Quantitative Association Rule Mining Using Information-Theoretic Approach

Dim En Nyaung
University of Computer Studies, Yangon
dimennyaung@gmail.com

Abstract

Quantitative Association Rule (QAR) mining has been recognized an influential research problem due to the popularity of quantitative databases and the usefulness of association rules in real life. Unlike Boolean Association Rules (BARs), which only consider boolean attributes, QARs consist of quantitative attributes which contain much richer information than the boolean attributes. To develop a data mining system for huge database composed of numerical and categorical attributes, there exists necessary process to decide valid quantization of the numerical attributes. One of the main problems is to obtain interesting rules from continuous numeric attributes. In this paper, the Mutual Information between the attributes in a quantitative database is described and normalization on the Mutual Information to make it applicable in the context of QAR mining is devised. It deals with the problem of discretizing continuous data in order to discover a number of high confident association rules, which cover a high percentage of examples in the data set. Then a Mutual Information graph (MI graph), whose edges are attribute pairs that have normalized Mutual Information no less than a predefined information threshold is constructed. The cliques in the MI graph represent a majority of the frequent itemsets.

Keywords: Quantitative Databases, Association Rules, Mutual Information

1. Introduction

Association rule mining is a significant research topic in the knowledge discovery area. Association analysis is a useful data mining technique exploited in multiple application domains [6]. One of the best known is the business field where the discovering of purchase patterns or associations between products that clients tend to buy together is used for developing an effective marketing [3]. Some examples of recent applications are finding patterns in biological databases, extraction of knowledge from software engineering metrics or obtaining user's profiles for web system personalization. Numerous methods for association rule mining have been proposed, however many of them discover too many rules, which represent weak associations and uninteresting patterns. The improvement of association rules algorithms is the subject of many works in the literature. The attributes considered in a relation are quantitative or categorical. In order to deal with the quantitative attributes in mining association rules, algorithms based on the generalized association rules that handle the continuous attributes as the Boolean vector by partitioning into several intervals are proposed [1].

Quantitative Association Rules (QARs) have served as a useful tool in discovering association relationships among sets of attributes in business and scientific domains. In a QAR, attributes are not limited to being boolean but can be either quantitative, which are numeric values (e.g., age, income), or categorical, which are enumerations (e.g., gender). Being able to represent a wide variety of real-life attributes, QARs are far more expressive and informative than Boolean Association Rules (BARs) [2], which are restricted to only boolean attributes. An example QAR in an employee database is {age [25, 40], gender [female]} ⇒ {income [13500, 18700]} (sup = 0.03, conf = 0.8). The values of the attribute are partitioned using an equidepth approach (that is, each interval resulting from the partition contains roughly the same number of tuples), and then adjacent intervals are combined as necessary.

In this paper describe a framework, called MIC (which stands for Mutual Information and Clique), to mine the set of QARs. The MIC framework has three phases. The first phase prepares the database by discretizing the quantitative attributes. In the second phase, investigate the mutual information between each pair of attributes firstly. Then, normalization on the mutual information is performed. A pair of attribute to have a strong informative relationship is defined if their normalized mutual information is no less than a predefined minimum information threshold, µ. A Mutual Information graph (MI graph) to represent attributes that have strong informative relationships is established. This paper shows that the MI graph can retain all or most of the information carried by the interaction graph. Since each frequent itemset is represented by a clique in the interaction graph, the cliques in the MI graph are used in the final phase to facilitate the computation of frequent itemsets as well as to guide the generation of QARs.

The following sections give some related work and preliminaries on QAR mining in Section 3. Then introduce the concept of interaction graphs in Section 4. In Section 5 describes the overall description of the MIC framework and describe the details in each phase of the framework.

2. Related Work

Due to the popularity of quantitative databases and the usefulness of association rules in real life, QAR mining has been identified as a long-standing research problem. Many studies [1, 3] have aimed at developing feasible approaches to mining QARs over the last decade. Mining simple association rules involves less complexity and considers only the presence or absence of an item in a transaction [5, 6]. Quantitative association mining denotes association with itemsets and their quantities.

The problem of QAR mining [1] is: given a database, a minimum support threshold and a minimum confidence threshold, find all QARs with support and confidence no less than the given thresholds. A common approach to the QAR mining problem is to transform it into a problem of conventional BAR mining [2, 4]. In order to deal with the quantitative attributes in mining association rules, algorithms based on the generalized association rules that handle the continuous attributes as the Boolean vector by partitioning into several intervals are proposed [7, 8].

