УДК 004.8

КВАНТОВЫЕ МОДЕЛИ И МЕТОДЫ ИНЖЕНЕРИИ ЗНАНИЙ В ЗАДАЧАХ DATA MINING

И.В. Шевченко

НАУ им. Н.Е. Жуковского «ХАИ»,
г. Харьков, Украина, ilona.shevchenko@gmail.com

Приведена постановка задачи интеллектуального анализа данных (Data Mining). Выделены классы методов Data Mining, а также классификация решаемых с их помощью задач. Описаны особенности квантовых моделей и методов представления знаний. Обоснована возможность применения методов инженерии квантов знаний как средств решения задач Data Mining.

DATA MINING, КЛАССИФИКАЦИЯ, ПРОГНОЗИРОВАНИЕ, КВАНТОВАЯ ИНЖЕНЕРИЯ ЗНАНИЙ

Введение

Data Mining представляет собой технологию, широко используемую в современном бизнес-анализе и способную при грамотном применении принести предприятию немалую прибыль.

Технология Data Mining – это технология углубленного анализа данных, которая предназначена для выявления в данных неочевидных, объективных и в то же время практически полезных закономерностей.

Data Mining – это комбинация различных подходов (статистики и искусственного интеллекта), которая дает принципиально новые свойства: возможность выявлять неочевидные закономерности.

Целью работы является обоснование возможности применения методов квантовой инженерии знаний при решении задач Data Mining.

Для достижения поставленной цели необходимо решить следующие задачи: 1) сформулировать особенность технологии Data Mining; 2) дать классификацию задач Data Mining; 3) рассмотреть классификацию методов Data Mining, рассмотреть их достоинства и недостатки; 4) сформулировать достоинства квантовой модели и методов инженерии знаний; 5) рассмотреть классы задач, решаемых с помощью квантовой инженерии знаний; 6) обосновать возможность применения квантовой методологии в решении задач Data Mining.

1. Особенности технологии Data Mining

Термин Data Mining получил свое название из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей.

Термин Data Mining часто переводится как «добыча», «раскопка», «промывание» данных, «извлечение зерен знаний из гор данных», интеллектуальный анализ данных, средства поиска закономерностей, извлечение (обнаружение) знаний, анализ шаблонов, информационная проходка данных [1-4]. Понятие «обнаружение знаний в базах данных» (Knowledge Discovery in Databases, KDD) можно считать синонимом Data Mining. Возникновение всех указанных терминов связано с новым витком в развитии средств и методов обработки данных.

В связи с совершенствованием технологий записи и хранения данных на людей обрушились колоссальные потоки информационной руды в самых различных областях. Деятельность любого предприятия (коммерческого, производственного, медицинского, научного и т.д.) стала сопровождаться регистрацией и записью всех подробностей его деятельности. Встал вопрос, а что делать с этой «свалкой» информации, ведь «сырые» данные могут содержать глубинный пласт знаний (совокупность фактов, закономерностей и эвристических правил, с помощью которых решаются поставленные задачи) и при грамотной его раскопке могут быть обнаружены настоящие самородки знаний. Ответ заключался в необходимости продуктивной переработки потоков «сырых» данных.

Итак, задачу анализа данных можно сформулировать так. Дана упорядоченная совокупность данных, представленная, например, в виде хранилища данных, требуется внутри этих данных найти скрытые закономерности, то есть сформулировать новые знания.

Однако методы математической статистики, базирующиеся на концепции усреднения по выборке, приводят к операциям над фиктивными величинами и оказались эффективными только для проверки заранее сформулированных гипотез и для «грубого» разведочного анализа [3, 4].

В основу современной технологии Data Mining положена концепция шаблонов, отражающих фрагменты многоаспектных взаимоотношений в данных. Шаблоны представляют собой закономерности, свойственные подвыборкам данных, которые могут быть компактно выражены в понятной человеку форме.

Информация, найденная в процессе применения методов Data Mining, должна быть нетривиальной и ранее неизвестной, например, средние продажи не являются таковыми. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других и т.д. Найденные знания должны быть применимы и на новых данных с некоторой степенью достоверности. Полезность заключается в том, что эти знания могут приносить определенную выгоду при их применении.

Важное положение Data Mining – нетривиальность разыскиваемых шаблонов, то есть найденные шаблоны должны отражать неочевидные, неожиданные закономерности в данных, составляющие так называемые скрытые знания.

Выделяют пять стандартных типов закономерностей, которые позволяют выявлять методы Data Mining: ассоциация, последовательность, классификация, кластеризация, прогнозирование [1-4].

