УДК 004.853

И.В. Бибиков, Р.Ю. Черевко, О.И.Федяев, И.Ю.Бондаренко
Донецкий национальный технический университет, г. Донецк
Кафедра прикладной математики и информатики

АВТОМАТИЗИРОВАННОЕ ИЗВЛЕЧЕНИЕ ЗНАНИЙ ИЗ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ

Аннотация

Бибиков И.В., Черевко Р.Ю., Федяев О.И., Бондаренко И.Ю. Автоматизированное извлечение знаний из реляционных баз данных. *Выполнен анализ работы алгоритма С4.5 автоматического извлечения знаний из реляционных баз данных. С помощью данного алгоритма автоматически строится дерево решений, из которого извлекаются знания в виде продукционных правил. Разработанная технология применяется для* выявления закономерностей из медицинских данных о клинико-психологических особенностях больных алкоголизмом.

Ключевые слова: *приобретение знаний, базы данных, дерево решений, алгоритм С4.5, технология Data Mining*.

Интеллектуальный анализ данных. Современные компьютерные технологии обрушили на людей колоссальные потоки информационной руды в самых различных областях. В результате накоплены огромные объёмы информации. Специалистам стало ясно, что без специальных продуктивных методов эти «сырые» данные образуют никому не нужную информационную свалку [1]. Эта проблемная ситуация разрешилась с появлением технологии Data Mining, которая автоматизирует процесс обнаружения в «сырых» данных ранее неизвестных, нетривиальных, практически полезных знаний (закономерностей), необходимых для принятия решений в различных сферах человеческой деятельности.

Методы Data Mining успешно применяются в первую очередь коммерческими предприятиями, медицинскими и государственными учреждениями, которые решают различные задачи на основе информационных хранилищ данных. Анализ сферы применения Data Mining показывает, что экономический эффект от использования этой технологии в некоторых случаях в 10-70 раз превышал первоначальные затраты.

Рынок систем Data Mining экспоненциально развивается в двух направлениях: многочисленные конкретные бизнес-приложения и универсальные инструментальные системы. Наиболее распространенными на данный момент инструментальными продуктами на рынке являются системы See5 и WizWhy [1].

Система See5 относится к наиболее представительному и популярному направлению, связанному с построением дерева решений. Результат работы выражается в виде деревьев решений и множества if-then-правил.

Другая система WizWhy является современным представителем подхода, реализующего ограниченный перебор. Этот алгоритм вычисляет частоты комбинаций простых логических событий в подгруппах (классах) данных. На основании сравнения вычисленных частот в различных подгруппах данных делается заключение о полезности той или иной комбинации для установления ассоциации в данных, для классификации, прогнозирования и пр. Одним из главных достоинств этой системы – более точные и быстрые вычисления, чем у других методов Data Mining.

Постановка задачи приобретения знаний из базы данных. Формально задача автоматического извлечения знаний из баз данных может быть описана следующим образом. Пусть предметная база данных представляется в виде реляционной модели данных, которая описывается отношением R, являющимся подмножеством кортежей декартового произведения

$$R(DX_1, \ldots, DX_n, DY_1, \ldots, DY_m) = \{<x_1,\ldots,x_n,y_1,\ldots,y_m> | \; x_i \in DX_i,\ y_j \in DY_j,\ i=1..n,\ j=1..m \wedge \; P(x_1,\ldots,x_n,y_1,\ldots,y_m)\}, \quad (1)$$

где x_i –значения входных атрибутов X_i из домена DX_i; y_i –значения выходных атрибутов Y_i из домена DY_i; $P(x_1,\ldots,x_n,y_1,\ldots,y_m)$– предикат, описывающий условия отображения конкретной предметной области в кортежи значений атрибутов $<x_1,\ldots,x_n,y_1,\ldots,y_m>$.

Необходимо сформировать отображение в виде набора правил

$$\{X_1,X_2,..,X_n\} \Rightarrow \{Y_1,Y_2,\ldots,Y_m\}, \quad (2)$$

ставящих каждому входному набору значений $\{x_i=DX_i,\ i=1..n\}$ в соответствие некоторый набор целевых значений $\{y_j=DY_j,\ j=1..m\}$.

Полученные функциональные зависимости

$$Y_j = F_j(X_1,X_2,\ldots,X_n), \qquad j=1..m$$

должны быть верны для кортежей отношения (1) и могут быть использованы при нахождении выходных атрибутов Y_j для новых значений входных атрибутов X_i $(i=1..n)$.

