Analysis of Rural's people by using clustering method

Nang Phyu Hnin Oo
Computer University(Lashio)
phyuninoo7@gmail.com

Abstract

Rural's people development plays an importance role in a country. The analysis of status of social, commercial of Rural's people is very main status. By using this analysis, fill can the requirement space. So, the statuses of social, commercial of rural region's are analyzed by using partition method. The education, economic, age and kinds of job from Kong Thar village have been checked. So the development of it can be easily known and analyzed. Data mining approach has emerged in order to exploit require information from Kong Thar information. Clustering a subcomponent of data mining process is a process of partitioning a set of data objects (information) into a set of meaningful subclass, called Clusters. This paper proposes the approach for clustering of Kong Thar people by using K-mean method. If K-means method is used to cluster information with similarities it can provide the set of required information to the user.

1. Introduction

It plays an important role to study the development of a country or a region, to Endeavour the development and to study the population. Economy can be developed together with the development of population. If there is an increase in population and it has proportionately increased in consumption and production, socio-economy of a country can be developed. There is a relationship between the development of population and socio-economy.

To study the development of a region, population and business are important. So, population, gender, business, education, job and income in Kong Thar village have been studied. It is easy to understand the situation of Kong Thar village by clustering and partitioning the similar object in its data

Data mining refer to extracting knowledge from large amount of data. There are many kinds of functionalities of data mining such as-class discrimination, association analysis, classification and prediction, cluster analysis, outlier analysis and evolution analysis. Among them, cluster analysis is used to study the data of Kong Thar village by dividing into the same groups.

Cluster is a collection of data objects that are similar to one another within the same cluster and one dissimilar to the objects in the objects in other cluster. A cluster of data objects can be treated collectively as one group in many applications.

There are various methods to extract similar data in cluster. These methods are partitioning method (similar), hierarchical methods (nested), density-based method (distance), grid-based method (cell) and model based method (construct model). In this paper, partition method which is a suited method to separate data is used. Data can be partitioned by using various kinds of algorithm. In this paper, K-means algorithm and its extension called Bisecting K-means are used. Bisecting K-means splits the set of all points into two clusters, selects one of these cluster to split and so on, until k cluster have been produced. Among data attribute age attribute and earning attribute are divided into two partitions and small partition is sub-divided. Age is arranged by ≥ 16 and earning by ≥ 30000.Age data are collected between 16 years and 65 years because of earning data, which are collected from 30000 to120000.

This paper is made up of 6 sections.. Firstly, the related work is presented. Then, theory of the system is described in section3 and the system design in section 4. Experimental result of system is shown in section 5 and section 6 concludes this paper.

2. Related work

Several approaches have been considered for information clustering until now. Clustering is the challenging process in the data mining approach. Various clustering methods are worked together with the different datasets [9].

The most commonly discussed distinction among different types of clustering is whether the set of clusters is nested or unnested, or in more traditional terminology, hierarchical or partitional. A partitional clustering is simply a division of the

set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset [1].

If we permit clusters to have sub clusters, then we obtain a hierarchical clustering, which is a set of nested clusters that are organized as a tree. Each node (cluster) in the tree is the union of its children (sub clusters), and the root of the tree is the cluster containing all the objects [2].

Moreover, J. Han and M. Kamber is described data mining, clustering and k-mean algorithm in [7]. P Berkhin present clustering Data Mining Techniques [2]. Marcos M. Campos describe the definition of Data Mining [8].

3. Data mining

Data mining refers to extracting knowledge from large amounts of data. Rapid advances in data collection and storage technology have enabled organizations to accumulated vast amounts of data. However, extracting useful information has proven extremely challenging. Often, traditional data analysis tools and techniques cannot be used because of the massive size of a data set. Sometimes, the non-traditional nature of the means that traditional approaches cannot be applied even if the data set is relatively small. In other situations, the questions that need to be answered cannot be addressed using existing data analysis techniques [7, 8].

3.1 Clustering

Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. The example below demonstrates the clustering of balls of same colour. There are a total of 10 balls which are of three different colours. We are interested in clustering of balls of the three different colours into three different groups. The balls of same colour are clustered into a group as shown below:

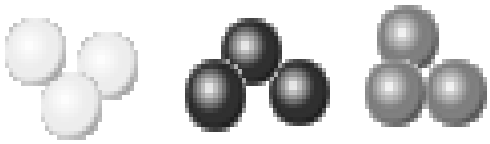

Figure 1. Clustering of balls of same colour

Thus, we see clustering means grouping of data or dividing a large data set into smaller data sets of some similarity [11, 12].

