Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра Математического и компьютерного моделирования

Применение технологий OLAP и Data Mining для анализа медико-социальных данных

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы

направления 09.03.03 – Прикладная информатика

механико-математического факультета

Брагиной Софьи Михайловны

Научный руководитель:
доцент, к.ф.-м.н _____ О.М. Ромакина

Зав. кафедрой:
зав. кафедрой д.ф.-м.н., _____ Ю.А. Блинков

Саратов 2017 г.

ВВЕДЕНИЕ

Данная работа посвящена вопросу интеллектуального анализа данных, так же известного под названием Data Mining. Поскольку «добыча данных» – это набор инструментов анализа, используемых для выявления закономерностей, было решено сосредоточиться на проблеме построения деревьев решений.

Впервые термин Data Mining был введен в 1989 году Григорием Пятецким-Шапиро как обозначение автоматизированного поиска правил, ускоряющим запросы к крупным базам данных. В общем случае постановка задача ставилась следующим образом: имеется достаточно крупная база данных и предполагается, что в ней есть какие-то скрытые значения. С помощью некоторого набора методов можно их обнаружить. Знания могут быть ранее неизвестные. Актуальность данной работы состоит в том, что зависимости, выявленные в результате интеллектуального анализа, могут быть использованы для решения важных социальных вопросов: так, например, можно разработать стратегию предотвращения появления алкогольной зависимости среди подростков, учитывая факторы, оказывающие самое сильное влияние на соответствующую характеристиу.

Целью данной выпускной квалификационной работы является анализ медико-социальных данных, хранящихся в хранилище данных некоторой предметной области, средствами Data Mining. В качестве данных для анализа предлагаются сведения об употреблении алкоголя подростками в одном из штатов Соединенных Штатов Америки. В рамках данной работы будут рассмотрены технологии OLAP и Data Mining, а именно понятия таких систем, их классификация, наряду с преимуществами и недостатками использования таких систем.

Практическая часть включает в себя анализ данных исследования алкогольной зависимости подростков в США. Для этого будет подробно описана выбранная предметная область в контексте созданной базы данных, создано и заполнено на основе базы данных хранилище данных.

В качестве главного инструмента анализа будет использован метод построения деревьев решений, так как он обеспечивает извлечение и анализ информации, помогающей принимать решения, представляя ее в наглядном

виде, удобном для ее интерпретации. Будут представлены и рассмотрены различные алгоритмы построения, перечислены их достоинства и недостатки.

Объем работы составляет 53 страницы. Основная часть данной работы состоит из двух разделов, каждый из которых разбит на несколько подразделов. Первый раздел «Теоретические основы интеллектуального анализа данных» включает в себя подразделы «OLAP-системы» («Понятия OLAP-систем», «Классификация OLAP-систем»), «Технология Data Mining. Классификация алгоритмов», «Деревья решений» («Построение деревьев решений с помощью Deductor», «Построение деревьев решений с помощью Microsoft»). Второй раздел «Анализ данных средствами Data Mining» включает в себя подразделы «Описание предметной области», «Анализ данных средствами Deductor» («Подготовка данных для анализа», «Построение деревьев решений»), «Анализ данных средствами Microsoft SQL Server» («Создание хранилища данных», «Подготовка данных для анализа», «Построение деревьев решений»), «Сравнительный анализ результатов». Список использованных источников включает в себя 26 наименований.

1 Основное содержание работы

Как было сказано выше, основная часть работы состоит из двух разделов. Первый раздел посвящен теоретическому аспекту задачи. Сначала в ней даются общие сведения о хранилищах данных - предметно-ориентированных системах, получающих данные из баз данных, используемых организацией, или других источников (таких как отдельные документы или наборы данных), главной целью разработки которых является подготовка отчетов и бизнес анализа с целью принятия управленческих решений. Приводится классификация хранилищ с наименованием компонент, присутствующих в них, а также приводится краткое описание операций, производимых в хранилищах:

1. извлечение;
2. преобразование;
3. загрузка;
4. анализ;
5. представление результатов анализа.