Решение поставленной задачи основывается на применении методов Data Mining [1]. Эти методы позволяют в базах данных выявить и сформировать причинно-следственные зависимости в условиях неполной и разнородной информации. Индуцированные знания из баз данных позволяют решать задачи

прогнозирования, конструирования функциональных зависимостей, диагностики и управления. Несмотря на обилие методов Data Mining, приоритет постепенно смещается в сторону логических алгоритмов поиска в данных if-then-правил [1]. Результаты таких алгоритмов эффективны и легко интерпретируются.

Цель данной статьи – провести анализ эффективности алгоритма C4.5 автоматического извлечений знаний на примере реальной реляционной базы данных в области медицины.

Алгоритм C4.5 автоматического извлечения знаний. Пусть задано множество примеров T (таблица), где каждый элемент (строка) этого множества описывается m условными атрибутами. Пусть целевой атрибут класса принимает следующие значения C_1, C_2 … C_k.

Задача заключаться в построении иерархической классификационной модели в виде дерева из множества примеров T. Процесс построения дерева происходит сверху вниз. Сначала создается корень дерева, затем потомки корня и т.д.

Рассмотрим подробнее критерий выбора условного атрибута, по которому целесообразно выполнить ветвление. Очевидно, что в нашем распоряжении m (по числу атрибутов) возможных вариантов, из которых мы должны выбрать самый подходящий.

Обозначим через $freq(C_j, S)$ – количество примеров из некоторого множества S, относящихся к одному и тому же классу Cj.

Тогда выражение

$$Info(T) = -\sum_{j=1}^{k} \frac{freq(C_j, T)}{|T|} * \log_2\left(\frac{freq(C_j, T)}{|T|}\right)$$

даёт оценку среднего количества информации, необходимого для определения класса примера (строки) из множества T. В терминологии теории информации это выражение называется энтропией множества T.

Ту же оценку, но только уже после разбиения множества T по условному атрибуту X, дает следующее выражение ($|\,T\,|$ - мощность множества):

$$Info_x(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} * Info(T_i)$$

Тогда критерий для выбора атрибута разбиения вычисляется по следующей формуле:

$$Gain(x) = Info(T) - Info_x(T)$$

Критерий Gain вычисляется для всех атрибутов. Для разбиения выбирается тот атрибут X, который имеет максимальное значение. Этот атрибут (графа таблицы) будет основой для проверки в текущем узле дерева. По этому атрибуту проводится сортировка таблицы. Затем в узле делается оценка результатов сортировки и дальнейшее построение дерева проводиться в зависимости от полученного результата классификации.

Такие же рассуждения применяются к получаемым подмножествам T_1, T_2 … T_n , если они не монотонны. Процесс построения дерева продолжается рекурсивно до тех пор, пока в узле не окажутся примеры из одного класса.

Результаты индуцирования знаний из медицинской БД.

Для анализа работы алгоритма C4.5 рассмотрим таблицу с данными о клинико-психологических особенностях больных алкоголизмом на начальном этапе формирования ремиссии. Данная таблица содержит 158 строк.

Таблица 1 – Данные о клинико-психологических особенностях больных алкоголизмом на начальном этапе формирования ремиссии

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	Группа
2	1	2	2	2	1	1	2	1	1	2	1	2	2	2	1
2	1	2	2	1	1	2	1	2	2	2	1	2	1	2	1
3	1	3	2	2	3	2	3	1	3	2	3	2	3	2	1
2	1	2	3	2	1	2	1	1	1	1	1	2	1	2	1
……..……..															
1	1	1	1	1	1	2	1	2	1	1	1	2	1	2	2
1	1	1	1	2	1	1	1	1	1	1	1	1	1	2	2

Расшифровка атрибутов таблицы:

a1 – установка на трезвость;
a2 – спонтанные ремиссии в прошлом;
a3 – влечение к алкоголю;
a4 – тревога;
a5 – внутреннее напряжение;
a6 – снижение настроения;
a7 – дисфория;
a8 – апатия;
a9 – эйфория;
a10 – дистимия;
a11 – астенические расстройства;
a12 – неврозоподобные расстройства;
a13 – психопадоподобные расстройства;
a14 – психоорганические нарушения;
a15 – критика к болезнию.

Расшифровка значений количественных атрибутов:

- для атрибута a1: 1 – нет, 2 – продолжительностью до 6 мес., 3 – до 1 года, 4 – более 2 лет;
- для атрибута a2: 1 – нет, 2 – эпизодическое, 3 – постоянное;
- для остальных атрибутов: 1 – нет, 2 – слабо-умеренно-выраженная, 3 – выраженная.

