Chapter 1

Introduction

This chapter gives a brief introduction to knowledge discovery process and basic data mining concepts. It also explains the motivation behind the research work and defines the objective of the research study. The contributions of the research work are also detailed in this chapter and it concludes with the organization of the thesis.
Development of Hierarchical Clustering Techniques for Gridded Data from Mixed Data Sequences
1. Introduction

1.1. Knowledge discovery

Knowledge discovery in databases is the non-trivial process of identifying valid, novel potentially useful and ultimately understandable patterns from data. The process predicts the future trends and behaviors from accumulated large volumes of data to make proactive, knowledge driven decisions. At an abstract level, the core of knowledge discovery process is to map low-level data which is too voluminous in nature to compact, predictive model for estimating the values of future cases. The phrase knowledge discovery in databases (KDD) was first coined at the first KDD workshop in 1989 [1].The term process implies that KDD comprises many steps, which involve data preparation, search for patterns, knowledge evaluation, and refinement, all repeated in multiple iterations. The steps involved in the KDD process is diagrammatically represented below [2].

![Figure 1-1: The Overview of the steps in the KDD process](image)

The knowledge discovery process starts with the understanding of the application domain and identifying the goal of the process from the user’s perspective. A target data which is relevant to the analysis task is selected on which
the discovery process is to be applied. The selected data may be considered dirty which may contain noisy values or missing values. The preprocessing step cleans the dirty data to a more consistent form. Next step in the process transforms the data into a more consolidated form appropriate for mining by performing operations like normalization, aggregation etc. Data reduction helps to find useful features which can represent the large dataset eliminating the redundant factors.

The next step, which is the actual data mining process does the exploratory analysis on the data and build some model based on known techniques of statistics, neural networks, machine learning and pattern recognition. Data mining is the step in KDD process that consists of applying data analysis and discovery algorithms on preprocessed subsamples, and transformed data. Data mining is the application of specific algorithms for extracting patterns from data. Data mining involves many different approaches to accomplish the pattern extraction tasks. The data mining process tries to fit a model to the data. The algorithm examines the data and determines a model that is closest to the characteristics of the data being examined. Data mining tasks can be classified into two categories, descriptive data mining and predictive data mining. The former describes the data in a concise and summarized manner whereas the latter constructs one or a set of models performs inference on the available set of data and attempts to predict the behavior of new data sets. The patterns of interest from the models are visually represented for better interpretation in the final step. These steps need not be mutually exclusive and may need iterations to fine tune the results.

A related field evolving from databases is data warehousing, which refers to the trend of collecting and cleaning transactional data to make them available for online analysis and decision support. Data warehousing helps set the stage for KDD in two important ways: (1) data cleaning and (2) data access. Data selection and preprocessing: the initial step produces a data store which is created by integrating
data from a number of databases. When integrating data, the problems like identifying data, missing data, data conflict and ambiguity may arise. An Extraction transformation and loading tool (ETL) is used to overcome these problems.

1.2. Data mining Approaches

1.2.1. Based on Statistical attributes

Research in Statistics has helped to produce many of the proposed data mining algorithms. Statistical concepts like determining a data distribution and calculating a mean and variance can be viewed as data mining techniques and each of these is a descriptive model for the data under consideration. Sampling is a technique which is often used in data mining. A subset of total population is examined and a generalization about the entire population is made from the subset.

1.2.2. Machine learning

Machine learning is the area of artificial intelligence that examines how to develop systems and algorithms that can learn, based on the feedback of prediction made by it, thereby correcting the domain knowledge also. When machine learning is applied to data mining tasks, a model is built to represent data. Samples from the entire database are used to train the model and generate the model parameter. Further samples are applied to the model to perform the task of data mining. There are two different types of machine learning:

(i) supervised learning where the model is trained from examples

(ii) Unsupervised learning, where the model learns on its own from the data stream.
1.2.3. **Classification**

Supervised classification works well, if data is known to have some predefined classes. Classification is a data mining technique used to predict group membership for data instances. Classification maps data into predefined groups or classes. Pattern recognition forms the basis of classification, where an input pattern is classified into one of the several classes based on similarity (or proximity) to the predefined classes.

1.2.4. **Regression**

Regression deals with the estimation of an output value based on a sequence of input values. It can be used to map a data item to a real-valued prediction variable. Regression involves in learning of the function that does this mapping. Regression tries to fit the target data to some known types of function like linear regression or logistic regression.

1.2.5. **Neural network based algorithms.**

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. An artificial neuron is a device with many inputs and one output. The neuron has two modes of operation; the training mode and the using mode. In the training mode, the neuron can be trained to fire (or not), for particular input patterns A firing rule determines how one calculates whether a neuron should fire for any input pattern. It relates to all the input patterns, not only the ones on which the node was trained. The ANN consists of three or more layers: a layer of "input" units is connected to one layer of "hidden" units; the output of a hidden layer is connected to either another hidden layer or to a layer of "output" units. The activity of the input units represents the raw
information that is fed into the network. The activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units. The behavior of the output units depends on the activity of the hidden units and the weights between the hidden and output units. The weights between the input and hidden units determine when each hidden unit is active, and so by modifying these weights, a hidden unit can choose what it represents. The neural network is trained by computing the weights using example cases in the supervised learning mode or the network learns on its own in the case of unsupervised learning mode. A trained neural network can thus represent a model in the case of machine learning or a regression function.