3.1.1. Clustering algorithms

A clustering algorithm attempts to find natural groups of components (or data) based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster [12].

3.1.2. Partitioning methods

Given a database of n objects and k, the number of clusters to form, a partitioning algorithm or organizes the objects into k partitions ($k \leq n$), where each partition represent a cluster. The clusters are formed to optimize an objective partitioning criterion, often called a similarity function, such as distance, so that the objects within a cluster are "similar", where as the objects of different clusters are "dissimilar" in terms of e database attributes. The most well known and commonly used partitioning methods are k-means, k - medoids, and their variations [2].

3.1.3. K-Means clustering

This method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart. Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated every time a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.

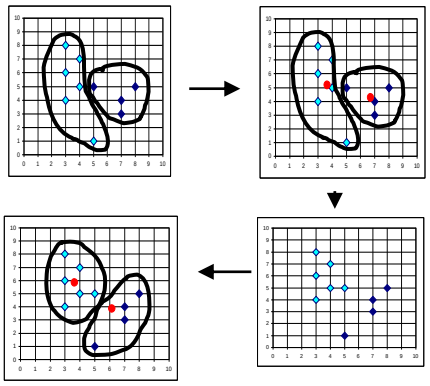

Figure 2. Clustering example

3.1.4. K-means algorithm

Algorithm : k-means algorithm for partitioning based on the mean value of the objects in the Cluster.
Input : The number of clusters k and a database containing n objects.
Output : A set of k clusters that minimizes the squared error criterion.
Method :

(1) arbitrarily choose k objects as the initial cluster centers;
(2) repeat
(3) (re) assign each objects to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means , i.e., calculate the mean value of the objects for each cluster,
(5) until no change [7].

3.1.5. Bisecting K-MEANS

The bisecting K-means algorithm is a straight forward extension of the basic K-means algorithm that is based on a simple idea: to obtain K cluster, split the set of all points into two clusters, select one of these cluster to split and so on, until K cluster have been produced. The detail of bisecting K-mean are given by Algorithm
Algorithm : Bisecting K- means algorithm

1. Initialize : the list of clusters to contain the cluster consisting of all points.
2. repeat
3. Remove a cluster from the list of clusters.
4. {Perform several "trial" bisections of the chosen cluster.}
5. for i = 1 to number of trials do
6. Bisect the selected cluster using basic K-means.
7. end for
8. Select the two clusters from the bisection with the lowest total.
9. Add these two clusters to the list of clusters.
10. Until the list of clusters contains K clusters [11].

4. Proposed system overview

There are five main stages in this paper. They are data entry, detail display of rural's people status, partition process, describes information by user choice and compare of rural's people income status.

Data entry

The analyzer of Rural's people input the information of one person such as name, father's name, age, job, education, income, sex and subgroup.

Detail display

All information of Rural's people is displayed in detail as name, father's name, age, job, education, income, sex and subgroup.

Partition process

The system partition the information by age, income, education, female, male, farm, garden, fruit seller, land and government and other by using K-mean algorithm. K-mean algorithm are partitioned the data until these data are empty. So we use the Bisection K-mean algorithm, the corollary of K-mean algorithm, which partitioned the data into two parts and then partitioned the data by using smaller parts.

User choice

When user input income, the system displays the information of input income. If user input age, the system displays the information of input age. If user input job, the systems displays the information of input job. When user input age and income, age and job, job and income, age-job-income, the respectively information is displayed.

