F-055

Contrasting Correlations Based on Double-Clique Search

Aixiang Li † Makoto Haraguchi †

Abstract

The studies on contrast set mining and correlation mining have been paid much attention in this decade. In this paper, we consider a problem of contrasting correlations measured by *k*-way mutual information. As the cost for enumerating item sets and computing their mutual information is high, we introduce a new graph theoretic technique to cut off useless itemsets. In fact, we make two graphs: one graph represents implicitly correlated pairs in one database, and the other shows mediately correlated pairs in another database. By using the graphs with the different types of edges, we experimentally show that the problem of contrasting correlations can be effectively solved.

1. Introduction

The study of correlation mining [1,2] has pointed out a simple fact that association rules are sometimes misleading from a statistical point of view. Therefore, Brin[1,2] has proposed a notion of correlation mining where the targets are correlated itemsets that can reject null hypotheses of independence based on the chi-square statistics. This paper also regards correlated itemsets, where, given two databases to be compared, we contrast correlations over the databases and try to find correlation changes among the two. It should be noted here that, unlike the preceding study of correlation change [6] in which only a positive correlation is regarded, our target itemsets may show negative correlation and positive one as well. Thus, each item is interpreted as a random variable throughout this paper.

The standard measure, correlation coefficient, is often used for correlation mining, as reported in [7]. However, correlation coefficient is too simple to measure the degree of correlatedness under some conditioning. The chi-square can be a candidate measure for the correlatedness even for such a case. However we have to compare two chi-square values over two databases with different sizes. For this reason, we measure the degree of correlation among itemsets by *k*-way mutual information so that the difference of database sizes does not make an influence on the values. In what follows, the term 'item' means a 'random variable taking Boolean values'.

The two databases are denoted as DB_1 and DB_2. Our goal is to find itemsets whose degree of correlation is low in DB_1 but is medium in DB_2. Particularly, negative correlations are also regarded. Both positive and negative correlations over two items, A and B, are needless to say calculated by the standard mutual information I(A;B). However correlation given some condition will be much more interesting. As is well known, the 3-way extended mutual information, I(A;B;C), shows higher value when there exists positive or negative dependence between A and B given the conditioning by the third variable C, even when there exists no dependency among paired items. This fact is easily verified by the formula:

$I(A;B;C) = I(C;A) + I(C;B) + I(A;B|C);$

where I(A;B); I(A;B;C) are mutual information and conditional mutual information, respectively. We use I(A;B;C) to measure the correlatedness among three items. When we have an itemset of more than three items, we use k-way information shown by the equation in 2.1.

Based on this correlation measure, our task is to find the itemsets whose correlation at DB_1 is low and the correlation at DB_2 is medium. As the number of all itemsets in the itemset lattice is very huge, we have to develop some pruning rules to reject useless ones. It is also a well known fact that the k-way mutual information increases monotonically as the itemsets grow to larger sets. More precisely, if $I(i_a,i_b)>\delta$ and then there must be $I(i_a,i_b,i_c) >\delta1$. Based on the property, to reduce the candidate sets that violate the constraints on the correlation in DB_1 (δ_1, implicitly) or DB_2 (δ_2, mediately), we construct two graphs G_1 (i_a,i_b connected, if I_{DB1} $(i_a,i_b)< \delta_1$) for DB_1 and G_2 (i_a,i_b connected, if I_{DB2} $(i_a,i_b)< \delta_2$)for DB2. And then we searched the double-cliques (cliques in both G1 and G2).

Another issue in considering the mutual informaion is the frequency problem. That is, highly frequent or lowly frequent items are not informative. Therefore we set standard support conditions, either given by minsup or maxsup. Another issue about the supports, discussed by Brin, has to be addressed to our case as well, by the nature of k-way mutual infomation. When an itemset has k items, the corresponding joint distribution tends to have very small probability values. When the number of events whose probability values are very low is large, it is often the case that some variables are redundant and the combination of k items is of no use. Therefore, we applied a refined form of support constraint based on Brin's.

