International Journal of Engineering Research ISSN:2319-6890)(online),2347-5013(print)
Volume No.3, Issue No.8, pp : 546-548 01 Sept 2014

A Study on Rural Health care Data sets using Clustering Algorithms

[1]Sathyendranath Malli,[2]Dr. Nagesh H R, [3]Dr. H G Joshi
[2] Department of Computer Science, MITE, Mangalore
[3]School Of Commerce, Manipal
[1] SOIS, Manipal
[1]sathya.malli@manipal.edu

Abstract: Rural healthcare datasets are often large, relational and dynamic. These datasets contain records related to child welfare, pregnant woman health information and socioeconomic status of family.Data mining is very popular and essential in the healthcare industry due to fact that huge amounts of heterogeneous data being generated through healthcare transactions. It is a processing procedure of extracting credible, novel, effective and understandable patterns from database. Additionally, database consists of inconsistent and noisy data. This paper focuses on pattern generated from rural healthcare datasets using clustering algorithms thus helps in decision making process. The result of the experiment shows the comparison between the cluster generated and also justifying the uniqueness of the cluster by the values of attributes of these patterns. These patterns are generated on socioeconomic status of the locality and the data sets used are from Rural Maternity and Child Welfare (RMCW) database. These clustering techniques are implemented and analysed using a clustering tool WEKA.

Keywords—Data Mining, Clustering algorithms, Rural Health care, Heterogeneous Data

I. Introduction

Data mining is very popular and essential in the healthcare industry due to fact that huge amounts of heterogeneous data being generated through healthcare transactions. These complex and voluminous data needs to be processed and analyzed using data mining methods. KDD (Knowledge discovery in databases) is the "non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data" [1]. In this process data mining is one of the steps, which produces the novel patterns. This can be achieved by applying pattern discovery algorithms and data analysis over the data [2] [3].

Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects. Clustering algorithms are used to organize data, categorize data, for data compression and model construction, for detection of outliers etc.Common approach for all clustering techniques is to find clusters centre that will represent each cluster. Cluster centre will represent with input vector can tell which cluster this vector belong to by measuring a similarity metric between input vector and all cluster centre and determining which cluster is nearest or most similar one [4].

A major component of rural healthcare systems is the identification, collection, analysis and interpretation of data, which further leads to take decision as well as action to effectively carry out the rural healthcare programs. The datasets used for the study are from eRMCWH database. eRMCW Home is an electronic version which takes care of the activities of Rural Maternity and Child Welfare (RMCW) Homes. This is designed to capture day-to-day activities performed by healthcare service provider known as the ANMs (Auxiliary Nurse and Midwife) and render these healthcare data when it is needed. These digital datasets can be used to develop decision support systems to assist ANMs in providing better care. Data mining techniques can be used for the large datasets of eRMCWH to extract useful patterns, which can assist in building the logic for decision support systems.

II. Methodology

K-Means Clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The algorithm is called k-means, where k is the number of clusters, since a case is assigned to the cluster for which its distance to the cluster mean is the smallest. It is a partition method technique which finds mutual exclusive clusters of spherical shape. It generates a specific number of disjoint, flat(non-hierarchical) clusters.The action in the algorithm centers around finding the k-means. We start out with an initial set of means and classify cases based on their distances to the centers. Next, we compute the cluster means again, using the cases that are assigned to the cluster; then, we reclassify all cases based on the new set of means. We keep repeating this step until cluster means don't change much between successive steps. Finally, we calculate the means of the clusters once again and assign the cases to their permanent clusters.

Algorithmic steps for k-means clustering Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, 'ci' represents the number of data points in ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

III. Results and Tables

The K-Means clustering technique is used and tested on a set of health care data related to socio-economic status of the

ISSN:2319-6890)(online),2347-5013(print)
01 Aug 2014

localities. The test conducted using WEKA clustering tool. The whole data set consists of 16 attributes and 1000 entries.