Далее в разделе рассматриваются и описываются OLAP-систем и технология Data Mining. OLAP (On Line Analytical Processing) - это технология комплексного многомерного анализа данных, являющаяся ключевым моментом организации хранилища данных. Такая технология используется для анализа деятельности компании и её прогнозирования OLAP технологии преобразовывают количественные показатели в качественные. Это позволяет сотрудникам организации, ответственным за принятие решений (или подготовку данных, как бизнез аналитикам), сформировать свое мнение о данных посредством быстрого доступа к различным способам представления информации. OLAP системы должны удовлетворять определенному набору требований (впервые сформулированы Эдгаром Коддом в 1993 году и реализованы в FASMI - Fast Analysis of Shared Multidimensional Information). Приводится описание достоинств и недостатков данной системы.

Несмотря на то, что невозможно полностью автоматизировать процесс принятия решения, OLAP-технологии - это один из самых удобных способов помощи принятия управленческих решений главе компании (или аналитикам компании), одним из ключевых моментов которого является представление данных, т.е. графический интерфейс. Внешнее отображение информации

в системе реализовано в виде электронных таблиц или графика (несколько экранных форм, включающих в себя данные компоненты).

Классификацию OLAP-систем можно выделить в отдельный список.

1. ROLAP, Relational OLAP - реляционный OLAP.
2. MOLAP, Multidimensional OLAP - многомерный OLAP.
3. HOLAP, Hybrid OLAP - гибридный OLAP.
4. DROLAP, A DENSE-Region BASED Approach to OLAP.
5. OOLAP, Object-relational OLAP – объектно - реляционный OLAP.
6. RTOLAP, Real-time ROLAP - ROLAP реального времени.
7. WOLAP, Web-based OLAP - OLAP, ориентированный на веб.

Data Mining («добыча данных») - это не конкретная технология, а процесс применения алгоритмов для поиска корреляций, тенденций, зависимостей внутри набора данных с помощью различных математических алгоритмов (а также алгоритмов математической статистики): кластеризация, создание выборок, регрессионный и корреляционый анализ. Поскольку данные технологии применяются корпорациями для анализа большого объема бизнес-информации, главной целью добычи данных является представление данных в виде, наглядно отображающем бизнес-процессы. На основе этого должна строиться модель, позволяющая прогнозировать бизнес-процессы. Основой для данного множества технологий стала концепция паттернов, в которых собраны закономерности, характерные для данной выборки из исходных данных.

Можно привести следующую классификацию алгоритмов Data Mining, поставляемых Microsoft SQL Analysis Services:

1. Microsoft Decision Tree («Деревья Решений Майкрософт»);
2. Naive Bayes («Наивный байесовский классификатор»);
3. Microsoft Clustering («Кластеризация Майкрософт»);
4. Sequence Clustering («Кластеризация последовательностей»);
5. Microsoft Clustering («Кластеризация Последовательностей»);
6. Microsoft Neural Networks («Нейронные сети Майкрософт»);
7. Time Series («Временные ряды»);
8. Microsoft Linear Regression («Логическая регрессия Майкрософт»);
9. Microsoft Logistic Regression («Логическая регрессия Майкрософт»).

Подробнее остановимся пункте «Деревья решений». Они представляют данные в иерархической структуре по правилу «если... то...». Более формально дерево решений можно представить как некоторое конечное множество T, состоящее из одного или нескольких узлов, таких что

1. Существует один специально обозначенный узел, называемый корнем данного дерева;
2. Остальные узлы (исключая корень) содержатся в $m > 1$ попарно непересекающихся множествах $T_1, ..., T_m$, каждое из которых в свою очередь является деревом. Деревья $T_1, ..., T_m$ называют поддеревьями данного корня.

