УДК 681.3.00:007

Д.Э. Ситников[1], Е.В. Титова[1], А.И. Коваленко[1], Ю.В. Банин[2], Р.О. Патык[2]

[1]*Харьковская государственная академия культуры, Харьков*
[2]*Харьковский национальный университет радиоэлектроники, Харьков*

МЕТОД ПОИСКА ЛОГИЧЕСКИХ ПРАВИЛ В ЭКСПЕРТНЫХ СИСТЕМАХ НА ОСНОВЕ АССОЦИАТИВНЫХ ЗАВИСИМОСТЕЙ МЕЖДУ ДИСКРЕТНЫМИ ПРИЗНАКАМИ

В данной работе рассмотрен метод формирования логических правил базы знаний экспертной системы на основе теории приближенных множеств. Логические ассоциативные правила формируются путем выделения из имеющегося покрытия, которое удовлетворяет определенному значению поддержки, наиболее значимых, информативных информационных признаков. Полученные результаты могут быть использованы при формировании базы знаний экспертной системы.

Ключевые слова: *экспертная система, ассоциативная зависимость, логические правила, приближенные множества.*

Введение

В настоящее время все большее внимание специалистов в области экспертных систем уделяется разработке математических средств и методов, на основе которых можно формировать базу знаний системы, включающую набор логических правил, которые позволяют автоматизировать работу эксперта. Одним из сравнительно новых направлений в разработке методов и алгоритмов формирования базы знаний в виде логических правил является генерация ассоциативных зависимостей между дискретными информационными признаками.

Поиск ассоциативных правил состоит в выводе правил импликативного вида, где каждое правило отражает набор взаимно сопутствующих событий. Первоначально эта задача возникла как прикладная в сфере маркетинга (так называемая "корзина покупок"). Выявлялись товары, которые покупатели стремились приобретать вместе. Впоследствии задача поиска ассоциативных правил была сформулирована в более общем виде как задача нахождения регулярностей (логических закономерностей) в данных. Эти правила могут использоваться для анализа данных и принятия решений, а также после соответствующей оценки служить исходным материалом для построения баз знаний [1].

Следует отметить принципиальное различие между задачей вывода правил ассоциации и правил классификации (в общепринятой постановке) [2]. Классификационные правила претендуют на предсказания (поэтому данные делятся на обучающие и тестовые), а правила ассоциации – нет.

В настоящее время имеется множество алгоритмов генерации ассоциативных зависимостей между бинарными и небинарными признаками. В основном, имеющиеся алгоритмы включают в себя поиск тех подмножеств исходного множества признаков, которые представляют интерес для эксперта. В данной работе предлагается метод формирования логических правил на основе теории приближенных множеств. При этом основное внимание авторов направлено не на поиск так называемых «покрытий», что составляет важную часть процедуры генерации ассоциаций, а на последующее выделение наиболее значимых логических зависимостей из уже имеющегося покрытия.

Цель работы – разработать метод получения логических зависимостей между дискретными признаками для формирования базы знаний экспертной системы на основе выделения наиболее важных признаков с помощью теории приближенных множеств.

Изложение основного материала

Установление логических зависимостей в данных в виде ассоциативных правил

Общее определение ассоциативного правила было дано в [3]:

Пусть $L = I_1, I_2, ..., I_m$ – множество признаков объектов. Пусть T – множество записей. Каждая запись t представлена бинарным вектором $t[k] = 1$, если t содержит признак I_k и $t[k] = 0$, если t не содержит признак I_k $(k = \overline{1, m})$. Пусть X – подмножество некоторых признаков из L, т.е. $X \subseteq L$. Мы говорим, что запись t удовлетворяет X, если $\forall I_k \subseteq X$, $t[k] = 1$.

Под ассоциативным правилом понимается выражение вида $X \to Y$, где $X \subseteq L$, $Y \subseteq L$. Причем $X \cap Y = \varnothing$.