Option 1: Clusters are generated based on the classes of the data sets.

a) By considering locality as cluster classes the experiment was conducted and following are the result

Number of iterations: 7

Within cluster sum of squared errors: 1541.4928981133683

Missing values globally replaced with mean/mode

TABLE 1: Cluster generated based on locality as cluster classes

Attribute	**1 (16)**	**2 (123)**	**3 (121)**	**4 (154)**	**5 (44)**	**6 (158)**	**7 (21)**	**8 (51)**	**9 (18)**	**10 (34)**	**11 (259)**
TypeOfHouse	Kutcha	Mixed	Pucca	Pucca	Pucca	Pucca	Kutcha	Pucca	Mixed	Mixed	Mixed
OwnerShip	Own	Own	Own	Own	Own	Own	Own	Own	Own	Own	Own
VBicycle	Yes	No	No	No	No	No	Yes	Yes	Yes	No	Yes
V2Wheeler	No	No	No	Yes	No	No	No	Yes	No	No	Yes
VAuto	No	No	No	No	No	No	No	Yes	No	No	No
V4Wheeler	Yes	No	No	No	No	No	No	Yes	Yes	Yes	No
Radio	Yes	No	No	No	No	Yes	No	Yes	Yes	Yes	Yes
TV	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Phone	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Goat	Yes	No	No	No	No	No	No	No	No	No	No
Cow	Yes	No	No	Yes	No	No	No	Yes	No	No	No
Buffalo	No	No	No	No	No	No	No	No	No	No	No
FamilyWorkingAbroad	No	No	No	No	No	No	No	No	Yes	No	No
TotalScore	55.25	26.80	35.52	53.34	36.72	47.70	24.85	75.09	63.88	56.05	49.95
SEStatus	**Middle**	**Low**	**Low**	**Middle**	**Low**	**Middle**	**Low**	**High**	**Middle**	**Middle**	**Middle**

Cluster 1<-- 16 (2%) Kodi

Cluster 2<-- 123 (12%) GuddeAngadi 5th

Cluster 3<--121 (12%) PaduAlevoor

Cluster 4<-- 154 (15%) Nailapadi

Cluster 5<-- 44 (4%) GuddeAngadi 1st

Cluster 6<-- 158 (16%) MooduAlevoor

Cluster 7<-- 21 (2%) Karval

Cluster 8<-- 51 (5%) Rampura

Cluster 9<-- 18 (2%) NaaduAlevoor

Cluster 10<-- 34 (3%) Doopadapadi

Cluster 11 <-- 259 (26%) GuddeAngadi

From the above table we can infer that cluster no 8 which has the highest score with Socio economic status (SEStatus) as high, in which most of the family in this locality has electronic gadgets, vehicles and cattle in their house compare to middle and low socio economic status area.

b) By considering socio economic status (SEStatus) as cluster classes the experiment was conducted and following are the result

Number of iterations: 4

Within cluster sum of squared errors: 2686.808000360232

Missing values globally replaced with mean/mode

TABLE 2: Cluster generated based on socio economic status as cluster classes

Attributes (999)	**Cluster 1 (369)**	**Cluster 2 (205)**	**Cluster 3 (425)**
TypeOfHouse	Mixed	Mixed	Pucca
OwnerShip	Own	Own	Own
VBicycle	Yes	No	No
V2Wheeler	Yes	No	No
VAuto	No	No	No
V4Wheeler	No	No	No
Radio	Yes	No	No
TV	Yes	No	Yes
Phone	Yes	No	Yes
Goat	No	No	No
Cow	Yes	No	No
Buffalo	No	No	No
FamilyWorkingAbroad	No	No	No
TotalScore	54.60	29.59	47.01
LocalityName	**MooduAlevoor**	**PaduAlevoor**	**MooduAlevoor**

Cluster 1<-- 369 (37%) High

Cluster 2<-- 205 (21%) Low

Cluster 3<-- 425 (43%) Middle

From the above table provides the information about socio economic status of locality with high, middle and low. We can distinguish the difference between these three values in terms of possession of electronic gadgets, vehicle and cattle.

