Abstract

Data mining is a process to find an interesting pattern from very large data warehouse. In data mining, many tasks can be done. Begin from classification, clustering and association. This paper discuss about clustering categorical data using LIMBO (scaLable InforMation Bottleneck) method.

Clustering is the process of grouping objects into a group (cluster) so that the object has a very great similiarity with other objects that are on the same cluster, but has a great dissimiliarity with objects that are in different clusters. Clustering has been extensively implemented in many fields such as market research, pattern recognition, customer segmentation etc. The problem of clustering becomes more challenging when the data is categorical, that is when there is no inherent distance measure between data values. Moreover many clustering algorithm take a long time so it is not suitable for large amount of data.

LIMBO clustering method uses tree-structure for the purpose of clustering sets of data. LIMBO clustering using the concept of distributional Cluster Feature (DCF) which stores information from the distribution of attribute values, and summarize information about the subcluster-subcluster in DCF Tree then form a cluster representative (centroid) which will then be used in the process of labeling data. From the analysis, we can see that tetha as a parameter that inputed by user influences software's accuracy. The smaller tetha value, the number of subcluster created is larger and F-measure accuracy tends to increasingly rise. Beside that, the increase the number of data influence construction of DCF Tree and clustering time, more data, so more time needed to execute the program because there are more subclusters created.

***Keywords**: data mining, clustering, LIMBO*