Секция «Дискретная математика и математическая кибернетика»

О качестве ассоциативных правил в зависимости от способа дискретизации числовых переменных

Научный руководитель – Рыжов Александр Павлович

Акшонов Андрей Петрович

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра математической теории интеллектуальных систем, Москва, Россия

E-mail: da.boss.up.in.here@gmail.com

Рассматривается известная проблема анализа рыночной корзины (market basket analysis, [1]). Пусть $I = \{i_1, i_2, ..., i_n\}$ — набор из n бинарных *признаков (items, товары)*. $D = \{t_1, t_2, ..., t_m\}$ — набор *транзакций (D — база данных)*. Каждая транзакция из D имеет уникальный ID и представляет собой некоторое подмножество товаров из I.

Ассоциативное правило определяется (согласно [1]) как импликация вида $X \Rightarrow Y$, где $X, Y \subseteq I$ и $X \cap Y = \emptyset$. Далее, согласно с [2], вводится обобщение понятия ассоциативного правила, допускающее наличие числовых переменных. Для этого их область значений разбивается на интервалы и в I добавляются признаки принадлежности переменных соответствующим интервалам.

Правило имеет следующие характеристики:

Поддержка правила $X \Rightarrow Y$ демонстрирует, как часто антецедент X встречается в D: $\text{supp}(X) = \frac{|\{t \in D; X \subseteq t\}|}{|D|}$.

Достоверность — показатель того, насколько часто ассоциативное правило оказывается верным: $\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} = \frac{|\{t \in D; X \subseteq t, Y \subseteq t\}|}{|\{t \in D; X \subseteq t\}|}$.

Лифт определяет интересность правила: $\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$.

Задача состоит в исследовании того, как меняются характеристики ассоциативных правил в зависимости от способа разбиения и числа интервалов разбиения. На практике интересными оказываются правила с большими значениями лифта и достоверности, поэтому мы пытаемся найти разбиения, которые дадут нам именно такие правила.

Теорема. Рассмотрим ассоциативное правило $X \Rightarrow Y$. Разобьём X на наборы элементов $X_1, ...X_s$ так, чтобы все X_j содержали равное число записей, и рассмотрим правила $X_j \Rightarrow Y, j = 1..s$. Тогда:

1) $\sum\limits_{i=1}^{s} \text{supp}(X_i) = \text{supp}(X)$.
2) $\sum\limits_{i=1}^{s} \text{conf}(X_i) = s \cdot \text{conf}(X)$.
3) $\sum\limits_{i=1}^{s} \text{lift}(X_i) = s \cdot \text{lift}(X)$.

Таким образом, статистически равномерное разбиение областей значения переменных на s интервалов гарантирует увеличение как суммарной достоверности, так и суммарного лифта полученных правил в s раз по сравнению с исходным.

Источники и литература

1) R. Agrawal, T. Imieliński, A. Swami. Mining Association Rules between Sets of Items in Large Databases. IBM Almaden Research Center, 1993

2) R. Srikant, R. Agrawal. Mining Quantitative Association Rules in Large Relational Tables. IBM Almaden Research Center, 1996