Available online at www.sciencedirect.com

ScienceDirect

Materials Today: Proceedings XX (2016) XXX–XXX

www.materialstoday.com/proceedings

PMME 2016

Analysis of Clustering and Classification Methods for Actionable Knowledge*

Arumugam P[a], Christy V[b*]

[a]*Associate Professor, Department of Statistics, Manonmanium Sundarnar University, Tirunelveli-627012,Tamilnadu, India*
[b]*Research Scholar,Department of Statistics, Manonmanium Sundarnar University, Tirunelveli-627012,Tamilnadu, India*

Abstract

Data Mining becomes a vital aspect in data analysis. Study on data mining using a synchronized Clustering, Neural based approach gives us the usage trend analysis and it is very much depends on the performance of the clustering of the number of requests. Clustering before classification is termed as cluster Classifier. Numerous classification techniques are there for data modeling. Recently knowledge based approached has become the key forces in data classification. Here performed a four way comparison of Logistic Regression (LR), Classification and Regression Trees (CART), Random Forest (RF) and Neural Network (NN) models using a continuous and categorical dependent variable for classification. A Customer relationship management (CRM) data set is used to run these models. Measurement of different classification accuracy methods are used to compare the performance of the models. Experimental results of test data of the model is used here to predict the accuracy. Based on the efficient method actionable knowledge is derived from the proposed methodology.

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

* Corresponding author. Tel.: +0-91-9900058913
E-mail address: christy.eben@gmail.com

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

Keywords: Data Mining, Random Forest, Actionable Knowledge, Clustering, Classification.

1. Introduction

Statistical methods like regression analysis, multivariate analysis and pattern recognition models have been applied to a wide range of decisions in many disciplines. Machine Learning algorithms have also been used in defining business insights. In this analysis CRM data set is used to analyze the methodology. Clustering before classification works well in defining the data classification [3]. Clustering algorithms has been surveyed and found self organizing map is the current researchers' context in the real world. The objective of the study is to define a hybrid method for clustering based on principle component analysis and neural gas combined with the self organizing map [4,7,9] and using the clusters in classification algorithms like logistic regression, classification and regression trees, neural network and random forest [5,8,6]. Although in this paper the classification method predictive accuracy is compared with CRM dataset. The main objective of this work is to provide an actionable insight to the business to obtain its productivity.

2. Related Works

Evans and Pfahringer introduced a new concept of clustering for classification [3]. This paper shows that clustering prior to classification is beneficial when using the sophisticated classifier. Ngai[5] Compared the data mining techniques like logistic regression, Linear discriminate Analysis and decision trees and found the decision trees as a appropriate model for customer relationship management. The paper [5] try to detection of outliers in multivariate data. There used various outlier techniques such as Mahalanobis distance, Cook‘s distance.[6] Parneet compared classification methods like Multilayer Preceptron from neural network, some tree methods and naïve bayes. [10] who extracted actionable knowledge from decision trees. Decision trees identify the features that are most discerning when it comes to identifying classes. Two well known methods are Boosting and Bagging. Breiman and Cutlers (2013) proposed random forests package using R programming, which add an additional layer of randomness to bagging. The combination of learning models increases the classification accuracy.

3. Proposed Methodology

The effectiveness of the proposed approach has been analyzed using the CRM data set. Almost 126 variables has been extracted and based on the data reduction methods like factor analysis, principle component analysis it has been derived to get some of 15 variables for the analysis [5].

3.1. Proposed Algorithm for Clustering

1. Initialise all weight vectors randomly
2. Chose a random data point from training data and present it to the self organizing map.

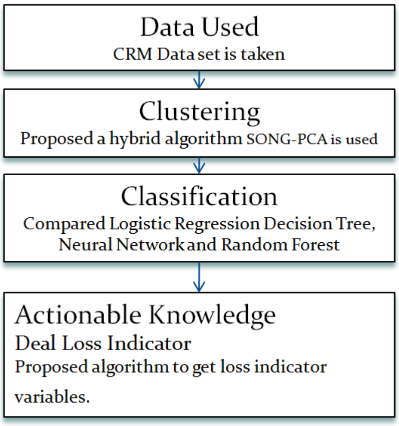
Data Used
CRM Data set is taken
Clustering
Proposed a hybrid algorithm SONG-PCA is used
Classification
Compared Logistic Regression Decision Tree, Neural Network and Random Forest
Actionable Knowledge
Deal Loss Indicator
Proposed algorithm to get loss indicator variables.

