International Research Journal of Computer Science (IRJCS)
Issue 08, Volume 4 (August 2017)

ISSN: 2393-9842
www.irjcs.com

PERFORMANCE COMPARISON OF CLUSTERING TECHNIQUES IN GAIN RATIO FEATURES REDUCED MCDR

Suban Ravichandran*
Information Technology,
Annamalai University, Chidambaram, India
rsuban82@gmail.com

Manuscript History
Number: **IRJCS/RS/Vol.04/Issue08/AUCS10089**
Received: 12, July 2017
Final Correction: 29, July 2017
Final Accepted: 06, August 2017
Published: August 2017
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Abstract— In today's world huge number of methods are available for data accumulation; numerous data are obliged from different databases. Data mining functionalities are depiction and discernment, mining recurrent patterns, correlation, association, classification and prediction, cluster analysis, outlier analysis and evolution analysis. Clustering analysis is one of the essential techniques in data mining, in which farthest-first algorithm is quick and appropriate for huge scale data mining applications. This paper examines some clustering algorithms like Canopy, Simple K-Means, Filtered cluster, Make Density Based cluster, Farthest First. Farthest-First algorithm is quick and the execution is too high than the rest of the algorithms. Clustering algorithms is applied to the dataset obtained from Gain ratio feature reduced MCDR and the execution is implemented in WEKA data mining tool. Our main aim of this process is to show the comparison of the different-different clustering algorithms available of WEKA and find out which algorithm will be most suitable for the users In PAKDD 2006 dataset, Farthest First algorithm performed well based on Accuracy, F-measure, and Time taken.

Keywords —Data Mining; Clustering Techniques; Canopy; Simple K-Means; Filtered cluster; Make Density Based cluster; Farthest First; PAKDD 2006; MCDR; Gainratio; WEKA;

I. INTRODUCTION

Data Mining (DM) is a method for doing data investigation done for discovering designs, uncovering shrouded regularities and connections in the data [1]. It is the process of finding useful hidden patterns in data and is known by different names in different communities. DM is the vital step in the KDD process for analysing the data. Knowledge Discovery in Databases (KDD) is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [2]. KDD is a broad area that integrates methods from several fields including statistics, databases, AI, machine learning, pattern recognition, machine discovery, uncertainty modelling, data visualization, high performance computing, optimization, management information system (MIS), and knowledge-based systems. The steps that are involved in KDD process are Data Gathering, Data Cleansing, Feature Extraction, Data Mining, Visualization of Data, and Verification and Evaluation of Results [3].

This paper is organized as follows. Section II describes an overview of PAKDD 2006 Dataset, Feature Selection method is given in Section III, Different Clustering Algorithms were discussed in Section IV, Experimental Setup is discoursed in Section V, Performance analysis and comparison are done in Section VI and Section VII concludes the carried-out research and possible future works.

II. DATASET DESCRIPTION

A Dataset is a collection of data that contains individual data units organized (formatted) in a specific way and accessed by a specific method based on the data set organization. A CDR contains detail such as calling number, called number, time of call, duration of call etc. It does not contain the call charge. A MCDR is a data record stored by a mediation system in network switch of format understandable by billing system.

Dataset was provided by an Asian Telco Operator for a Competition called PAKDD 2006 [4]. This company has launched 3G Technology and would like to know the details of customer interested in switching from 2G to 3G. The dataset contains 20K 2G customers and 4K 3G customers with 251 attributes for each instance. The 251 attributes contain information like personal details, call details, message details, GPRS details, WAP details, application details like games, videos, etc., payment details and handset specification details. Training set contains 18000 instances which have been taken for this mining process and the Test set contains 6000 instances. Training set contains 15K 2G customers and 3K 3G customers and Test set contains 5K 2G customers and 1K 3G customers respectively [5].The information is summarized in Table I.

