International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 5, 2017 ISSN: 2348-6600
Reference ID: IJCS-179 PAGE NO: 1089-1093
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

Comparison of Feature Selection Methods for Credit Risk Assessment

G. Arutjothi[1], Dr. C. Senthamarai[2]
Ph.D. Research Scholar[1], Assistant Professor[2]
Department of Computer Applications,
Govt. Arts College (Autonomous),
Salem-7, Tamil Nadu, India[1,2]

ABSTRACT- Fund is the greatest variable of the Banking Industry. In Banking Industry achievement and disappointment depends on the credit. Keeping money Industries are focused today with increment in volume, speed and assortment of new and existing information. Managing and analyzing the massive data is more difficult. The credit scoring databases are often large and characterized by redundant and irrelevant features. With this features, classification methods become more difficult. This difficulty can be solved by using feature selection methods. The main objective of the feature selection is to reduce the size of dimensions, costs and increase the classification accuracy. This research paper uses a filter feature selection model for finding the optimal feature subset to evaluate the credit risk. The filter model is implemented using WEKA tool. Comparison study is made to find the credit risk assessment.

Key Terms *- Classification, Data Mining, Credit Risk, Feature Selection, Filter .*

1. INTRODUCTION

Financial data analysis is used in many financial organizations for accurate analysis and assessment of the loan proposal**.** Credit risk is the greatest challenge for the Banking Industry. Granting credit approval depends on borrower's credit risk. Credit risk which encompasses the borrower's ability and willingness to pay and it is one of the main factors for defining a lenders credit policy [9]. The credit scoring databases are often large and characterized by redundant and irrelevant features. Managing and analyzing the massive data is more difficult. With this features, classification methods become more difficult. This difficulty can be solved by using feature selection methods. Datasets with high dimensional features have more complexity and spend larger computational time for classification [1]. Feature selection can be defined as a process that chooses a minimum subset of M features from the original set of N features, so that the feature space is optimally reduced according to a certain evaluation criterion [6]. Feature selection methods can be used to create

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 5, 2017 ISSN: 2348-6600
Reference ID: IJCS-179 PAGE NO: 1089-1093
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

an accurate predictive model and to identify and remove unneeded, irrelevant and redundant attributes from data. Feature selection methods are classified as three categories, such as Filter methods, Wrapper methods and embedded methods.

2. LITERATURE REVIEW

Feature selection techniques improve the classification accuracy and develop a new method in combining the sample domain filtering, resampling and feature subset evaluation methods [1]. Many proposed feature selection methods have been compared and studied it is very difficult to find out the effectiveness of the feature selection methods, because data sets may include many challenges such as the huge number of irrelevant and redundant features, noisy data, and high dimensionality in term of features or samples [2]. Feature extraction carries out a transformation of the original features while variable selection selects a subset from the original features. Variable selection reduce the dimension of data without transforming data into a new set [3]. Janecek[4] showed the relationship between feature selection and data classification and the impact of applying Principle Component Analysis (PCA) on the classification process. Feature selections methods are used to improve the classification accuracy and reduce the subset of data analysis. It can be classified into two major groups such as, filter and wrapper [5].

3. MODEL and RESULT

This research work focus on reducing the feature set using the filter feature selection model. This research work uses feature selection techniques such as Relief, correlation based, entropy filter methods to evaluate the feature set. Managing and analyzing the massive data is more difficult. Financial industries comprise on huge amount of data. It is very difficult to handle the dimensionality data which uses huge number of features. In this study, we have analyzed the classification of credit risk data set using filter feature selection techniques. The bank dataset was taken from UC Irvine Machine Learning Repository [7]. It comprises of 21 attributes and 1000 instances. The feature selection methods used are Relief, Entropy and Correlation Feature selection (CFS). The three classification algorithms Decision Tree, Logistic Regression and Support Vector Machine are used to find the accuracy in the optimal feature subset. The WEKA software is used for analysis and evaluation. The metrics used for analysis are accuracy, recall and precision. Table 4.1 shows the result for three algorithms without using feature selection. Among the three algorithms the highest classification percentage is found in SVM.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 5, 2017 ISSN: 2348-6600
Reference ID: IJCS-179 PAGE NO: 1089-1093
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

Classifier	Classification percentage	Classification accuracy	Precision	Recall
C4.5	72.4	0.724	0.715	0.724
SVM	75.2	0.752	0.749	0.752
LR	73	0.73	0.727	0.73

Table 4.1.Results for Classification Accuracy without Feature Selection

The table 4.2 shows the classification accuracy for three algorithms using filter feature selection methods that is, entropy, relief and CFS.