Ассоциативное правило $X \to Y$ поддерживается во множестве записей с уровнем доверия (Confidence – сокр. Conf) c, если c% записей в T, которые содержат X, также содержат Y. Правило $X \to Y$ имеет поддержку (Support – сокр. Sup) s во множестве записей T, если s% записей в T содержат $X \cup Y$.

Проблема поиска ассоциативных правил состоит в генерации всех ассоциативных правил, которые преодолевают заданные пороги поддержки и доверия, т.е. имеют Support ≥ minSupport и Confidence ≥ minConfidence.

Отметим, что Support и Confidence могут выражаться как в процентном соотношении, так и числом, заданным и интервале от 0 до 1. В этом случае Support определяется как соотношение количества объектов, обладающих признаками X и Y к общему количеству объектов в базе данных, а Confidence – как отношение количества объектов, обладающих признаками X и Y к количеству объектов, обладающих признаком X.

Процесс вывода правил ассоциации состоит из двух алгоритмов – нахождения всех "больших наборов" признаков и собственно оформления правил.

Нахождение "больших наборов" признаков (так называемых покрытий) имеет своей целью выделение подмножеств признаков, имеющих удовлетворительную поддержку. Собственно оформление ассоциативных правил сводится к разбиению покрытия на подмножества: если S – покрытие, то для любых непустых $S_1 \subset S$ и $S_2 = S - S_1$ ассоциативным правилом является $S_1 \to S_2$, если имеет необходимый уровень доверия.

Для нахождения покрытий (самая трудоемкая часть генерации ассоциативных правил) часто используется метод ограниченного перебора, предложенный М.М. Бонгардом [4], который позволяет исключать из анализа события с низкой частотой. Ограничения накладываются на длину комбинаций признаков входящих в ассоциативную зависимость, на размер записи, на частоту анализируемых логических комбинаций в зависимости от типа и размера БД и поставленных задач.

Алгоритмы, реализующие этот метод, используют различные достаточно оригинальные варианты искусственного ограничения перебора. Наиболее известными являются алгоритмы AIS [5], Apriori, AprioriTid, AprioriHybrid [6], Sampling [7], PARTITION [8].

Алгоритм AIS [45] позволяет строить ассоциативные правила в виде $X \to I_j$, где X – множество признаков объектов (записей в БД), I_j – единичный признак $(X \cap I_j = \varnothing)$. Кроме того, на признаки, входящие в ассоциативные правила накладываются определенные синтаксические ограничения: перед началом работы задается признак (подмножество признаков), который будет входить в правую (левую) часть ассоциативного правила. Алгоритм использует многократное сканирование БД, формируя при этом множество так называемых кандидатов (candidate itemsets), которые потенциально являются покрытиями. Кандидаты формируются путем расширения граничного множества (frontier set).

Во время следующего сканирования БД для полученных кандидатов считается их уровень Support. Для ограничения перебора во время текущего сканирования используется отбрасывание тех кандидатов, которые не могут набрать нужный уровень поддержки. Для этого по мере сканирования БД для каждого кандидата X подсчитывается количество оставшихся записей. Если данный кандидат на оставшихся записях не может набрать Support > minSupport, то он перестает рассматриваться.

Кандидат, набравший нужный уровень Support, считается покрытием, включается в граничное множество для последующего сканирования БД и расширения.

Алгоритмы Apriori, AprioriTid [6] отличаются от AIS способом генерации кандидатов. В отличие от AIS, в Apriori, AprioriTid кандидаты формируются из кандидатов на предыдущем шаге.

К недостаткам данных алгоритмов следует отнести многократное сканирование БД и как следствие резкое увеличение времени выполнения при увеличении количества записей в БД, работа только со статичными БД (при удалении старых или добавлении новых записей требуется заново запускать программу на выполнение). Несмотря на вышеуказанные недостатки, алгоритм Apriori является одним из самых распространенных в системах поиска ассоциативных правил. На его базе создана целая группа алгоритмов, улучшающих его характеристики: Sampling, PARTITION и т.д.

