УДК 519.248

ЭВЕНТОЛОГИЧЕСКИЙ АНАЛОГ МЕТОДА АССОЦИАТИВНЫХ ПРАВИЛ

Комарова О. А.,
научный руководитель канд. физ.-мат. наук, доцент Баранова И.В.
Сибирский Федеральный Университет

1. Введение

Интеллектуальный анализ данных (Data Mining) - мультидисциплинарная область, возникшая и развивающаяся на базе прикладной статистики, искусственного интеллекта, теории баз данных и др. Оригинальное англоязычное название Data Mining было предложено Григорием Пиатецким-Шапиро в 1989. Название происходит 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 - это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Таким образом, Data Mining представляет собой технологию, предназначенную для поиска в больших объемах данных неочевидных, объективных и полученных на практике закономерностей.

К методам интеллектульного анализа данных относятся всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования, ассоциативной памяти, нечеткой логики. Также методами Data Mining считаются статистические методы: дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов.

В интеллектуальном анализе данных существуют следующие стадии:

1. выявление закономерностей (свободный поиск),
2. прогностическое моделирование,
3. анализ исключений.

На первой стадии (стадии выявления закономерностей) 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.

Прогностическое моделирование использует результаты работы первой стадии. К действиям данной стадии относятся предсказание неизвестных значений и прогнозирование развития процессов.

Третья стадия представляет собой анализ исключений или аномалий, выявленных в найденных закономерностях. Действия этой стадии сводятся к выявлению отклонений. Для этого необходимо определить норму, которая рассчитывается на стадии свободного поиска.

В дополнение к этим стадиям иногда вводят еще одну стадию - валидацию, следующую за стадией свободного поиска. Целью валидации является проверка найденных закономерностей.

2. Основные задачи и области применения интеллектуального анализа данных

Перечислим основные задачи интеллектуального анализа данных:

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 в любых сферах человеческой деятельности, в которых существуют статистические данные:

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

3. Основы ассоциативных правил в интеллектуальном анализе данных

Как было сказано выше, в ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных. Рассмотрим алгоритм работы метода ассоциативных правил в интеллектуальном анализе данных. Рассмотрим метод ассоциативных правил в интеллектуальном анализе данных.

Приведем основные понятия, связанные с данным методом.

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, являющихся потенциально частыми.

Алгоритм имеет следующий вид:

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

Шаг 2. Увеличить размерность набора 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 необходимо 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.

4. Метод ассоциативных правил на языке эвентологии

Теперь сформулируем метод ассоциативных правил на языке эвентологии. Приведем несколько основных определений из эвентологии.

1. Вероятностным пространством называется тройка (Ω, F, P), где Ω - пространство элементарных событий, F - алгебра событий и $\boldsymbol{P}$ – вероятность, определенная на элементах множества X.

2. Конечное множество избранных событий X={a,b...}, выбранных из алгебры вероятностного пространства и состоящее из n=|X| событий, называется множеством случайных событий.

3. Случайное множество событий K определяется как измеримое отображение $K:(\Omega, F, P) \to \left(2^X, 2^{2^X}\right)$, где 2^X - множество всех подмножеств множества X.

4. Полной характеристикой множества случайных событий X служит его эвентологическое распределение. Приведем ниже вид для одной из форм эвентологического распределения – распределения вероятностей пересечений событий множества: $p(X) = \mathrm{P}(K = X),\ X \in 2^X$.

4. Сет-среднее случайного множества представляет собой множество элементов из X следующего вида: EK= {x: p(x) ≥ h}, для которого его мера μ (EK) наиболее близка к числу λ= E μ (K) - средней мере случайного множества K.

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

Алгоритм эвентологического метода ассоциативных правил выглядит следующим образом:

1. Пусть задано X={a,b...} - множество случайных событий; для каждого подмножества заданы вероятности распределения p(x)
2. Вычисляем индивидуальную вероятность моноплетов
3. Находим сет-среднее по формуле из определения.
4.Отбрасываем моноплеты, у которых индивидуальная вероятность меньше меры вычисленного сет-среднего.
5. Вычисляем вероятности дуплетов.
6. Находим сет-среднее.
7. Отбрасываем дуплеты, у которых индивидуальная вероятность меньше меры вычисленного сет-среднего.
8. И так далее повторяем пункты 2-4 по всем слоям C_{X^m}, $X^m = \{X \subseteq X, |X| = m\}$, m=3,...n.

Рассмотрим поиск ассоциативных правил по описанному методу на простом примере.

Пусть имеется следующая тразакционная база данных покупок продуктов:

1. Хлеб, молоко, печенье;
2. Хлеб, молоко, печенье;
3. Молоко, сметана;
4. Молоко, сметана;
5. Молоко, сметана;
6. Молоко, хлеб, сметана, печенье;
7. Молоко, хлеб, сметана, печенье;
8. Колбаса, сметана;
9. Колбаса, сметана;
10. Колбаса, сметана;

11. Конфеты;
12. Конфеты;
13. Конфеты;
14. Колбаса, молоко;
15. Колбаса, молоко;
16. Хлеб, колбаса, молоко;
17. Хлеб, колбаса, молоко.
Необходимо найти наиболее встречающиеся наборы товаров.

Обозначим приведенные выше товары следующими переменными.
a={Хлеб},
b={Молоко},
c={Печенье},
d={Сметана},
e={Колбаса},
f={Конфеты}.

Можно переписать приведенную выше базу данных в виде набора подмножеств покупки товаров:
{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}.

Вычисляем индивидуальные вероятности покупки товаров, путем оценивания частоты встречаемости каждого товара в покупках:
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 ЕК=0,179. Далее в соответствии с пунктом 4 алгоритма отбрасываем моноплеты {a}, {c}, {f}. В итоге остаются моноплеты {b}, {d}, {e}.

Из них формируем дуплеты:
p(bd)= 0,15625
p(be)=0,125
p(de)=0,093

По формуле находим сет-среднее и определяем, что для него ЕК=0,125.
Получаем {b, d,e} - часто встречающийся набор товаров.

Таким образом, с помощью предложенного эвентологического метода можно получать не только парные правила, в которых могут участвовать два события, но и более сложные их модификации – для множеств событий.