Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Engineering Applications of Artificial Intelligence 19 (2006) 419–428
Engineering Applications of
ARTIFICIAL
INTELLIGENCE
www.elsevier.com/locate/engappai

Discovery of hidden correlations in a local transaction database based on differences of correlations

Tsuyoshi Taniguchi*, Makoto Haraguchi

Division of Computer Science, Hokkaido University, N-14 W-9, Sapporo 060-0814, Japan

Received 5 January 2006; accepted 5 January 2006
Available online 7 March 2006

Abstract

Given a transaction database as a global set of transactions and its local database obtained by some conditioning of the global database, we consider pairs of itemsets whose degrees of correlation are higher in the local database than in the global one. A problem of finding paired itemsets with high correlation in one database is already known as discovery of correlation, and has been studied as the highly correlated itemsets are characteristic in the database. However, even noncharacteristic paired itemsets are also meaningful provided the degree of correlation increases significantly in the local database compared with the global one. They can be implicit and hidden evidences showing that something particular to the local database occurs, even though they were not previously realized to be characteristic. From this viewpoint, we have proposed measurement of the significance of paired itemsets by the difference of two correlations before and after the conditioning of the global database, and have defined a notion of DC pairs, whose degrees of difference of correlation are high. In this paper, we develop an algorithm for mining DC pairs and apply it to a transaction database with time stamp data. The problem of finding DC pairs for large databases is computationally hard in general, as the algorithm has to check even noncharacteristic paired itemsets. However, we show that our algorithm equipped with some pruning rules works successfully to find DC pairs that may be significant.

Keywords: Data mining; Correlation mining; Itemset pairs; Contrast sets; Time series

1. Introduction

In studies of data mining from transaction databases, many researchers have concentrated on finding itemsets with high support, i.e., paired itemsets appearing in association rules with high confidence (Agrawal and Srikant, 1994), or paired itemsets with strong correlation (Aggarwal and Yu, 1998; Brin et al., 1997a,b; Morishita and Sese, 2000). These concepts are considered useful for distinguishing characteristic paired itemsets with strong correlation in a single transaction database. A similar strategy, based on the concept of change of support, known as emerging patterns (Dong and Li, 1999), is successful for finding itemsets characterizing either of two databases. All of these concepts regarding itemsets are therefore proposed to extract paired itemsets with required characteristics in a given database or in one of two or more databases.

Some characteristic paired itemsets with high correlation are useful for finding a general tendency. However, there may exist such paired itemsets with high correlation independently of time stamps and multiple times. Some users may regard this as trivial because they already know the relation. On the other hand, as is indicated in the study of chance discovery (Ohsawa and Nara, 2002), some itemset pairs that are not characteristic may also be useful because they are *potentially significant* under some conditions.

For example, a supermarket manager who tries to set up shop in an urban area may wish to know a general tendency for the area. In this case, a general knowledge about the standard goods in the area or about a specified

*Corresponding author.
E-mail addresses: tsuyoshi@kb.ist.hokudai.ac.jp (T. Taniguchi), makoto@kb.ist.hokudai.ac.jp (M. Haraguchi).

0952-1976/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2006.01.006

customer is useful for the success of his or her shop. However, when the number of customers begins to increase, the manager cannot meet the needs of the customers by using only the previous knowledge gained. That is, the manager has to react responsively to the changing needs of their customers. However, the needs are not always characterized in the process at the beginning of a change in tendency; rather, in many cases, the needs may not appear as a characteristic one.

To detect such uncharacteristic itemset pairs as *implicit evidence*, we consider an itemset pair as soon as it is seen that its correlation becomes high. If it continues to show high correlation after the change, it can be found by some existing method. On the other hand, it may not always show high correlation if the relation between the itemsets begins to change. As a result, the itemset pair with a large change of correlation but still with a lower correlation is hidden by itemset pairs with high correlation. It is worth marking such itemset pairs as implicit evidences showing that something occurred in recent time because their changes of correlation are very high.

