Asian Journal of Information Technology 5 (2) : 169-171, 2006

Discovering Class-bridge Rules Within Conceptual Classes

[1]Ze Qin and [2]Feng Chen
[1]Nanning Prefecture Education College, Nanning, 530000
[2]Department of Computer Science, Guangxi Normal University, Guilin, 541004

Abstract: We design efficient algorithms for mining those association rules that the antecedent and action of such a rule belong to different conceptual classes, referred to class-bridge rules. Class-bridge rules are useful in applications, such as cross-sale in marketing and engraftation in bioengineering.

Key words: Clustering, association rule, class-bridge

INTRODUCTION

There are over 144,000 documents containing the well-known story of I llustrating what is data mining with the great example is the association rule: ‘beer → diaper’, searched by Google. This rule fully displays the power of data mining in discovering those patterns that are comprehensible, surprising, previously unknown, interesting and useful. And such a rule assists in augmenting the sales of these two products through cross-sale. Consequently, it has boosted the data mining research deeply and won huge investments to the development data mining applications. However, other association rules might not be good examples at attracting business investments to data mining. ‘bread ∧ butter → milk’ may be such an example. This is because bread, butter and milk all are everyday foods, and they are frequently purchased together everyday. And this rule should be referred to as commonsense knowledge. It is not believable that data mining techniques can discover new and hidden patterns in databases.

Different from ‘bread ∧ butter → milk’, in ‘beer → diaper’, ‘beer’ and ‘diaper’ belong to two different conceptual classes: drink and tissue. Data marketers must be surprised at the rule because they should never imagine that ‘beer’ does associate with ‘diaper’. And the unexpected rule can make the marketers believe that data mining is worthy of investing. We refer this rule to class-bridge rule. Therefore, it is highly desired to develop new techniques for identifying class-bridge rules in databases.

This poster briefly presents our preliminary strategies for class-bridge rule discovery that are based on the clustering algorithm ‘Chameleon’[1] and the weight of itemsets.

Approach description: Association rule mining is an active research topic in data mining and has been rooted in market basket analysis. By the definition in[1], an association rule, A → B is interesting if it satisfies two constraints: minimum support (minsupp) and minimum confidence (minconf), where A and B are frequent itemsets in a transaction database. In practical applications, the rule A → B can be used to predict that ‘if A occurs in a transaction, then B will likely also occur in the same transaction’, and we can apply this association rule to predict the customer behavior that the presence of A implies the presence of B in marketing. Therefore, mining association rules in databases has received much attention recently[2,3,4].

Association rule mining does not discover the true correlation relationship, because high minimum support usually generates commonsense knowledge, while low minimum support generates huge number of rules, the majority of which are uninformative[3]. Generally a mining algorithm can generate thousands of association rules. Users feel difficult to browse all the rules and say nothing of identifying which rules are really useful in applications. This leaves a large gap in the machinery available to association rule discovery. Mining class-bridge rules in this research assists in filling this large gap.

A class-bridge rule is an implication of the form X → Y, where it satisfies the constraints for association rules of interest; and X and Y belong to two different conceptual classes. Class-bridge rules are useful in applications, such as cross-sale in marketing, engraftation in bioengineering, and chemical synthesis.

In the followings, we describe the main ideas of our two algorithms: agglomeration based algorithm and weighting based algorithm.

Agglomeration based strategy: Agglomeration based method is an agglomerative hierarchical algorithm. We design the algorithm by improving the clustering technique ‘Chameleon’[1]. This algorithm is specifically developed for discovering class-bridge rules in relation databases.

Corresponding Author: Feng Chen, Department of Computer Science, Guangxi Normal University, Guilin, 541004

Chameleon is a hierarchical clustering algorithm. It takes into account inter-connectivity of the clusters as well as closeness of items within the clusters.

Let D be a relation database containing T records with k attributes. The Chameleon discovers those clusters that satisfy constraints: relative inter-connectivity RI(Ci, Cj) and relative closeness RC(Ci, Cj) as follows.

$$\mathrm{RI(Ci,Cj)} = \frac{|\mathrm{EC}_{\{Ci,Cj\}}|}{\frac{|\mathrm{EC}_{Ci}| + |\mathrm{EC}_{Cj}|}{2}}$$

The inter-connectivity between a pair of clusters Ci and Cj denoted by $\mathrm{EC}_{\{Ci,Cj\}}$ is the sum of the weight of the edges that connect vertices in Ci to vertices in Cj. EC_{Ci} is the weighted sum of edges that partition the graph into two roughly-equal parts. This function can deal with the problems of differences in shapes of the clusters as well as differences in the connectivity of different clusters.

$$\mathrm{RC(Ci,Cj)} = \frac{\bar{S}_{EC_{\{Ci,Cj\}}}}{\frac{|Ci|}{|Ci|+|Cj|}\bar{S}_{EC_{Ci}} + \frac{|Ci|}{|Ci|+|Cj|}\bar{S}_{EC_{Cj}}}$$

$\bar{S}_{EC_{\{Ci,Cj\}}}$ is the average weight of the edges that connect vertices in Ci to vertices in Cj. And $\bar{S}_{EC_{\{Ci,Cj\}}}$ is the average weight of the edges that belong to the min-cut bisector of cluster Ci.

To find interactions (class-bridge rules) within conceptual classes, our strategy is based on the following principle. For user specified thresholds: $T1_{RI}$, $T2_{RI}$, $T1_{RC}$, and $T2_{RC}$, the agglomeration between (two groups) C_i and C_j is measured by

$$T2_{RI} < RI(C_i, C_j) < T1_{RI} \wedge T2_{RC}(C_i, C_j) < RC(C_i, C_j) < T1_{RC}$$

The algorithm mainly includes three steps as follows:

- Constructing a K-nearest neighbor graph;
- Partitioning the graph to get the sub-clusters ready to be merged; and
- Merging sub-clusters to get the final clusters.