При поиске ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных. Задача отыскания последовательности отличается тем, что устанавливаются закономерности между событиями, связанными во времени. В задаче классификации обнаруживаются признаки, которые характеризуют группы (классы) объектов исследуемого набора данных; по этим признакам можно отнести новый объект к тому или иному классу. Особенность кластеризации заключается в том, что классы объектов изначально не известны. В задаче прогнозирования на основе исторических данных оцениваются пропущенные или будущие значения целевых показателей.

Применение Data Mining оправданно при наличии достаточно большого количества данных, в идеале – содержащихся в корректно спроектированном хранилище данных (собственно, сами хранилища данных обычно создаются для решения задач анализа и прогнозирования, связанных с поддержкой принятия решений).

Data Mining является развитием нескольких научных направлений: классической статистики, искусственного интеллекта и машинного обучения. Сейчас Data Mining – это множество как классических, так и современных математических и статистических методов, которые развивались в этих направлениях.

Различают две группы методов Data Mining: статистические и кибернетические.

Статистические методы включают:

– дескриптивный анализ и описание исходных данных;

– анализ связей (корреляционный, регрессионный, факторный и дисперсионный анализы);

– многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.);

– анализ временных рядов (динамические модели и прогнозирование).

Кибернетические методы включают:

– нейронные сети;

– эволюционное программирование;

– генетические алгоритмы;

– ассоциативная память;

– нечеткая логика;

– деревья решений.

Обратим свое внимание именно на деревья решений, которые являются одним из наиболее популярных подходов к решению задач Data Mining. Они создают иерархическую структуру классифицирующих правил типа “ЕСЛИ ... ТО ...”, имеющую вид дерева. Популярность подхода связана как бы с наглядностью и понятностью. Но деревья решений принципиально не способны находить “лучшие” (наиболее полные и точные) правила в данных. Они реализуют наивный принцип последовательного просмотра признаков и “цепляют” фактически осколки настоящих закономерностей, создавая лишь иллюзию логического вывода. Вместе с тем, большинство коммерческих систем Data Mining используют именно этот метод.

Рассмотрим более подробно наиболее популярные задачи Data Mining.

Классификация является наиболее простой и одновременно наиболее часто решаемой задачей Data Mining. Задачей классификации является определение закономерностей, позволяющих делать вывод относительно определения характеристик конкретной группы объектов.

Для проведения классификации с помощью математических методов необходимо иметь формальное описание объекта, которым можно оперировать, используя математический аппарат классификации. Таким описанием может выступать база данных. Каждый объект несет информацию о некотором свойстве объекта.

Набор исходных данных (или выборку данных) разбивают на два множества: обучающее и тестовое. Обучающее множество – множество, которое включает данные, использующиеся для обучения (конструирования) модели. Такое множество содержит входные и выходные (целевые) значения примеров. Выходные значения предназначены для обучения модели. Тестовое множество также содержит входные и выходные значения примеров. Однако здесь выходные значения используются для проверки работоспособности построенной классификационной модели.

Прогнозирование направлено на определение тенденций динамики конкретного объекта или события на основе ретроспективных данных, то есть анализа его состояния в прошлом и настоящем. Таким образом, решение задачи прогнозирования также требует некоторой обучающей выборки данных.

Прогнозирование – установление функциональной зависимости между зависимыми и независимыми переменными.

Прогнозирование является распространенной и востребованной задачей во многих областях человеческой деятельности. В результате прогнозирования уменьшается риск принятия неверных, необоснованных или субъективных решений.

Различие задач классификации и прогнозирования состоит в том, что в первой задаче предсказывается класс зависимой переменной, а во второй – числовые значения зависимой переменной, пропущенные или неизвестные (относящиеся к будущему).

2. Особенности квантовой инженерии знаний

Квантовый метод инженерии знаний (метод алгоритмических квантов знаний – РАКЗ-метод) [5-7], предложенный проф. Сироджей И.Б. и развитый в работах его учеников [8-10], обладает рядом преимуществ по сравнению с известными ориентированными на знания методами принятия решений:

– РАКЗ-метод представляет собой индуктивный метод представления и вывода знаний с целью моделирования процесса принятия решений;

– РАКЗ-метод обеспечивает строгую формализацию порций знаний как содержательных алгоритмических структур данных (квантов), машинное манипулирование ими средствами алгебр конечных предикатов и векторно-матричных операторов;

– важным преимуществом РАКЗ-метода является обеспечение экстраполяционных свойств квантовых моделей принятия решений на основе принципа внешнего дополнения, что позволяет решить центральную проблему всех индуктивных методов, состоящую в адекватном соотнесении сложности модели с объёмом обучающей выборки.