1.3. **Scope of the Thesis**

Clustering is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes. Since no classes are known prior for most of the real world data, substantial research is going on in unsupervised learning techniques, to evolve the classes. The present study specifically focuses on clustering of multidimensional mixed category data. Most of the clustering algorithms act on a dataset with uniform format, since the similarity or dissimilarity between the data points is a significant factor in finding out the clusters. If a dataset consists of mixed attributes, i.e. a combination of numerical and categorical variables, then a preprocessing step is done distinctly for different data types. A preferred approach is to convert different formats into a uniform format. The present research study explores the various techniques to convert the categorical attributes to a numerical equivalent, so as to make it equipped for applying the common clustering algorithms. The techniques developed in the thesis are then applied to the converted data types and the results are illustrated.
Moreover, another issue, with the clustering process, which merits attention, is the visualization of output. Visualization of higher dimensional data on a lower dimension is an attractive feature in evaluating properties of clusters. Different geometric techniques like scatter plot, or projection plots are available; but most of the techniques display the result suitable for pair wise analysis of attributes. On the other hand, projecting the whole database to two dimensions, based on importance of the chosen dimension, often gives impressive visualization. In-order to represent a multi-dimensional dataset into a plane or a space, dimensionality reduction techniques are explored and implemented. The clustered data set then becomes gridded in two dimensions.

With the gridded representation of a multidimensional dataset, the present study further progresses on spatial clustering techniques, on the resultant spatial image (or gridded data). The Quad tree, a prevalent spatial data structure for representing two dimensional data based on regional homogeneity, is investigated for suitability in representing the gridded data in two dimensions. The homogeneity of the region is considered as the factor to decompose the quadrants into sub-level quadrants. Though the homogeneity can be measured using statistical measures like mean and variance or entropy, the present work uses fuzzy rules to decompose the quad tree, to avoid crisp boundaries of mean and variance. The information dense regions can be extracted from the leaf nodes of the quad tree, using quadrant merging. The research also explores the possibility of using the information content of the regions to identify the edges of clusters from the merged quadrants.

In all the phases of research, the record level identifications of the observations in the dataset are maintained and this helps to remap the cluster index into the original dataset. The clusters identified can be indexed back to the original data set for subjective interpretation and assimilation. These clusters will help to infer useful patterns out of the dataset based on the domain of data.
1.4. Objective of the thesis

The objective of this research study is to construct a general framework which will

- Convert a mixed attributed data set into a uniform format to make it equipped for general clustering algorithms, which can be extended to completely categorical dataset,
- Find out a gridded representation of high-dimensional normalized uniform formatted dataset, after reducing the dimensionality,
- Spatially cluster the gridded representation and extract the cluster edges after merging the neighbor dense regions using spatial data structures,
- Map the cluster index back into the original dataset, without any information loss, and establish grouping of records based on some leading traits.

1.5. Structure of the thesis

The thesis is organized as follows. Chapter 2 provides the literature survey done as part of this research, which helped the author to focus on the research topic, by deriving the hints and directions from the published results, at the same time maintaining the individually of the approach. The chapters to follow describe work carried out and the contributions of the thesis. Toward this, the Chapter 3 explains the preprocessing steps applied on the mixed attributed dataset to convert it into a uniform numerical format. The chapter also suggests an extension of the algorithm to a completely categorical dataset. The algorithm is applied on a mixed attributed crime dataset and a combined hierarchical-k-means clustering is done on the numerical equivalent of the crime dataset to demonstrate the influence of the procedure. Chapter 4 explores the technique of visually representing the high-dimensional dataset obtained from the first phase using singular value decomposition.
of the space corresponding to the data set and projecting the data set to chosen reduced dimensions. The chapter concludes with the experimentation with standard data sets from UCI repository like Iris, Wine, Yeast, Thyroid, and Breast Cancer. The crime dataset used in the chapter 3 is also experimented with in order to ascertain the outcome of the technique proposed. The dimensionality reduction and thereby projection to a lower dimension of the high-dimensional mixed attributed dataset, results in a spatial image, which is then used for spatial clustering. Chapter 5 details the application of fuzzy techniques for the decomposition of quad tree, merging the neighbor dense quadrants using, fuzzy rules. This chapter also deals with the final contribution of the thesis viz. the remapping of clusters into the original dataset, to infer useful patterns from the clusters of data elements. Though every chapter is provided with case studies on different domain, a complete demonstration of the framework of the technique developed as part of the thesis is done on FARS (Fatal Accident Reposting System) dataset, which has got more than 37 thousand observations defined with 16 variables of mixed type and the case study is presented as the chapter 6. Chapter 7 summarizes the conclusion and outlines the future scope of the research study.