**CHAPTER 1**

# INTRODUCTION

## Overview

With the increasing progress of both computer software and hardware, a large amount of data is generated and collected from different sources daily. Data collected contains useful information which is hidden so one can extract useful information from this data by discovering patterns in the data. However, the major problem for obtaining useful knowledge from data lies in the data itself. Therefore, Data mining plays an important role in extracting knowledge and useful patterns from the data.

Recently, data management and processing for wireless sensor networks (WSNs) has become a topic for research in several fields of computer science, such as distributed system, the database system, and the data mining [1]. We have proposed Clustering a data mining approach for making efficient clusters in wireless Sensor Network. Different sensor receives data and form clusters using k-prototype clustering algorithm discussed in chapter 2 and transmit this data to the base stations.

The nature of sensor data, certain special characteristics and limitations of WSNs prevent them from being directly applied to the traditional data mining techniques. Advances in wireless communication led to the development of low-power sensors and the deployment of large-scale sensor networks. The areas such as business, society, engineering and health care are generating data at an exponential rate. This has become possible due to tremendously rapid development of information technology. Data mining satisfies the need of extracting the useful information and knowledge from thus large scale data. [1], [2] and [3].

In data mining, clustering is important technique. The clustering algorithms are widely used in image processing, customer segmentation, gene expression analysis [4] and text documents analysis [4] and text documents analysis [5] etc. The aim of clustering is to divide a set of data objects into clusters such that data objects in the same cluster are more similar to each other than those in other clusters [6], [7], [8] and [9].

In real world, data sets usually contain both numeric and categorical attributes [10] and [11]. However, most existing clustering algorithms assume all attributes are either numeric or categorical, examples of which include the k-mean [12], k-modes [13], fuzzy k-modes [14], TGCA [15], COOLCAT [16], and G-ANMI algorithms [17]. When mixed data are encountered, most of them usually exploit transformation approaches to convert one type of the attributes to the other and then apply traditional single-type clustering [3]. However, in most cases, transformation scheme may result I loss of information, leading to undesired clustering outcomes [11].

A Technique known as Clustering is well established in data mining. This is used for grouping the data based on its similarity. It is used in various applications. One such application is Wireless Sensor Networks (WSNs). The WSNs usually consist of energy processing and storage limited devices known as sensor nodes. Sensor nodes are usually distributed over a certain area covered by the WSN to observe phenomena (light, temperature, humidity) by collecting data. Physical clustering of sensor nodes aims at grouping together sensor nodes that are sensing correlated data and selecting one of them as a representative, while turning others off [2]. This minimizes the energy consumption and thus extends the lifetime of the nodes.

In sensor networks, data mining is the process of extracting patterns and application-oriented models with acceptable accuracy from a rapid, continuous and possibly non ended flow of data streams from sensor networks. In this case, whole data cannot be stored and must be processed immediately. Data mining algorithm has to be sufficiently fast to process high-speed arriving data. The conventional data mining algorithms are meant to use the multistep techniques and multi-scan mining algorithms and handle the static data for analyzing static data-sets. Therefore,

Conventional data mining techniques could not be used for handling the high dimensionality, massive quantity and distributed nature of the data generated by the WSNs.

The k-means algorithm is famous for its efficiency in clustering large data sets but usually when working only on numeric values it is not used to cluster real world data containing categorical values. In this project we have used k-prototype algorithm which extend the k-means algorithm to categorical domains and domains with mixed numeric and categorical values. The k-modes algorithm uses a simple matching dissimilarity measure to deal with categorical objects, replaces the means of clusters with modes, and uses a frequency-based method to update modes in the clustering process to minimize the clustering cost function [3]. With the help of these extensions the k–modes algorithm enables the clustering of categorical data same as done in k-means algorithm. Further, the k-prototypes algorithm integrates the k-means and k -modes algorithms to allow for clustering objects described by mixed numeric and categorical attributes.