В данном методе база квантов знаний (*k*-знаний) строится по обучающей выборке – таблице эмпирических данных (ТЭД) – и представляет собой совокупность импликативных и функциональных закономерностей. При этом база квантов знаний (БкЗ) является одновременно и механизмом принятия идентификационных (классификационных), а также прогнозных решений.

Рассмотрим, что понимается под закономерностью в квантовом методе инженерии знаний.

Устойчивая связь между r характеристиками объекта принятия решений (ОПР) из общего числа n $(r \le n)$, выражающая недопустимость хотя бы одной комбинации их значений на множестве *k*-знаний, называется импликативной закономерностью или запретом *r*-го ранга.

Функциональной закономерностью *r*-го ранга на множестве *k*-знаний называется устойчивая связь между r $(r < n)$ признаками ОПР и некоторым $(r+1)$-м признаком, позволяющая по значениям признаков-аргументов, однозначно определить значение признака-функции.

Приведем пример гипотетической обучающей выборки, которая представлена матричным точным (*t*-) квантом знаний 2-го уровня $tk_2\Sigma_{\text{обуч}}$, а каждый объект обучающей выборки (строка в $tk_2\Sigma_{\text{обуч}}$) описывается *t*-квантом знаний 1-го уровня:

$$tk_2\Sigma_{\text{обуч}} = \begin{bmatrix} \underbrace{10}_{x_1}:\underbrace{010}_{x_2}:\underbrace{10}_{x_3}:\underbrace{100}_{x_4=x_{\text{ц}}} \\ 10:100:01:\ 010 \\ 01:001:10:\ 001 \\ 10:001:10:\ 010 \end{bmatrix}.$$

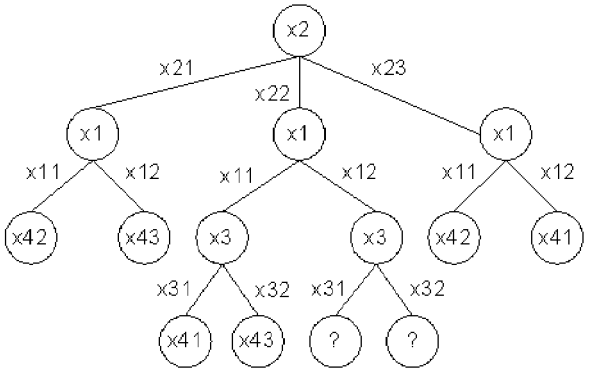
Видим, что ОПР в данной обучающей выборке описывается с помощью 4 признаков. Признак x_1 может принимать 2 значения (x_{11} и x_{12}), признак x_2 – 3 значения (x_{21}, x_{22} и x_{23}), признак x_3 – 2 значения (x_{31} и x_{32}). Признак x_4 является целевым ($x_{\text{ц}}$), то есть признаком, по которому объекты из $tk_2\Sigma_{\text{обуч}}$ относятся к одному из трех классов x_{41}, x_{42} и x_{43}.

Построим, согласно методу точных квантов знаний, импликативную (запретную) базу t-квантов знаний, соответствующую приведенной обучающей выборке:

$$tk_2\Sigma_{\text{запр}} = \begin{bmatrix} \underbrace{01}_{x_1}:\underbrace{010}_{x_2}:\underbrace{--}_{x_3}:\underbrace{---}_{x_4=x_{\text{ц}}} \\ 01:---:--:\ 010 \\ --:001:01:--- \\ --:100:10:--- \\ --:001:--:\ 100 \\ --:010:--:\ 010 \\ --:100:--:\ 001 \\ 10:001:--:\ 001 \\ 10:100:--:\ 100 \\ 01:---:01:\ 001 \\ 01:---:10:\ 100 \\ 10:---:01:\ 100 \\ 10:---:10:\ 001 \\ --:010:01:\ 100 \\ --:010:10:\ 001 \\ \end{bmatrix}.$$

Приведем пример семантического значения каждой строки кванта $tk_2\Sigma_{\text{запр}}$. Так, например, первая строка запретного кванта знаний читается так: «не существует объекта, который бы имел следующую комбинацию признаков: x_1 принимает значение x_{12}, а признак x_2 – значение x_{22}».

