УДК 004.93

НЕЧЕТКИЕ КЛАССИФИКАЦИОННЫЕ МОДЕЛИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА БАНКОВСКИХ ДАННЫХ

С.И. Ватлин

Национальный банк Республики Беларусь, Расчетный Центр, vatlin_si@nbrb.by

Введение. Успешное развитие банка напрямую зависит от его способности адекватно и оперативно реагировать на изменения внешней среды, а также умения прогнозировать результаты тех или иных внешних воздействий.

Для достижения указанных целей многие банки внедряют системы интеллектуального анализа данных (data mining systems), обеспечивающие возможности решения следующих основных задач: анализ кредитного риска, прогнозирование остатков на счетах клиентов, управление портфелем ценных бумаг, повышение качества архивной финансовой информации, верификация данных по курсам валют и других.

Классификационные модели (алгоритмы) обработки информации (классификаторы) являются одним из базовых типов моделей интеллектуального анализа данных [1]. На базе таких моделей определяется с объектами какого типа (из некоторого заранее определенного множества типов объектов $L = (w_1, w_2, ..., w_L)$, аналитик оперирует в фиксированный момент времени.

Широкий спектр существующих классификационных моделей наиболее полно проанализирован в [1,2].

Одно из возможных направлений в развитии классификационных моделей связано с теорией нечетких множеств. Как отмечает Л.Заде, глубинная связь между такими моделями и моделями теории нечетких множеств основана на том факте, что большинство реальных классов размыты по своей природе и переход от принадлежности к непринадлежности им скорее непрерывен, чем скачкообразен [3].

Принцип Бритвы Оккама для нечетких классификационных моделей анализа данных. Пусть $X(CardX = m)$ и $L(CardL \le m)$ – фиксированные дискретные множества произвольной природы, а Y – фиксированная совокупность функций, переводящих L в отрезок [0,1] ($\mu \in Y \Leftrightarrow \mu : L \to [0.1]$ и μ – обладает некоторой совокупностью свойств, характеризующих конкретную предметную область**).** Множества X и Y будем называть множествами начальных и финальных символов в задаче нечеткой классификации. Пусть также f соответствует целевой классификационной функции (модели) из X в Y, $f : X \to Y$. Обозначим посредством Θ совокупность упорядоченных пар из $X \times Y$ такую, что $\Theta = \{(x_i, y_i)\}_{i=1}^{i=n}, x_i \in X, y_i \in Y, y_i = f(x_i), \forall i = \overline{1,n}$. В дальнейшем будем называть Θ – обучающим множеством (для функции f), а n – порядком обучающего множества Θ.

Обозначим посредством G произвольный (нечеткий) классификационный алгоритм (классификатор), переводящий Θ в гипотетическую классификационную функцию $h_{G\Theta}, h_{G\Theta} = G(\Theta), h_{G\Theta} : X \to Y$.

Построение приемлемой гипотетической функции (модели), как правило, осуществляется в рамках некоторой языковой структуры, которая обеспечивает символическое представление множества потенциальных классификационных моделей H. Обозначим множество всех нечетких классификаторов, индуцирующих множество моделей (функций) H, посредством $\{G(\Theta)\}$. Определим также фиксированную языковую структуру (язык) Z как упорядоченную пару (I,T), где T – есть множество предложений в языке и I – интерпретатор языка $I : T \to H$.

На каждом $t \in T$ определим меру сложности $C: T \to R$ языка Z, которая может характеризовать как синтаксические, так и семантические аспекты предложения t.

При фиксированной мере сложности C языка Z определим меру сложности $C(h)$ модели (функции) $h_{G\Theta}: X \to Y$ по следующему правилу:

$$C(h) = \frac{\min}{t: I(t) = h}\{C(t)\}.$$

Иными словами, сложность нечеткой классификационной модели h зависит от языка Z и определяется как сложность наипростейшего предложения $t \in T$, которое представляет эту модель.

Пусть $Z = (I, T)$ – фиксированная языковая структура, $C: T \to R$ – фиксированная мера сложности языка Z, X и Y – множества начальных и финальных символов в фиксированной задаче нечеткой классификации (интеллектуальной обработки данных). Пусть также f соответствует фиксированной (но неизвестной) целевой классификационной функции и $\Theta = \{(x_i, y_i)\}_{i=1}^{i=n}, x_i \in X, y_i \in Y, y_i = f(x_i), \forall i = \overline{1,n}$ – обучающее множество целевой функции f порядка n.

Тогда, принцип Бритвы Оккама (принцип простоты) для нечетких классификационных моделей интеллектуального анализа данных может быть сформулирован следующим образом: при прочих равных условиях следует выбирать тот элемент $G^* \in \{G(\Theta)\}$, который порождает наипростейшую (по мере сложности C) нечеткую классификационную модель $h^* = h_{G^*\Theta}$,

$$G^* \Leftrightarrow \frac{\min}{G \in \{G(\Theta)\}}\{C(h_{G\Theta})\} \tag{1}$$

Основной проблемой, возникающей при использовании приведенной формулировки, является проблема неоднозначности в определении сложности (простоты) нечеткой классификационной модели $h \in H$. Модель, простая в одном языке, может оказаться сложной в другом.

Однако, по крайней мере, для одного типа нечетких классификаторов принцип простоты интеллектуального анализа данных может быть сформулирован строго однозначным образом независимо от используемой языковой структуры.

