Rough Set Based Clustering Using Active Learning Approach

Rekha Kandwal, Ministry of Earth Sciences and Science and Technology, India
Prerna Mahajan, Guru Gobind Singh Indraprastha University, India
Ritu Vijay, Bansthali University, India

ABSTRACT

This paper revisits the problem of active learning and decision making when the cost of labeling incurs cost and unlabeled data is available in abundance. In many real world applications large amounts of data are available but the cost of correctly labeling it prohibits its use. In such cases, active learning can be employed. In this paper the authors propose rough set based clustering using active learning approach. The authors extend the basic notion of Hamming distance to propose a dissimilarity measure which helps in finding the approximations of clusters in the given data set. The underlying theoretical background for this decision is rough set theory. The authors have investigated our algorithm on the benchmark data sets from UCI machine learning repository which have shown promising results.

Keywords: Active Learning, Clustering, Data, Hamming Distance, Rough-Set

1. INTRODUCTION

Clustering is one of the major techniques in the data mining as well as an active research topic. It involves identification of clusters (groups) in the data that are similar to each other. The cluster usually corresponds to points that are more similar to each other than they are to points from another cluster (Bagirov, Rubinov, Soukhoroukova, & Yearwood, 2003). Clustering is an unsupervised technique which naturally exploits the subset that has points with greater similarity and it doesn’t have a priori information about the structure of cluster. It is basically data driven and it leads to the discovery of the knowledge structures (clusters) which provides information about the data in the cluster. The clustering technique has been extensively studied in many fields such as pattern recognition, image segmentation, data visualization and similarity search, signal processing and trend analysis (Jain, Murthy, & Flynn, 1999; Dy & Brodley, 2004). Many algorithms on clustering exist in literature (Gibson, Kleinberg, & Raghavan, 1998; Zhang, Fu, Cai, & Heng, 2000; Huang, 1997; Wang, Xu, & Liu, 1999; Huang, 1999).

Rough sets have been introduced as a tool to deal with inexact, imprecise or vague knowledge in the real world data (Gibson, Kleinberg, & Raghavan, 1998). It uses the concept of lower and upper approximation of rough sets.
to handle the uncertainty in information. It has been applied in many fields such as machine learning, document classification and so on (Xu, Xuedong, Sen, & Bin, 2006; Widzlzak & Revett, 2004; Chengdong, Mengxin, Zhonghua, Zhang, & Yong, 2004). Chen et al. have applied rough set theory to clustering analysis and introduced the decision attribute to better define attribute membership matrix. They have used similarity measure based on Euclidean distance (Chen, Cui, Wang, & Wang, 2006). BWang has proposed a clustering technique based on Olary code to transform nominal attributes into integer. It provides a useful way to estimate the number of underlying clusters (Wang, 2010). Chengdong. Wu et al. have given a hierarchical clustering method for attribute discretization. It gives the significant clusters by determining the best classes for discretization picked from scatter plots of several statistics (Chengdong, Mengxin, Zhonghua, Zhang, & Yong, 2004).

Mitra et al have introduced a novel clustering architecture which integrates advantages of both fuzzy sets and rough sets. The algorithm aims to find a common structure revealed at global level which is determined by exchanging prototypes of the subsets of data & by moving prototypes of the corresponding clusters toward each other (Mitra, Banka, & Pedrycz, 2006). Hirano and Tsumoto have proposed a method which secluded the influence of initial equivalence relations by introducing perfect initial equivalence relations defined by the original class of each object and their randomly mutated variants (Hirano & Tsumoto, 2006).

2. BACKGROUND

2.1. Rough Set Theory

Rough set theory proposed by Pawlak (1982), is an extended approach to the classical set theory to represent vagueness or uncertainty. It extends the crisp notion of set by using the concept of approximation. One of the main objectives of Rough set theory is to reduce data size which makes it successful in broad domains such as machine learning, knowledge discovery in databases, pattern recognition. It introduces the notions of indiscernibility, rough set lower and upper approximations and reducts used to approximate inconsistent information and to exclude redundant data (Magnani, 2003).

For a given set of objects U called universe and an indiscernibility relation $R \times U$. Let $X$ be a subset of $U$ characterized with respect to R, as shown in Figure 1 (Pawlak, 1982).

![Figure 1. Rough set model (Pawlak, 1982)](image)

The lower approximation of a set $X$ with respect to $R$ is the set of all objects, which can be for certain classified as $X$ with respect to $R$ (certainly)

$$R(X) = \bigcup_{x \in U} \{ R(x) : R(x) \subseteq X \}$$

The upper approximation of a set $X$ with respect to $R$ is the set of all objects which can be possibly classified as $X$ with respect to $R$ (are possibly $X$ in view of $R$)

$$\bar{R}(X) = \bigcup_{x \in U} \{ R(x) : \bar{R}(X) \cap X \neq \emptyset \}$$
Related Content

Box-Counting Dimension of Fractal Urban Form: Stability Issues and Measurement Design
Shiguo Jiang and Desheng Liu (2012). *International Journal of Artificial Life Research* (pp. 41-63).
[www.igi-global.com/article/box-counting-dimension-of-fractal-urban-form/81213?camid=4v1a](www.igi-global.com/article/box-counting-dimension-of-fractal-urban-form/81213?camid=4v1a)

Basics for Olfactory Display
Yasuyuki Yanagida and Akira Tomono (2013). *Human Olfactory Displays and Interfaces: Odor Sensing and Presentation* (pp. 60-85).
[www.igi-global.com/chapter/basics-olfactory-display/71919?camid=4v1a](www.igi-global.com/chapter/basics-olfactory-display/71919?camid=4v1a)

Prediction of Epigenetic Target Sites by Using Genomic DNA Sequence
[www.igi-global.com/chapter/prediction-epigenetic-target-sites-using/52316?camid=4v1a](www.igi-global.com/chapter/prediction-epigenetic-target-sites-using/52316?camid=4v1a)
Multi-Objective Evolutionary Algorithm NSGA-II for Variables Selection in Multivariate Calibration Problems

[www.igi-global.com/article/multi-objective-evolutionary-algorithm-nsga-ii-for-variables-selection-in-multivariate-calibration-problems/93623?camid=4v1a](www.igi-global.com/article/multi-objective-evolutionary-algorithm-nsga-ii-for-variables-selection-in-multivariate-calibration-problems/93623?camid=4v1a)