МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра Математического и компьютерного моделирования

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

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 247 группы

направление 09.04.03 – Прикладная информатика

механико-математического факультета

Брагиной Софьи Михайловны

Научный руководитель
зав.каф., д. ф. – м. н., доцент Ю.А. Блинков

Зав. кафедрой
зав.каф., д. ф. – м. н., доцент Ю.А. Блинков

Саратов 2019

Введение

Любая деятельность человека связана с накоплением больших объемов информации, данных, необходимых для бесперебойного и эффективного функционирования компании. Накапливание знаний, получение экспертизы в определенной предметной области, увеличением объемов данных - это естественный процесс, который свидетельствует о нормально протекающей работе. Однако накопление знаний не является конечным результатом - это всего лишь один из этапов, необходимых для адекватной экспертной оценки.

В данной магистерской работе рассматриваются методы обработки и анализа данных для извлечения новой информации. Главными инструментами анализа настоящей работы являются Data Mining и NoSQL. В настоящее время технологии NoSQL становятся все более развитыми, обеспечивают удобное хранение данных, быстрое их извлечение. А Data Mining помогает анализировать данные для дальнейшего использования полученных знаний в принятии решений. Таким образом, комбинация этих двух технологий становятся эффективным инструментом анализа.

Благодаря разнообразию предоставляемых алгоритмов Data Mining каждая конкретная задача может быть решена оптимальным способом. В качестве примера анализа проводится исследование данных употребления алкоголя среди подростков Соединенных Штатов Америки алгоритмом построения решения «Случайный Лес».

В качестве целей данной работы можно выделить следующее:

- Исследование существующих инструментов Data Mining.
- Сравнение нескольких методов построения деревьев решений и выбор подходящего для анализа данных.
- Исследование существующих типов баз данных NoSQL.
- Разработка программы по построению дерева решений.
- Анализ полученных результатов.

Итоговой задачей исследования является построение классификации входных данных по признаку «употребления алкоголя в течение недели». Научная новизна исследования заключается в применении комбинации алгоритма интеллектуального анализа данных и хранения данных в NoSQL базе данных с целью повышения достоверности получаемых результатов и ско-

рости обработки. В ходе работы было разработано приложение для анализа медико-социальных данных.

Магистерская работа состоит из введения, трех разделов, заключения, списка используемых источников и трех приложений. В первом разделе «Технология Data Mining и NoSQL», состоящем из двух подразделов, приводится теоретическое обоснование используемых концепций и алгоритмов. Во втором разделе «Алгоритмы построения деревьев решений» проводится детальный разбор и исследование существующих алгоритмов, выбирается оптимальный для поставленной задачи. Третий раздел «Анализ алкогольной зависимости среди подростков США» , состоящем из пяти подотделов, содержит реализацию приложения, описание процесса извлечения данных, непосредственный анализ полученных результатов.

В ходе исследования был представлен доклад на ежегодной научной конференции «Представляем научные достижения миру. Естественные науки».

Основное содержание работы

В первом разделе даются определения всем используемым инструментам, описываются алгоритмы интеллектуального анализа данных, предоставляется историческая сводка.

Сначала описывается Data Mining - 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 корреляционный анализ. Это исследование и обнаружение «машиной» (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком.

Далее описываются стандартные задачи, решаемых с помощью добычи данных, обусловленные следующими типами связей между данными: ассоциация, последовательность, классификация, кластеризация и временная закономерность.

Классификация - задача распределения множества объектов по заранее заданным группам (классам), так чтобы внутри одного класса сформиро-

валась группа похожих объектов. «Похожесть» определяется одинаковыми свойствами и признаками (атрибутами). По аналогии с графами/деревьями, при наличии всего двух классов имеет место бинарная классификация. Типичным примером является банковская сфера, а именно - определение рисков кредитования.

Кластеризация - закономерное развитие (и усложнение) задачи Классификации: аналогичный процесс разбиения объектов на группы (кластеры), с единственным изменением - классы не являются заранее предопределенными. Примером такой задачи может служить самоорганизующаяся карта Кохонена.