TABLE I. DATASET DESCRIPTION

Properties	Value
Domain	Call Detail Record
File Type	DAT
Data Type	Text
Class	Binomial {2G,3G}
Number of Records	18000
2G Records	15000
3G Records	3000
No.of Attributes	251

III. FEATURE SELECTION METHOD

Data Mining models have been raised and executed utilizing distinctive clustering algorithms for distinguishing 2G/3G clients. To improve performance, lower computational complexity, build better optimized models, and reduced memory storage feature selection strategy is applied to the dataset before implementing Clustering. In this work the selection of most suitable attributes from the dataset, was carried-out using Gain Ratio Attribute Evaluation. This feature selection strategy estimates the quality of every feature with the criteria, and rank the highlights in light of the measures.

Gain Ratio

Gain ratio (GR) is a modification of the information gain that reduces its bias [5]. Gain ratio takes number and size of branches into account when choosing an attribute. It corrects the information gain by taking the intrinsic information of a split into account. Intrinsic information is entropy of distribution of instances into branches (i.e. how much info do we need to tell which branch an instance belongs to). Value of attribute decreases as intrinsic information gets larger.

$$\text{GainRatio(Attribute)} = \frac{\text{Gain(Attribute)}}{\text{Intrinsic\_info (Attribute)}}$$

The attribute with the maximum gain ratio is selected as the splitting attribute. On Applying this Attribute Evaluator to the dataset containing 251 attribute and 18000 instances, 24 attributes have been selected as shown in Table II.

TABLE II. SELECTED ATTRIBUTE BY GAIN RATIO ATTRIBUTE EVALUATOR

Attribute Evaluator	No. of Attributes Selected
Gain Ratio Attribute Evaluator	24

IV. CLUSTERING ALGORITHMS

Cluster analysis or clustering is the errand of collection an arrangement of items such that articles in a similar gathering (called a Cluster) are more comparative (in some sense or another) to each other than to those in different gatherings (Clusters).A best clustering strategy will create great clusters in which the intra-class (that is, Intra-Cluster) closeness is high. The Inter-Class likeness is low. The nature of a clustering result additionally relies upon both the closeness measure utilized by the technique and its execution. The nature of a clustering technique is likewise estimated by its capacity to find a few or the greater part of the hidden patterns. Be that the target assessment is complicated: normally done by human/experts' examination.

There are various clustering techniques available in weka. They are,

- Canopy
- Farthest First
- Filtered
- Make Density Based
- Simple K-Means

A. Canopy

Canopy clustering algorithm is an unsupervised pre-clustering algorithm introduced by Andrew McCallum, Kamal Nigam and Lyle Ungar in 2000 [6]. Usually utilized as pre-processing venture for the K-means algorithm or the Hierarchical clustering algorithm. It is expected to accelerate clustering tasks on huge data sets, where utilizing another algorithm straightforwardly might be unrealistic because of the extent of the data set.

B. Farthest first

Farthest first is a heuristic based method of clustering. It is a variation of K-Means that likewise picks centroids and doles out the items in cluster yet at the point furthermost from the current cluster focus existing in the data zone. Quick clustering is given by this algorithm in the vast majority of the cases since less reassignment and alteration is required.

C. Filtered

Filtered is the process of running an arbitrary clusterer on data that has been gone through an arbitrary filter. Like the clusterer, the structure of the filter is construct only in light of the training data and test instances will be prepared by the filter without changing their structure.

D. Make Density Based

Density Based Clustered algorithm is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996 [7]. Density based clustering algorithm has assumed an essential part in discovering non-straight shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most broadly utilized Density Based algorithm. It utilizes the idea of density reachability and density connectivity.

E. Simple K-Means

K-Means is the basic and generally utilized strategy of clustering. It depends on dividing system. It segments information things into k-groups where k demonstrates the quantity of clusters determined by a client. Clusters are framed with the end goal that everything in the cluster has least distance from the centroid. For computing distance between a thing and the centroid, K-Means algorithm utilizes the Euclidean distance estimation.

V. EXPERIMENTAL SETUP

A. Data Preparation

1) Replace Missing Values

Missing value [8] is the number (percentage) of instances in the data for which the attribute is unspecified. We use a filter called “Replace Missing Values”, which replaces all missing values for all nominal and numeric attributes in a dataset.

