Использование подходов Data Mining в развитии систем ЛИМС

Г.Т. Маракаева (МГУ им. М.В. Ломоносова)

Информационные технологии необходимы, прежде всего, там, где имеется большой объем ручной рутинной работы. В лабораториях предприятий производится огромное число испытаний, результаты которых обрабатываются и записываются в журналы. Полная или частичная автоматизация лабораторий обеспечивается системами ЛИМС (Лабораторные Информационные Системы). В функции ЛИМС входят разнообразные задачи, автоматизирующие процессы лаборатории, начиная от планирования работы лаборатории и регистрации проб в системе до формирования паспорта продукции, включая все промежуточные задачи [1]. Кроме того, часто эти системы интегрируются в общее информационное пространство предприятия. Со временем эксплуатации системы накапливается большой объем данных, к которым могут применяться алгоритмы исследования данных и поиск скрытых закономерностей Data Mining [2, 3]. В статье рассмотрена одна из задач Data Mining - классификация химических проб на основе обучающей выборки, то есть определение класса неизвестной пробы по значениям некоторых атрибутов.

Необходимость в решении такой задачи была сформулирована заводской лабораторией, занимающейся анализом химических проб. Наряду с анализом проб на предмет соответствия требованиям качества, в лабораторию поступали и неизвестные пробы для определения принадлежности к тому или иному классу. При этом в большинстве случаев работники лаборатории вспоминали, что какое-то время назад им приходилось анализировать похожие пробы, но точного результата они не могли вспомнить.

Работа лаборатории заключается в проведении экспериментов над различными пробами по определению их свойств (например, содержание серы, вязкость и т.д.), а также по определению класса пробы (например, питьевая вода, природная вода и т.д.). Каждый полученный результат фиксируется в лабораторном журнале. Запись о пробе в системе состоит из значений ее параметров (атрибутов пробы) и класса пробы, а ключом к этой записи является идентификационный номер. Со временем работы системы ЛИМС в лаборатории накапливается достаточно большое число записей о пробах, содержащие значения параметров и класс. Эта информация может быть использована как экспертные знания для классификации проб, для которых неизвестны классы, а заданы лишь некоторые значения атрибутов. Такая задача, например, может возникнуть, когда лаборатория занимается исследованием неизвестных проб, а для определения класса пробы требуется серия дорогостоящих экспериментов. Классификатор позволяет определить класс пробы по неполному набору параметров, то есть провести лишь некоторые эксперименты, и на основе исторической экспертной информации будет сделан вывод о принадлежности пробы.

Итак, задача классификации состоит в определении класса новой пробы по неполному набору значений атрибутов. В лабораторию поступает новая проба неизвестного класса. В общем случае сотрудник должен по внешним факторам экспертно определить возможный класс и выполнить полный набор экспериментов, призванных подтвердить или опровергнуть его гипотезу. Если гипотеза не подтвердится, то проводится следующая серия экспериментов для исследования другой гипотезы.

В общем случае очень часто классификация рассматривается как один из способов выявление закономерностей в объеме данных (Data Mining) [2]. Существует достаточно много различных алгоритмов классификации. Каждый из них может давать хорошие результаты для одного класса задач и плохие для другого. Например, работа классификатора на основе нейронной сети будет давать хороший результат, если имеется достаточно большое число данных, на которых проводится обучение, и каждый класс представлен примерно одинаковым числом тестовых объектов. Если же во входных данных находится очень много объектов одного класса, то результат работы сети будет часто склоняться именно к этому классу, независимо от значимости остальных объектов [5].

Рис. 1. Пример функциональных модулей ЛИМС и возможностей интеграции с другими информационными системами предприятия

Некоторые алгоритмы классификации используются в пакетных решениях таких, как системы SAS, SPSS, STATISTICA и др. [3, 6, 7]. Все пакеты можно разделить на три категории: профессиональные, универсальные и специализированные. Профессиональные пакеты рассчитаны на то, что пользователем является специалист по статистике. Все универсальные пакеты имеют много пересечений по составу статистических процедур. Специализированные пакеты ориентированы на одну или несколько смежных предметных областей, их алгоритмы "заточены" под определенные данные, а также под наиболее актуальные задачи предметной области.

Рассмотрим алгоритм, ориентированный на работу в специализированных системах класса ЛИМС. На рис. 1 приведен пример функциональных модулей ЛИМС и возможности его интеграции с другими системами.

Рынок лабораторных систем в России только формируется, и список российских предприятий, где используется ЛИМС, на сегодняшний день достаточно короткий. Но, ориентируясь на западных коллег и принимая во внимание стремление повышения качества продукции за счет автоматизации процессов, можно однозначно говорить о том, что в ближайшие годы многие лаборатории задумаются о внедрении системы. Как правило, в функционал ЛИМС входят стандартные действия по автоматизации рабочего места лаборанта, начальника лаборатории и составлению отчетов. В то же время более сложные аналитические задачи и задачи по выявлению закономерностей и классификации в стандартных модулях ЛИМС не решены. Однако спустя некоторое время после начала повсеместной эксплуатация ЛИМС будет накоплено достаточно информации для построения классификатора.

