УДК 61:004.651 (075.8)

ПРО КЛІНІЧНУ ЕКСПЕРТНУ СИСТЕМУ, ЩО ГРУНТУЄТЬСЯ НА ПРАВИЛАХ, НА ОСНОВІ ТЕХНОЛОГІЇ DATA MINING

В. П. Марценюк, О. О. Стаханська

ДВНЗ "Тернопільський державний медичний університет імені І. Я. Горбачевського МОЗ України"

У роботі розглянуто питання програмної реалізації методу індукції правил на основі алгоритму послідовного покриття. Такий підхід дозволяє розробити систему підтримки клінічних рішень. Проект реалізовано в середовищі Netbeans на основі Java-класів.

Ключові слова: диференційна діагностика, прийняття рішень, data mining, індукція правил, Java, SQL.

О КЛИНИЧЕСКОЙ ЭКСПЕРТНОЙ СИСТЕМЕ, ОСНОВАННОЙ НА ПРАВИЛАХ, НА ОСНОВЕ ТЕХНОЛОГИИ DATA MINING

В. П. Марценюк, О. А. Стаханская

ГВУЗ "Тернопольский государственный медицинский университет имени И. Я. Горбачевского МОЗ Украины"

В работе рассмотрены вопросы программной реализации метода индукции правил на основе алгоритма последовательного покрытия. Такой подход позволяет разработать систему поддержки клинических решений. Проект реализован в среде Netbeans на основе Java-классов.

Ключевые слова: дифференциальная диагностика, принятие решений, data mining, индукция правил, Java, SQL.

ABOUT CLINICAL EXPERT SYSTEM BASED ON RULES USING DATA MINING TECHNOLOGY

V. P. Martsenyuk, O. O. Stakhanska

SHEI "Ternopil State Medical University by I. Ya. Horbachevsky of MPH of Ukraine"

In the work the topics of software implementation of rule induction method based on sequential covering algorithm are considered. Such approach allows us to develop clinical decision support system. The project is implemented within Netbeans IDE based on Java-classes.

Key words: differential diagnostics, decision making, data mining, rule induction, Java, SQL.

Вступ. В медицині поняття диференційної діагностики означає системний підхід, що грунтується на доказовості, для визначення причини наявних симптомів, у випадку, коли є кілька альтернативних пояснень, а також для зменшення переліку можливих діагнозів.

Сьогодні медичне діагностування може виконуватися автоматично з використанням комп'ютеризованих систем та алгоритмів. Такі системи переважно називаються діагностичними системами підтримки прийняття рішень або медичними діагностичними системами. Вони належать до загальнішого класу клінічних систем підтримки прийняття рішень [9-11]. Метою таких систем є системний супровід лікаря в процесі диференційної діагностики. Багато з таких систем можуть надавати результати навіть коли не вистачає даних, тобто в умовах невизначеності, і, що найважливіше, - вони не обмежені щодо кількості інформації, яку можуть зберігати та обробляти [3-8].

У даній роботі ми представимо класифікатор, що грунтується на правилах, в якому модель знань представляється множиною правил IF-THEN. Спершу ми покажемо, як такі правила можуть використовуватися для класифікації. Далі представимо метод ге-

нерації таких правил на основі алгоритму послідовного покриття.

Зазначимо, що правила можуть генеруватися як з дерева рішень, так і безпосередньо з навчальних даних, використовуючи алгоритм послідовного покриття [1, 2].

Означення класифікаційних правил

Традиційне означення IF-THEN-правила наведене в роботах [1, 5]. Математично задача індукції класифікаційних правил формулюється таким чином. Маємо множину *D,* що містить *N* наборів навчальних даних. При цьому кожен *i*-й набір $(A_1', A_2', ..., A_p', C')$ складається з вхідних даних - атрибутів $A_1, \ldots, A_p$ та вихідних даних - атрибуту класу С. Можна припустити, що атрибути $A_1, \ldots, A_p$ приймають лише категоріальні значення. Атрибут класу С приймає одне з К дискретних значень: $C \in \{1, ..., K\}$. Метою є прогнозування класифікаційним правилом значення атрибуту класу *C* на основі значень атрибутів $A_1, \ldots, A_p$.

Класифікаційним правилом *R* називається імплікація вигляду: $R: \underset{i \in S}{\wedge}(A_i isa a_j^*) \Rightarrow C = c^*$. Тут $S \subseteq \{1, ..., p\}$ – деяка під множина індексів атрибутів.