This paper describes to mine rules from quantitative databases using an information-theoretic approach [9].

3. Quantitative Association Rule (QAR)

Let I = $\{x_1, x_2, . . . , x_m\}$ be a set of distinct attributes or random variables. An attribute can be either quantitative or categorical. Let dom(x_j) be the domain of an attribute x_j, for $1 \leq j \leq m$. An item, denoted as x $[l_x, u_x]$, is an attribute x associated with an interval $[l_x, u_x]$, where $x \in I$ and $l_x, u_x \in$ dom(x) and if $l_x = u_x$, x is categorical and if $l_x \leq u_x$,x is quantitative. An itemset is a non-empty set of items with distinct attributes. Given an itemset X, its attribute set as attr(X) = $\{x \mid x\ [l_x, u_x] \in X\}$ are defined.

A transaction T is a sequence $<v_1, v_2, . . . ,v_m>$, where $v_j \in$ dom(x_j), for $1 \leq j \leq m$. A transaction T supports an itemset X if $\forall\ x_i\ [l_i, u_i] \in X$, $l_i \leq v_i \leq u_i$, where $i \in \{1, . . . , m\}$. Let D denote a quantitative database, which consists of a set of transactions. The frequency of X in D, denoted by freq(X), is the number of transactions in D that supp(X). The support of X, denoted by supp(X), is the probability that a transaction Tin D supports X, and is defined as supp(X) = freq(X)/|D|. X is a frequent itemset if supp(X) $\geq \sigma$, where σ ($0 \leq \sigma \leq 1$) is a predefined minimum support threshold.

A Quantitative Association Rule (QAR), r, is an implication of the form X $\Rightarrow$Y, where X and Y are itemsets, and attr (X) $\cap$ attr (Y) = $\emptyset$.X and Y are called the antecedent and the consequent of r, respectively. The attribute set of r are defined as attr (r) = attr (X) $\cup$ attr (Y). The support of r is defined as supp(X $\cup$ Y). The confidence of r is defined as conf(r) = supp(X$\cup$Y)/supp(X), which is the conditional probability that a transaction T supports Y, given that T supports X. The QAR mining problem is to find all the QARs with support and confidence no less than σ and c where a minimum support threshold σ ($0 \leq \sigma \leq 1$), and a minimum confidence threshold c ($0 \leq c \leq 1$). For example, Given σ = 0.3 and c = 0.6, sup (age[25,30] gender[M,M]) = 0.3 $\geq \sigma$ and conf (age[25, 30] $\Rightarrow$ gender[M,M]) = 0.3/0.4 = 0.75 $\geq$ c.

4. Interaction Graph

The interaction graph represents the relationships between attributes in a QAR mining problem. Given a QAR mining problem P, the interaction graph is defined as an undirected graph $G_I = (V_I, E_I)$, where the set of vertices V_I = I, and the set of undirected edges $E_I = \{(x_i, x_j) \mid \exists\ r \in$Rules (P) such that $x_i, x_j \in$ attr(r)$\}$. Thus, the interaction graph is a graph representation of Rules (P). Thus, if this can obtain the interaction graph prior to performing QAR mining, the search space can be restricted to a much smaller one that encompasses all QARs. More specifically, by finding the cliques in the interaction graph, the set of attributes which is the attribute set of some QARs are derived. Based on the attribute sets, the qualified interval sets are found to produce the QARs. This paper shows that most of the relationships of the attributes reflected in the interaction graph can be acquired by establishing a mutual information graph.

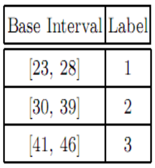
5. The MIC framework

The MIC framework seamlessly incorporates the mutual information concept from information theory into the context of QAR mining. There are three main phases in the MIC framework.

