A Model for Association Rules Based on Clustering

Carlos Ordonez
Teradata, NCR
San Diego, CA, USA

ABSTRACT

Association rules and clustering are fundamental data mining techniques used for different goals. We propose a unifying theory by proving association support and rule confidence can be bounded and estimated from clusters on binary dimensions. Three support metrics are introduced: lower, upper and average support. Three confidence metrics are proposed: lower, upper and average confidence. Clusters represent a simple model that allows understanding and approximating association rules, instead of searching for them in a large transaction data set.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

General Terms

Algorithms, Theory

Keywords

Association rules, clustering, model, support, bound

1. INTRODUCTION

Clustering [2, 7] finds groups of similar points according to some similarity metric, generally distance. Association rules describe frequent patterns found in large transaction data sets. Research on association rules has become extensive since their introduction in the seminal paper [1]. Unfortunately, little has been done on building a model for them. This article proposes a simple, yet comprehensive, model for association rules based on clustering. The relationship between clustering and association rules is interesting because clustering looks for global patterns, producing in general disjoint subsets of points, whereas association rules describe local patterns referring to overlapping subsets of points. Therefore, the model unifies two seemingly unrelated techniques.

2. DEFINITIONS

Definitions commonly used in clustering [7] are combined with definitions used for association rules [1]. The intuition behind definitions is that item identifiers act as subscripts to access dimensions from a cluster centroid. Clusters represent a descriptive model summarizing a transaction data set.

This paragraph introduces clustering definitions. The input is a data set D having n d-dimensional points and k, the desired number of clusters. Let $\mathcal{S} = [0,1]^d$ be a d-dimensional binary space. Let $D = \{t_1, t_2, \ldots, t_n\}$ be a data set of n points in $\mathcal{S}$. That is, t_i is a binary vector. The model is represented by matrices C, R, W, containing the means, the variances and the weights respectively for each cluster and a partition of D into k subsets. Matrices C and R are $d \times k$ and W is a $k \times 1$ matrix. To refer to a column of C or R we use the j subscript (i.e. C_j, R_j). We use the following convention for subscripts. For transactions we use i; $i \in \{1, 2, \ldots, n\}$. Notice that i alone is a subscript to index transactions, whereas i_j refers to item j defined below. For cluster number we use j; $j \in \{1, 2, \ldots, k\}$ and to refer to one dimension we use l: $l \in \{1, 2, \ldots, d\}$. Let $D_1, D_2, \ldots, D_k$ be the k disjoint subsets of D induced by clusters s.t. $D_j \cap D_g = \emptyset$ for $j \neq g$.

This paragraph provides definitions for association rules. Let $T_i = \{l | D_{li} = 1, l \in \{1, 2, \ldots, d\}, i \in \{1, 2, \ldots, n\}\}$. That is, T_i is the set of non-zero coordinates of t_i; T_i can be understood as a transaction or an itemset to be defined below. Since transactions represent sparse vectors $|T_i| << d$. This fact is crucial for fast distance computation to efficiently cluster high dimensional binary data [7]. Let $\mathcal{I}$ be a set of d integers identifying d items, $\mathcal{I} = \{1, 2, \ldots, d\}$. Each dimension of $\mathcal{S}$ corresponds to one item out of d items and vice-versa. The presence/absence of an item in a transaction is indicated by the presence/absence of its identifier. A set of items is called an *itemset*. An itemset $A = \{\mathrm{i}_1, \mathrm{i}_2, \ldots, \mathrm{i}_p\}$, containing p items, is called a p-itemset. We use the vectorial notation $(C_j)_{\mathrm{i}_l}$ meaning "access the mean of dimension i_l in cluster j". In other words, items are used as an index for dimensions. An itemset has a measure of statistical significance or relative frequency called support. For an itemset $X \subseteq \mathcal{I}$, $support(X) = |\{T_i | X \subseteq T_i\}|/n$ for $i = 1 \ldots n$ (i.e. support is the fraction of transactions in D containing X). An *association rule* is an implication of the form $X \Rightarrow Y$, where $X, Y \subset \mathcal{I}$, and $X \cap Y = \emptyset$, where X and Y are itemsets. Itemset X is called the antecedent and itemset Y is the consequent of the rule. The association rule $X \Rightarrow Y$ has a measure of strength called confidence defined as $confidence(X \Rightarrow Y) = support(X \cup Y)/support(X)$.

3. BOUNDING AND ESTIMATING SUPPORT AND CONFIDENCE

http://doi.acm.org/10.1145/1066677.1066800

We use a fast K-means algorithm that clusters D in one pass [7]. Its most salient features are the use of sufficient statistics, sparse matrix operations and efficient distance computation for sparse binary data. Clusters are used to compute bounds and to estimate support and confidence for any potential association rule. Thus association rules can be mined from clusters instead of the transaction data set. Proofs are omitted due to lack of space.

