Chapter 1
Uncertainty-Based Clustering Algorithms for Large Data Sets

B. K. Tripathy
VIT University, India

Hari Seetha
Vellore Institute of Technology – Andhra Pradesh, India

M. N. Murty
IISC Bangalore, India

ABSTRACT

Data clustering plays a very important role in Data mining, machine learning and Image processing areas. As modern day databases have inherent uncertainties, many uncertainty-based data clustering algorithms have been developed in this direction. These algorithms are fuzzy c-means, rough c-means, intuitionistic fuzzy c-means and the means like rough fuzzy c-means, rough intuitionistic fuzzy c-means which base on hybrid models. Also, we find many variants of these algorithms which improve them in different directions like their Kernelised versions, possibilistic versions, and possibilistic Kernelised versions. However, all the above algorithms are not effective on big data for various reasons. So, researchers have been trying for the past few years to improve these algorithms in order they can be applied to cluster big data. The algorithms are relatively few in comparison to those for datasets of reasonable size. It is our aim in this chapter to present the uncertainty based clustering algorithms developed so far and proposes a few new algorithms which can be developed further.

DOI: 10.4018/978-1-5225-2805-0.ch001

An intelligent being cannot treat every object it sees as a unique entity unlike anything else in the universe. It has to put objects in categories so that it may apply its hard-won knowledge about similar objects encountered in the past, to the object at hand. – Steven Pinker, How the Mind Works, 1997

1. INTRODUCTION

We are living in a world full of data. Every day, people deal with different types of data coming from all types of measurements and observations. Data describe the characteristics of a living species, depict the properties of a natural phenomenon, summarize the results of a scientific experiment, and record the dynamics of a running machinery system. More importantly, data provide a basis for further analysis, reasoning, decisions, and ultimately, for the understanding of all kinds of objects and phenomena. One of the most important of the myriad of data analysis activities is to classify or group data into a set of categories or clusters. Data objects that are classified in the same group should display similar properties based on some criteria. Actually, as one of the most primitive activities of human beings (Anderberg, 1973; Everitt et al., 2001), classification plays an important and indispensable role in the long history of human development. In order to learn a new object or understand a new phenomenon, people always try to identify descriptive feature and further compare these features with those of known objects or phenomena, based on their similarity or dissimilarity, generalized as proximity, according to some certain standards or rules. As an example, all natural objects are basically classified into three groups: animal, plant, and mineral. According to the biological taxonomy, all animals are further classified into categories of kingdom, phylum, class, order, family, genus, and species, from general to specific. Thus, we have animals named tigers, lions, wolves, dogs, horses, sheep, cats, mice, and so on. Actually, naming and classifying are essentially synonymous, according to Everitt et al. (2001), with such classification information at hand, we can infer the properties of a specific object based on the category to which it belongs. For instance, when we see a seal lying easily on the ground, we know immediately that it is a good swimmer without really seeing it swim.

Basically, classification systems are either supervised or unsupervised, depending on whether they assign new data objects to one of a finite number of discrete supervised classes or unsupervised categories, respectively (Bishop, 1995; Cherkassky and Mulier, 1998; Duda et al., 2001).

A cluster is a collection of data elements that are similar to each other but dissimilar to elements in other clusters. A vast amount of data is generated and made available across multiple sources. It is practically impossible to manually