5.1. Phase I : Discretization

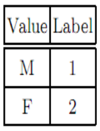
The domain of each quantitative attribute is partitioned into a set of base intervals. The base intervals are then labeled with a set of consecutive integers, {1, 2, . . , n}, such that the order of the base intervals is preserved. The values of a categorical attribute are mapped to a set of consecutive integers. The equidepth discretization technique is proved to minimize the information loss caused by discretization in [1]. Any discretization technique can be applied in this phase of the MIC framework. In this paper, equidepth discretization technique is used. The equidepth discretization technique is proved to minimize the information loss caused by discretization in [1]. Equidepth partitions the domain of a quantitative attribute into n base intervals so that the number of transactions in each base interval is roughly the same. The number of base intervals n is an important factor since it determines the degree of information loss due to discretization. The larger the n, the information loss may be less but the computational cost may be high to mine QARs. A smaller n results in more information loss. The following example illustrates the idea of equidepth discretization.

Table1. age



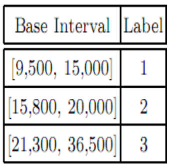
Base Interval	Label
[23, 28]	1
[30, 39]	2
[41, 46]	3

Table2. gender



Value	Label
M	1
F	2

Table3.income



Base Interval	Label
[9,500, 15,000]	1
[15,800, 20,000]	2
[21,300, 36,500]	3

Table4. The Discretized Transactions of Sales Data

age	gender	income	education	service yrs
1	2	1	1	2
1	1	2	2	2
1	1	2	3	1
2	1	3	3	2
2	2	1	1	1
3	1	3	4	1
3	2	1	4	3

5.2. Phase II : Mutual Information Graph Construction

This section describes how to apply the concepts of entropy and mutual information that originates from information theory in the context of QAR mining.

5.2.1. Entropy and Mutual Information

Entropy. Entropy is a central notion in information theory [10], which measures the uncertainty in a random variable. The entropy of a random variable x, denoted as H(x), is defined as

$$H(x) = -\sum_{v_x \epsilon dom(x)} p(v_x) \log p(v_x) \qquad (1)$$

The conditional entropy of a random variable y given another variable x, denoted as H (y|x), is defined as

$$H(y/x) = -\sum_{v_x \epsilon dom(x)} \sum_{v_x \epsilon dom(y)} p\,(v_x, v_y) \log p\,(v_y/v_x) \qquad (2)$$

Mutual Information. Mutual information describes how much information one random variable tells about another one. The mutual information of two random variables x and y, denoted as I(x; y), is defined as

$$I(x, y) = \sum_{v_x \epsilon dom(x)} \sum_{v_x \epsilon dom(y)} p(v_x, v_y) \log \frac{p(v_x, v_y)}{p(v_x)p(\,v_y)} \qquad (3)$$

An important interpretation of mutual information comes from the following property.

Property 1 I(x; y) = H(x) − H(x|y) = H(y) − H(y|x).

From Property 1, the information that y tells us about x is the reduction in uncertainty about x due to the knowledge of y, and similarly for the information that x tells about y. When the greater the value of I(x; y), x and y tell the more information about each other.

Property 2 I(x; y) = I(y; x).

Property 2 suggests that MI is symmetric, which means the amount of information x tells about y is the same as that y tells about x.

Property 3 I(x; x) = H(x).

Property 3 states that the MI of x with itself is the entropy of x. Thus, entropy is also called self-information.

Property 4 I(x; y) ≥ 0.

Property 4 gives the lower bound for MI. When I(x; y) = 0, there have $p(v_x, v_y) = p(v_x)p(v_y)$ for every possible values of v_x and v_y, which means that x and y are independent, that is, x and y tell us nothing about each other.

Property 5 I(x; y) ≤ H(x) and I(x; y) ≤ H(y).

Property 5 gives the upper bound for MI.

5.2.2. Normalized Mutual Information

The normalized mutual information of two attributes x and y, denoted as Ĩ(x, y), is defined as

$$\tilde{I}(x, y) = \frac{I(x;\, y)}{I(x;\, x)} \qquad (4)$$

This idea is to normalize the mutual information between x and y by the maximal value of mutual information between x and another attribute, which is I(x; x) = H(x). The following represents some useful properties of normalized mutual information:

Property 6 I(x, y) = Ĩ(y,x) if I(x, y) = I(y, x)

Proof. It follows directly from Property 2.

Property 6 shows that, the same as mutual information, normalized mutual information is also symmetric.

Property 7 $0 \leq \tilde{I}(x; y) \leq 1$

Proof. Since $I(x; x) \geq 0$ and $\tilde{I}(x; y) \geq 0$, $\tilde{I}(x; y) \geq 0$. It follows by Properties 3 and 5 that $\tilde{I}(x; y) \leq 1$

This property ensures that the value of normalized mutual information falls within the unit interval [0, 1].