Fig. 1. Flowchart of Proposed work.

3. Find the “Bes Matching Unit” (BMU) in map –the most similar node based on Mahalanobis distance Δ.
4. Determine the nodes within the neighbourhood of the BMU
    1. The size of the neighbourhood decreases with each iteration.
5. Adjust Weights of nodes in the BMU neighbourhood towards the chosen data point
    1. Learning Rate decreases with each iteration based on neural gas
    2. The magnitude of the adjustment is proportional to the proximity of the node to the BMU
6. Repeat steps 2 to 5 for N iterations or convergence

Here the self organized algorithm is used with the mahalanobis distance measure as follows. The standardized Mahalanobis distance depends on estimates of the mean, standard deviation, and correlation for the data. A classical approach for is to compute the Mahalanobis distance (MD$_i$) for each observation x_i is,

$$MD_i = \sqrt{(\mathrm{x}_i - \mu)^T \Sigma^{-1} (\mathrm{x}_i - \mu)} \qquad i = 1, 2, \ldots, n \tag{1}$$

Where μ and ∑ are mean and vector, covariance matrix respectively.

3.2. Comparison of Classification Methods

Regression analysis is used to fit data where the relation between independent and dependent variables is nonlinear when the specific form of the nonlinear relationship is unknown. Logistic regression has the binary outcome Y, and the conditional probability $\Pr(Y = 1|X = x)$ is a function of x, any unknown parameters in the function are to be estimated by maximum likelihood. The most obvious idea is to let p(x) be a linear function of x. Every increment of a component of x would add or subtract so much to the probability. The conceptual problem here is that p must be between 0 and 1, and linear functions are unbounded. Finally, the easiest modification of log p which has an unbounded range is the logistic transformation, log (p /1−p).

Formally, the model logistic regression model is that

$$log\frac{p(x}{1-p(x)} = \beta_0 + x.\beta \qquad (2)$$

The decision trees methods are widely accepted and it also has some weak points like data fragmentation in data mining. CART procedure derives conditional distribution of sales **y** given **x** independent variables. The partitioning procedure searches from beginning to end all values of dependent variables to find the independent variable that provides best partition into child nodes. Each and every node spilt is formed based on the conditional probability.The best partition is the one that minimizes the weighted variance and the distribution *f(***y***│Øi)* of **y│x** represents the situation that **x** reside in the side corresponding to the ith terminal node.

Neural network works with both categorical and continuous variables. Neural network has many advantage over classical models used to analyzed data. Neural Network methods handles nonlinearity associated with the data well. NNs method imitates the structure of biological neural network. Processing elements (PE) are the neurons in a Neural Network. The neuron receives one or more inputs, processes those inputs, and generates a single or more output. The main components of information processing in the Neural Networks are Inputs, Weights, Function (weighted average of all input data going into a processing element, Transformation function and Outputs.

Random Forest uses a large number of decision trees are generated randomly for the same data set, and used simultaneously for prediction. A random forest is a classifier consisting of a collection of tree form of classifiers $\{h(\mathbf{x},\Theta k), k=1, ...\}$ where the $\{\Theta k\}$ are iid (independent identically distributed) random vectors and each tree casts a unit vote for the most likely class at input x. Given an ensemble of classifiers $h1$(x), $h2$(x), ..., hK (x), and with the training of random set drawn from the distribution of the random vector Y, X, which defines the margin function as

$$mg(\mathrm{X},Y) = avk\ I(hk\ (\mathrm{X})=Y) - \mathrm{max}j \neq Y\ avk\ I(hk\ (\mathrm{X})=j\ ) . \qquad (3)$$

where $I(\cdot)$ is the indicator function and the margin measures the extent to which the average number of votes at X,Y for the right set exceeds the average vote for any other set . In random forests, hk (X) = h(X, Θk) . The large number of trees follows the Strong Law of Large Numbers and the tree structures used.

In this section the classification analysis is done based on clustering prior to classification is beneficial when using the sophisticated classifier. Based on this concept of classification after clustering is used is our analysis. Random forest has given the efficient classification accuracy based on the clusters we derived from the proposed hybrid algorithm compared with LR, CART and NN.

3.3. Proposed Algorithm for Actionable Knowledge

The random forest algorithm builds multiple decision trees using a concept called bagging. Bagging is the idea of collecting a random sample of observations into a bag. Multiple bags are made up of randomly selected observations obtained from the dataset.