3.1 Bounding and Estimating Support

The following propositions hold regardless of how D is partitioned, but bounds are tighter if D is partitioned minimizing squared error. That is, when K-means converges to a locally optimal solution, Let the upper bound of support of A be defined as follows:

$$upper(A) = \sum_{j=1}^{k} W_j min((C_j)_{\mathrm{i}_1}, (C_j)_{\mathrm{i}_2}, \ldots, (C_j)_{\mathrm{i}_p}). \quad (1)$$

Theorem 3.1. *Let D be a set of n transactions. Let C, W represent k clusters from D. Let A be any p-itemset given by $A = \{i_1, i_2, \ldots, i_p\}$. Then $support(A) \leq upper(A)$.*

The following bound is the counterpart of $upper()$.

$$lower(A) = \sum_{j=1}^{k} W_j [max(0, 1 + \sum_{l=1}^{p} ((C_j)_{\mathrm{i}_l} - 1))], \quad (2)$$

where $max()$ avoids negative lower bounds per cluster.

Theorem 3.2. *Let D, C, W be as stated in Theorem 3.1. Let A be a p-itemset $A = \{i_1, i_2, \ldots, i_p\}$. Then $lower(A) \leq support(A)$.*

Support can be estimated by:

$$average(A) = \frac{lower(A) + upper(A)}{2}. \quad (3)$$

It is evident Theorems 3.1 and 3.2 also hold for Eq. 3.

3.2 Bounding and Estimating Confidence

Let A and B be two itemsets s.t. $A \cap B = \emptyset$. Let the lower confidence of a rule be defined as

$$lowerconfidence(A \Rightarrow B) = \frac{lower(A \cup B)}{upper(A)} \quad (4)$$

Let the upper confidence of a rule be defined as

$$upperconfidence(A \Rightarrow B) = min[\frac{upper(A \cup B)}{lower(A)}, 1] \quad (5)$$

Notice the symmetry between Eq. 4 and Eq. 5. These equations lead to the following results that estimate tight bounds for association rule confidence.

Theorem 3.3. *Let D be a set of transactions. Let C, W represent k clusters from D. Let A and B be two itemsets s.t. $A \cap B = \emptyset$. Then $lowerconfidence(A \Rightarrow B) \leq confidence(A \Rightarrow B) \leq upperconfidence(A \Rightarrow B)$.*

Finally, rule confidence can be estimated by:

$$averageconfidence(A \Rightarrow B) = \frac{average(A \cup B)}{average(A)} \quad (6)$$

Theorem 3.4. *Let D be a set of transactions. Let C, W represent some clustering model of D into k clusters. Let A and B be two itemsets. Then $lowerconfidence(A \Rightarrow B) \leq averageconfidence(A \Rightarrow B) \leq upperconfidence(A \Rightarrow B)$.*

4. RELATED WORK

There has been a lot of work on both scalable clustering [2] and efficient association mining [6, 4], but little has been done finding relationships between association rules and other data mining techniques. One such work is [3] where association rules and classification rules obtained from decision trees are contrasted. There has been work on how to cluster transactions from itemsets [8]. Clustering association rules, rather than transactions, once they are mined, is analyzed in [5].

5. CONCLUSIONS

Clusters on binary dimensions are used as a simple and comprehensive model for association rules. We proposed novel metrics based on binary clusters that provide upper bounds, lower bounds and estimations for both association support and rule confidence. We presented theoretical results for the proposed metrics.

Using clusters to understand and approximate association rules can be exploited for deeper research. We want to study other association rule metrics based on clusters including lift, entropy and gain. We intend to investigate the asymptotic properties of the model.

6. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Conference*, pages 207–216, 1993.

[2] P. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In *Proc. ACM KDD Conference*, pages 9–15, 1998.

[3] A. Freitas. Understanding the crucial differences between classification and association rules - a position paper. *SIGKDD Explorations*, 2(1):65–69, 2000.

[4] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proc. ACM SIGMOD Conference*, pages 1–12, 2000.

[5] B. Lent, A. Swami, and J. Widom. Clustering association rules. In *Proc. IEEE ICDE Conference*, pages 220–231, 1997.

[6] S. Morishita and J. Sese. Traversing itemsets lattices with statistical pruning. In *ACM PODS Conference*, 2000.

[7] C. Ordonez. Clustering binary data streams with K-means. In *Proc. ACM SIGMOD Data Mining and Knowledge Discovery Workshop*, pages 10–17, 2003.

[8] K. Wang, C. Xu, and B. Liu. Clustering transactions using large items. In *ACM CIKM Conference*, pages 483–490, 1999.