Ассоциация - задача поиска ассоциативных правил (закономерностей) между событиями в наборе данных, другими словами - поиск повторяющихся образцов. В качестве примера можно рассмотреть задачу анализа рыночной корзины с цель нахождения устойчивых связей в корзине покупателя.

Последовательность (так же называемая последовательной ассоциацией) - задача установления связей (закономерностей) между событиями, связанными во времени, по правилу «наступление события X приводит к наступлению события Y».

Регрессия - задача прогнозирования пропущенных или будущих численных показателей на основе исторических данных. Пример - оценки экономических кризисов. Эта задача на самом деле похожа на задачу Классификации, различие состоит в том, что классификация подразумевает предсказывание класса зависимой переменной, а регрессия - предсказывание значения.

Отдельно описывается используемый в дальнейшем исследовании алгоритм – дерево решений. Структурно дерево является набором «узлов», «листьев» и «веток». Ветки дерева содержат записи об атрибутах, от которых зависит целевая функция, листья хранят значения целевой функции, а остальные узлы — атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение.

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

- Выбирается атрибут A из множества заданных атрибутов и помещается в промежуточный корень
- 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 строится дерево решений.

Далее рассматривается концепция NoSQL - это термин, который обозначает ряд подходов, используемых для реализации баз данных, не использующих традиционные для реляционных СУБД моделей. Поскольку не существует общепринятого определения термина NoSQL, говоря о NoSQL можно лишь обсуждать некоторый набор свойств, характерных для БД, выделенных в данную категорию:

1. Не используется NoSQL. Многие пытаются разработать свой язык запросов (пример - CQL для Cassandra), однако не был реализован ни один настолько же гибкий.
2. Открытый исходный код. Спорный пункт, так как термин NoSQL часто применяется к системам закрытым кодом.
3. Работа с кластерами. Первый семинар 2009 года повлиял на модель данных и подход к обеспеченности их согласованности. И изначально базы данных NoSQL были ориентированы на работу именно с кластерами (но далеко не все, пример - графовые БД).
4. Неструктурированные (schemaless). Базы работают без схем, предоставляя возможность свободно добавлять поля без предварительных изменений структуры. Однако у неструктурированной схемы есть свои недостатки: накладные расходы в коде приложения при смене модели данных, отсутствие таких ограничений со стороны базы как not null, unique, check constraint, сложность в понимании, восприятии и контроле структуры данных. Тем не менее подобного рода гибкость является все-таки преимуществом. Самый яркий пример - Твиттер, который сейчас в дополнение к самому сообщению в базе сохраняется еще несколько килобайт метаданных.
5. Представление данных в виде агрегатов. Как известно, реляционная модель сохраняет логическую бизнес-сущность приложения в различные физические таблицы в целях нормализации, тогда как NoSQL хра-

нилища оперируют с этими сущностями как с целостными объектами. Стоит отметить, что работа с большими и/или денормализованными объектами может привести к многочисленным проблемам при попытках произвольных запросов к данным, когда запросы не укладываются в структуру агрегатов. К сожалению, это компромисс, на который приходится идти в распределенной системе: здесь нельзя проводить нормализацию данных как в обычной односерверной системе, так как это создаст необходимость объединения данных с разных узлов и может привести к значительному замедлению работы базы.

Теперь рассмотрим конкретную NoSQL базу данных – MongoDB. Она активно развивается и пока еще далека от совершенства. Но уже поддерживает тысячи приложений, использующих крупные и мелкие кластеры базы данных. Это хранилище JSON-документов, и Благодаря отсутствию структурированной схемы, Mongo может видоизменяться вместе с моделью данных.