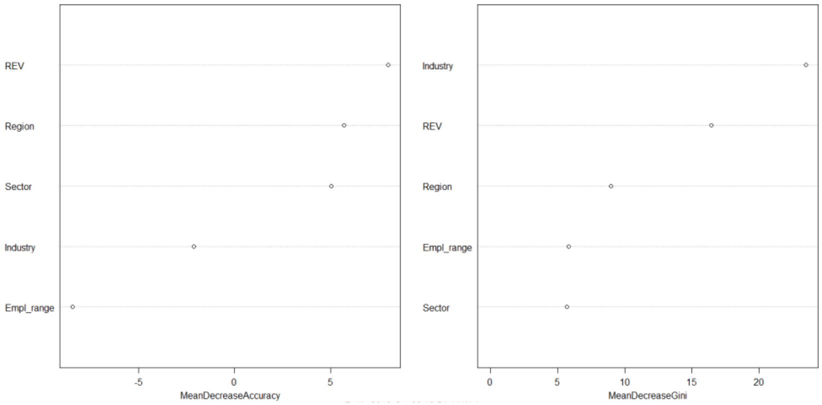

Fig. 2. Random Forest Variable Importance

- Step.1 Create Random Subset with random values.
- Step.2 Build 100 Decision Trees with random subsets.
- Step.3 Classify the error of each (100) decision tree rule and find the min rule.
- Step.4 Define the individual Variable Loss probability value and concentrate only on high probability value >0.8
- Step.5 Sort the corresponding rule variable loss probability.
- Step. 6 Extract the top 3 Variable with high loss probability.

Even though consider the large number of R Random forest packages. Randomforest R package is used in this analysis. Initially 100 trees are running using Random Forest package. Then the proposed methodology is coded using R programming to get the deal loss indicators.

4. Results

The CRM data set is used and it has taken as input with 87 variables in the proposed clustering method. The grasping error is a weighted sum of different translational and rotational deviations from the ideal grasping posture and the lower values are better. Here the values around 1.0 indicate a very good performance. Table 1 reports the similarity results between SOM and proposed algorithm. The Proposed algorithm manages to achieve overall lowest error as 0.96 and it is better than the classic SOM algorithm.

Table. 1. Proposed Clustering Method -Error Results, m is number of principle component

Methods	m=3	m=6	m=9
SOM	1.2	1.3	1.5
Proposed	**0.96**	1.0	1.6

Three principle components with .96 error rate is defined when compared to the SOM cluster with the proposed methodology. After using the proposed methodology in the CRM data three clusters were formed. Based on the profile of these data it named as top, middle, low clusters.

Table. 2. Classification Methods - Receiver Operating Characteristic (ROC) Result

Data Set	LR	CART	RF	NN
Top - Cluster	72%	73%	83%	81%
Medium - Cluster	78%	79%	82%	79%
Low -Cluster	68%	61%	81%	76%
CRM	75%	78%	83%	80%

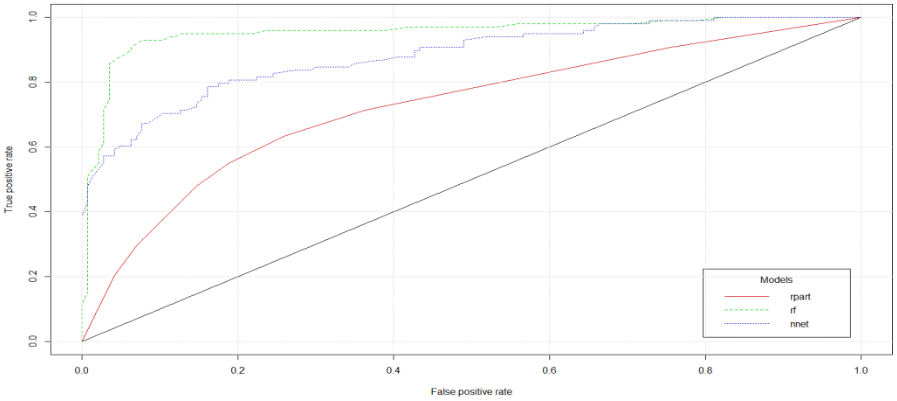

Fig. 3. CRM data set ROC curve for CART, RF and NN.

