Datamining: Clustering (Information from Rural Villages of Sivagangai District)

Dr.S.S.Dhenakaran [#1], M.Sathish Kumar [*2]
[#1] Assistant professor, Department of Computer science and Engineering, Alagappa University, Karaikudi. India.
[*2] Department of Computer science and Engineering, Alagappa University, Karaikudi. India.

Abstract— Research work is aimed to mining the rural villages of sivagangai district. Key factors to incorporated for mining information useful to village peoples and government are number of villages, number of families, number of schools based on government and private, number of colleges based on government and private, number of universities, educated level of study up to elementary schools, Secondary Schools, Higher Secondary Schools, Under graduation, Post graduation, Research, Drought hit area in villages, Availability of waste land, frequent causes of diseases, etc… From this data, useful information is proposed to generate the benefit to the peoples, as well as provided useful reporting to the government for sanction various beneficial schemes for rural village peoples in sivagangai district.

Keywords— Government, Schools, Colleges, Village.

I. INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).

Knowledge Discovery and Data Mining (KDD)

Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the Internet and the widespread use of databases have created an immense need for KDD methodologies. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions.

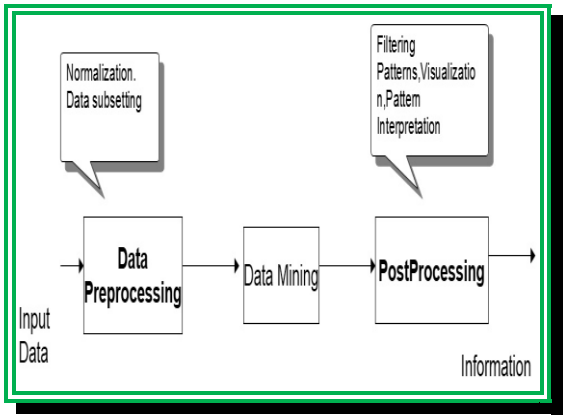

Fig. 1.1 Knowledge Discovery and Data Mining (KDD)

II. DATA MINING TECHNIQUES

Clustering is often followed by a stage in which a decision tree or rule set is inferred that allocates each instance to the cluster in which it belongs. Then, the clustering operation is just one step on the way to a structural description. Data Mining is automated extraction of previously unknown data that is interesting and potentially useful. Data mining techniques are applied to discover new trends and patterns of behavior that are hidden. These can be used for prediction in a variety of applications.

Data Mining Techniques include techniques for

- Association analysis
- Classification
- Prediction
- Cluster Analysis
- Evolution and Deviation analysis

Association Analysis:

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association rule mining finds interesting association

or correlation relationships among a large set of data items. The discovery of interesting association relationships among huge amounts of business transaction records can help catalogue design, cross-marketing, loss-leader analysis, and other business decision making processes.

Classification:

Classification is a data mining (machine learning) technique used to predict group membership for data instances. Classification is used to extract models describing important data classes or to predict future data trends. Classification predicts categorical labels. Data classification is a twostep process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, called the class label attribute. In the context of classification, data tuples are also referred to as samples, examples, or objects. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population.

Prediction:

Classification predicts categorical labels (or discrete values), while prediction models continuous-valued functions. For example, a classification model may be built to categorize bank loan applications as either safe or risky, while prediction model maybe built to predict the amount of loan that can be safely disbursed.Thus prediction model predicts continuous values (using regression techniques). Classification and prediction have numerous applications including credit approval, medical diagnosis, performance prediction, and selective marketing.

Cluster Analysis:

A process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. Objects within one cluster are similar to each other (homogeneous). Objects belonging to one cluster differ from the objects belong to another cluster. Cluster analysis is used in many applications like Pattern Recognition, Data Analysis, Image Processing and Marketing Research.

As a data mining function it can be used as a standalone tool to gain insight into the distribution of data to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Alternatively it may serve as a pre-processing step for other algorithms like classification and characterization operating on the detected clusters.

Evolution and Deviation analysis:

Outlier Analysis: Outlier is a data object that does not comply with the general behavior of the data. A set of data objects that are grossly different from or inconsistent with the remaining set of data are called outliers of the data set. It can be used in fraud detection for finding unusual usage of credit cards or telecommunication services, in customized marketing for finding spending

behavior of extremely rich or poor people or in medical analysis for finding unusual response to certain medicine or treatment.

III. SIVAGANGAI DISTRICT: CENSUS 2011 DATA

Sivagangai District Overview

An official Census 2011 detail of Sivagangai, a district of Tamil Nadu has been released by Directorate of Census Operations in Tamil Nadu. Enumeration of key persons was also done by census officials in Sivagangai District of Tamil Nadu.

Sivagangai District Population 2011

In 2011, Sivagangai had population of 1,339,101 of which male and female were 668,672 and 670,429 respectively. In 2001 census, Sivagangai had a population of 1,155,356 of which males were 566,947 and remaining 588,409 were females.

Sivagangai Literacy Rate 2011

Average literacy rate of Sivagangai in 2011 were 79.85 compared to 72.18 of 2001. If things are looked out at gender wise, male and female literacy were 87.92 and 71.85 respectively. For 2001 census, same figures stood at 83.14 and 61.74 in Sivagangai District. Total literate in Sivagangai District were 959,744 of which male and female were 526,304 and 433,440 respectively. In 2001, Sivagangai District had 738,000 in its district.

Sivagangai Sex Ratio 2011

With regards to Sex Ratio in Sivagangai, it stood at 1003 per 1000 male compared to 2001 census figure of 1038. The average national sex ratio in India is 940 as per latest reports of Census 2011 Directorate. In 2011 census, child sex ratio is 960 girls per 1000 boys compared to figure of 952 girls per 1000 boys of 2001 census data.