By applying the pruning technique mentioned as in the above, we experimentally show the contrasting correlation problem can be solved efficiently.

2. Methods

2.1 *k*-way mutual information

Itemset $X = \{i_1,i_2,\ldots,i_k\}$, i_j is considered as a binary variable with *'0'* and *'1'* values, the correlation of the set of items is defined as *I(X)*:

$$\sum_{j=1}^{k}\sum_{v_j=0}^{1} p(i_1 = v_1,\ldots,i_k = v_k)\log\frac{p(i_1=v_1,\ldots,i_k=v_k)}{p(i_1=v_1)\ldots p(i_k=v_k)}$$

p: the support in percent.

It is proved that $I(sup(X)) \geq I(X)$, *sup(X)*: supersets of *X*

2.2 Support constraint:

For a set X= $\{i_1, i_2, \dots, i_k\}$, there are 2^k support values for 2^k kinds of events, for example, corresponding to $\{ia, ib, ic\}$, there are 8 supports: $p(ia=1, ib=1, ic=1)$, $p(ia=1, ib=1, ic=1)$, $p(ia=0, ib=1, ic=1)$, $p(ia=0, ib=0, ic=1)$,$p(ia=1, ib=0, ic=0)$,$p(ia=1, ib=0, ic=0)$, $p(ia=0, ib=1, ic=0)$, $p(ia=0, ib=0, ic=0)$. To avoid more rare events (noises) and ensure efficient partition by adding items, we applied a varied support constraints: when the 2^k supports are partitioned into *1* and *0* halves by $k+1_{th}$ item, at each half, at least *p%* supports of the *k+1* itemset are equal or greater than *minsup*. In the above example, if we set *p% >25%*, at least 2 supports have *minsup* in both $p(, , ic=1)$ and $p(, , ic=0)$ half . The support constraint is upward closed, and then it is used as an efficient pruning rule.

2.3 Problem definition

Input:

DB_1, DB_2, correlation constraints: δ_1 in DB_1, δ_2 in DB2, and $\delta_2 > \delta_1$, increase rate: *β%*, support constraint: *p%* and *minsup.*

Output:

Itemsets *X*, s.t. $I_{DB1} < \delta_1$, $I_{DB2} < \delta_2$ and $(I_{DB2} - I_{DB1}) / I_{DB1} >= \beta\%$

2.4 Double graph construction

Based on the property: $I(sup(X)) \geq I(X)$,

if $I(\{ia, ib\})$ violate the constraint δ_1 (or δ_2), $I(\{ia, ib\}) > \delta_1$ (or δ_2), then $I(\{ia, ib, ic\}) > \delta_1$ (or δ_2).

Graph G_1 for DB_1: edge *(ia,ib) is drawn*, if $I_{DB1}(ia, ib) < \delta_1$,

Graph G_2 for DB_2: edge *(ia,ib) is drawn*, if $I_{DB2}(ia, ib) < \delta_2$,

2.5 Double clique search

double-clique: clique in G1 and also cliques in G2.

Based on the standard depth-first clique search algorithm, we developed the double-clique search method. From the result of experiments, this strategy reduces the candidates for extending branch substantially

The method is illustrated in Figure 1.

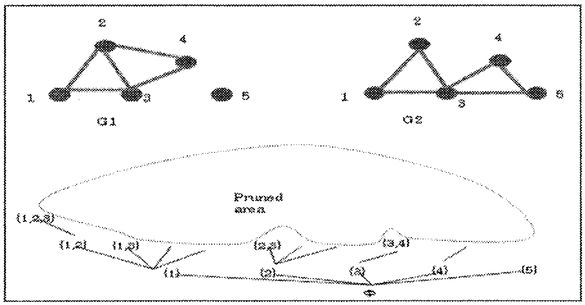

Figure1: The searched area is reduced more by double-clique search

3. Results

Experiment data: Nikkei convenience store Pos data,

DB_1: One month transactions of one store in Dec 2007,

DB_2: One month trans. of the same store in Nov 2008.