Метод генерации логических правил базы знаний экспертной системы на основе теории приближенных множеств

Теория приближенных множеств, разработанная польским ученым Здиславом Павлаком [9, 10], послужила базой для развития целого направления интеллектуального анализа данных, основанного на этой теории. Одним из основных понятий теории приближенных множеств является понятие "неразличимости" (или "неразличимого" отношения). Утверждается, что с разными элементами множества может ассоциироваться одинаковая информация, что делает невозможным точное определение принадлежности этих элементов к некоторому множеству (такое множество называют приближенным). Приближенное множество характеризуется двумя аппроксимациями – нижней, которая определяет элементы, которые однозначно принадлежат данному множеству, и верхней, которая определяет элементы, которые могут принадлежать данному множеству.

Пусть Y – некоторое множество, являющееся подмножеством универсума U и I – отношение "неразличимости". Нижней аппроксимацией Y является:

$$I_*(Y) = \{y \in U : I(y) \subseteq Y\},$$

а верхней аппроксимацией:

$$I^*(Y) = \{y \in U : I(y) \cap Y \neq 0\},$$

где I(y) определяет множество объектов, "неразличимых" с y (с точки зрения имеющейся информации) [9]. Под граничным регионом (boundary region) понимается $BN_I = I^*(Y) - I_*(Y)$. Таким образом, граничный регион содержит элементы, которые принадлежат верхней аппроксимации множества и не принадлежат нижней. Если $BN_I \neq 0$, то множество Y является приближенным.

Для приближенных множеств вводится также понятие функции принадлежности:

$$\mu_Y^I(y) = \frac{|Y \cap I(y)|}{|I(y)|}.$$

Очевидно, что $0 \le \mu_Y^I(y) \le 1$.

Функция принадлежности может быть использована для определения верхней, нижней аппроксимаций приближенного множества, а также граничного региона:

$$I_*(Y) = \{y \in U : \mu_Y^I(y) = 1\};$$

$$I^*(Y) = \{y \in U : \mu_Y^I(y) > 0\};$$

$$BN_I(Y) = \{y \in U : 0 < \mu_Y^I(y) < 1\}.$$

Наборы "неразличимых" элементов, с которыми ассоциирована одинаковая информация, Павлак называет "гранулами". В своей работе [11] он также вводит понятие редакта (от англ. reduction – уменьшение, сокращение). Под редактом понимается минимальный набор признаков, позволяющий различать "гранулы". Минимальность этого набора подразумевает невозможность его сокращения без потери различимости "гранул". Таким образом, редакт определяет набор признаков, адекватно описывающий приближенное множество. Нахождение редакта (редактов) для произвольного множества Павлак называет интересной, но довольно сложной задачей [11] и не приводит никаких способов ее решения. Один из подходов к решению этой задачи описан в [12].

Метод нахождения верхней и нижней аппроксимаций приближенного множества с использованием алгебры конечных предикатов был предложен в [13, 14]. Он позволяет находить аппроксимации заданного множества за один просмотр набора данных (т.е. имеет линейную сложность), при этом используя только булевы операции.

В [15] авторами предложен метод нахождения локальных редактов, т.е. таких наборов атрибутов, которые наиболее близко описывают некоторое множество. Мерой близости описания является сравнение исходной мощности граничного региона с полученной. Этот подход к определению редакта (редактов) представляется более гибким, в отличие от классического определения по Павлаку. Если для некоторого приближенного множества может не существовать классического редакта (иными словами, количество признаков не может быть уменьшено и весь набор признаков является редактом), то при использовании предлагаемого метода могут быть получены минимизированные наборы, с достаточной точностью описывающие множество.

Представляется интересной связь между ассоциативными правилами и логическими правилами, полученными на основе теории приближенных множеств. Ассоциативные зависимости выявляются на основе нахождения покрытий (covers) как наборов признаков, удовлетворяющих необходимым значениям Support (поддержки правил). После того, как покрытие найдено, необходимо рассмотреть возможные логические зависимости между этими признаками. В общем виде это – задача нахождения некой булевой функции, структура которой может быть достаточно сложной. В более простом виде это – некое логическое правило, в правой части которого находится один целевой признак, а в левой части – остальные признаки покрытия. В этом случае предлагается после нахождения покрытия выделить целевой признак, который будет фигурировать в правой части ассоциативного правила, на основе разработанного в [16] алгоритма.