При цьому слід максимізувати точність прогнозування атрибуту класу, а саме $P\{C = c\}$ для довільного $c \in \{1, ..., K\}$. В результаті ми повинні отримати множину правил для кожного $c \in \{1, ..., K\}$ відповідно, що в антеседенті містять умови включення для категоріальних атрибутів, а в консеквенті значення $c \in \{1, ..., K\}$.

Метою роботи є розробити метод індукції класифікаційних правил з можливістю програмної реалізації у вигляді клінічної експертної системи.

Алгоритм послідовного покриття

Використаємо алгоритм послідовного покриття, описаний в роботі [Han, 2001]. Зауважимо ще раз, що припускаємо, що усі атрибути - категоріальні.

Алгоритм послідовного покриття

Вхідні дані:

D - множина навчальних наборів даних $(A_1', A_2', ..., A_p', C')$

Att_vals - множина всіх атрибутів $A_1, \ldots, A_p$ та їх можливих значень $A_i \in (a_i^1, a_i^2, ..., a_i^{k_i})$.

Вихідні дані: *Rule_set* - множина класифікаційних правил.

Метод:

1. Множина класифікаційних правил *Rule_set* = {};
2. Для кожного класу c;
3. Розпочати цикл «до»;
4. Побудувати нове класифікаційне правило;
Rule = Добути_одне_правило (D, Att_vals, c);
5. Вилучити набори навчальних даних з *D,* що покриваються правилом *Rule,*
6. Виконувати цикл з кроку 3 до настання термінальної умови;
7. Додати нове правило до множини класифікаційних правил:
Rule_set = Rule_set + Rule;
8. Кінець циклу з кроку 2;
9. Множина навчальних правил в *Rule_set*.

В основу методу *Добути_одне_правило (D, Att_vals, c)* покладена міра приросту інформації для побудови правил логіки першого порядку FOIL (First Order Inductive Learner). Метод є ітераційною процедурою по усіх атрибутах $A_1, \ldots, A_p$.

Припусгимо, що ми вже маємо класифікаційне правило:

R: IF *condition* THEN *class = c.*

Метою кожного кроку $i = \overline{1, p}$ є кон'юнкція умови condition за рахунок умови condition" вигляду $(A_i = a_i^j)$. Тут $j \in \{1, ..., K\}$. Тобто нове правило матиме вигляд:

R': IF *condition* AND *condition'* THEN *class = c.*

Згідно з методом FOIL *condition'* вибирається з умови мінімізації міри:

$$FOIL\_Gain = pos' \times (\log_2 \frac{pos'}{pos' + neg'} - \log_2 \frac{pos}{pos + neg}) \quad (1)$$

Тут *pos (neg)* - число позитивних (негативних) навчальних наборів, що покриваються правилом *R, pos'(neg')* - число позитивних (негативних) навчальних наборів, що покриваються правилом *R'*. Під позитивними (негативними) навчальними наборами для певного правила маємо на увазі навчальні набори з умовою юнсеквенту, які задовольняють (не задовольняють) умови антеседенту правила.

Міра (1) сприяє побудові правил, що мають більшу точність і при цьому покривають якомога більше позитивних навчальних наборів.

Програмна реалізація алгоритму

База даних mysql складається з двох таблиць - таблиці attribute, призначеної для зберігання інформації про атрибути, та таблиці categorized_data - для наборів навчальних даних. Структура таблиць на мові SQL для Прикладу наведена нижче:

```
CREATE TABLE mysql.attribute (
        id integer not null unique,
        attribute_name varchar(25),
        attribute_field_name varchar(25),
    primary key (id)
  ) ENGINE=InnoDB;
CREATE TABLE mysql.categorised data (
        id integer not null unique,
        A1 varchar (12),
        A2 varchar (8),

        A21 varchar (7),
        class varchar (68),
    primary key (id)
  ) ENGINE=InnoDB;
```

Програмні класи проекту включено до пакету rule.model. Сюди входять beans-класи Attribute, Attribute_for_list для роботи з даними відповідних таблиць та Rule - для представлення правил. SQL-запити щодо отримання відповідних даних реалізовано в класах AttributeListPeer та TuplesPeer.

У класі Ruleset зберігається набір навчальних правил. До того ж, даний клас безпосередньо реалізує алгоритм послідовного покриття. Клас містить члени: менеджер даних m_dataManager, хеш-таблиці наборів навчальних даних mhtTuples, усіх атрибутів з їх можливими значеннями m_htAtt_vals та безпосередньо множину правил mhtRuleset.

У конструкторі класу Rule set здійснюється побудова хеш-таблиць m htTuples та m htAtt vals, а також застосування алгоритму послідовного покриття - через виклик методу Sequential_covering (m htTuples, m htAtt vals). Отримана множина правил виводиться в текстовий файл.