Property 8 $\tilde{I}(x, y) = \dfrac{H(x) – H(x/y)}{H(x)}$

Proof. By Properties 1 and 3, there have $I(x; y) = H(x) − H(x|y)$ and $I(x; x) = H(x)$.

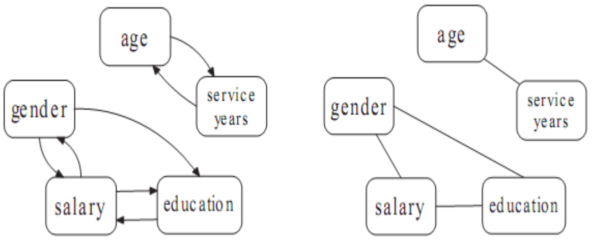
Property 8 suggests the semantics of the normalized mutual information between x and y, which is the percentage of reduction in uncertainty about x due to the knowledge of y.

Thus, normalized mutual information gives the threshold μ an intuitive meaning and makes it relatively independent of specific attributes. Now the threshold μ indicates the minimum percentage of reduction in uncertainty about an attribute due to the knowledge of another attribute.

5.2.3. Mutual Information Graph Construction

Given a predefined minimum information threshold μ, a pair of attributes, x_i and x_j, have a strong informative relationship with each other if $I(x_i; x_j) \geq \mu$.

Given a QAR mining problem, a Mutual Information graph (MI graph) is constructed, which is a directed graph, $G_{MI} = (V_{MI}, E_{MI})$, where the set of vertices $V_{MI} = I$ and the set of directed edges $E_{MI} = \{(x_i, x_j) \,|\tilde{I}(x_i; x_j) \geq \mu\}$. Thus, the MI graph retains and represents the strong informative relationships between the attributes in a QAR mining problem.



(a) G_{MI}

(b) $\hat{G}_{MI}$ (Undirected Graph from G_{MI})

Figure1. G_{MI} and $\hat{G}_{MI}$ of Table 4

5.3. Phase III : Clique Computation and QAR Generation

In this final phase of MIC, all the cliques in G_{MI} are found and simultaneously compute the set of frequent itemsets based on the cliques. And then generate the QARs from the frequent itemsets.

5.3.1. Clique Computation and Frequent Itemset Generation

A clique in G_I represents the set of attributes in a potential frequent itemset. Since $\hat{G}_{MI}$ is constructed to recover the edges in G_I that represent strong informative relationships, most of the attribute sets are obtained that potentially form frequent itemsets by finding all the cliques in $\hat{G}_{MI}$.

Essentially, this approach utilizes $\hat{G}_{MI}$ to do the pruning at the attribute level. Only the attribute sets, which form a clique in $\hat{G}_{MI}$, are considered to generate frequent itemsets. Then compute all the cliques in $\hat{G}_{MI}$ and generate frequent itemsets using a prefix tree structure. Each node at level is labeled with the corresponding attribute name and is attached with a set of intervals whose support is no less than σ. Consecutive base intervals are combined and also attached to the node as long as the support of the combined intervals are no less than σ.

To avoid the occurrence of too general combined intervals, a maximum support threshold σ_m [1] is specified as an upper bound of the support of a combined interval.

Algorithm 1 *CliqueMine*(u)

```
1. if (|RightSibling(u)| > 0)
2.   for each node v ∈ RightSibling(u) do
3.     if ((u, v) ∈ Ĝ_MI)
4.       Add a new node w, with the same label as v, as u's child;
5.       Join the sets of frequent itemsets associated with u and v;
6.       for each itemset, X, obtained from the join do
7.         if (supp(X) ≥ σ)
8.           Attach X to the node w;
9.   Output the set of frequent itemsets associated with u;
10. if (|Child(u)| > 0)
11.   for each node w ∈ Child(u) do
12.     CliqueMine(w);
13. else
14.   Output the set of frequent itemsets associated with u;
```

In the prefix tree constructed by Algorithm 1, each path from a child of the root at Level 1 to a node at Level k represents a k-clique in $\hat{G}_{MI}$, where a k-clique is a clique that consists of k nodes.