Главная цель деревьев - создание модели, предсказывающей значение целевого атрибута по некоторому набору атрибутов-переменных, задаваемых на входе. Традиционно дерево отрисовывается сверху вниз, где в качестве «корня» дерева может быть использовано какое-то зарезервированное служебное слово или словосочетание, также некоторые сервисы отображают там название предсказываемого атрибута или же просят пользователя ввести название дерева. Все чаще программные продукты отрисовывают деревья слева направо, т.е. корень находится слева, а не сверху, и дерево «разворачивается» в горизонтальной плоскости. Считается, что расположенные таким образом деревья решений лучше поддаются анализу человеком.

Обобщенный алгоритм построения дерева решений очень прост и состоит из двух пунктов:

1. выбирается атрибут A из множества заданных атрибутов и помещается в промежуточный корень;
2. для всех его значений i_k, k = $1, ..., K$ оставляются только те тестовые («обучающие») примеры, значение атрибута A которых равно i_k; в этом потомке рекурсивно строиться дерево решений.

Выделим следующие алгоритмы построения дерева решений:

1. алгоритм ID3 - атрибут выбирается на основании прироста информации, разработан Джоном Р. Квинланом;
2. алгоритм C4.5 является усовершенствованной версией атрибута ID3, поддерживающей возможность отсечения ветвей, работы с числовыми и неполными данными;

3. алгоритм CART (Classification and Regression Tree) - предназначен для построения бинарных деревьев.

В подразделе «Построение деревьев решений с помощью Deductor» подробно описывается алгоритм $C4.5$ (так как именно этот алгоритм использует программный продукт), выделяются требования к данным, приводится его математическое обоснование. В подразделе «Построение деревьев решений с помощью Microsoft» подробно описывается алгоритм, используемый Microsoft SQL Server для построения деревьев решений, описываются рабочие типы данных, приводится список уникальных функций, присущих этому алгоритму.

Во втором разделе приводится практическая часть выпускной квалификационной работы. Приводится описание предметной области. В качестве данных для анализа были взяты данные анонимного опроса из открытого источника. Была исследована проблема алкогольной зависимости среди подростков от 15 лет до 21 года на территории Соединенных Штатов Америки. Для описания предметной области дадим перечень атрибутов, которыми характеризуется каждый подросток:

1. пол;
2. возраст;
3. наличие романтических отношений;
4. наличие личного времени;
5. время, которое подросток проводит с друзьями;
6. есть или нет доступ к сети «Интернет» дома;
7. здоровье;
8. уровень употребления алкоголя в течение рабочей недели и на выходных.

Кроме этого для каждого студента определяется школа и семья. Семья включает в себя такие понятия как: состав семьи, проживают ли родители вместе, оценка отношений в семье по пятибалльной шкале, а также образование, которое получили родители и их профессии.

Школа определяется временем, которое ученик тратит на дорогу, временем, которое ученик проводит на занятиях, причиной, по которой подросток ходит в школу, наличием или отсутствием денежной поддержки от школы,

курсов, оплачиваемых родителями, дополнительных внеклассных активностей.

Сначала анализируем данные с помощью Deductor. Для этого все данные помещаются в файл с расширением .txt. Будем строить два дерева решений. В первом будем искать зависимости, обуславливающие употребление подростками алкоголя в течение рабочей недели. С помощью второго будем выявлять причины алкогольной зависимости в целом (для этого воспользуемся понятием среднего показателя, вычисляемого как просто арифметическое среднее между «Dalc» и «Walc»).

Первое дерево решений оказалось информативным и достаточно точным (по таблице сопряженности): если студент старше 20, 5 лет, то его уровень алкогольной зависимости является «средним». Если же студент моложе указанного возраста, то надо обращать внимание на его пол. Для девушек моложе 20 лет значение наблюдается самое низкое пристрастие к алкоголю. Самая «ветвистая» ситуация у молодых людей в возрасте меньше 20 лет. Их пристрастие к алкоголю определяется причиной по которой они посещают школу. Если это "репутация школы то ожидается умеренное употребление алкоголя в будние дни. Очень низкий уровень у студентов, которые посещают школу, расположенную близко к дому, и низкое, если в графе причина было указано «другое». Если же главной причиной является набор предоставляемых курсов, то все зависит от того, есть ли дома интернет. Однако и в этом случае ожидается слабая зависимость.