Клас Rule призначений для зберігання окремих правил. Його членами класу є дві хеш-таблиці: m_htAntecedent - для зберігання антеседенту правила та m htConsequent - для консеквенту. За допомогою методу

```
public void conjunctCondition(Attribute_for_list
        attribute, String sAttribute_value)
```

здійснюється кон'юнкція нової умови до правила. За допомогою методу

```
public Rule copy()
```

створюється «глибока» копія правила. При цьому використовується протокол JOS (Java Object Serialization).

Підрахунок кількості позитивних та негативних навчальних наборів здійснюється у методах класу TuplesPeer.

Приклад. Для прикладу використано експериментальну базу даних біохімічних аналізів залежно від виду політравми. Навчальні набори містять 21 категоріальний атрибут та 6 різних значень атрибуту класу. Нижче наведено побудовані класифікаційні правила:

```
IF Il-10 = normal AND TNF-a = high THEN
class=craniocerebral_injury+
orthopedic_trauma_12_hours

IF Ig M = normal AND TNF-a = high AND
Il-2 = high THEN
class=craniocerebral_injury+
orthopedic_trauma+bleeding_2_hours

IF Il-10 = normal AND TNF-a = high THEN
class=craniocerebral_injury+
orthopedic_trauma+bleeding_12_hours

IF Ig G = low AND TNF-a = high AND
Il-2 = normal THEN
class=craniocerebral_injury+
orthopedic_trauma_2_hours

IF Il-6 = high AND TNF-a = high THEN
class=craniocerebral_injury+
orthopedic_trauma_24_hours

IF Il-6 = high AND TNF-a = normal THEN
class=craniocerebral_injury+
orthopedic_trauma+bleeding_24_hours
```

Час побудови множини класифікаційних правил - 10207 мілісекунд. Зазначимо, що у випадку класів `craniocerebral_injury+ orthopedic_trauma_12_hours` та `craniocerebral_injury+ orthopedic_trauma+bleeding_12_hours` антеседенти правил співпадають. Це підтверджується з думкою експертів про складність діагностування даного виду травм через 12 годин. Для уточнення правил потрібні додаткові навчальні набори.

Висновки. У роботі розглянуто питання програмної реалізації методу послідовного покриття з метою побудови класифікаційних правил.

На прикладі продемонстровано, що такий підхід дозволяє розробити систему підтримки клінічних рішень.

За раху нок використання Java-класів дана реалізація методу послідовного покриття є веб-інтегрованою.

Перспективами досліджень є аналіз продуктивності програмного продукту залежно від кількості атрибутів та обсягу наборів навчальних даних.

Література

1. Han J., KamberM. DataMning: Concepts and Techniques, Morgan Kaufmann, San Francisco, 1st edition, 2001.
2. Hastie T., Tibshirani R, Friedman J. H. The Elements of Statistical Learning, Springer, New York, 1st edition. 2001.
3. Ordonez C. Comparing association rules and decision trees for disease prediction. In Proc. ACM NIKM Workshop - 2006. - P. 17-24.
4. Ordonez C. Integrating K-means clustering with a relational DBMS using SQL, IEEE / C. Ordonez // Transactions on Knowledge and Data Engineering (TKDE). - 2006. - Vol. 18 (2).-P. 188-201.
5. Ordonez C. Models for association rules based on clustering and correlation / C. Ordonez // Intelligent Data Analysis. - 2009. - Vol. 13 (2). -P. 337-358.
6. Quinlan J. R. Induction of decision trees. / J. RQuinlan // Machine Learning. -1986. -Vol. 1. -P. 81-106.
7. Quinlan. J. R. Programs for Machine Learning /J. R. Quinlan // Morgan Kauftnann, 1993.
8. ClassificationandRegressionTrees/L.Breiman, J.Friedman, R.Olshen, C. Stone//WadsworthInternational Group, 1984.
9. Марценюк В. П. О программной среде проектирования интеллектуальных баз данных/В. П. Марценюк, Н. О. Кравец // Клиническая информатика и телемедицина - 2004. - №1, - С. 47-53.
10. Математичні моделі в системі підтримки прийняття рішень страхового забезпечення лікування онкологічних захворювань: підхід на основі динаміки Гомперца / В. П. Марценюк, І. Є. Андруніак, І. С. Гвоздецька, Н. Я. Климук // Доповіді Національної академії наук України. -2012. -№ 10.-С. 34-39.
11. Марценюк В. П. О модели онкологического заболевания со временем пребывания на стадии в соответствии с распределением Гомперца / В. П. Марценюк, Н. Я. Климук // Проблемы управления и информатики. Международный научно-технический журнал . - 2012. - № 6. - С. 137-143."