При определении наиболее важных ассоциативных зависимостей логично искать наиболее важные признаки, которые будут фигурировать в правой части логических правил. Для этого после нахождения покрытия предлагается следующий метод получения целевых признаков, а также тех признаков, которые будут фигурировать в левой части правил:

1. Путем последовательной проверки каждого признака покрытия выбирается признак, исключение которого приводит к максимальному увеличению граничного региона (boundary region).
2. Выбранный признак исключается из рассмотрения, и среди остальных признаков находится редакт (несократимый набор признаков).
3. Полученный редакт ставится в левую часть правила.
4. Найденный на первом этапе самый важный признак ставится в правую часть правила.
5. Проверяется значение Confidence для полученного правила. Если данное значение устраивает эксперта, то правило вносится в выходной набор ассоциативных зависимостей.
6. Процедура повторяется для следующего признака в порядке убывания значимости.

Предложенный метод позволяет получать логические правила базы знаний экспертной системы в порядке уменьшения важности признаков, фигурирующих в правых частях правил.

Выводы

В данной работе предложен подход к генерации логических ассоциативных зависимостей на основе теории приближенных множеств. Этот подход предполагает наличие покрытия, удовлетворяющего заданному заранее значению поддержки логических правил, которые должны формировать базу знаний экспертной системы.

Описанный метод позволяет получать логические правила, в правых частях которых находятся наиболее важные, значимые, информативные признаки. Важность признака, фигурирующего в имеющемся покрытии, определяется с помощью теории приближенных множеств путем последовательного исключения признаков и рассмотрения граничного региона (boundary region), который увеличивается с увеличением важности исключенного признака. Предложенный метод включает в себя поиск редактов приближенного множества с целью определения тех признаков, которые необходимо включить в левую часть правила. Данный метод основан на более ранних работах авторов. Полученные результаты можно использовать при формировании базы знаний экспертной системы.

В перспективе предполагается провести вычислительные эксперименты с обработкой больших информационных массивов на основе предложенного метода. Также авторы считают необходимым продолжить теоретические исследования перспектив применения теории приближенных множеств к генерации ассоциативных правил. При этом предполагается рассмотрение правил с более сложной структурой.

Список литературы

1. Городецкий В.И. Современное состояние технологии извлечения знаний из баз и хранилищ данных / В.И. Городецкий, В.В. Самойлов, А.О. Малов // Новости искусственного интеллекта. – 2002. – № 3. – С. 2-19.

2. Балабанов А.С. Выделение знаний из баз данных – передовые компьютерные технологии интеллектуального анализа данных / А.С. Балабанов // Математичні машини і системи. – 2001. – № 1, 2. – С. 40-54.

3. Agrawal R. Mining association rules between sets of items in large databases / R. Agrawal, T. Imielinski, A. Swami // Proc. of the ACM SIGMOD Conference. – Washington DC, USA, May 1993. – P. 207-216.

4. Бонгард М.М. Проблема узнавання / М.М. Бонгард. – М.: Наука, 1967. – 319 с.

5. Agrawal R. Fast algorithms for mining association rules / R. Agrawal, R. Srikant // Proc. of the 20th VLDB Conference Santiago. – Chile, September 1994. – P. 487-499.

6. Savasere A. An Efficient Algorithm for Mining Association Rules in Large Databases / A. Savasere, E. Omiecinski, S.B. Navathe. // Proc. of the 21nd International Conference on Very Large Databases. – Zurich, Switzerland, 1995. – P. 432-444.

7. Toivonen H. Sampling Large Databases for Association Rules / H. Toivonen // Proc. of the 22nd International Conference on Very Large Databases. – Mumbai, India, 1996. – P. 134-145.