Language: Java,

Machine: CPU: intel core2 duo E8500, 3.16GHz,
3.25GB RAM

One of the outputted itemset : { *Kent super lighter5, JT seven star soft 20, JT seven star box 20* }

To check the efficiency of double-clique search, we compared with No-clique search (standard itemset enumeration) and Single-clique search (one DB clique enumeration) at different δ_1 and δ_2 (other parameters: *p% = 25%*, minsup = 0.01, *β% =20%*) . The computation time is showed in Figure 2.

4. Discussion

The focus of our research was to contrast the correlation of a group of items (or attributes) and search the itemsets whose items are correlated (positively or negatively) implicitly in the first dataset and mediately correlated in the second datasets. By our developed method, the significant itemsets were extracted at different correlation constraints. Their items are correlated positively or negatively.

The results of experiment showed that double-clique search method is applicable for searching this kind of itemsets.

Because the number of the outputted itemsets is some large, it is difficult to analyze which itemset is interesting to users. In the future work, we will apply additional strategy (e.g Top N) to improve the understandability of the result based on the double-clique method.

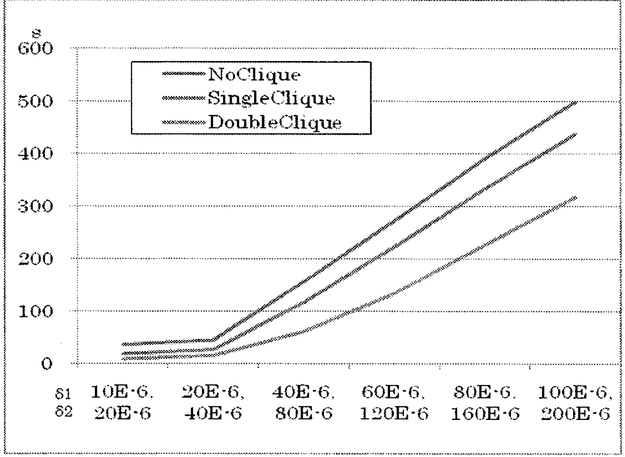

Figure 2: Double-clique method is more efficient than the other two methods

Reference

[1] S. Brin, R.Motwani and C.SilverStein, "Beyond Market: Generalizing Association Rules to Correlations", Proc. of ACM/SIGMOD'97, pp.265-276, 1997.

[2] C.SilverStein, S. Brin and R.Motwani, "Beyond Market Baskets: Generalizing Association Rules to Dependence Rules", Data Mining and Knowledge Discovery,2, pp.39-68, 1998.

[3] S.D.Bay, and M.J. Pazzani, "Detecting Change in Categorical Data: Mining Contrast Set", Proc. of the fifth ACM and SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.302-306, 1999.

[4] S.D.Bay, and M.J. Pazzani. "Dectecting Group Difference: Mining Contrast Set," Data Mining and Knowledge Discovery, 5(3), pp. 213-246, 2001.

[5] T.Taniguchi, "A Study on Correlation Mining Based on Contrast Set", Doctoral Dissertation, Graduate School of Information Science and Technology, Hokkaido University, 2007.

[6] X. Wu, Y.Ye, K.R. and Subramanian, "Interactive Analysis of Gene Interactions Using Graphical Gaussian Model", Proc. of BIOKDD03, 2003.

[7] H.Xiong, W.Zhou, M.Brodie and S.Ma "Top-k ΦCorrelation Computation", Informs Journal on Computing , Vol. 20, No. 4, pp. 539-552, Fall 2008.

P.K.N. Nada Lavrač, G.I. Webb, "Supervised Descriptive Rule Discovery: a unifying survey of contrast set, emerging pattern and Subgroup mining", Journal of Machine Learning Research,10, pp.377-403, 2009.