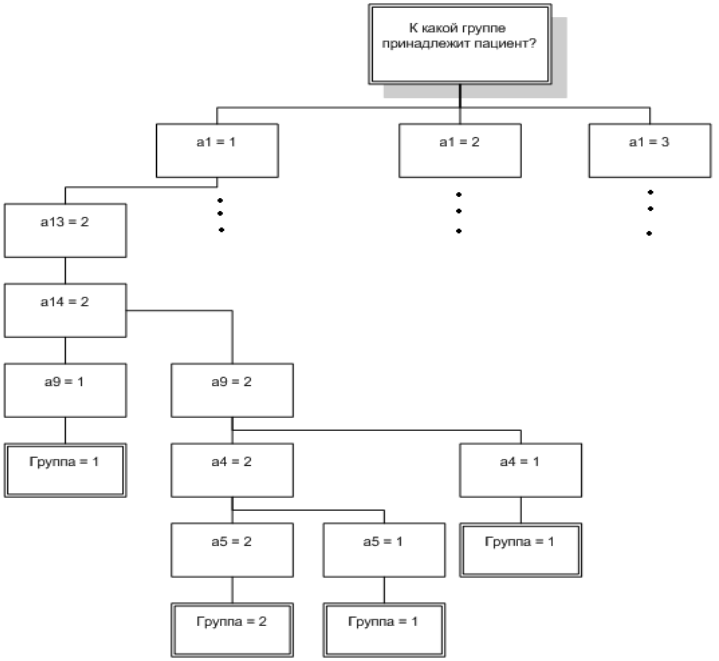

Рисунок 1 – Фрагмент построенного дерева решений

Извлечение знаний позволяют ответить на вопрос: “К какой группе принадлежит диагностируемый пациент?”.

После обработки табл.1 алгоритмом C4.5 было получено дерево решений (рис.1), из которого извлечены продукционные правила (табл.2).

Таблица 2 –Извлечённые знания в виде продукционных правил

№	Продукционное правило
1	**If** Установка на трезвость = Продолжительностью до 6 мес., Психопаиоподобные расстройства = Слабо-умеренно-выраженные, Психоорганические нарушения = Слабо-умеренно-выраженная, Эйфория = Нет **Then** Группа = 1

2	**If** Установка на трезвость = Продолжительностью до 6 мес., Психопаиоподобные расстройства = Слабо-умеренно-выраженные, Психоорганические нарушения = Слабо-умеренно-выраженная, Эйфория = Слабо-умеренно-выраженная, Тревога = Слабо-умеренно-выраженное, Внутреннее напряжение = Слабо-умеренно-выраженное **Then** Группа = 2
3	**If** Установка на трезвость = Продолжительностью до 6 мес., Психопаиоподобные расстройства = Слабо-умеренно-выраженные, Психоорганические нарушения = Слабо-умеренно-выраженная, Эйфория = Слабо-умеренно-выраженная, Тревога = Слабо-умеренно-выраженное, Внутреннее напряжение = Нет **Then** Группа = 1
4	**If** Установка на трезвость = Продолжительностью до 6 мес., Психопаиоподобные расстройства = Слабо-умеренно-выраженные, Психоорганические нарушения = Слабо-умеренно-выраженная, Эйфория = Слабо-умеренно-выраженная, Тревога = Нет **Then** Группа = 1
5	**If** Установка на трезвость = Продолжительностью до 6 мес., Психопаиоподобные расстройства = Слабо-умеренно-выраженные, Психоорганические нарушения = Нет, Эйфория = Нет, Влечение к алкоголю = Слабо-умеренно-выраженная, Дисфория = Нет **Then** Группа = 2
	……
63	**If** Установка на трезвость = Нет, Эйфория = Слабо-умеренно-выраженная, Спонтанные ремиссии в прошлом = Постоянное **Then** Группа = 2

Выводы. Проведен анализ методики извлечения знаний из медицинской реляционной базы данных с помощью алгоритма C4.5. Он автоматически строит дерево решений и формирует знания в виде продукционных правил. Результаты показали, что количество полученных правил более чем в два раза меньше числа записей в таблице и полностью отражают её реляционные отношения. Поэтому алгоритм C4.5 может использоваться в экспертных системах для пополнения её базы знаний из различных информационных источников данных.

Список литературы

1. Дюк В., Самойленко А. Data Mining: учебный курс.– СПб: Питер, 2001. – 368 с.