International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2763
Issue 07, Volume 3 (July 2016) www.ijirae.com

Performance Comparison of Dimensionality Reduction Methods using MCDR

Suban Ravichandran* *Computer Science & Engineering* *Annamalai University*

Chandrasekaran Ramasamy *Computer Science & Engineering* *Annamalai University*

Abstract*— The recent blast of dataset size, in number of records and in addition of attributes, has set off the improvement of various big data platforms and in addition parallel data analytic algorithms. In the meantime however, it has pushed for the utilization of data dimensionality reduction systems. Mobile Telecom Industry competition has become more and more fierce. In order to improve their services and business in the competitive world, they are ready to analyse the stored data by several data mining technologies to retain customers and maintain their relationship with them. Mobile Call Detail Record (MCDR) comprises diversity and complexity information containing information like Voice Call, Text Message, Video Calls, and other Data Services usages. It is proposed to evaluate and compare the performance of different dimensionality reduction methods such as Chi-Square (Chi^2) Method, Principal Component Analysis (PCA), Information Gain Attribute Evaluator, Gain-Ratio Attribute Evaluator (GRAE), Attribute Selected Classifier (ASC) and Quantile Regression (QR) Methods.

Keywords*— Data Mining, Dimensionality Reduction, PAKDD 2006, MCDR, Chi-Square, PCA, Quantile Regression, Attribute Selected Classifier, Information Gain Attribute Evaluator, Gain Ratio Attribute Evaluator, WEKA.

I. INTRODUCTION

Data Mining (DM) is a way of doing data analysis aimed at finding patterns, revealing hidden regularities and relationships in the data [1]. It is the process of finding useful patterns in data and is known by different names in different communities. DM is the central 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].

Mobile operators have collected massive data of users of their network services. They are willing to mine 100 TB and more of data for optimizing network, strategically positioning service personnel, perform customer service requests, analyse the behavioural aspects of customers and more. The analysis provides information such as human behaviour, communication patterns, geographical behaviour, mode of service usage, economic consuming behaviour, etc.

This paper is organized as follows. Section II describes an overview of PAKDD 2006 Dataset, Different feature selection methods were discussed in Section III, Experimental work is given in Section IV, Performance analysis and comparison are done in Section V and Section VI 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 contains information like personal details, call details, message details, GPRS details, WAP details, application details like games, videos, etc., payment details and handset specification details.

© 2014- 16, IJIRAE- All Rights Reserved

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. 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 METHODS

Machine learning models have been constructed and implemented using different algorithms for identifying 2G/3G customers. To improve classification performance, lower computational complexity, build better optimized models, and reduced memory storage feature selection strategy is applied to the dataset before classification stage. In this work the selection of most appropriate attributes from the dataset in hands, was carried out using various feature ranking methods such as chi square, information gain, Gain Ratio, Attribute Selected Classifier, Quantile Regression and PCA. These feature selection methods measure the strengths of each feature for classification with the criteria in their names, and rank the features based on the measures. The feature selection methods are evaluated and Chi-square technique is identified as optimal feature reduction method.

A. Chi-Squared

Feature Selection via chi-square (χ 2) [5] test is another, very commonly used method Chi-squared attribute evaluation evaluates the worth of a feature by computing the value of the chi-squared statistic with respect to the class. The initial hypothesis H0 is the assumption that the two features are unrelated, and it is tested by chi-squared formula:

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

where Oij is the observed frequency and Eij is the expected (theoretical) frequency, asserted by the null hypothesis. The greater the value of χ 2, the greater the evidence against the hypothesis H0 is. On Applying Chi-Square Attribute Evaluator to the dataset containing 251 attribute and 18000 instances, 13 attributes has been selected.

TABLE II. SELECTED ATTRIBUTE BY CHI – SQUARE ATTRIBUTE EVALUATOR

Attribute Selected	Selected Attributes
Selected Attributes ID by Chi-Square	11, 23, 26, 77, 85, 88, 89, 98, 102, 103, 132, 219, 251
Selected Attributes Names	SUBPLAN HS_AGE HS_MODEL AVG_BILL_AMT AVG_CALL AVG_MINS_OB AVG_CALL_OB AVG_MINS_OBPK AVG_CALL_MOB AVG_MINS_MOB AVG_VAS_GAMES STD_VAS_GAMES Customer Type

The resultant dataset with selected 13 features from Chi – Square Attribute Evaluator as shown in Table II.