Therefore, in this paper, we wish to find itemset pairs with large changes of correlation by conditioning some time stamp compared with a global database with all time stamps. To find such itemset pairs, a notion of DC pairs, which we have already defined in (Taniguchi et al., 2005), can be used. A DC pair is a pair of itemsets such that their degree of difference in correlation before and after some conditioning is very large. We have also improved the algorithm in (Taniguchi and Haraguchi, 2005). Using these algorithms, we try to find useful DC pairs from a family of databases with time stamp data.

In the experiment described in the latter section, we examine census data (Ruggles et al., 2004) for 1980, 1990, 2000. In order to simplify our problem, we consider a database containing two time stamps. That is, the 1980 census is combined with the 1990 census, and the global database is the combined databases, the 1980–1990 census. Similarly, the 1990 census is combined with the 2000 census as a global database, the 1990–2000 census. A condition is given the latter time stamp; i.e., the 1990 census in the combined 1980–1990 census and the 2000 census in the combined 1990–2000 census. We show that useful itemset pairs, which may be potentially significant, can be found using the algorithm for finding DC pairs in each combined database.

Generally, it is computationally a hard task to find DC pairs, because the degrees of difference of correlations are *nonmonotonic* with the expansion of the itemsets. For this reason, we consider a restricted problem, given two parameters, ζ and ε. More precisely, we evaluate the degrees of difference between correlations by a function defined with ζ and ε, and restrict the DC pairs we are endeavoring to find. Note that a DC pair is syntactically regarded as a compound itemset consisting of two component itemsets. Although various means of mining DC pairs may be considered, we first find candidates for component itemsets, then combine one component with another component, based on our algorithm in (Taniguchi and Haraguchi, 2005). In the first phase, the search space for finding the candidates can be reduced by using a pruning rule depending on ε.

On the other hand, in the second phase, we generate a space of paired itemsets by combining candidate component itemsets found in the first phase. As the size of database increases, the number of possible components and their possible combinations becomes larger. So, we need to have some pruning rules to cut off useless combination. For this purpose, we use a simple fact that all the combined itemsets as well as their component itemsets must have non-zero support in the local database. We present a sufficient condition, stated in terms of cooccurrences of items, for combinations violating the support condition to be excluded as useless ones.

This paper is an extended version of our work presented at the International Conference on Data Mining and Machine Learning in Pattern Recognition MLDM 2005 (Taniguchi et al., 2005). In Section 2 we describe related work. In Section 3 we define some of the terminology used throughout this paper. In Section 4, we introduce the concept of DC pairs and define our problem of mining DC pairs. An algorithm for finding DC pairs is described in Section 5. Section 6 presents our experimental results. In the final section, we summarize our study.

2. Related work

Many studies in the field of data mining are based on a strategy of contrasting two or more databases in order to extract significant properties, or patterns, from a huge data set. Some particular data mining techniques, known as contrast-set mining (Bay and Pazzani, 2001; Dong and Li, 1999; Webb et al., 2003), have been designed specifically to identify differences between databases to be contrasted.

For example, in the study of emerging patterns (Dong and Li, 1999) for two transaction databases, itemsets whose supports are significantly higher in one database compared to another are considered significant, as they can be candidate patterns for distinguishing the former from the latter. A similar strategy is also used in STUCCO (Bay and Pazzani, 2001) to obtain characteristic itemsets in one database based on the χ^2 test. In addition, Magnum Opus (Webb et al., 2003) examines relations between itemsets and a database from several databases. On the other hand, this paper seeks paired itemsets whose correlations radically increase in one database. Therefore we can say that the subject of this paper is a variety of "contrast-set mining of correlations between itemsets".

Second, many methodologies have been proposed to detect characteristic correlations in a single database (Aggarwal and Yu, 1998; Brin et al., 1997a,b). In those studies, strongly correlated itemsets in a given database, or in one database from two given databases, were examined