающихся двумерных кластеров

Таблица 2. Результаты классификации с применением нечеткой логики

№	Правильно классифицировано	Ошибка	Невозможно классифицировать
1	147	2	1
2	145	1	4
3	148	1	1
4	145	1	4
5	147	2	1
Среднее, %	**97.6**	0.93	**1.47**

Сравнение результатов классификации без применения нечеткой логики и с ее применением представлены в таблицах 1 и 2. При классификации использовались все правила, порожденные кластерами с количеством элементов 5 и более. Кластеры с меньшим количеством элементов отбрасывались как случайные выбросы.

Заключение

В статье рассмотрены преимущества и недостатки широко распространенного метода классификации Hruschka E., Ebecken N. на основе применения кластерного генетического алгоритма, который использован для получения классифицирующих правил из обученной нейронной сети. *Основным результатом данной статьи является* подход к повышению точности классификации многомерных данных, использующий аппарат нечеткой логики и позволяющий снизить число объектов, которые

не удается классифицировать. Результаты вычислительных экспериментов на наборе данных *Iris.data* показали, что применение предложенного подхода и процедуры принятия решений с учетом всего множества активированных классифицирующих правил существенно повышает точность классификации до уровня 97.6 % и в 2.47 раза (до 1.47%) снижает число объектов, которые невозможно было классифицировать согласно методу Hruschka E., Ebecken N.

Дальнейшие исследования предполагают проведение тестирования предложенного подхода на реальных данных и проверку возможности его применения для классификации медицинских данных в задачах прогнозирования и выбора протокола лечения в детской лейкозологии.

Список литературы

1. *Lu H., Setiono R., Liu H.* Effective Data Mining Using Neural Networks // IEEE Transactions on Knowledge and Data Engineering. – 1996. – Vol.8, №6. – P. 957-961.
2. *Craven M.W., Shavlik J.W.* Using Neural Networks for Data Mining // Future Generation Computer Systems. – 1997. – Vol.13, №2. – P. 211-229.
3. *Fu L.* Neural Networks in Computer Intelligence. – McGraw-Hill Inc., USA, 1994.
4. *Hruschka E., Ebecken N.* A Clustering Genetic Algorithm For Extracting Rules From Supervised Neural Network Models In Data Mining Tasks // International Journal of Computers, Systems and Signals. – 2000. – Vol.1, №1. – P. 17-29.
5. *Gaines B.R.* Transforming Rules and Trees into Comprehensible Knowledge Structures // In Advances in Knowledge Discovery and Data Mining. – MIT Press. – 1996. – P. 205-229.

Статья поступила в редакцию 27.04.2008