Второе дерево получилось слишком ветвистое, присутствует больше количество зависимостей «если... то...», что является первым признаком некорректности построенного дерева. Эту теорию также подтверждает и таблица сопряженности. Даже небольшую часть дерева трудно описать и прочитать. Запутанные связи, многократное использование атрибутов для разбиения в подмножествах, а также очень маленькое количество примеров в каждой категории говорит о том, что построенное дерево является неудачным и не стоит опираться на него, чтобы объяснить эффект алкоголизма среди подростков. Т.е. есть необходимость пересмотреть способ, с помощью которого вычисляется предсказываемый атрибут.

Так как показатели «Dalc» и «Walc» описывают употребление алкоголя в течение недели, выведем атрибут, характеризующий средний уровень употребления студентом алкоголя в течение недели. Для этого снова переведем «Dalc» и «Walc» в числовую шкалу, для каждого студента сложим их значения, но теперь будем делить на 7, по количеству дней в неделе. Округлим получившиеся результаты до 1 знака после запятой, так что все значения находятся в диапазоне от 0 до $1,4$. Затем для каждого примера переведем полученное числовое значение в текстовый формат с использованием более гибкой шкалы.

Полученное в результате такого преобразования атрибута дерево гораздо компактнее, чем предыдущее, его легче описать: для людей моложе $16,5$ лет уровень употребления алкоголя низкий. Если пол мужской и старше указанного возраста, то разбиение идет по атрибуту, отвечающему за образование отца. Результат весьма неожиданный, потому что логично было предположить, что образование матери также должно играть роль. Это дерево отличается от первого дерева - единственный атрибут в «пересечении» - возраст студента.

Перейдем к анализу данных с помощью Microsoft SQL Server. Для анализа данных об употреблении алкоголя подростками с помощью Microsoft SQL Server хранилище данных создается таким образом, чтобы достичь оптимально быстрого извлечения данных из хранилища при анализе данных средствами Data Mining. Созданное ХД «AlcoholAnalysisTable» содержит одну таблицу фактов (StudentTable) и четырнадцать таблицы измерений (реализована схема «звезда»).

Таблица фактов «StudentTable» содержит информацию о конкретном студенте. В ней отражается информация об идентификационном номере студента, его поле, причине, адресе, размере семьи, отношениям внутри нее, по которой подросток ходит в щколу, времени, затрачиваемом на школу и уроки, свободном времяни, образовании и профессии родителей, пристрастию к алкоголю. Все таблицы измерений построены единообразно: включают в себя два атрибута «id» и какое-то соответствующее значение.

Для того, чтобы анализировать данные средствами Data Mining, т.е. создать модель данных на основе представления, необходимо создать представ-

ление источника данных, которое содержит таблицы, на основе которых будет строиться модель. Представление StudentView извлекает информацию из таблиц в AlcoholAnalysisTable. Она содержит полную информацию о студентах, которые употребляют алкоголь, а также полную информацию о его семье и школе (присутствует только идентификатор студента).

Были построены несколько деревьев решений, в которых предсказывались значения атрибутов пристрастия к спиртному. При этом зависимости искались в различных входных параметрах. В отличие от алгоритма $C4.5$, регрессионный алгоритм Майкрософт не отсекает незначимые атрибуты, поэтому выбор входных данных становится процессом творческим. По данным, полученным в данном подразделе получается, что алгоритм построения деревьев решений Майкрософт не подходит для анализа медико-социальных данных.

ЗАКЛЮЧЕНИЕ

В данной работе был проведен анализ медико-социальных данных об употреблении алкоголя подростками в Соединенных Штатах Америки. Был применен один из алгоритмов Data Mining под названием «Дерево решений». Были рассмотрены два алгоритма построения дерева решений ($C4.5$ и регрессионный алгоритм Майкрософт), для чего было создано хранилище данных.