При выборе подхода к классификации химических проб была рассмотрена специфика предметной области и выбран регрессионный метод классификации на основе разделяющей функции [4, 5]. Затем на основе предложенного метода был разработан алгоритм, учитывающий особенности поставленной прикладной задачи.

Прежде, чем перейти к описанию алгоритма, определим некоторые термины.

Признаком назовем пару $x = <n, T>$, где n – имя признака, T – множество значений для признака x, которое определяет тип значений признака: целый или плавающий, перечисление или диапазон, признаки с булевским множеством значений.

Количественными признаками назовем признаки, значения которых заданы целыми или вещественными числами. *Качественными признаками* назовем признаки, на которых не задано отношение порядка. Например, признак "цвет" является качественным признаком, если множество его значений описывается названиями цветов. Если же цвет определяется длиной волны, то этот признак будет количественным.

Будем считать, что признаки независимы и из общего числа признаков N в признаковом пространстве: N_1 – число количественных признаков, N_2 – число качественных признаков.

Пространство количественных признаков представляет собой декартово произведение множеств значений признаков T_i, $i = 1, ..., N_1 : \tau = T_1 \times ... \times T_{N_1}$.

Объект $\omega = <имя, (x_1, ..., x_N)>$, где $(x_1, ..., x_N)$ – набор значений признаков. Объект представляет собой точку в признаковом пространстве. Множество всех объектов W является конечным или счетным, и все объекты могут быть занумерованы.

Пусть на множестве объектов W определено m классов $\Omega_1,...,\Omega_m$. Для каждого класса Ω существует характеристическая функция:

$$\chi(\omega) = \begin{cases} 1, \omega \in \Omega, \\ 0, \omega \notin \Omega. \end{cases} \quad (1)$$

Таким образом, класс представляет собой пару $\Omega = <имя, \chi>$, где χ – характеристическая функция, определенная формулой (1). Если для всех $j = 1,..,m$ заданы характеристические функции χ_j, то для каждого объекта $\omega \in W$ можно определить его класс Ω_j.

Обучающая выборка – это множество объектов $V = \{\omega_1,...,\omega_{l_1}, \omega_{l_1+1},..., \omega_{l_2},..., \omega_{l_{m-1}},...,\omega_{l_m}\}$, для которых известны классы, к которым они принадлежат: $\omega_1,...,\omega_{l_1} \in \Omega_1$, $\omega_{l_1+1},..., \omega_{l_2} \in \Omega_2,..., \omega_{l_{m-1}},..., \omega_{l_m} \in \Omega_m$.

Помимо характеристических функций χ_j будем рассматривать разделяющие функции $f_j(x,V)$, про которые известно, что на объектах обучающей выборки $x \in V \Rightarrow f_j(x,V)=1$, если $x \in \Omega_j$, $f_j(x,V) = 0$, если $x \notin \Omega_j$, и они каким-либо образом расширены на все признаковое пространство.

Если задана обучающая выборка, то по ней можно построить разделяющие функции.

Классификация – это сопоставление каждому объекту определенного класса, то есть определение множества пар <*объект*, *класс*>.

Тестовый объект – это объект, у которого заданы значения признаков, но неизвестен его класс. Входными данными для классификатора являются данные тестового объекта.

Чтобы с помощью разделяющих функций определять класс любого объекта признакового пространства, необходимо доопределить разделяющие функции на признаковом пространстве таким образом, чтобы они приближались к характеристическим функциям. Если разделяющая функция недостаточно приближена к характеристической функции, например, для заданного значения максимума невязки, то необходимо пополнить обучающую выборку новыми объектами и переобучить систему.

Качественные признаки используются для предварительной классификации тестового объекта, то есть с помощью качественных признаков ищется подмножество возможных классов этого объекта в пространстве классов. Имея иерархию классов в виде дерева, ребрами которого являются селекторы по качественным признакам (условия перехода на следующую вершину дерева), можно проводить анализ при-

надлежности тестового объекта к тому или иному классу только для выбранного поддерева.

Задача обучения системы ставится следующим образом: заданы множество имен классов и обучающая выборка. Требуется построить дерево классов и разделяющие функции для этих классов.

Точки признакового пространства по количественным признакам образуют множество всех объектов предметной области. В этом признаковом пространстве N_1-мерными параллелепипедами можно задать области, в которые будут входить объекты обучающей выборки, принадлежащие одному классу, и такие, что объединение всех параллелепипедов даст полное признаковое пространство. Задача разделяющих функций – описать границы параллелепипедов, другими словами задать границы классов для количественных признаков. Область признакового пространства, ограниченную таким параллелепипедом, будем называть *кластером*. Каждый класс может иметь несколько кластеров.