2) Data Preprocessing

The dataset PAKDD 2006 is available in “dat” for in default. This format is imported in Excel and converted into Comma Separated Value (CSV) format. The converted file is then fed into the Pre-processing Tool developed [9] in PHP and MySQL as shown in Fig 1. The developed tool removes the duplicate instances and attributes values available in the dataset. The tool is developed in five stages namely.

i. Importing Data.
ii. Remove Duplicate Instances Values.
iii. Remove Duplicate Column (Attribute) and Null or Zero Values.
iv. Conversion of Text into Numeric Values.
v. Export preprocessed data in CSV format.

Once the data is pre-processed the data is exported in the CSV format as a file and fed for further modeling. Each and individual steps ii, iii, iv are designed in such a way they are not interrelated. Those steps can be executed separately without each other. The end result of the pre-processing contains the same number of attributes and instances as original since no duplicates are available. The resultant dataset contains 251 attributes with 18000 instances.

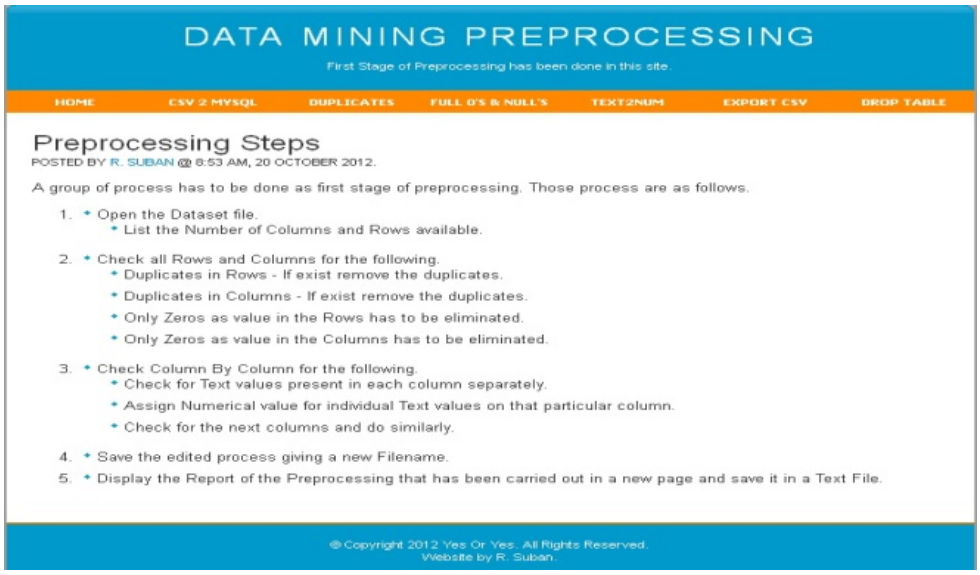

Fig 1. Developed Data Preprocessing Tool

B. Feature Extraction & Clustering

The feature extraction extracts a set of features which represents customer retention prediction attributes. Gain Ratio Attribute Evaluator is used to effectively handle the large dimensionality of the training set, so that it is convenient to perform effective learning. In order to evaluate a subset of features, the accuracy of a predictor, which use the feature subset, have to be considered. In the experiments, based on the training dataset mentioned above, feature selection method is employed to select the features, independently.

Based on the selected feature subset, various Clustering Algorithms like Canopy, Simple K-Means, Filtered cluster, Make Density Based cluster, Farthest First were experimented and compared for better performance. An effective analysis is done only on the dataset which shows correct clustered instances. The dataset with highest accuracy, f-measure and lowest time taken for clustering will give an effective analysis. The final result demonstrates that the proposed approach revealed the high value customers. It can help to achieve better distinguishable groups, so that marketing managements can design more suitable marketing strategy.

