Интеллектуальные системы
2013. №1(35)

УДК 681.3.06(075.8)

© 2013 г. **Н.С. Безруков,** канд. техн. наук
(Дальневосточный научный центр физиологии и патологии дыхания, Благовещенск),
А.Д. Плутенко, д-р техн. наук
(Амурский государственный университет, Благовещенск)

ПОСТРОЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ СРЕДСТВАМИ СУБД И DATA MINING

Рассматривается вопрос построения интеллектуальных систем с помощью СУБД и Data Mining. Анализируются основные проблемы при использовании СУБД и Data Mining, и предлагаются способы их решения.
Ключевые слова: интеллектуальные системы, СУБД, SQL, Data Mining, нейронные сети, деревья решений.

Введение

С переходом от бумажного к цифровому хранению информации человек осуществил переход на новый этап качественного развития. Ранее, используя свои познания и опыт, он все более качественно использовал вещество (углеводороды), чтобы получить полезную работу, затем научился вырабатывать ее из окружающего мира (ветряные и водяные мельницы). Накапливая бумажную информацию, человек все более эффективно использовал окружающий мир, тогда как технология накопления информации качественно не менялась [1].

С появлением компьютерной техники процесс накопления и использования информации стал меняться качественно. Из-за большого объема информации, которую ручными средствами не обработать, возникла задача программного анализа данных, решаемая с помощью системы поддержки принятия решений (СППР). В общем случае СППР реализует две основные задачи: во-первых, запись/изменение базы данных (БД), их хранение и анализ, во-вторых, построение интеллектуальных систем (ИС), их хранение и последующее использование/изменение. Только решение каждой задачи позволяет в совокупности создать эффективную СППР, способную обрабатывать информацию в различных областях с целью получения качественно новых знаний. Так сложилось, что эвристический подход построения ИС закрепился за слабоформализованными областями человеческой деятельности (например, в медицине), где невозможно получить обоснованную модель, хотя сама модель способна давать приемлемое решение задачи в большинстве практически значимых случаев.

Задачу организации БД в СППР принято решать средствами системы

управления базами данных (СУБД) [2]. Тогда как задачу построения ИС принято реализовывать методами и алгоритмами Data Mining [3]. При совместном использовании СУБД и Data Mining возникают проблемы, требующие решений разработчика СППР. Эти решения должны быть оформлены так, чтобы у пользователя из сферы медицины не было необходимости стать специалистом в информатике и математическом моделировании, например, средствами ППП Matlab.

Построение обучающей выборки из БД

СУБД в общем случае позволяет решать задачи ввода, хранения и первичного анализа данных. Структура БД менялась вместе с ростом объема обрабатываемой информации. В начале 60-х гг. это были иерархические базы данных в виде деревьев, но такая структура не позволила описать многие объекты и на сегодня применяется только в специализированных областях. В начале 70-х гг. были предложены реляционные БД, получившие широкое распространение. В них данные представлены в виде таблиц с рядами и колонками. Ряд представляет собой набор значений, относящихся только к одному объекту, а колонка характеризует одну переменную для всех объектов. Доступ к данным в такой БД осуществляется через идентификаторы таблицы, колонки и ряды. Идентификаторами таблицы и колонки являются уникальные имена, а ряда – первичный ключ, в качестве которого в динамических процессах может выступать время, а в статических – уникальный параметр объекта (фамилия или шифр карты пациента в поликлинике). Формируя запросы на языке SQL к такой БД, пользователь получит таблицу, по которой затем сможет построить ИС средствами Data Mining. Однако такой подход может иметь ряд проблем для пользователя СППР, ведь СУБД создавалась как независимая среда со своими правилами функционирования, отличающимися от правил в СППР. Также пользователь может не знать языка запросов SQL, и для запроса ему необходим интуитивно понятный интерфейс.

В Matlab это реализуется с помощью пакета Database Toolbox, причем функция querybuilder позволяет напрямую строить запрос к источникам данных ODBC. На базе этой функции можно реализовать модуль запросов, результатом которого будет матрица, ее необходимо проверить на ошибки в модуле обработки. В матрице могут находиться неправильно введенные данные, и это не влияет на функционирование СУБД, тогда как построенная по таким данным ИС может иметь недопустимую ошибку. Если выборка небольшая, то специалист визуально способен определить неверные значения, тогда как для большой выборки следует разрабатывать фильтры проверки данных на логичность, – например, по сильному отклонению от среднего значения. Фильтр обработки отклонений должен либо выводить пользователю ошибочные данные в наглядной форме, либо автоматически обрабатывать их.

