УДК 735.29.(32)

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ В ЭВЕНТОЛОГИИ

Комарова О.А.

Научный руководитель – доцент Баранова И.В.

Сибирский федеральный университет

1. Введение

Термин Data Mining получил свое название из двух понятий: поиска ценной информации в большой базе данных(Data) и добычи горной руды (Mining). Термин переводится как "добыча" или "раскопка" данных. Data Mining- мультидисциплинарная область, возникшая и развивающаяся на базе прикладной статистики, искусственного интеллекта, теории баз данных и др.Технологию Data Mining достаточно точно определяет Григорий Пиатецкий Шапиро.

Data Mining - это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полученных на практике закономерностей.

2. Стадии Data Mining

Data Mining состоит из стадий:

1. Выявление закономерностей (свободный поиск)

На этой стадии осуществляется исследование набора данных с целью поиска скрытых закономерностей. Свободный поиск представлен такими действиями: выявление закономерностей условной логики, выявление закономерностей ассоциативной логики, выявление трендов и колебаний.

2. Прогностическое моделирование

Использует результаты работы первой стадии. Действия: предсказание неизвестных значений, прогнозирование развития процессов. В дополнение к этим стадиям иногда вводят валидацию, следующую за стадией свободного поиска. Цель валидации - проверка найденных закономерностей.

3. Анализ исключений

Анализ исключений или аномалий, выявленных в найденных закономерностях. Действия: выявление отклонений. Для выявления необходимо определить норму, которая рассчитывается на стадии свободного поиска.

3. Основные задачи и области применения Data Mining

1. Классификация (Classification)

2. Кластеризация (Clustering)

3. Ассоциация (Associations)

4. Прогнозирование (Forecasting)

5, Оценивание (Estimation)

6. Визуализация (Visualization, Graph Mining)

7. Подведение итогов (Summarization) .

Сфера применения Data Mining ничем не ограничена, он применим везде, где существуют данные:

- Розничная торговля
- Банковское дело
- Телекоммуникации
- Страхование
- Приложения в бизнесе
- Медицина
- Молекулярная генетика и генная инженерия
- Прикладная химия

4. Основы ассоциативных правил в Data Mining

В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных. Отличие ассоциации от других задач Data Mining: поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно. Наиболее известный алгоритм решения задачи поиска ассоциативных правил – алгоритм Apriori.

Рассмотрим метод ассоциативных правил в интеллектуальном анализе данных.

Основные понятия:

1. Транзакция --- это множество событий, которые произошли одновременно.

2. Транзакционная (операционная) база данных представляет собой двумерную таблицу, которая состоит из номера транзакции(TID) и перечня событий, происходящих во время этой транзакции

3. Поддержка --- количество или процент транзакций, содержащих определенный набор данных

Обозначения используемые в алгоритме:

Lk - множество k-элементных наборов, чья поддержка не меньше заданной пользователем.

Ck - множество k-элементных наборов потенциально частых.

Алгоритм:

Шаг 1. Присвоить k = 1 и выполнить отбор всех 1-элементных наборов, у которых поддержка больше минимально заданной пользователем Suppmin.

Шаг 2. k = k + 1.

Шаг 3. Если не удается создавать k-элементные наборы, то завершить алгоритм, иначе выполнить следующий шаг.

Шаг 4. Создать множество k-элементных наборов кандидатов из частых наборов. Для этого необходимо объединить в k-элементные кандидаты (k-1)-элементные частые наборы. Каждый кандидат будет формироваться путем добавления к (k-1) - элементному частому набору - p элемента из другого (k-1)-элементного частого набора - q. Причем добавляется последний элемент набора q, который по порядку выше, чем последний элемент набора p.

При этом все k-2 элемента обоих наборов одинаковы.

Шаг 5. Для каждой транзакции T из множества D выбрать кандидатов Ct из множества Ck, присутствующих в транзакции T. Для каждого набора из построенного множества Ck удалить набор, если хотя бы одно из его (k-1) подмножеств не является часто встречающимся т.е. отсутствует во множестве Lk-1.

Шаг 6. Для каждого кандидата из Ck увеличить значение поддержки на единицу.

Шаг 7. Выбрать только кандидатов Lk из множества Ck, у которых значение поддержки больше заданной пользователем Suppmin. Вернуться к шагу 2.

Результатом работы алгоритма является объединение всех множеств Lk для всех k.

5. Метод ассоциативных правил на языке эвентологии

Теперь сформулируем метод ассоциативных правил на языке эвентологии.

Основные понятия:

1. X={a,b...} - множество случайных событий;

2. p(x) - вероятность распределения

3. Сет - среднее случайного конечного абстрактного множества

EK= {x: $p(x) \geq h$}, для которого его мера μ (EK) наиболее близка к числу λ= E μ (K) - средней мере случайного множества K.

Средняя мера случайного множества K - математическое ожидание случайной величины - вычисляется по теореме Роббинса.

Алгоритм:

1. Пусть задано X={a,b...} - множество случайных событий; для каждого подмножества заданы вероятности распределения p(x)

2. Вычислить индивидуальную вероятность моноплетов

3. Найти сет-среднее

4.Отбрасываем моноплеты, у которых индивидуальная вероятность меньше, чем вычисленное сет-среднее

5. Вычислить индивидуальную вероятность дуплетов

6. Определить сет-среднее

7. Отбрасываем дуплеты, у которых индивидуальная вероятность меньше вычисленного сет-среднего

8. И так далее по всем слоям

Рассмотрим пример, нам необходимо найти наиболее встречающиеся наборы товаров

Есть X = {a,b,c...}, где

Хлеб = a, Молоко = b, Печенье = c, Сметана = d, Колбаса = e , Конфеты = f. Задано распределение вероятности.

{a,b,c}; {a,b,c}; {b,d}; {b,d}; {b,d}; {a,b,c,d}; {a,b,c,d}; {d,e}; {d,e}; {d,e}; {f}; {f}; {f}; {b,e}; {b,e} ;{ a,b,e}; {a,b,e}.

Считаем индивидуальные вероятности:

p(a)=0.1538

p(b)=0.282

p(c)=0.1026

p(d)=0.2051

p(e)=0.1795

p(f)=0.076

Вычисляем сет – среднее EK=0,179. У нас остаются моноплеты {b}, {d}, {e}.

p(bd)= 0,15625

p(be)=0,125

p(de)=0,093

EK=0,125.

Получаем {b,c,d}- часто встречающийся набор товаров.