Sivagangai Child Population 2011

In census enumeration, data regarding child under 0-6 age were also collected for all districts including Sivagangai. There were total 137,235 children under age of 0-6 against 132,891 of 2001 census. Of total 137,235 male and female were 70,022 and 67,213 respectively. Child Sex Ratio as per census 2011 was 960 compared to 952 of census 2001. In 2011, Children under 0-6 formed 10.25 percent of Sivagangai District compared to 11.50 percent of 2001. There was net change of -1.25 percent in this compared to previous census of India.

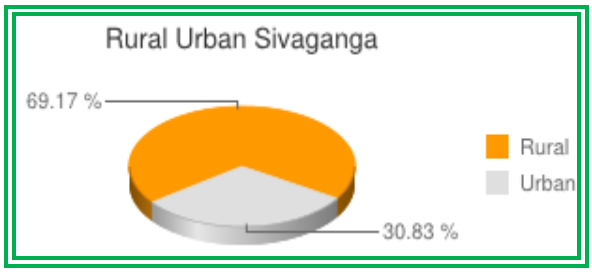

Fig.3.1 Rural urban sivagangai

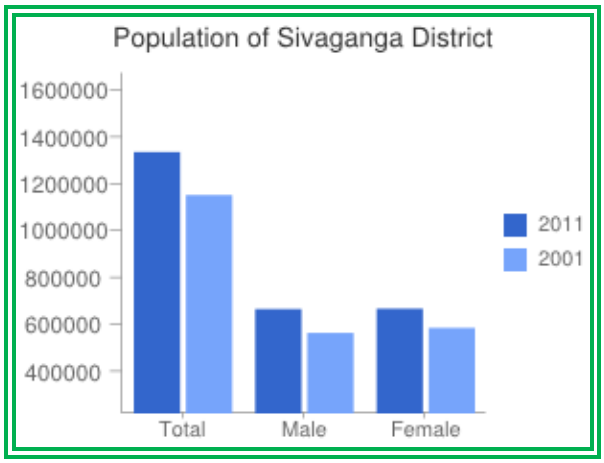

Fig.3.2 Population of Sivagangai District

IV. SAMPLE DATA IN SIVAGANGAI DISTRICT

According to the 2011 census Sivagangai district has a population of 1,341,250, roughly equal to the nation of Swaziland or the US state of Maine. This gives it a ranking of 360th in India (out of a total of 640). The district has a population density of 324 inhabitants per square kilometer (840 /sq mi). Its population growth rate over the decade 2001–2011 was 16.09%. Sivagangai has a sex ratio of 1,000 females for every 1,000 males, and a literacy rate of 80.46%.

The district had a population of 1,150,753 with male population 565,594 and female 585,159 (as of 2001). The rural population is 826,427 and the urban population is 324,326. It is 28.22% urbanized. It has a population density of 274.7. The district has a literacy of 52.5%, below the average for the state. Tamil is the principal language spoken in the district. Hindus formed the majority of the population.

Revenue Division	**Taluks**	**No. of revenue villages**
Sivagangai	4 (Sivagangaii, Manamadurai, Ilayankudi, Thiruppuvanam)	267
Devakottai	3 (Devakottai, Karaikudi, Tiruppattur)	255
Total	7	521

List of Taluks in Sivagangai

List of Taluks in Sivagangai	
S.No	Name
1	Devakottai Taluk
2	Ilayankudi Taluk
3	Karaikudi Taluk
4	Manamadurai Taluk
5	Sivagangai Taluk
6	Tirupathur Taluk

V.CONCLUSION

PERFORMANCE ANALYSIS OF CLUSTERED THE DATASETS

Clustering high dimensional data is the cluster analysis of data with anywhere from a few dozen to many thousands of dimensions. Multiple dimensions are hard to think in, impossible to visualize,and, due to the exponential growth of the number of possible values with each dimension, impossible to enumerate. Hence to improve the efficiency and accuracy of mining task on high dimensional data, the data must be preprocessed by efficient dimensionality reduction methods such as Principal Component Analysis (PCA). Cluster analysis in high-dimensional data as the process of fast identification and efficient description of clusters.

The clusters have to be of high quality with regard to a suitably chosen homogeneity measure.

K-means is a well known partitioning based clustering technique that attempts to find a user specified number of clusters represented by their centroids. There is a difficulty in comparing quality of the clusters produced Different initial partitions can result in different final clusters. Hence in this paper we proposed to use the Principal component Analysis method to reduce the data set from high dimensional to low dimensional. The new method is used to find the initial centroids to make the algorithm more effective and efficient. By comparing the result of original and proposed method, it was found that the results obtained from proposed method are more accurate.

ACKNOWLEDGMENT

I am greatly indebted to my parents and department faculties for their great encouragement and co-operation in all aspects to develop this paper.

I wish to thank everyone who helped us directly or indirectly for the successful completion of this paper.

REFERENCES

[1] Data mining techniques for customer relationship management Technology in Society **24 (2002) 483–502.**

[2] DISCOVERING KNOWLEDGE IN DATA .An Introduction to Data Mining, DANIEL T. LAROSE, Director of Data Mining, Central Connecticut State University. Copyright **©2005 by John Wiley & Sons, Inc.** All rights reserved.

[3] Department of Homeland Security Privacy Office 2012, Data Mining Report to Congress, **February 2013.**

[4] Data Mining Practical Machine Learning ,Tools and Techniques, **Edition Ian H. Witten . Eibe Frank. Mark A. Hall**

[5] Data Mining Techniques in CRM: Inside Customer Segmentation, **Konstantinos Tsiptsis Antonios Chorianopoulos**