БД может содержать данные в разных форматах, тогда как средства Data Mining работают с числовыми переменными. Поэтому категориальные значения (маленький, средний, большой) следует кодировать числами. Фильтр строковых переменных может справиться с этим при условии, что в тексте нет ошибок.

В полученной из БД выборке могу быть переменные типа Null, что в СУБД воспринимается как неизвестное значение. Такие пустоты могут получиться в

связи с изменением контролируемой среды (пациентам проводили один анализ, а затем заменили его на другой, как итог – в БД две колонки, которые говорят об одном и том же) или отсутствует необходимость заполнять ряд в колонке (больному в клинике делают три анализа, а для подтверждения диагноза достаточно двух, тогда третья колонка может быть пустой, хотя ее можно восстановить по аппроксимации первых двух). При операциях над колонкой в СУБД средствами SQL (поиск среднего, максимального или минимального) такая переменная опускается. Тогда как в ИС подобное недопустимо, данные должны быть полными или избыточными, поскольку будущая модель должна отражать закономерности между данными, а они должны быть представлены в явном виде. Разработчику ИС в таком случае необходим фильтр пустых переменных, который может выбирать: либо заполнить пустоты специально разработанным алгоритмом (небольшой ИС) из других данных, либо отбросить ряды, где есть незаполненные колонки, но тогда данных может оказаться недостаточно для создания ИС.

Если данных недостаточно, то можно поступить, как показано в работе [4]. При построении ИС для n признаков необходим набор из k примеров в зависимости от количества настраиваемых коэффициентов системы, т.е. на практике сложно реализовать одну большую ИС со всеми возможными признаками на входе, так как это требует большого количества примеров. Поэтому предлагается создавать небольшие подбазы для подгруппы признаков. Признаки можно разбивать на подгруппы, к примеру, в зависимости от их метода получения, тогда на каждой подбазе можно строить отдельную ИС и тем самым получить множество параллельных подсистем. Средствами Matlab можно реализовать структуру (рис. 1), содержащую модуль запросов, обработки и деления данных.

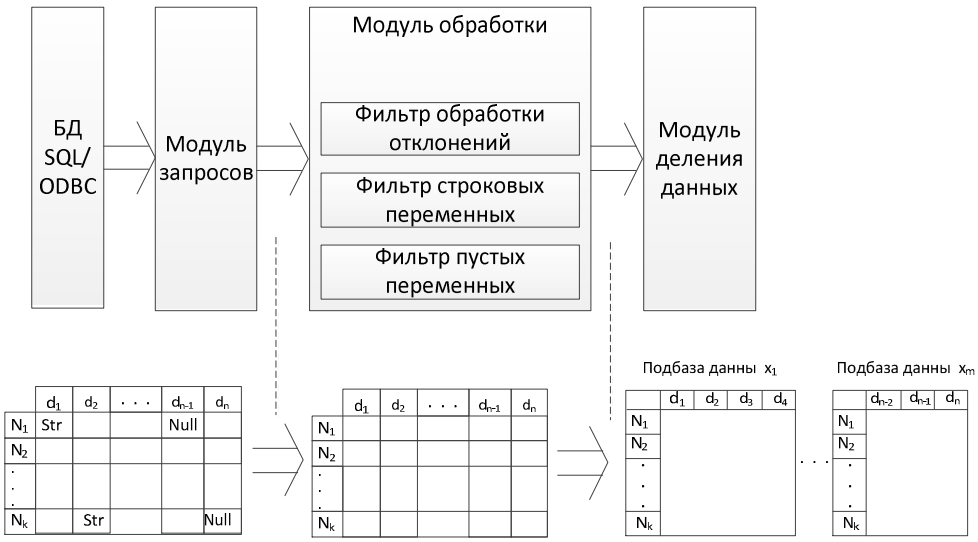

Рис.1. Структура обработки данных, полученных из БД.

Она позволит пользователю СППР сформировать из БД массив с данными, которые можно будет обработать средствами Data Mining.

Этапы построения интеллектуальной системы средствами Data Mining

Развитие компьютерных технологий привело к бурному росту количества собираемой информации, объем которой в Интернете удваивается каждые 2-3 го-

да. Поэтому в СППР для проведения автоматического анализа данных все чаще применяют инструментарий Data Mining [1].

Data Mining – это исследование и обнаружение в "сырых" данных скрытых закономерностей, ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации. Поскольку Data Mining развивался на стыке разных дисциплин (таких как математическая статистика, теория информации и теория баз данных), то большинство алгоритмов получилось с различной структурой, что затруднило их реализацию в рамках одной прикладной программы, или перенос из одной СППР в другую. Эту проблему может решить стандарт PMML (Predictive Model Markup Language), который описывает форму представления модели в виде XML-документа.