Feature selection / Classifier		Entropy	Relief	CFS
C4.5	%	70.7	70.6	71.6
	Accuracy	0.7	0.706	0.716
	Precision	0.689	0.701	0.688
	Recall	0.7	0.74	0.726
SVM	%	76.4	74.4	72.2
	Accuracy	0.764	0.744	0.722
	Precision	0.757	0.738	0.71
	Recall	0.764	0.744	0.722
LR	%	75.2	74.8	74.8
	Accuracy	0.752	0.748	0.748
	Precision	0.746	0.743	0.743
	Recall	0.752	0.748	0.748

The table 4.2 shows the classification accuracy for three algorithms using filter feature selection methods that is, entropy, relief and CFS.

Table 4.2 Results for Classification Accuracy with Filter Feature Selection

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 5, 2017 ISSN: 2348-6600
Reference ID: IJCS-179 PAGE NO: 1089-1093
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

The result shows that the highest accuracy is found in entropy filter based SVM.

Fig 4.1 shows the comparison of three algorithms. The graph is drawn with the classification algorithms on x-axis and percentage of classification accuracy on y-axis.

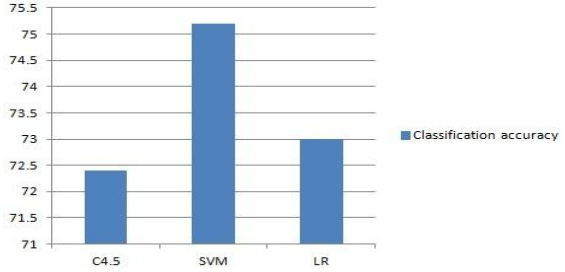

Fig 4.1 Classification accuracy for three algorithms (without using feature selection)

Fig 4.2 shows the comparison of three algorithms with using filter feature selection techniques such as, entropy, CFS and relief. The graph is drawn with the feature selection techniques on x-axis and percentage of classification accuracy on y-axis.

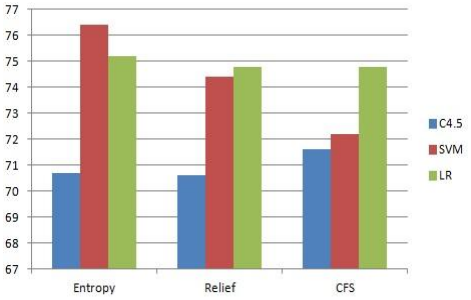

Fig 4.2 Comparison of filter feature selection techniques.

In Fig 4.2 the experimental results shows that the percentage of classification accuracy for bank credit data is high when using feature selection methods as compared to fig 4.1 where the classification accuracy is low by using without feature selection methods.

4. CONCLUSION

The comparison study was made by using three classification algorithms such as C4.5, SVM and LR using with feature selection and without feature selection methods for credit risk assessment. The credit scoring databases are often large and characterized by redundant and irrelevant features. Thus the comparison study proves that classification accuracy is highest for support vector machine algorithm compared to C4.5 and LR. Thus the filter model helps banking Industry to make decision whether the loan can be approved or rejected for the new potential customer.

5. REFERENCES

A. Mehdi Naseriparsa, Amir, Touraj, "A Hybrid Feature Selection Method to Improve Performance of a Group of Classification Algorithm" , International Journal of Computer Applications, Volume 69, No.17,May 2013.

B. Vipin Kumar and Sonajhania minz, " Feature selection: A Literature Review" , Smart Computing Review, Volume 4, no.3, June 2014.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 5, 2017 ISSN: 2348-6600
Reference ID: IJCS-179 PAGE NO: 1089-1093
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

C. Silvia Cateni, Colla and Vannucci, " A Hybrid Feature Selection Method for Classification Purposes", 8th European Modeling Symposium, IEEE, 2014.

D. Janecek, G.K, Gansteres, N, Demel, A and Euker, " On the Relationship between Feature Selection and Classification Accuracy" Journal of Machine Learning and research, Workshop and Conference Proceedings 4, P 90-105.

E. Cheng, Li,Cho and Chen-hong, " A Hybrid Feature Selection Method for Microarray Classification", International Journal of Computer Science, 2008.

F. M.Dash, H.Liu, " Feature Selection for Classification", Intelligent Data Analysis, p.131-156, 1997.

G. http://mlr.cs.umass.edu/ml/datasets.htm lhttp://mlr.cs.umass.edu/ml/datasets.html