B. Principal Component Analysis

Principal component analysis (PCA) is a scientific procedure in which orthogonal change is costumed with a specific end goal to change over an arrangement of potentially related variables into an arrangement of estimations of directly uncorrelated foremost segments [6].

© 2014- 16, IJIRAE- All Rights Reserved

PCA procedure is delicate to the relative scaling of the first variables and is firmly identified with factor analysis. It is the most straightforward of the genuine Eigen vector-based multivariate examination. The PCA operation is useful to uncover the accurate inward structure of the information through the subtle elements in the fluctuation in the information.

C. Quantile Regression

Quantile Regression (QR) is a most promising process to provide useful information in biomedicine, finance, data mining, econometrics, and also in environmental studies [7], [8]. High efficient computational tools are needed for bringing this technique more generally applicable, reliable, and optimal. Advanced algorithms are essential in the state of the art lies in their implementation with contemporary computing techniques. Main purpose of the work is to evaluate the implemented advanced algorithms in combination with the contemporary computational techniques.

By considering Quantile $\tau \in (0,1)$, the Regression Quantile $\beta(\tau)$ can defined as

$$\hat{\beta}(\tau) = \underset{\beta \epsilon R^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(\mathcal{Y}_i - \chi_i' \beta)$$

Where $\mathcal{Y}_i = \{\mathcal{Y}_1, \dots \dots, \mathcal{Y}_n\}$.

D. Attribute Selected Classifier

Attribute selection techniques [9] can be categorized according to a number of criteria. One popular categorization has coined the terms “filter” and “wrapper” to describe the nature of the metric used to evaluate the worth of attributes. Wrappers evaluate attributes by using accuracy estimates provided by the actual target learning algorithm. Filters, on the other hand, use general characteristics of the data to evaluate attributes and operate independently of any learning algorithm. Another useful taxonomy can be drawn by dividing algorithms into those which evaluate (and hence rank) individual attributes and those which evaluate (and hence rank) subsets of attributes.

Dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier. Some of the important options in attribute selected classifier are as follows

- Classifier -- The base classifier to be used.
- Debug -- If set to true, classifier may output additional info to the console.
- Evaluator -- Set the attribute evaluator to use. It is used during the attribute selection phase before the classifier is invoked.
- Search -- Set the search method. This method is used during the attribute selection phase before the classifier is invoked.

E. Information Gain

Information gain (IG) [9], [10] is attributed evaluator used in feature selection when information gain chooses then default the ranker search method gets selected. Information gain is biased towards multivalued attributes, the attribute select measure information gain select the attribute with the highest information gain Let p_i be the probability that an arbitrary tuple in D belongs to class C_i , estimated by $|C_{i,D}|/|D|$ Expected information (entropy) needed to classify a tuple in D:

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

Information needed (after using A to split D into v partitions) to classify D:

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(Dj)$$

Information gained by branching on attribute A

$$Gain(A) = \text{Info}(D) - \text{Info}_A(D)$$

F. Gain Ratio

Gain ratio (GR) is a modification of the information gain that reduces its bias [9], [11]. 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.

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

The attribute with the maximum gain ratio is selected as the splitting attribute.

© 2014- 16, IJIRAE- All Rights Reserved

IV. EXPERIMENTAL SETUP

A. *Data Preparation*

1) *Replace Missing Values*

Missing value [12] 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 Preprocessing Tool developed [12] 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.