Во втором разделе производится разбор различных алгоритмов построения деревьев решений. В основном они отличаются способом выбора наилучшего разбиения, проходят несколько итераций разработки для достижения этого:

1. Алгоритм $ID3$ - выбор атрибута происходит на основании прироста информации, либо на основании критерия Джини (мера того, насколько часто случайно выбранный элемент из набора неверно помечается, если он случайным образом помечается согласно распределению меток в подмножестве.
2. Алгоритм $C4.5$ (улучшенная версия $ID3$) - добавлено отсечение ветвей, возможность работы с числовыми атрибутами, а также возможность построения дерева из неполной обучающей выборки, в которой отсутствуют значения некоторых атрибутов.
3. Алгоритм построения решающих деревьев CART (Classification and Regression Tree) - решает задачи классификации и регрессии построением дерева решений, разработанный в том числе Лео Брейманом (Беркли). Предназначен для построения бинарного дерева решений.
4. Random Forest (случайный лес) - предложен Лео Брейманом и Адель Катлер, заключается в использовании ансамбля решающих деревьев,

реализует сочетание двух идей: метода бэггинга Бреймана и метод случайных подпространств. Используется для задач классификации, регрессии и кластеризации.

В третьем разделе описывается предметная область – данные об алкогольной зависимости среди подростков США, реализуется алгоритм «Случайный лес» и производится анализ результатов классификации данных.

Общий алгоритм построения дерева решений представлен в качестве псевдокода:

```
GrowTree(D, F) - grow a feature tree from training data.

Input : data D; set of features F.
Output : feature tree T with labelled leaves.
 if Homogeneous(D) then return Label(D) ;
 S = BestSplit(D, F) ;
 split D into subsets Di according to the literals in S;
 for each i do
 if Di not empty then Ti = GrowTree(Di, F)
else Ti is a leaf labelled with Label(D);
 end
 return a tree whose root is labelled with S
and whose children are Ti

BestSplit(D, F) - find the best split for a decision tree.

Input : data D; set of features F.
Output : feature f to split on.
 Imin  =1;
 for each f \subset F do
 split D into subsets D1 ,..., Dl
according to the values Vj of f;
 if Impurity({D1 ,..., Dl}) < Imin then
 Imin = Impurity({D1 ,..., Dl});
 fbest =  f;
```

```
end
end
return fbest
```

Ниже предоставлен список признаков (атрибутов), характеризующих каждого студента:

- Пол
- Возраст (в диапазоне от 15 до 19 лет)
- Наличие романтических отношений
- Наличие личного времени (разбивается на два интервала: меньше или равно 3 часам, больше 3 часов в день)
- Время, проводимое с друзьями (аналогично предыдущему пункту разбивается на два интервала, однако подсчет времени ведется за неделю)
- Есть или нет доступ к сети «Интернет» дома
- Здоровье (допустимые значения: «очень плохое», «плохое», «приемлемое», «плохое» и «очень плохое», также возможно переложение на пятибалльную шкалу)
- Уровень употребления алкоголя в течение рабочей недели и на выходных («очень редко», «редко», «иногда», «часто» и «очень часто»)
- Состав семьи (больше или меньше 3 человек)
- Статус родителей (проживают вместе или по-отдельности (предположительно, в разводе))
- Образование родителей. Выделяют следующие уровни образования:
    1. Никакого образования
    2. Окончил (-а) младшую школу
    3. Окончил (-а) 9 классов
    4. Получил(-а) среднее образование
    5. Получил (-а) высшее образование

Основа работы – создание приложения для анализа данных. Диаграмма классов изображена в соответствии с рисунком 1:

