International Journal of Computational Intelligence Systems, Vol.2, No. 1 (March, 2009), 10-16

REDUCT DRIVEN PATTERN EXTRACTION FROM CLUSTERS

SHUCHITA UPADHYAYA
Deptt. of Computer Science and Application, Kurukshetra University
Kurukshetra-136119, India.
shuchita_bhasin@yahoo.com

ALKA ARORA
Indian Agricultural Statistics Research Institute
Library Avenue, Pusa, New Delhi-110012, India
alkak@iasri.res.in, alka27@yahoo.com

RAJNI JAIN
National Center for Agricultural Economics and Policy Research
Library Avenue, Pusa, New Delhi-110012, India
rajni@ncap.res.in

Received: 27-05-2008
Revised: 10-10-2008

Abstract

Clustering algorithms give general description of clusters, listing number of clusters and member entities in those clusters. However, these algorithms lack in generating cluster description in the form of pattern. From data mining perspective, pattern learning from clusters is as important as cluster finding. In the proposed approach, reduct derived from rough set theory is employed for pattern formulation. Further, reduct are the set of attributes which distinguishes the entities in a homogenous cluster, hence these can be clear cut removed from the same. Remaining attributes are then ranked for their contribution in the cluster. Pattern is formulated with the conjunction of most contributing attributes such that pattern distinctively describes the cluster with minimum error.

Keywords: Clustering, Cluster description, Data mining, Knowledge discovery, Pattern, Rough set theory, Reduct

1. Introduction

Fast developing computer science and engineering techniques has made the information easy to capture process and store in databases. Discovery of knowledge from this huge amount of data is a challenge indeed. Knowledge discovery in databases (KDD), popularly known as data mining is an attempt to make sense of the information embedded in large databases[1]. Clustering is a key area in data mining. The underlying assumption of clustering in data mining is to find out the hidden patterns in data, which can be revealed by grouping the entities into clusters. Clustering algorithms partitions a given dataset into clusters such that entities in a cluster are more similar to each other than entities in different clusters. Clustering algorithms in literature are broadly classified into hierarchical and partitional methods (See Refs. 1, 2, 3 and 4 for details on different clustering algorithms). Hierarchical algorithms construct a tree like structure (dendogram) combining all the entities. Description is subjective in case of dendogram. Partitional method divides the entities into k non overlapping clusters, where k is the number of clusters specified by the user as input. K-means and Expectation Maximization (EM) algorithms are the widely known partitional algorithms. These clustering algorithms only generate general description of the clusters depicting number of clusters and member entities of each cluster. However, it lacks in generation of underlying pattern in the dataset, as this approach has no mechanism for selecting and evaluating the attributes in the process of