Работа системы разбивается на два основных этапа (рис. 2): обучение и непосредственно классификация.

На этапе обучения классификатора на вход поступают объекты обучающей выборки. По входной информации формируется дерево классов.

Далее для каждого класса поочередно строятся разделяющие функции, построение которых базируется на следующих принципах:

- на обучающей выборке значения характеристических и разделяющих функций совпадают;
- разделяющие функции строятся для каждого класса из условия минимизации функционала разности характеристических и разделяющих функций и имеют вид:

$$f(x,V)=\begin{cases}1, x\in[w_1^1,k_1-w_1^1],[k_j-w_1^j,k_{j+1}-w_1^{j+1}],\ j=3,5,...,g,\\ 0, x\in(k_j-w_1^j,k_{j+1}-w_1^{j+1}),\ j=2,4,...,g,\end{cases}$$

где x – тестовый объект, представленный в виде вектора, w_1^j – объекты обучающей выборки соответствующего класса (в данном случае класс 1), g – число кластеров для рассматриваемого класса.

Таким образом, с помощью полученных дерева и функций будет осуществляться классификация: на вход классификатору подается вектор значений признаков некоторого объекта неизвестного класса; по значениям качественных признаков определяется подмножество потенциальных классов; затем определяются значения разделяющих функций для полученного подмножества классов и в зависимости от этого либо однозначно определяется класс объекта, либо объекту присваивается класс по умолчанию Ω_0.

Описанный алгоритм применялся для классификации химических проб, где проба являлась объектом признакового пространства и описывалась некоторыми параметрами.

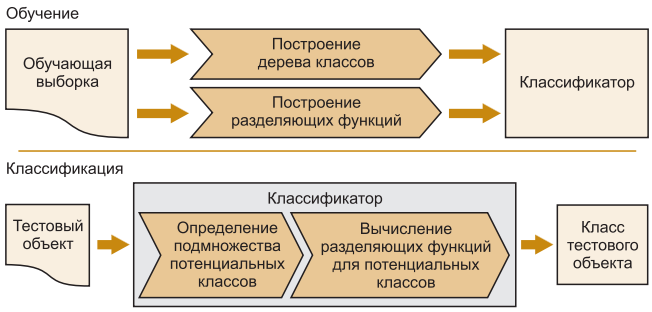

Рис. 2. Подход к классификации химических проб

Алгоритм реализован на платформе .Net, на языке C# с применением технологии XML.

Данные хранятся и обрабатываются в формате XML. Формат хранения позволяет конвертировать данные в формат любой реляционной БД. Таким образом, алгоритм классификации может быть интегрирован практически в любую систему ЛИМС. Отчетные документы формируются с использованием механизма генерации отчетов SharpShooter. Возможен импорт информации в MS Office.

Алгоритм дает хорошие результаты при достаточно большом наборе обучающей выборки. Непосредственно из подхода построения классификатора следует, что алгоритм будет более точным, если обучающая выборка будет содержать объекты с граничными допустимыми значениями признаков.

Важной особенностью предметной области является то, что данные, накопленные в одной лаборатории, могут быть использованы в другой лаборатории, так как описание проб с помощью признаков является унифицированным для всех экспертов. Таким образом, если классификатор получит широкое распространение, то возможно будет выполнять консолидацию данных из многих лабораторий, то есть создавать единую обучающую выборку. Это существенно повысит качество работы классификатора и расширит область его применения.

Список литературы

1. Лабораторные Информационные Системы. Сб. статей. Под редакцией Матвейко П.Е., ООО "Маркетинг. Информационные технологии". 2006.
2. *Mehmed Kantardzic.* Data Mining. Concepts, Models, Methods and Algorithms. Wiley-Interscience. 2003.
3. *Дюк В., Самойленко А.* Data Mining: учебный курс. СПб.: Питер. 2001.
4. *Горелик А.Л., Скрипкин В.А.* Методы распознавания. Издание третье. М.: Высшая школа. 1989.
5. *Маракаева Г.Т.* Применение методов выявления закономерностей для классификации химических соединений. Сб. статей ИСП РАН. 2006.
6. *Paolo Guidici.* Applied Data Mining: Statistical Methods for Business and Industry, Statistics in Practice Ser. Wiley. 2005.
7. *Georg Zangl.* Data Mining: Application to the Petroleum Industry. Round Oak. 2001.

Маракаева Галина Тимировна *– инженер кафедры Системного программирования факультета Вычислительной математики и кибернетики Московского государственного университета им. М.В. Ломоносова, консультант компании Accenture.*

Контактный телефон (495) 991-56-84. E-mail: galina.marakaeva@accenture.com