8. Pawlak Z. Rough set approach to knowledge-based decision support / Z. Pawlak // Proc. of the 14 European Conference on Operational Research Jerusalem, Israel, July 1995.

9. Pawlak Z. Vaguenes and uncertainty: a Rough set perspective / Z. Pawlak // Computational Intelligence. – Vol. 11 (Issue 2). – May 1995. – Pp. 227-232.

10. Pawlak Z.: Rough set approach to knowledge-based decision support / Z. Pawlak // European Journal of Operational Research. – 1997. – No. 99. – Pp. 420-432.

11. Определение минимального набора признаков, адекватно описывающих нечеткое множество / Д.Э. Ситников, О.С. Вильчинская, Н.С. Кравец, Е.В. Титова // Вестник национального технического университета "ХПИ". – Х.: ХНИУ ХПИ, 2002. – № 20. – С. 65-70.

12. Sitnikov D. An algebraic approach to defining rough set approximations and generating logic rules / D. Sitnikov, O. Ryabov // Zanasi, A.; Ebecken, N.; Brebbia, C.; (eds), Data Mining V. – Malaga, Spain, 2004. – P. 179-188.

13. Метод нахождения аппроксимаций приближенных множеств и построения логических правил на основе алгебры конечных предикатов / Д.Э. Ситников, О.А. Рябов, Е.В. Титова, О.А. Романенко // Системи обробки інформації. – Х.: ХУПС, 2007. – № 4 (62). – С. 144-149.

14. Ситников Д.Э. Метод нахождения минимизированных наборов признаков в базах данных с использованием теории приближенных множеств / Д.Э. Ситников, О.А. Романенко, С.В. Титов, Е.В. Титова // Системи обробки інформації. – Х.: ХУПС, 2007. – № 7 (65). – С. 91-95.

15. Sitnikov D. A method for finding minimal sets of features adequately describing discrete information objects / D. Sitnikov, E. Titova, O. Ryabov // Zanasi, A.; Ebecken, N.; Brebbia, C.; (eds), Data Mining X – Greece, 2009. – P. 22-30.

Поступила в редколлегию 15.05.2014

Рецензент: д-р техн. наук, проф. И.В. Гребенник, Харьковский национальный университет радиоэлектроники, Харьков.

МЕТОД ПОШУКУ ЛОГІЧНИХ ПРАВИЛ В ЕКСПЕРТНИХ СИСТЕМАХ НА ОСНОВІ АСОЦІАТИВНИХ ЗАЛЕЖНОСТЕЙ МІЖ ДИСКРЕТНИМИ ОЗНАКАМИ

Д.Е. Ситніков, О.В. Тітова, А.І. Коваленко, Ю.В. Банін, Р.О. Патик

Запропоновано метод формування логічних правил експертної системи з використанням теорії наближених множин. Цей метод базується на аналізі інформативності ознак, що входять до покриття. Ознака з найбільшою інформативністю розміщується в правій частині логічного правила, а редакт приблизної множини, побудований на базі ознак, що залишились, розміщується в лівій частині правила. Таким чином, логічні правила бази знань експертної системи можуть бути впорядковані за важливістю цільових ознак.

Ключові слова: *експертна система, асоціативна залежність, логічні правила, наближені множини.*

A METHOD FOR SEARCHING LOGIC RULES IN EXPERT SYSTEMS BASED ON ASSOCIATION DEPENDENCIES BETWEEN DISCRETE FEATURES

D.E. Sitnikov, E.V. Titova, A.I. Kovalenko, U.V. Banin, R.O. Patik

In this paper a method for forming logic rules of the knowledge base of an expert system has been considered based on rough sets theory. This method is based on the analysis of the importance of the features included in a cover. The most informative feature is placed on the right side of the logic rule, and the rough set reduct built on the basis of the remaining features is placed on the left side. Thus the logic rules of the knowledge base of an expert system can be ordered by the importance of target features.

Keywords: *consulting model, associative dependence, logical rules, close great numbers.*