SELF-TUNING CLUSTERING: AN ADAPTIVE CLUSTERING METHOD FOR TRANSACTION DATA

HALEY NUGENT, TAYLOR HEILMAN

1. Introduction

Database mining has wide applications for improving marketing strategies. To improve these marketing strategies, we use data clustering. Data clustering divides a set of data items into separate groups such that items in the same group are as similar to one another as possible. Clustering large datasets can uncover useful patterns among the data.

Market-basket data is known to have high dimensionality, sparsity, and to have massive outliers. The *small-large (SL) ratio* is the ratio of the number of small items to large items in the data. A *large item* is an item which occurs frequently in transactions. A *small item* is an item that occurs infrequently in transactions. Smaller SL ratios indicate more similarity between the items in the cluster. This paper develops a *Self-Tuning Clustering algorithm (algorithm STC)* to efficiently cluster the market-basket data by adaptively tuning the SL ratio. The algorithm consists of three phases:

(1) *pre-determination*: calculates the *minimum support S* and the *maximum ceiling E* according to a given parameter called *SL distribution rate β*.
(2) *allocation*: Algorithm STC uses the minimum support S to identify the large items. It uses the maximum ceiling E to identify the small items. It accomplishes this by scanning the database and allocating each transaction to a cluster for minimizing the SL ratio.
(3) *refinement*: Each transaction is evaluated to minimize its SL ratio in its corresponding cluster.

The algorithm uses two different kinds of SL ratio thresholds to evaluate the quality of the clustering, *output SL ratio threshold α^o* and *input Sl ratio threshold α^i*. A transaction is moved from one cluster to the excess pool if its SL ratio is larger than α^o and moved from the excess pool to one cluster is the SL ratio is smaller than α^i.

Algorithm STC significantly improves the clustering quality for synthetic and real market-basket data.

2. Problem Description

The market-basket data is represented by a set of transactions $D = t_1, t_2, ..., t_h$, where D is the database holding the set of transactions. Each transaction t_i is a set of items $i_1, i_2, ..., i_h$. A clustering $U =< C_1, C_2, ..., C_k >$ is a partition of transactions where C_i is a cluster consisting of a set of transactions. The minimum support S and the maximum ceiling E are determined according to the SL distribution rate β in the pre-determination phase.

2.1. Large Items and Small Items.

$Sup_C(i)$ is the support of an item i in a cluster C. If $Sup_C(i)$ is larger than the minimum support S, the item i in a cluster C is called a *large item.* If an item j in a cluster C has a $Sup_C(j)$ that is smaller than the maximum ceiling E the item j is called a *small item.* An item is called a middle item if it is neither large or small.

2.2. Small-Large (SL) Ratio.

There are three kinds of SL ratios that need to be calculated in the data clustering procedure.

(1) **SL Ratio of a Transaction:** The SL ratio for a transaction t in cluster C_i is defined as:

$$R_{SL}(C_i, t) = \frac{|S(C_i, t)|}{|L(C_i, t)|},$$

$|S(C_i, t)|$ represents the number of small items in t and $|L(C_i, t)|$ represents the number of large items in t.

(2) **SL Ratio of a Clustering:** The SL ratio for a clustering $U =< C_1, ..., C_p >$ is defined as:

$$R_{SL}(U) = \sum_{i=1}^{p} \sum_{j=1}^{N^T(C_i)} R_{SL}(C_i, t_j),$$

$N^T(C_i)$ is the number of transactions and t_j is the jth transaction in the cluster C_i.

(3) **Average SL Ratio:** The average SL ratio for a clustering $U =< C_1, ..., C_p >$ is defined as:

$$\alpha(u) = \frac{R_{SL}(U)}{N^T(U)}$$

$N^T(U)$ is the number of transactions in clustering U.

2.3. The Objective of Clustering Market-Basket Data.

The objective clustering market basket data is *Given a database of transactions, determine a clustering U such that the average SL ratio $\alpha(U)$ is minimized.* The large items in each cluster are the products which are sold frequently. The clustering technique aims to maximize both the *intra-cluster similarity* and the *inter-cluster dissimilarity* of the data in order to minimize the average SL ratio.

Intra-Cluster Similarity: Intra-cluster similarity of transactions is achieved by maximizing the number of large items in each cluster while minimizing the number of small items in each cluster. We know that an item i is large if there are relatively many transactions containing i, while an item j is small if there are relatively few transactions containing j. Transactions are similar to one another if the contain many common large items and fewer small items.

Inter-Cluster Dissimilarity: We achieve inter-cluster dissimilarity of transactions by maximizing the number of large items and minimizing the number of small items so dissimilar transactions will be allocated to the different clusters. If an item i is large in a cluster C_a, i should be small in cluster C_b.

3. Conclusion

Algorithm STC utilizes a self-tuning technique for adaptively tuning the input and output SL ratio thresholds to minimize the SL ratios of transactions in the clusters efficiently.