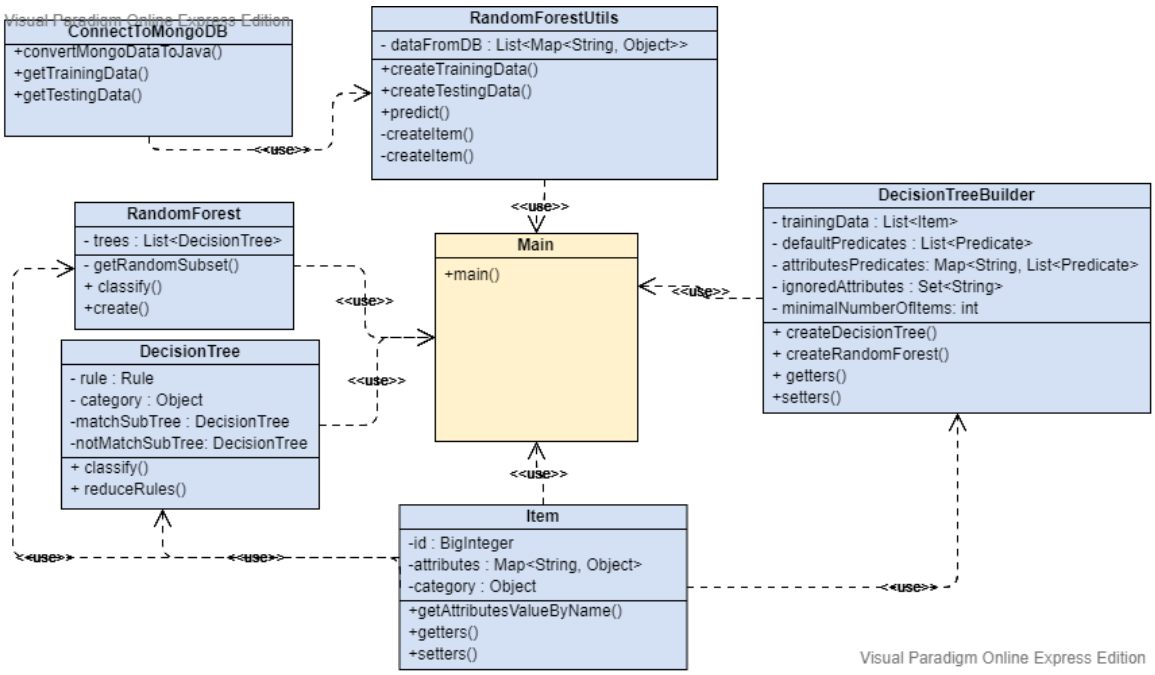

Рисунок 1 — Диаграмма классов программы «RandomForest»

Видно, что входной точкой программы является класс Main с единственным методом - main. В этом методе данные из базы данных разделяются на 2выборки: тренировочная и тестовая. На основе тренировочных данных строится лес, который используется для классификации тестовых данных. В результате выполнения программы получается построенный лес (из 100 деревьев) и структура с предсказаниями, содержащая айди студента и категорию употребления алкоголя (возможные значения перечислены в классе Category: RARELY, SOMETIMES, OFTEN, VERYOFTEN).

Заключительным этапом работы является анализ полученных результатов. «Лес», построенный в результате работы программы содержит 1000 деревьев, каждое из которых является уникальным, построенным на основе случайного подмножества множества входных данных. Тем не менее, главным результатом работы приложения является файл с предсказаниями, содержащими уникальный идентификатор студента и его категорию (уровень употребления алкоголя). Была получена доля верных ответов, равная 91%, что является хорошим показателем.

Заключение

Широкое использование NoSQL баз данных и развитие алгоритмов интеллектуального анализа данных закономерно. Построенные приложения или

экспертные системы для анализа данных, основанных на NoSQL и Data Mining, используемых в данной работе, достигают хороших результатов.

В рамках магистерской работы был рассмотрен конкретный алгоритм интеллектуального анализа данных «Случайный лес». Были рассмотрены основные термины и понятия. Для понимания всех возможных структурных решений были рассмотрены разные алгоритмы построения деревьев решений, приведена математическая основа, описаны особенности их реализации. Особенности технологии Data Mining были разобраны на основе алгоритма построения деревьев решений. Были перечислены преимущества выбранного метода.

Была приведена реализация алгоритма на языке Java. Данная реализация использовалась в приложении для анализа данных об алкогольной зависимости. Был проведен анализ полученных результатов и сравнительный анализ с результатами предыдущих исследований.

Результат исследования, проведенного в рамках настоящей работы, могут быть использованы для составления программы по борьбе с алкогольной зависимостью среди подростков.

Поставленные цели считаются достигнутыми.