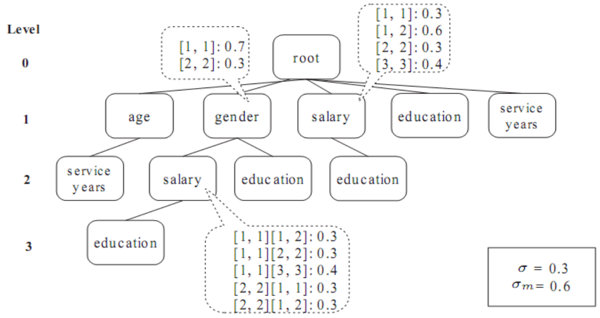


Figure 2. Prefix Tree for $\hat{G}_{MI}$

In this figure, the set of intervals attached with gender are joined that attached with salary. The set of intervals attached with gender is {([1,1]:0.7), ([2,2]: 0.3)} and that attached with salary is {([1,1]: 0.3), ([1,2]: 0.6), ([2,2]: 0.3), ([3,3]: 0.4)}, where the number following colon symbol ":" is the support of the corresponding itemset. The join of gender and salary produces five frequent 2-itemsets. Since these five 2-itemsets have the same attribute set, {gender, salary}, then attach their intervals, ([1,1][1,2]:3), ([1,1][2,2]:3), ([1,1][3,3]:4), ([2,2][1,1]:3) and ([2,2][1,2]:3) with the child node salary of gender. Similarly, the node education is created as the second child of gender, with the set of intervals, {([1,1][3,3]:3), ([2,2][1,1]:3)}, that are obtained by joining the intervals of gender and education.

After the set of frequent itemsets is derived, each frequent itemset is simply mapped into a boolean itemset. Then, the algorithm for Boolean Association Rule generation like Apriori Algorithm can be applied to generate the QARs. By enumerating the cliques in $\hat{G}_{MI}$ with a prefix tree structure, the search space of the frequent itemset computation to the prefix tree representation of all cliques in $\hat{G}_{MI}$ can be limited. The clique enumeration limits the mining process to a smaller but more relevant search space, thereby significantly improving the mining efficiency. This approach ensures that MIC speeds up the mining process for up to orders of magnitudes on both synthetic and real datasets. Most importantly, MIC obtains most of the QARs that have high confidence. This approach ensures that the QARs that are not returned by MIC are insignificant by a formal measure of interestingness for association rules.

6. Conclusion

This approach formalizes the connections between the normalized mutual information, and the support and confidence of QARs. The significance of this approach is twofold. First, this approach guarantee that any pair of attributes pruned by normalized mutual information cannot form a QAR with a confidence greater than the maximum information threshold. Second, this approach ensures that the attributes retained in the MI graph generate QARs with confidence greater than the minimum information threshold. Therefore, this method is not an approximation technique that improves the efficiency at the expense of accuracy.

References

[1] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables", In Proc. of SIGMOD, 1996.

[2] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases", In Proc. of SIGMOD, 1993.

[3] MARÍA N. MORENO, SADDYS SEGRERA, VIVIAN F. LÓPEZ and M. JOSÉ POLO, "A method for mining quantitative association rules". In Porc. of the 6th WSEAS, 2006.

[4] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases", In Proc. of VLDB, 1994, pp. 487–499.

[5] D. Sujatha and Naveen CH, "Quantitative Association Rule Mining on Weighted Transactional Data", International Journal of Information and Education Technology, Vol. 1, No. 3, August 2011.

[6] S. Prakash and R.M.S. Parvathi, "An Enhanced Scaling Apriori for Association Rule Mining Efficiency", European Journal of Scientific Research ISSN 1450-216X Vol.39, No.2, 2010, pp.257-264.

[7] PAURAY S. M. TSAI and CHIEN-MING CHEN, "Mining Quantitative Association Rules in a Large Database of Sales Transactions", Journal of Information Science and Engineering 17, 2001, pp. 667-681.

[8] W. Toshihiko and T. Hirokazu, "Study on Quantitative Association Rules Mining Algorithm Based on Clustering Algorithm", Biomedical Soft Computing and Human Sciences, Vol.16, No.2, 2010, pp.59-67.

[9] Y. Ke, J. Cheng, W. Ng, "Mining Quantitative Correlated Patterns Using an Information-Theoretic Approach", KDD'06, August 20–23, 2006.

[10] C. Shannon, "A mathematical theory of communication". The Bell System Technical Journal, 1948, pp. 379–423, 623–656.