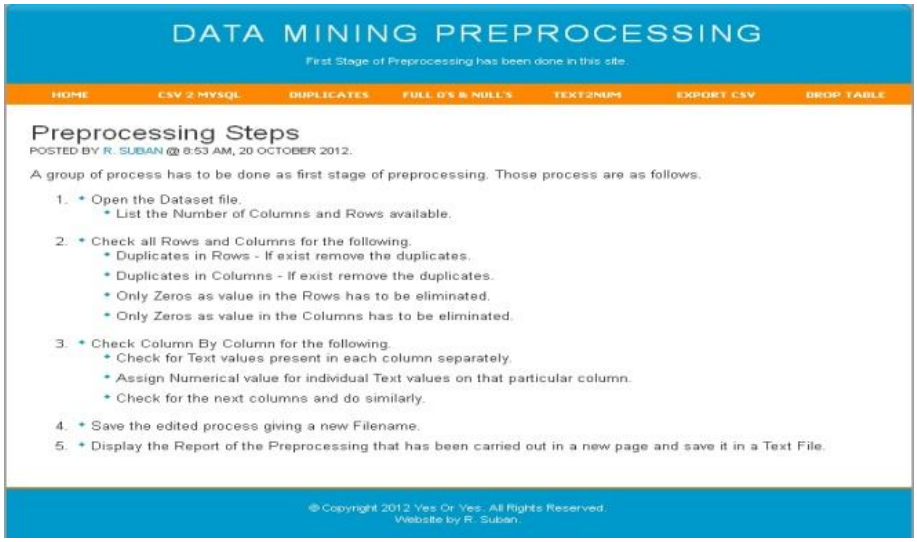

Fig 1. DEVELOPED DATA PREPROCESSING TOOL

Once the data is preprocessed 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 preprocessing contains the same number of attributes and instances as original since no duplicates are available. The resultant dataset contains 251 attribute with 18000 instances.

B. *Feature Extraction & Classification*

The feature extraction extracts a set of features which represents customer retention prediction attributes. Various feature selection methods like Principal Component Analysis, Information Gain, Gain Ratio, Quantile Regression, Attribute Selected Classifier and Chi-Square Attribute Evaluator are 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 methods were employed to select the features, independently. Then based on each selected feature subset, J48 is used to predict the performance of the feature selection. Accuracy is the performance metrics used to compare the different results of J48 on the various reduced features by different selection methods.

© 2014- 16, IJIRAE- All Rights Reserved

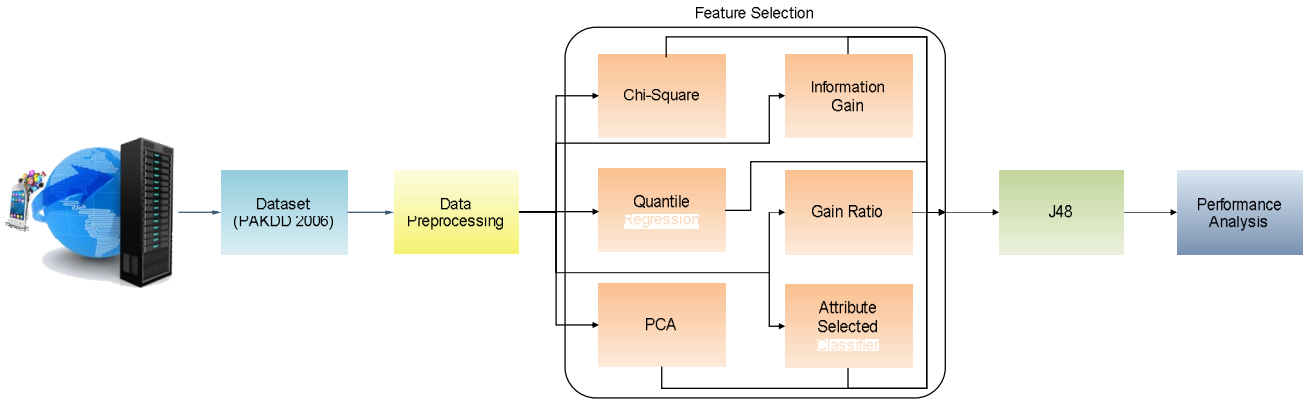

Fig 2. PROPOSED METHOD FOR COMPARING FEATURE SELECTION METHODS USING J48 CLASSIFIER

V. RESULTS AND DISCUSSION

The systematic method for analyzing and comparing various feature selection method is by training the same classifier with the various feature sets obtained by the implementation of various feature selection methods. J48 is taken as the common classifier for comparison since it is performing better for major feature selection methods. Table III will show the performance of the various feature reduction method.