Compare process

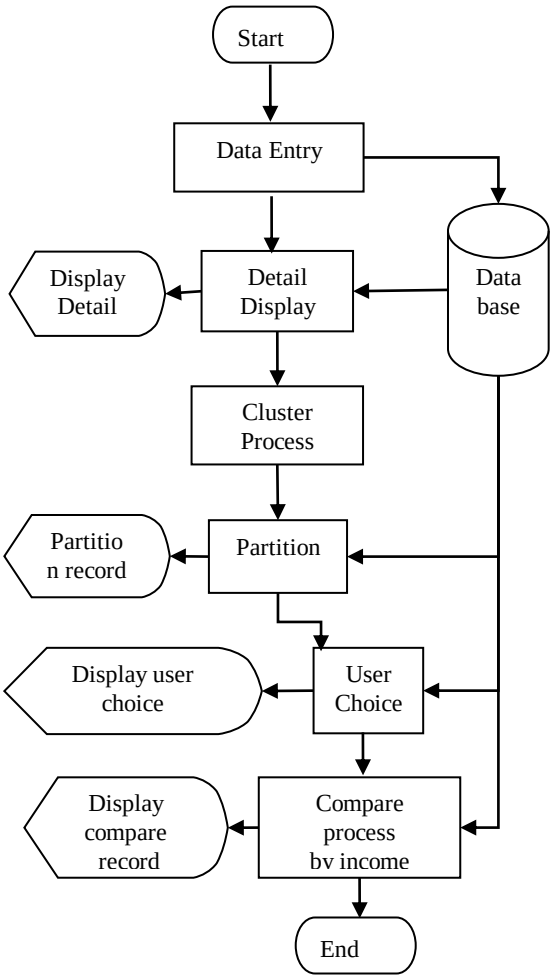

Figure 3. Proposed system overview

In this portion, the system analyzes the subgroup of Rural's people by income. The most income and the less income of sub group information are displayed.

5. Experimental result

In this section, the results of the experiment are reported. Figure-5.1 shows the data entry of Kong Thar people data. It includes name, father's name, gender, age, job, income, education and group name of Kong Thar people. There are many output result of partition table such as age, gender, job, education and income. Figure-5.2 shows the partition of income that is partitioned until income>=30000. And also the partition of age that is partitioned until age>=16. In gender, partition male table and female table. In job, partition farm table, land table, fruit seller table, garden table, government table and other job table. In education, partition graduate table, ungraduate table and other education table.

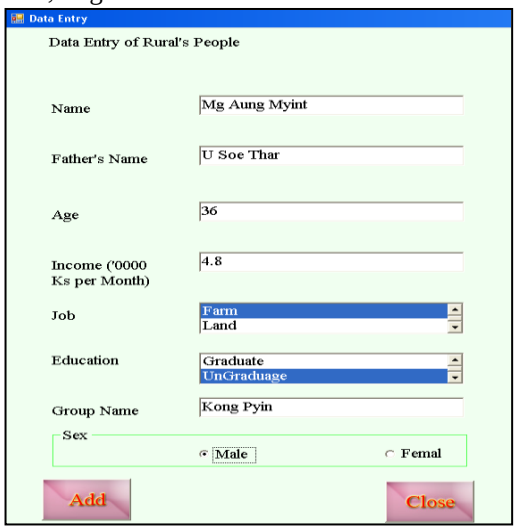

Figure 4. Data Entry Form

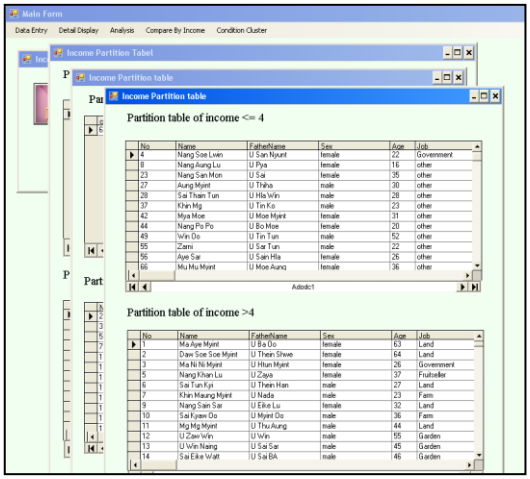

Figure 5. Experiment result of Income Clustering

6. Conclusion

By using this system, the benefits of database approach will be affected on the Rural's people information. The system briefly describes the analysis of Rural's people by using clustering method.

Current system may have limitations, but high gain in efficiency can nevertheless often be achieved. In the application, the trial data are used from all rural' people data. The age is also used greater than 16 and less than 65. Such as - incomes are limited between 30000 and 120000. This system can be used by only Rural's people analyzer.

The system describes the status of Rural's people by age, by income, by education and by kinds of job. The system can extend not only populations of specific region but also the status of possess and expense of farms.

Reference

[1] A. K. Jain, M.N.Murty, and P,J.Flynn Data clustering a review ACM computing Surveys,

[2] C. Jeong-Ho, *"Cluster Analysis"*, BioIntelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Korea.

[3] F. Giannoti, C. Gozzi, and G.manco, "Clustering Transactional Data." in proc. PKDD, pages 175-187, 2002.

[4] International *Journal* "Information Theories & Applications" Vol.14/2007

[5] J. Han and M. Kamber, *"Data Mining Concepts and Techniques",* Morgan Kanfmann, San Francisco, 2001.

[6] Marcos M.Campos, *"What is data mining?"* January, 2006.

[7] P.Berkhin Survery of Clustering Data Mining Techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[8] Teknome, K-Means Clustering Tutorial, PhD December 2004.

[9] http://people.revoledu.com/kardi/tutorial/kMean /index.html

[10] www.db.book.com for conditions on re.use