**1.2 Wireless Sensor Network**

A wireless sensor network (WSN**)** (sometimes called a wireless sensor and actor network(WSAN)are spatially distributed autonomous sensors to monitor physical or environmental conditions, such as pressure, sound, temperature etc. and to cooperatively pass their data to a main location. The more advanced networks are bi-directional which enables the control of sensor activity. The development of wireless sensor networks was motivated by military applications such as battlefield surveillance; today such networks are used in many consumer and industrial applications, such as machine health monitoring, industrial process monitoring and control and so on.

**1.3 Clustering**

Clustering can be considered the most important unsupervised learning problem. So, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.  
The definition of clustering can be “the process of organizing objects into groups whose members are similar in some way”. A collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters is known as clusters.

Clustering is mainly used to determine the intrinsic grouping in a set of unlabeled data. But what constitutes a good clustering? How can we can we decide that? It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, the user must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outliers detection).

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point we need to re-calculate k new centroids as barycenter’s of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function.

**1.4 K-mode**

The cause that the k-means algorithm cannot cluster categorical objects is its dissimilarity measure. These barriers can be removed by making the following modifications to the k-means algorithm:

1. Using a simple matching dissimilarity measure for categorical objects,

2. Replacing means of clusters by modes, and

3. Using a frequency-based method to find the modes

Like the k-means algorithm the k-modes algorithm also produces locally optimal solutions that are dependent on the initial modes and the order of objects in the data set

The k-modes algorithm extends the k-means algorithm for clustering categorical data by using a simple matching dissimilarity measure for categorical objects, modes instead of means for clusters and a frequency based method to update the modes using the k-means method to minimize the cost of clustering cost function [ref1].

**1.5 K-prototype**

In our project we used a k-prototypes algorithm for clustering mixed data. In our method, we first named the distribution centroid to represent the prototype of categorical attributes in a cluster [ref2]. Then we integrate the mean with distribution centroid to represent the prototype of a cluster with mixed attributes, and propose a new dissimilarity measure, which takes account of the significance of each attribute, to evaluate the dissimilarity between data objects and prototypes.

The algorithm is used to present an effective representation for the categorical attribute part in a mixed prototype since the mean is good enough for the numeric attribute part, and on the other hand to consider the significance of different attributes towards the clustering process.

It is straightforward to integrate the k-means and k-modes algorithms into the k-prototypes algorithm that is used to cluster the mixed-type objects. The k-prototypes algorithm is practically more useful because frequently encountered objects in real world databases are mixed-type objects.

**1.4 Motivation**

As the data collected from different sources consists of both categorical and numerical values.

So, the conventional approaches like k-means clustering does not fit well for data containing both kind of attributes. So we have used new approaches which are extensions for existing clustering algorithms. K-prototype algorithm is used for data containing both kind of attributes categorical and numerical. These new approaches uses data mining for finding similarity measures between the attribute values for different data objects.

The main objective of using these algorithm is to minimize the cost function discussed in later section.

As the wireless sensor network collects incoming data like temperature, humidity and rainfall on the sensors and sends the collected data to the base stations but the collected data varies a lot.

So efficient clusters are formed so that a cluster contains homogeneous data. Therefore data objects with attribute values which have the most similarity measure are put in the same cluster.

## 1.6 Organization of the Report

The report is divided into 5 chapters. Explanation about each chapter is given below.

**Introduction chapter** gives an overview about the algorithms used for clustering in wireless sensor network such as k-means, k-modes and k-prototype clustering.

**Literature Survey** contains introduction about the basic idea about the data mining techniques like Data preprocessing and its associated techniques , Classification, association , clustering and different types of clustering like k-means , k-mode, k-prototype, LEACH , WLEACH.

**Proposed Method** chaptercontains the methodology used for the application of k-prototype algorithm for clustering data set based on similarity measure of data objects.

**Experimental Result and Analysis** chaptercontains the sample data set used for the application of k-prototype clustering algorithm.

**Conclusion and Future Work** chapter gives overall concluded results which shows that k-prototype being a good clustering algorithm for numerical and categorical values suffers from drawbacks like mapping of categorical values to binary values.