На основе выявленных запретных закономерностей можно построить дерево принятия решений (см. рис.). Главным достоинством деревьев в качестве решающих правил является то, что для вывода решения (определения значения целевого признака, в данном случае $x_{\text{ц}} = x_4$) не требуется знать значения всех признаков ОПР.



Дерево принятия решений

В квантовом методе инженерии знаний задача соотнесения сложности квантовой модели объекта принятия решений (ОПР) с объёмом обучающей выборки сведена к обеспечению разработчиков знаниеориентированных систем расчетной методикой для формирования и оценивания обучающей выборки с целью гарантированного индуктивного построения базы квантов знаний как системы закономерностей предметной области требуемого качества. Для этого формулируется индуктивный принцип внешнего дополнения при оценке и поиске импликативных/функциональных закономерностей по обучающим *k*-знаниям на основе формулировки и доказательства соответствующих теорем [5]. Заметим следующее, чтобы принять или отвергнуть гипотезу, нужно оценить её достоверность, то есть вероятность того, что гипотеза ошибочна. Достоверными принято считать гипотезы, вероятность ошибочности которых настолько мала, что ею можно пренебречь. Таким образом, чем меньше вероятность того, что гипотеза о существовании закономерностей между *r* признаками ОПР ошибочна, тем больше вероятность того, что данная связь существует.

Автором в работе [8] приведен вывод уточненных формул для вычисления оценки вероятности гипотезы об отсутствии закономерности между признаками, что позволило повысить адекватность индуктивно синтезируемой базы *k*-знаний объему обучающих знаний. Данные формулы используются для расчета максимального ранга (числа переменных, участвующих в связи) искомых в обучающей выборке закономерностей, путем задания порогового значения для оценки вероятности отсутствия закономерностей между признаками.

В РАКЗ-методе, для того чтобы принять классификационные решения по обучающей выборке, необходимо синтезировать классификационную БкЗ, тогда как при необходимости принимать прогнозные решения – прогнозную БкЗ.

Выводы

В статье были рассмотрены задачи интеллектуального анализа данных, среди которых особое внимание было уделено задаче классификации и прогнозирования. Также был рассмотрен квантовый метод инженерии знаний, особенность которого заключается в формализации знаний порциями (квантами). В методе квантов знаний база знаний строится как совокупность найденных в обучающей выборке импликативных и функциональных закономерностей. По синтезированной базе квантов знаний могут быть приняты классификационные и прогнозные решения. Таким образом, можно сделать вывод о возможности использования моделей и методов инженерии квантов знаний для эффективного решения задач Data Mining.

Список литературы: 1. *Дюк В.А., Самойленко А.П.* Data Mining: учебный курс. – СПб.: Питер, 2001. – 368 с. 2. *Киселев М., Соломатин Е.* Средства добычи знаний в бизнесе и финансах // Открытые системы. – 1997. – №4. – С. 41-44. 3. *Чубуков И.А.* Data Mining. Учебное пособие. – М.: Интернет-Университет Информационных технологий; Бином. Лаборатория знаний, 2006. – 382 с. 4. Технологии анализа данных: Data Mining, Visual Mining,Text Miining, OLAP / Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. – СПб.: БХВ-Петербург, 2007. – 384 с. 5. *Сироджа И.Б.* Квантовые модели и методы искусственного интеллекта для принятия решений и управления. – К.: Наук. думка, 2002. – 427 с. 6. *Сироджа И.Б. Петренко Т.Ю.* Метод разноуровневых алгоритмических квантов знаний для принятия производственных решений при недостатке или нечёткости данных. – К.: Наук. думка, 2000. – 247 с. 7. *Сироджа И.Б.* Квантовые модели и методы инженерии знаний в задачах искусственного интеллекта // Искусственный интеллект. – 2002. – №3. – С.161–171. 8. *Варфоломеева И.В.* Развитие принципа внешнего дополнения РАКЗ-моделей знаниеориентированного принятия решений // Проблемы бионики: Всеукр. межвед. науч.-техн. сб. – Харків: Харківський національний університет радіоелектроніки, 2004. – Вып. 60. – С. 48–57. 9. *Варфоломеева И.В.* Проблема информативности системы признаков при построении идентификационной квантовой базы знаний // Открытые информационные и компьютерные интегрированные технологии: Сб. науч. трудов. – Х.: Нац. аэрокосмический ун-т «ХАИ», 2004. – Вып. 24. – С. 299 – 304. 10. *Варфоломеева И.В.* Квантовый метод принятия решений на основе запретной логической сети // Радиоэлектроника и информатика. – 2004. – №4. – С. 93–99.

Поступила в редколлегию 16.10.2007