Любую ИС можно представить из этапов ее реализации и функционирования – жизненных этапов.

Первое, что должен решить разработчик ИС, – это выбрать тип модели в зависимости от решаемой задачи: классификация – отнесение объектов к заранее известному классу; регрессия – установление зависимости непрерывных выходных от входных переменных; ассоциация – выявление закономерностей между связанными событиями в виде правил «если... то…»; кластеризация – поиск независимых групп и их характеристик в данных (обучение без учителя).

Решение этой задачи помогает лучше понять данные, а в ряде случаев сократить их число (из множества переменных сформировать один класс, как показано в [4, 5]).

Также на выбор модели оказывает влияние количество данных в БД, их тип или необходимый вид полученной в итоги модели. Например, если взять нейронные сети и построить с их помощью модель, то для эксперта она будет неявна, он не сможет из нее выделить ассоциативные правила, объяснить правильность работы. Тогда как в дереве решений все представлено в явной и понятной для эксперта форме. Разработчик может взять несколько алгоритмов и построить по ним несколько моделей, а затем выбрать лучшую модель.

Построенную модель следует проверить и качественно оценить. При создании модели на нейронных сетях принято общие данные делить на обучающую и проверочную выборки и по ошибкам для данных выборок оценивать качество модели. С точки зрения статистики, точность такой модели возрастает с увеличением количества обучающих данных, однако на практике не всегда имеется возможность работать с большой БД. Поэтому в работе [6] для небольших баз был предложен способ деления выборки на обучающую и проверочную, что позволяет объективно оценивать результат работы созданной модели.

Следующим этапом идет эксплуатация ИС в рамках СППР, где с ней начинает работать пользователь. Здесь разработчику требуется навыки опытного программиста для создания дружественного интерфейса. После внедрения ИС работа разработчика с ней не заканчивается. Ее периодически необходимо проверять на новых данных, которые будут формироваться при эксплуатации ИС. Если она успешно применяется в данный момент, то это не гарантирует успех в будущем, поскольку среда, в которой работает ИС, может меняться как количественно, так и

качественно. И если в первом случае ИС достаточно переобучить, то при изменении качества приходится вводить новые параметры или удалять устаревшие, что приводит к созданию новой ИС.

Рис. 2. Этапы жизненного цикла ИС.

Matlab является эффективной инструментальной средой для реализации этапов жизненного цикла ИС (рис. 2). В нем уже реализованы: линейная регрессия, нейронные сети с учителем, нейронные сети без учителя, деревья решений и множество других. Однако в нем нет механизмов визуализации, значительно облегчающих использование полученной модели и интерпретацию результатов. Но разработчик всегда может создать средствами Matlab независимый код, который затем подключить к СППР сторонних производителей, при условии соблюдения стандарта создания систем Data Mining.

Заключение

Предложена структура, состоящая из модуля запросов, обработки и деления данных, позволяющая пользователю СППР сформировать из БД массив с данными, которые можно будет обработать средствами Data Mining.

Рассмотрены проблемы, возникающие при совместном использовании СУБД и Data Mining, и приведены способы их решения. Предложены этапы жизненного цикла ИС на основе Data Mining.

ЛИТЕРАТУРА

1. *Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.* Технологии анализа данных: Data Mining, VisualMining, TextMining, OLAP. – СПб.: БХВ-Петербург, 2007.
2. *Vincent Rainardi*. Building a Data Warehouse: With Examples in SQL Server, 2008.
3. *David Taniar.* Data Mining and Knowledge Discovery Technologies.
4. *Безруков Н.С., Еремин Е.Л.* Интеллектуальная система поддержки принятия решений для диагностики заболеваний на основе адаптивных нейро-нечетких сетей // Научно-техническая информация. – Сер. 1. Организация и методика информационной работы. – 2007. – №6. – С.30-34.
5. *Безруков Н.С., Еремин Е.Л.* Построение и моделирование адаптивной нейро-нечеткой системы в задаче медицинской диагностики // Информатика и системы управления. – 2005. – №2(10). – С. 36-46.
6. *Безруков Н.С., Колосова Е.В.* Способы региональной кластеризации по параметрам человеческого капитала на основе самообучающихся нейронных сетей // Информатика и системы управления. – 2008. – №1(15). – С. 96.102.

E-mail:
Безруков Николай Сергеевич – bezrukow@mail.ru;
Плутенко Андрей Долиевич – plutenko@bk.ru.