The improvements on the CHAMELEON algorithm are as follows.

- Using new thresholds: $T1_{RI}$, $T1_{RC}$, $T2_{RI}$, and $T2_{RC}$, for measuring the agglomeration when merging two groups (sub-clusters) C_i and C_j in step (3), where $T1_{RI}>T2_{RI}$, $T1_{RC}>T2_{RC}$.
- Recording the similarity between any two objects across classes pair by pair, so as to identify interactions of interest between objects. These interactions are the class-bridge rules what we want.

Weighting based strategy: This strategy is a post-process specifically designed for mining class-bridge rules in transaction databases. Our weighting based algorithm takes into account some properties of the Apriori algorithm and the features of class-bridge rules. The algorithm mainly includes three steps as follows:

- Identifying frequent itemsets across classes;
- Finding correlative itemsets on the basis of frequent itemsets; and
- Calculating the importance of itemsets.

Our weighting based strategy uses the Chi-squared test to judge the correlation of itemsets. The premise of chi-squared test is to construct the null hypothesis that all items are independent and the preparation hypothesis that all items are not. The null hypothesis is though to be wrong when the chi-squared value is higher than a cutoff value. So we reject the null hypothesis and accept the preparation hypothesis, i.e., items are correlated. Generally in statistics, this cutoff value is 3.84 at the 95% significance level, which our algorithm conforms to.

There may be one more class-bridge rules between two conceptual classes. We identify class-bridge rules of interest based on the importance 'IMPOR' of itemsets that generate these class-bridge rules. The importance impor is defined as follows.

Let support(S) be the support of itemset S, the chi-squared value X^2 (S), where S = $\{i_1, i_2, \ldots, i_m\}$ and the weight of i_j be w_j. We define

$$\mathrm{Impor(S) = support(S)} * X^2(S) * \max\{w_1, w_2, .., w_j\}$$

The weight of an item can be determined by, for example, its sales.

EXPERIMENTAL EVALUATION

To evaluate our algorithms, we have conducted extensive experiments on a DELL Workstation PWS650 with 2G main memory, 2.6G CPU, and WINDOWS 2000.

One of datasets used for our agglomeration based algorithm is the Zoo from the Internet which includes 101 records. Every record includes 18 attributes. For the sake of identifying class-bridge rules, we delete two attributes: name and category. Our algorithm discovers 2 class-bridge rules: 'seasnake → bass' and 'seasnake → catfish'. The experiment results have shown that our agglomeration based algorithm can not only learn the clusters, but also identify class-bridge rules desired efficiently (Fig. 1).

In addition to this mentioned above, we have performed comprehensive experiments on CMC database

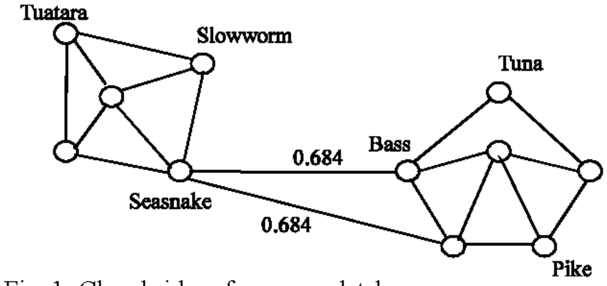

Fig. 1: Class bridges from zoo database

Table 1: Items and classes

Male aids	cigarette, lighter, western-style clothes, beer, tie
Female aids	pregnant woman’s clothes, perfume, lipstick, hairpin, high-heeled shoes, silk scarf, earring
Baby aids	Toy, diaper, nursing bottle

and mushroom database from the Internet. The class bridges we expected have bee found too. Moreover, we found that bridges would be different with different clustering.

One of datasets used for our weighting based algorithm is supermarket basket data from the Synthetic Classification Data Sets on the Internet. This dataset includes 50 transactions with 15 items that are partitioned into 3 classes (Table 1). Our algorithm discovers 5 class-bridge rules. From the results, these class-bridge rules really look like ‘beer → diaper’. For practical applications, all class-bridge rules identified are ranked by their importance.

Some interesting class bridges have been found:

- cigarette, lipstick, hairpin
- cigarette, lipstick, high-heeled shoes
- lighter, lipstick, hairpin
- beer, diaper
- perfume, western-style clothes

Agglomeration based algorithm is suitable for mining relationship database. It can get class bridges without increasing the cost of time and space. Further more, the reasons of generating class bridges could be explored for every record including some attributes. Weighting based algorithm is a good choice for transaction database. And in fact it can be considered to be a kind post-process.

CONCLUSIONS

Rules that the antecedent and action of such a rule belong to different conceptual classes are referred to class-bridge rules. Class-bridge rules are useful in applications, such as cross-sale in marketing and engraftation in bioengineering. In this study, we design two kinds of algorithms to find class-bridge which are agglomeration based and weighting based. Our experiments illustrates that they are efficient.

REFERENCES

1. Karypis, G., E. Han and V. Kumar, 1999. Chameleon: A hierarchical clustering algorithm using dynamic modeling. IEEE Computer, pp: 68-75.
2. Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association rules between sets of items in large databases. SIGMOD: pp: 207-216.
3. Kim, W., Y. Lee and J. Han, 2004. CCMine: efficient mining of confidence-closed correlated patterns. PAKDD: pp: 569-579.
4. Wu, X., C. Zhang and S. Zhang, 2004. Efficient mining of both positive and negative association rules. ACM Trans. Inf. Sys., 22: 381-405.