Критерий качества для вырождено самоотгадывающих нечетких классификационных моделей, основанный на инвариантной мере простоты. Пусть $\Theta = \{(x_i, y_i)\}_{i=1}^{i=n}$ – обучающее множество порядка n целевой классификационной функции $f: X \to Y$, G – нечеткий классификатор, переводящий Θ в гипотетическую классификационную функцию $h_{G\Theta}$ из X в Y, $h_{G\Theta}: X \to Y$. Мы будем говорить, что G согласован с обучающим множеством Θ в том и только в том случае, когда выполняется соотношение: $h_{G\Theta}(x_i) = y_i, \forall i = \overline{1,n}$. При фиксированных X, Y и Θ обозначим посредством $Con(\Theta)$ – совокупность всевозможных нечетких классификаторов, согласованных с Θ.

Пусть G – фиксированный элемент совокупности $Con(\Theta)$. Будем называть G – вырождено самоотгадывающим (для Θ) нечетким классификатором в том и только в том случае, когда выполняются следующие условия:

1. G согласован с любым подмножеством $\theta^* \subset \Theta$ таким, что $Card\Theta^* = n^* > n_0 - 1$ (n_0 – фиксированный параметр, характеризующий структуру множества X. В случае, когда $X = R^m$, $n_0 = m + 1$).
2. Существует подмножество $\Theta' \subset \Theta, Card\Theta' = n' \geq n_0 - 1$, являющееся функцией от G и Θ, $\Theta' = \Phi(G, \Theta)$ такое, что:

2.1. $h_{G\Theta'}(x) = h_{G\Theta}(x), \forall x \in \Theta \setminus \Theta'$;

2.2. если $(x, y) \notin \Theta'$, то $\Phi(G, \Theta) = \Phi(G, \Theta \setminus (x, y))$.

При фиксированных X, Y и Θ обозначим посредством $DSg(\Theta)$ совокупность всевозможных нечетких классификаторов вырождено самоотгадывающих для Θ. Пусть $\Theta_1 (Card\Theta_1 = n_1)$ и $\Theta_2 = \Theta_1 \cup \{(\bar{x}, \bar{y})\} (Card\Theta_2 = n_2 = n_1 + 1)$ соответствуют произвольным обучающим множествам для целевой классификационной функции f, $f : X \to Y$.

Предположим, что в $Con(\Theta_1)$ существует, как минимум, два различных элемента G_1 и G_2. Мы будем говорить, что нечеткий классификатор G_2 более эффективен, чем нечеткий классификатор G_1 (при переходе от обучающего множества Θ_1 к обучающему множеству Θ_2) и обозначать этот факт как $G_2 \succ G_1$, в том и только в том случае, если выполняется следующее соотношение:

$$P(h_{G_2\Theta_1}(\bar{x}) = (\bar{y})) > P(h_{G_1\Theta_1}(\bar{x}) = (\bar{y})), \tag{2}$$

где $P(\bullet)$ – вероятность события $(\bullet)$.

Предположим, что в $DSg(\Theta_1)$ существует, как минимум, два различных элемента G_1 и G_2. Обозначим $n_1 = Card(\Theta_1')$, $\Theta_1' = \Phi(G, \Theta_1)$ и $n_2 = Card\Theta_2'$, $\Theta_2' = \Phi(G, \Theta_2)$. Справедлива следующая

Теорема 1 [4]. Вероятность события $h_{G_2\Theta_1}(\bar{x}) = (\bar{y})$ больше либо равна $1 - \frac{n_2'}{n_1 - 1}$.

Будем говорить, что вырождено самоотгадывающий нечеткий классификатор G_2 экономнее (проще) вырождено самоотгадывающего нечеткого классификатора G_1 (при переходе от обучающего множества Θ_1 к обучающему множеству Θ_2) и обозначать этот факт как $G_2 \rhd G_1$, в том и только в том случае, если выполняется следующее соотношение:

$$n_2' > n_1'. \tag{3}$$

Справедлива следующая

Теорема 2. (Критерий качества вырождено самоотгадывающих нечетких классификаторов)

$$G_2 \rhd G_1 \Leftrightarrow G_2 \succ G_1.$$

Доказательство непосредственно вытекает из определений (2) и (3) и теоремы 1.

Заключение. При использовании принцип Бритвы Оккама (принципа простоты) в моделях интеллектуального анализа банковских данных (базирующихся на нечетких классификаторах) возникает проблема выбора адекватной меры простоты (сложности) таких моделей. Модель анализа данных являющаяся простой в одной языковой структуре может оказаться достаточно сложной в другой подобной структуре.

В работе показано, что как минимум для одного типа нечетких классификаторов принцип простоты может быть сформулирован строго однозначным образом независимо от типа используемой языковой структуры.

Использование указанного типа нечетких классификаторов в системах интеллектуального анализа данных позволяет повысить эффективность обработки информации в банках.

Список использованных источников

1. Holmes, D.E. Data mining: Foundations and intelligent paradigms. Volume 1. Clustering, association and classification / D.E. Holmes, L.C. Jain. – Berlin: Springer, 2011. – 352 p.

2. Michalski, J.G. Carbonnel, T.M. Mitchell. Machine learning: An artificial learning approach / R.S. Michalski, J.G. Carbonnel, T.M. Mitchell. – Palo Alto, CA: M. Kaufmann, 1983. – 482 p.

3. Zadeh, L. Fuzzy sets and their application to pattern classification and cluster analysis. In J. van Ryzin (Editor). Classification and Clustering / L. Zadeh. – New York: Academic Press, 1977. – P. 148–223.

4. Vatlin, S.I. Strongly selfguessing fuzzy classifiers / S.I. Vatlin // Informatica (Vilnius). – 1995. – vol. 6. – P. 85–93.