TABLE III. PERFORMANCE OF FEATURE REDUCED DATASET

Feature Selection Methods	Number of Attributes	Accuracy (%)	Time (in sec)
Original	250	81.65	7.2
PCA	110	82.62	4.6
Chi-square	13	83.81	3.1
Information Gain	18	83.33	3.5
Gain ratio	24	82.90	3.9
Quantile	24	82.21	3.4
Attribute selector	49	81.94	5.6

From the Table III and Figure 3, it is clear the Chi-Square Attribute Evaluator performs better than all the other feature selection methods. Chi-Square Attribute Evaluator performed better with highest accuracy of 83.81%, with less time taken of 3.1 secs for evaluation and over all with 13 the least number of attribute.

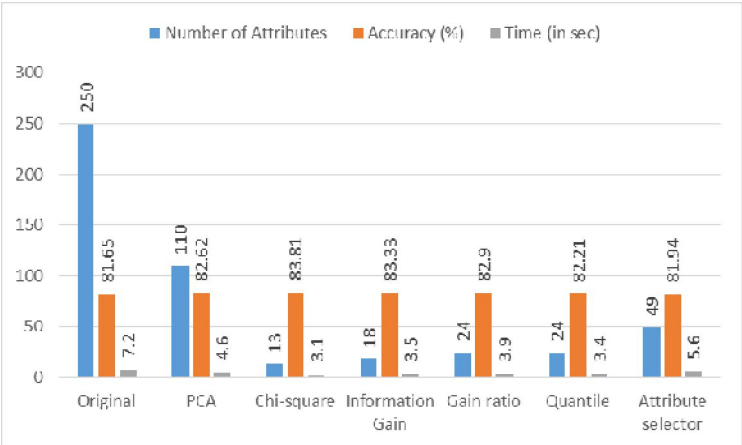

Fig 3. COMPARISON OF FEATURE SELECTION METHODS

VI. CONCLUSIONS

The evaluation of various feature selection methods like Principal Component Analysis, Information Gain, Gain Ratio, Quantile Regression, Attribute Selected Classifier and Chi-Square Attribute Evaluator are done for dimensionality reduction purpose in Telecommunication Industry. Feature selection techniques illustrate that more information is not always good in machine learning applications. Different algorithms can be applied for the data at hand and with baseline classification performance values that can select a final feature selection algorithm. 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.

© 2014- 16, IJIRAE- All Rights Reserved

Overall applying feature selection will always provide benefits such as providing insight into the data, better classifier model, enhance generalization and identification of irrelevant variables. It is found the Chi-Square method performed better than all the other reduction methods in all the aspects like high accuracy, low time taken, and at least number of attribute. This proves that with reduced feature set with less attribute takes less time for processing and with best attribute set selected produce better performance based on accuracy. The attributes selected by Chi-Square Attribute Evaluator plays a vital role on the prediction set and in future based on a perspective clustering method or classification method behavioral analysis of the customers for 2G and 3G services can be predicted.

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] Novaković, Jasmina, Perica ŠTRBAC, and Dušan Bulatović. "Toward optimal feature selection using ranking methods and classification algorithms." *The Yugoslav Journal of Operations Research* 21, no. 1 (2011).
[6] Karamizadeh, Sasan, et al. "An overview of principal component analysis." *Journal of Signal and Information Processing* 4.3B (2013): 173.
[7] Chen, Colin. "An introduction to quantile regression and the QUANTREG procedure." *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc., 2005.
[8] Chernozhukov, Victor, Iván Fernández-Val, and Amanda E. Kowalski. "Quantile regression with censoring and endogeneity." *Journal of Econometrics* 186.1 (2015): 201-221.
[9] Dinakaran, S., and P. Ranjit Jeba Thangaiah. "Role of Attribute Selection in Classification Algorithms." *the International Journal of Scientific & Engineering Research* 4.6 (2013): 67-71.
[10] Pereira, Rafael B., et al. "Information Gain Feature Selection for Multi-Label Classification." *Journal of Information and Data Management* 6.1 (2015): 48.
[11] Karegowda, Asha Gowda, A. S. Manjunath, and M. A. Jayaram. "Comparative study of attribute selection using gain ratio and correlation based feature selection." *International Journal of Information Technology and Knowledge Management* 2.2 (2010): 271-277.
[12] 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.

© 2014- 16, IJIRAE- All Rights Reserved