Three clusters in the CRM data set as well as the entire CRM data are used to run the classification methods. Comparison of all classifiers using R programming is shown in Fig 3 using ROC curve. ROC Curve is a plot of the true positive rate against the false positive rate for the different possible cut points of test data for the model. The area under the curve (AUC) is a measure of model accuracy. It is related to the Gini index coefficient (G_1) by the formula $G_1 = 2AUC - 1$. The performance comparison on the basis of accuracy among methods are shown in Table2.

Table.3. Snap shot of actionable knowledge - Deal Loss Indicators

ID -Cluster	PB	1.Loss Indicator	2.Loss Indicator	3.Loss Indicator
S90345 – C1	.02	Business Line	Business unit	Price
S67846 –C2	.03	Account WR	Product WR	Business Unit
S65793 –C3	.01	Parent Loss	Service	Business Line

Random forest method is used to give the actionable knowledge to the end users. The proposed methodology enhances to extract the loss indictor variable from the selected random forest rules. Finally the loss deal indicators using the R programming is shown in Table 3. It gives us the detailed representation about the deal loss. For each and every deal it can describe the deal loss indicators.

5. Conclusions and Future Enhancement

In this paper classification methods are used to analyze the CRM data set. Prior to classification the proposed hybrid clustering method is used for clustering the data set. The derived clusters are used for classification analysis.

Random forest classification method performs effectively with 83% accuracy. Therefore the proposed the deal loss indicator algorithm is based on the random forest algorithm. The framework consists of algorithms extracting and selecting conditions/rules, and extracting frequent variable interactions/conditions from tree ensembles. Note the methods here can be applied to both classification and regression problems. The proposed algorithm has been implemented using R defined functions. . However, the low error loss rule has been extracted and analyzed for deal indictors. This conclusion can be valuable to the rule mining area. Further extracted deal loss indicators which leveraging the random forest rule and gives the actionable insights for the end users. Also the proposed methodology is easy and understandable and it is interfaced with various techniques. Proposed methodology can be applied to any kind of business to get actionable knowledge to improve the business process. Hence the future is promising for other research areas like sports, healthcare, based on the availability of huge dataset.

Acknowledgements

This research paper is truly contribution and guidance that belong to the list of author. The authors pay homage to Manonmanium Sundarnar University for continuous encouragement, guidance and support during their work.

References

[1] P.Arumugam. and V. Christy, A Hybrid Method for Data Mining. International Journal of Research and Scientific Innovation – IJRSI 3(7), 2016, pp 87-94.

[2] Breiman, Cutler's, Random forests for classification and regression. R Package 'random Forest' version 4.6-7, 2013.

[3] R. Evans, B. Pfahringer, Clustering for classification. In Proceedings of 2011 7th International Conference Information Technology in Asia (CITA 11), IEEE Publications, 12-13 July 2011, pp. 1-8).

[4] P. Hanafizadeh, M. Mirzazadeh, Visualizing market segmentation using self organizing maps and fuzzy Delphi method-ADSL market of a telecommunication company. Expert systems with Applications, 2011, 38(1), pp198-205.

[5] E.W.T. Ngai, Li. Xiu, D.C.K. Chau, Application of data mining techniques in customer relationship management: A literature review and classification, *Expert Systems with Applications. Elsevier,* 36 (2009) : 2592–2602

[6] Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan, Classification and Prediction based data mining algorithms to predict slow learners in education sector, 3rd International Conference on recent trends in computing (ICRTC) 57(2015) , 500-508.

[7] K. Senthamarai Kannan, K. Manoj, Outlier detection in Multivariate data‖, Proceedings of the international conference on Mathematics and its applications – 2014, University College of Engineering, Villupuram, 2014, pp1150-1159

[8] S.Vijayarani, S. Deepa, Protein Sequence Classification In Data Mining– A Study, International Journal of Information Technology, Modeling and Computing (IJITMC), 2 (2014), pp 1 – 8.

[9] Xiufen Fang, Guisong Liu,Ting-zhu Huang, Principal Components Analysis Neural Gas Algorithm for Anomalies Clustering. *WSEAS TRANSACTIONS on SYSTEMS.* 9 (2010).

[10] Qiang Yang, Jie Yin, Charles Ling, Rong Pan, Extracting Actionable Knowledge from Decision Trees, IEEE Transactions on Knowledge and Data Engineering, 19(1), 2007,pp 43-56.