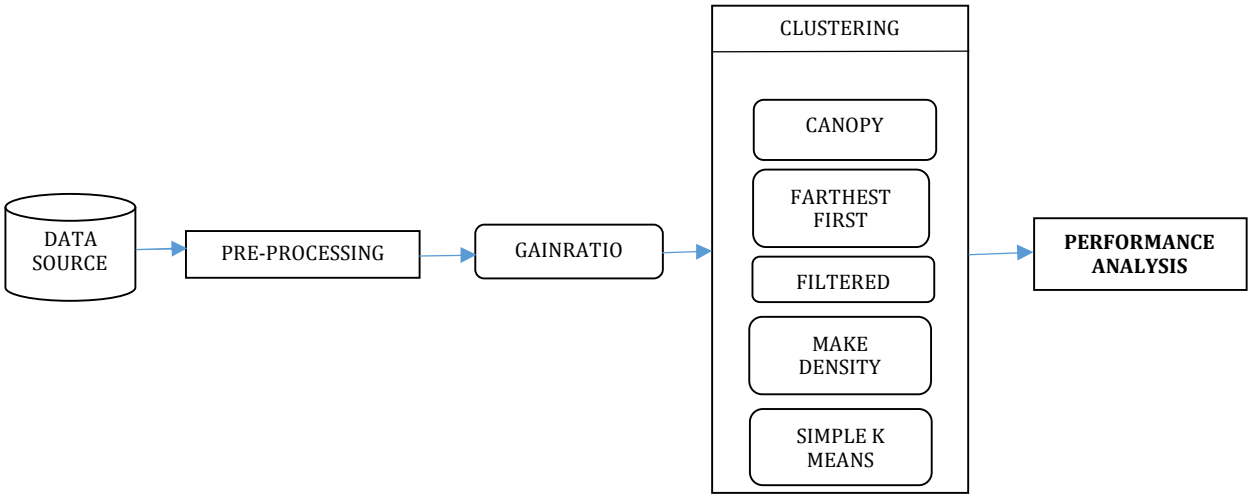

Fig 2. Proposed Method For Comparing Clustering Algorithms

VI. RESULTS AND DISCUSSION

PAKDD 2006 dataset is being used in this proposed process, in which 24 attributes were carefully chosen through the GainRatio Attribute Evaluator. The attained feature set is then assessed by different clustering algorithms like Canopy, Simple K-Means, Filtered cluster, Make Density Based cluster, Farthest First. Table III shows the performance of the various clustering method.

TABLE III. Performance of Clustering Algorithms

Metrics	Canopy	Filtered	Farthest First	Make Density	Simple K Means
Accuracy	86.6444	47.9500	87.5723	76.7945	47.9500
F-Measure	92.6712	47.9500	93.3732	86.1179	59.7603
Time Taken	0.78	3.89	0.31	4.11	3.77

From the Table III and Figure 3, it shows the results of different clustering types. Among all the clustering types Make Density Based Clustering technique gives the highest accuracy value, highest f-measure value for the Gain Ratio dataset. The log likelihood value of this clustering type is also the high of all the other clustering type. Hence Make Density Based Clustering technique can be used in Gain Ratio dataset for efficient results.

