Contrasting Correlations Based on Mutual Information

Aixiang Li Makoto Haraguchi Yoshiaki Okubo

Graduate School of Information Science and Technology, Hokkaido University

Contrasting linearly ordered datasets by correlation mining is well-studied. This paper proposes a method to search out Top N itemsets with significant change of correlation on two datasets under support constraints. We consider an item as a variable taking values of absence and presence. The correlation is defined as mutual information among the items of an itemset. The mutual information consists of positive and negative dependence of items. It is a convex function to the joint probability and the product of marginal probabilities of items. It takes maximum value at the extreme point of feasible region. By contrasting the maximum value of mutual information of an itemset and its supersets on the two datasets, we can find out aimed itemsets efficiently.

1. Introduction

Detecting a change pattern from a series of datasets of some event has been studied extensively in recent years [Bay 1999, Bay 2001, Taniguchi 2007]. One way of exploring the pattern is tracking the correlations of some attribute-values (called *items*) and contrasting them in the linear order. The correlation represents the positive dependence (items co-occur) and negative one (items occur exclusively) among items. The mined out itemsets with significant change of correlation are informative for people to understand the development of the investigated event definitely. For example, in the analysis of basket data, we found that item A and B were purchased almost independently in last years because their correlation was very lower. But in recent years, the two items are purchased together more frequently. Thus, their correlation becomes higher. From the mined result, store managers would make a new adaptable sales plan.

About the relation between items, it has been generalized from association rule to correlation covering the dependence of absence and presence of items [Brin 1997]. Brin et al. applied chi-square statistic to judging whether items in an itemset are correlated or not in a dataset. For comparing several datasets, Dong and Li. have proposed emerging pattern mining [Dong 1999]. They compared the support of an itemset in two datasets. Extensively Bay and Pazzani developed contrast set mining on a group of datasets [Bay 1999, Bay 2001]. They aimed at itemsets with significant difference of support on the categorical datasets. To contrast the correlation through several datasets, Taniguchi has given a proposal of mining pair of itemsets that have higher change of correlation in two datasets [Taniguchi 2007]. But he contrasted only the positive dependence (co-occurs) of itemsets.

Naturally the complete relation of items should include both positive and negative dependence of them. Extensively not limited to judgment that items are correlated or not, in this paper, we give an evaluation on the degree of correlation among items based on mutual information. Further, we contrast the correlation of an itemset through all the given datasets to detect a change pattern. Additionally, to avoid more frequent and specific itemsets, we give a constraint on their support. In result, the number of itemsets is numerous, mining out all the possible itemsets is not a feasible execution; therefore, only Top N itemsets with higher correlation are extracted in our proposal.

In the present study, we consider an item as a variable with two values of presence (1) and absence (0). And then we defined the mutual information among the variables (two or more). As we know, the mutual information can be considered as the KL divergence between the actual joint probability and the supposed independent probability. Basing on the convex property of KL divergence, we proposed a heuristic method of searching Top N itemsets efficiently. By this method, firstly, we will contrast two datasets in this paper.

Contact: Aixiang Li, Makoto Haraguchi, graduate school of information science and technology, Hokkaido University, N-14, W-9, Sapporo 060-0814, Japan. Phone: 011-706-7575, Fax: 011-706-7161, E-mail address: aixiang@kb.ist.hokudai.ac.jp, mh@ist.hokudai.ac.jp

2. Correlation based on mutual information

In this paper, *item* means a sales item in a store (e.g coffee in basket data) or an attribute-value pair (e.g age >20 in census data). Let S be the itemset including all items in a dataset *DB*, $Z = \{i_1, i_2, \cdots, i_m\} \subseteq S$ be a set of m items in itemset space, $cand = \{c_1, c_2, \cdots, c_k\} \subseteq S$ the candidate set of Z under the constraint of support: $minsup <= Support$ (Z or >Z) <= $maxsup$, $>Z = \{Z \cup sc_i \mid sc_i \subseteq cand\}$, called superset of Z.

Mutual Information: every item is considered as a *variable* in statistic. It has two values 1(presence) or 0 (absence). For the itemset including two items, for example, $Z = \{i_1, i_2\}_{m=2}$, the mutual information of the set is defined as follows:

$$I(Z_{m=2}) = \sum_{i_1=0,1} \sum_{i_2=0,1} p(i_1, i_2) \log_2 \frac{p(i_1, i_2)}{p(i_1)\,p(i_2)} = D(p(i_1, i_2) \,\|\, p(i_1)\,p(i_2)) \quad (1)$$

D: KL divergence.