Option 2: By specifying the number (random) of clusters

Number of iterations: 6

Within cluster sum of squared errors: 2704.2674248679295

Missing values globally replaced with mean/mode

TABLE 3: Cluster generated by specifying the number of clusters

Attribute	**1 (295)**	**2 (162)**	**3 (175)**	**4 (367)**
Type of House	Mixed	Mixed	Pucca	Pucca
Ownership	Own	Own	Own	Own
VBicycle	Yes	No	No	No
V2Wheeler	Yes	No	No	Yes
VAuto	No	No	No	No
V4Wheeler	No	No	No	No
Radio	Yes	No	No	Yes
TV	Yes	No	Yes	Yes
Phone	Yes	No	No	Yes
Goat	No	No	No	No
Cow	Yes	No	No	No
Buffalo	No	No	No	No
FamilyWorking Abroad	No	No	No	No
TotalScore	54.2949	28.0494	36.7429	52.327
SEStatus	**Middle**	**Low**	**Low**	**Middle**
LocalityName	**Kodi**	**PaduAlevoor**	**MooduAlevoor**	**MooduAlevoor**

ISSN:2319-6890)(online),2347-5013(print)
01 Aug 2014

Clustered Instances
Cluster 1 ← 295 (30%)
Cluster 2 ←162 (16%)
Cluster 3 ← 175 (18%)
Cluster 4 ← 367 (37%)

IV.Conclusion

K-Means algorithm is faster and also produces quality clusters when using huge datasets. This paper provides information about the cluster being generated using k-Means clustering technique. The results are displayed in the table 1,2,3 represents results under different factors and situations. Here the results are focused on socio economic status of the localities which are taken from eRMCWH database. The cluster generated from this technique clearly provides the information with regard to comparison between the clusters. Results are very useful in decision making process for primary healthcare centres.This paper focused only on one set of data i.e socio economic status of the localities from eRMCWH database. As a future work the clusters can be generated on child welfare, pregnant woman health information on localities wise. The results from these data sets will give more information to officials of primary health care center to take proper decision. There is a scope for using normalized data which will give better and different results. As a result performance and the quality cluster.

References

i. Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996) Data Mining to Knowledge Discovery in Databases. AI Magazine, 17, 37-54.

ii. Han, J., Kamber, M. and Pei, J. (2011) Data Mining: Concepts and Techniques. 3rd Edison, Morgan Kaufmann Publishers, Burlington.

iii. Khosla, R. and Dillon, T. (1997) Knowledge Discovery, Data Mining and Hybrid Systems. Engineering Intelligent Hybrid Multi-Agent Systems, Kluwer Academic Publishers, Norwell, 143-177.

iv. Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Reserch and Applications (IJERA), Vol. 2, Issue 3, pp.1379-1384, 2012.

v. J.A.Hatigan and M.A.Wong,‖a K-Means Clustering Algorithm, ―Applied statistics, 28:100—108, 1979

vi. Azuaje, F., Dubitzky, W., Black, N., Adamson, K., ―Discovering Relevance Knowledge in Data: A Growing Cell Structures Approach,‖ IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics, Vol. 30, No. 3, June 2000 (pp.448)

vii. Tapankanungo, senior member IEEE David M. Mount, member IEEE ―An Efficient K–means algorithm: analysis and implementation‖

viii. Chakraborty, S. and Nagwani, N.K., ―Analysis and study of Incremental K-Means clustering Algorithm‖, Communication in Computer and Information Science, 1, Volume 169, High Performance Architecture and Grid Computing, Part 2, Pages 338-341, 2011.

ix. VITO M. LOGRILLO, Director of Health Statistics, New York State Department of Health; Technical papers on "Health Data Issues for Primary Health Care Delivery Systems in Developing Countries"

x. *S Pushpalatha, Dr Jagdesh Pandya, "Implementing Data mining in Primary HelathCenter- A Review", IOSR Journal of Engineering, ISSN:2250-3021 Volume 2, Issue 8 (August 2012), PP 183-186*