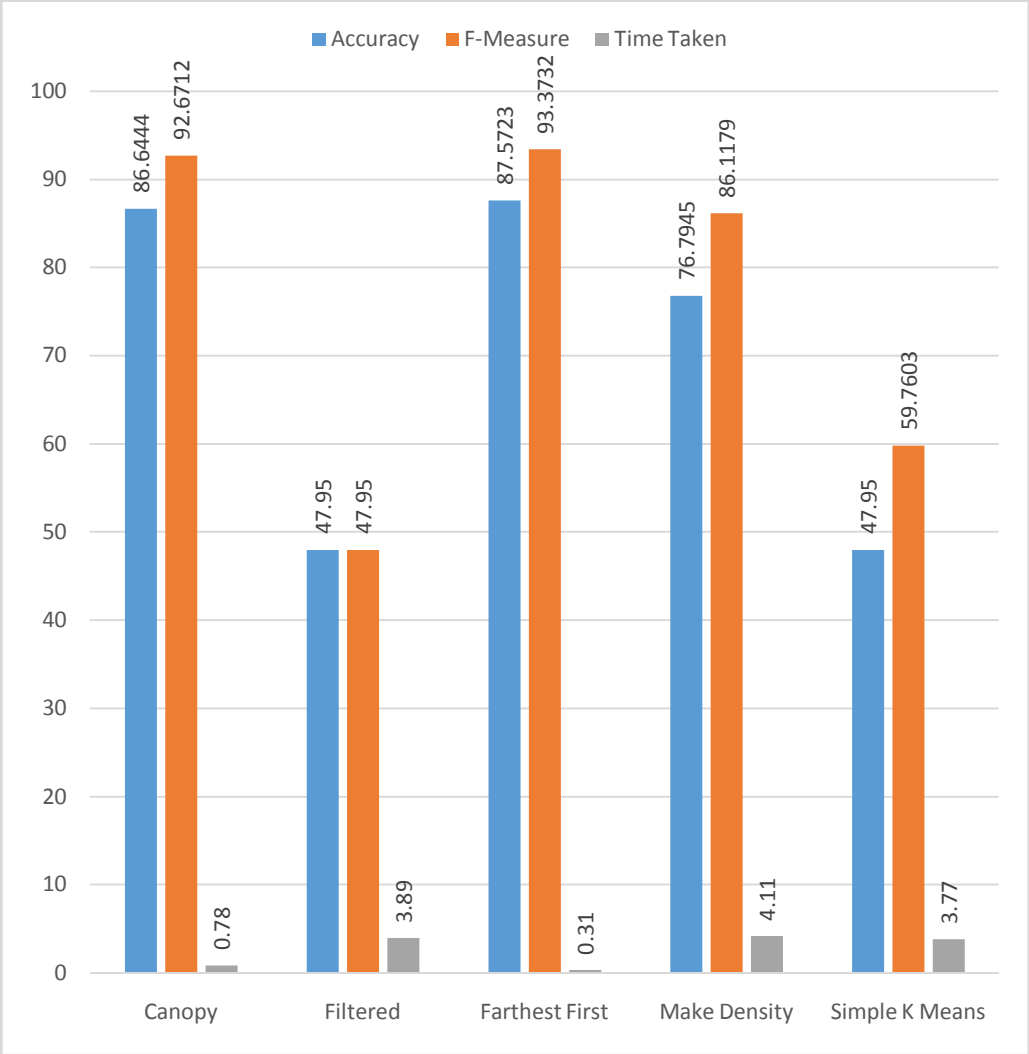

Fig 3. Comparison of Clustering Algorithms

VII. CONCLUSIONS

In the proposed process, the feature selection method Gain RatioAttribute Evaluator was performed for dimensionality reduction to illustrate that more information is not always good in machine learning applications. For the application at hand, a feature selection algorithm can be selected based on the following considerations: simplicity, stability, number of reduced features, classification accuracy, storage and computational requirements. Among various clustering technique used, Farthest Firstclustering gives the highest accuracy value of 87.5723. The F-Measure value of FF is 93.3732 which is the highest of all the other clustering techniques. The time taken by the FF technique is the lowest than all the other Clustering Technique. Thus, Gain Ratio feature reduced dataset with Farthest FirstClustering Technique performed well than all other Clustering Techniques which might help better in the analysis of the customer; based on their usage behaviour.

REFERENCES

1. J. Han and M. Kamber, "Data Mining: Concepts and Techniques", San Francisco, Morgan Kauffmann Publishers, 2001.
2. Dzeroski, Saso, and Nada Lavrač, eds. Relational data mining. Springer, 2001.
3. Hafez, Alaaeldin M. "Knowledge Discovery in Databases." (2008).
4. The 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006), http://www3.ntu.edu.sg/SCE/pakdd2006/competition/overview.htm.
5. Suban Ravichandran and Chandrasekaran Ramasamy, "Performance Comparison of Dimensionality Reduction Methods using MCDR.", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 3, Number 7, Jul 2016, pp. 64-69, 2016.
6. McCallum, Andrew, Kamal Nigam, and Lyle H. Ungar. "Efficient clustering of high-dimensional data sets with application to reference matching." Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000.
7. Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. No. 34. 1996.
8. Suban Ravichandran and Chandrasekaran Ramasamy, "Customer Retention of MCDR using 3SCDM Approaches.", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Volume 5, Number 8, Aug 2016, pp. 218-222. 10.17148/IJARCCE.2016.5841.
9. Ravichandran, Suban, and Chandrasekaran Ramasamy. "Customer Retention of MCDR using Three-Stage Classifier based DM Approaches." International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET) 5.6 (2016): 11190-11196.