Extensively, for the itemset including three or more items, the mutual information can be defined as:

$$I(Z_{m\geq 3}) = \sum_{i_1=0,1}\cdots\sum_{i_m=0,1} p(i_1,i_2,\ldots i_m)\log_2\frac{p(i_1,i_2\ldots,i_m)}{p(i_1)p(i_2)\ldots p(i_m)}$$
$$= D(p(i_1,i_2\ldots,i_m)\,\|\,p(i_1)p(i_2)\ldots p(i_m)) \qquad (2)$$

In abbreviation, $p = p(i_1, i_2,\cdots, i_m)$, $q = p(i_1)\ p(i_2)\ \ldots\ p(i_m)$, thus $D(p(i_1, i_2,\cdots, i_m)\,\|\,p(i_1)\,p(i_2)\ \ldots\ p(i_m)) = D(p\|q)$.

We notice that p and q must decrease as we extend branches from an itemset Z with its candidate set *cand*. Therefore (p,q) has the upper bound (p_u,q_u)={ $p(i_1, i_2,\cdots, i_m)$, $p(i_1)\ p(i_2)\ldots\ p(i_m)$} and the lower bound (p_l,q_l)={ $p(i_1, i_2,\cdots, i_m, c_1,c_2,\cdots,c_k)$, $p(i_1)\ p(i_2)\ldots\ p(i_m)\ p(c_1)\ p(c_2)\ldots\ p(c_k)$}.

Important Property of Mutual Information: mutual information can be represented by KL divergence between p and q. It is proved that $D(p\|q)$ is a convex function of (p,q). Therefore it takes the maximum value at an extreme point, which is a corner of the feasible region of (p,q). The corners are four (2^2) combinations of the upper bound and lower bound of p and q. That is:

$$D_{\max} = \max_{\substack{p\in\{p_u,p_l\}\\ q\in\{q_u,q_l\}}} D(p\,\|\,q) \qquad (3)$$

It is illustrated by a space of two dimensions (Figure 1).

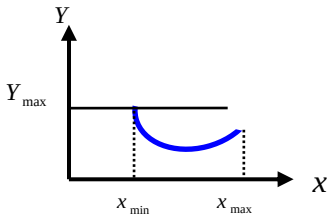

Figure 1: Y(x) takes maximum value at bound of *x*

3. Pruning rules

Given two datasets *DB1* and *DB2*, and support constraint: {*minsup*, *maxsup*}, Heuristically, we search only top N itemsets *W* with higher correlation in the first dataset and also $|I_{DB2}(W)$-$I_{DB1}(W)|>=$ **β** ($I_{DBi}(W)$: the mutual information of itemset *W* in *DBi*, *i*=1 or 2 here).

Basically a branch and bound algorithm is applied. For the super extended space (include *Z* and >*Z*) of itemset *Z* with *cand*, the maximum value of the correlations of the sets is known (Figure 2).

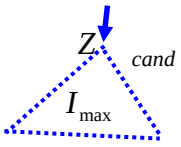

Figure2: the super space of Z, the I_{max} is known.

Pruning rule 1: under the support constraint of itemsets, the candidates of an itemset are selected, thus the number of extended branches is reduced.

Pruning rule 2: in one dataset, only Top N itemsets with higher correlations are captured. For an itemset Z and its *super space*(the space of superset), if $I_{max}(Z,>Z)<=$ the smallest values in Top N list, the search space will not be extended.

Pruning rule 3: to contrast two ordered datasets, in this paper, we only aim at the itemsets that have an increase of correlation, $|I_{DB2}(W)$- $I_{DB1}(W)|>=$ **β** . As branches of the search tree are extended, the correlations of itemsets are calculated in two datasets correspondingly and compared. For itemset *Z* and >Z, if $I_{DB2}(Z, >Z)< =I_{DB1}(Z, >Z)$, the Z and its super space are pruned heuristically, because of the lower possibility of the increasing of correlation.

By the algorithm with defined pruning rules, the Top N itemsets with a significant increasing of correlations are searched out under the support constraints.

4. Concluding remarks

Contrasting the correlation of items in a set is another helpful tool for detecting a change pattern of the development of an event. In this study, we proposed a measure of correlation that consists of negative and positive dependence based on the information theory. Not limited to the difference of correlation in different datasets, we aimed to find out a trend of continuous increasing (or decreasing) of correlation of some itemsets.

To expose the change of correlation in essence, we kept the support in a relatively stable interval. Instead of all the possible itemsets, we only search out the Top N ones efficiently.

Based on the convex property of mutual information, we found the maximum value of correlation in a subspace of all itemsets space; therefore, we created some heuristic pruning rules to improve the efficiency of the algorithm.

As discussed in [brin 1997], when the number of items in a mined out itemset becomes larger, the correlated items in a set are not easy to be understood. In this situation, we need to propose an efficient way for the partition of itemsets.

In this paper, we compared two datasets basically. For a series of linearly ordered datasets, comparing two datasets step by step is a basic method. We will extend the method in the next work.

In this paper, we have not presented a result of experiments. In the next step we will implement the proposed using the experimental data. Basing on results of experiments, we will improve the method.

References

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

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

[Dong 1999] Dong G. and Li J.: Efficient Mining of Emerging Patterns: Discovering Trends and Differences, In Proceedings of the fifth ACM and SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.43-52, 1999.

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

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