International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

A Strategic Framework for Feature Selection in Banking Sector for Credit Risk Analysis

Femina Bahari T, Sudheep Elayidom M

Abstract*: *Feature selection in data mining is critical as it generates optimum subset of features which are relevant for any classification model study. Selecting the optimum subset of features helps to improve the performance of the classifier. A large volume of data gets accumulated on a daily basis in banking industry as part of its various operations. Data is collected and stored in data warehouses for further processing. Information obtained from this data related to customers, transactions, services etc, if analyzed closely can contribute to the growth of industry to a large extent. But the vastness of data and the large number of features in the database makes this analysis a tedious process. Feature selection helps to remove the irrelevant features that adversely affect the performance of learning process in a classifier. A framework combining both filter and wrapper approaches are proposed in our work and an aggregate ranking strategy is followed to select the best features for classification. Naïve Bayes classifier is used in the wrapper method. A district bank dataset is used for testing both approaches and optimum features are selected for further experimental studies in classification.

Keywords*: *Credit risk, Feature selection, Filter approach, Wrapper method, Naïve Bayes Classifier.

I. INTRODUCTION

Credit Risk Analysis is an important topic in the banking industry. The general approach in credit risk analysis is to use the credit history of the existing and previous customers to compute the default risk associated with any new applicant [1], [2]. Data associated with these customers are thoroughly analyzed and studied to generate the best and optimal features that contribute to credit risk analysis. But this task is not only about cutting down the features but retaining the relevant features that are significant to the credit risk analysis model study [3]. In the banking sector, credit worthiness of an applicant can be accessed from their profile, demographic features, transaction history and other details available to the firm as a result of their interaction with the bank. These features can be carefully monitored and analyzed for further processing and obtaining relevant information. Accurate classification increases the creditors profit or reduces his loss and this is in fact beneficial to the borrower as it avoids any kind of over commitment from his side [4].

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Femina Bahari T*, Department of Computer Science & Engineering, Cochin University of Science and Technology, Kerala, India.
Sudheep Elayidom M, Department of Computer Science & Engineering, Cochin University of Science and Technology, Kerala, India.

Reducing the number of features without affecting the outcome of the study is one common method to reduce dimensionality. Feature reduction not only reduces the dimensionality of the data but also reduces the training time required for the induction algorithm, computational cost, improves accuracy , makes the outcome of data mining more knowledgeable and understandable[5], [6]. A feature subset most relevant to the classification is obtained in feature selection [7]. Reducing the feature set has improved the performance of most classification algorithms.

Credit risk is associated with any dealing of a bank in lending to corporate, other banks, individuals, financial institutions [8]. To identify the risks in lending situation is the major task in credit analysis. Apart from identifying the risks, the risk analysis should assess the repayment ability and draw major conclusions regarding the nature of loan, financial needs and risks and make recommendations based on the analysis [9]. The five C's relevant to the borrower forms the major components of a credit analysis. The Character, Capacity, Capital, Condition and Collateral are the major features of the borrower to be analyzed in a credit risk analysis. These terms expose the borrower's morale values, business ability, financial

In industry any technique or strategy that selects the optimum and relevant features, to be used in the learning process of large amount of data is considered as profitable. Such strategies are preferred in the analysis of data as it removes unwanted features from the processing list. Applying a strategic framework to feature selection in banking for credit risk analysis will select the most relevant features the lender has to focus upon performing a decision in business.

Study proposes a strategic framework in selecting the feature subset with application of both filter and wrapper approaches on the selected features. Section II explains the major steps in the Feature Selection Process. Section III explains the feature selection methods. The general strategic framework for the feature selection process and the algorithm for the method are proposed in section IV. Experimental analysis with results and conclusion are covered in section V and VI. The dataset for the experimental study is collected from a Rural District Bank and Weka tool is used for the experimental analysis.

II. FEATURE SELECTION STEPS

The major steps defined in the feature selection process include Subset Generation, Subset Evaluation, Stopping Criteria and Validation [10]. The steps followed in feature selection process are shown in Fig.1.

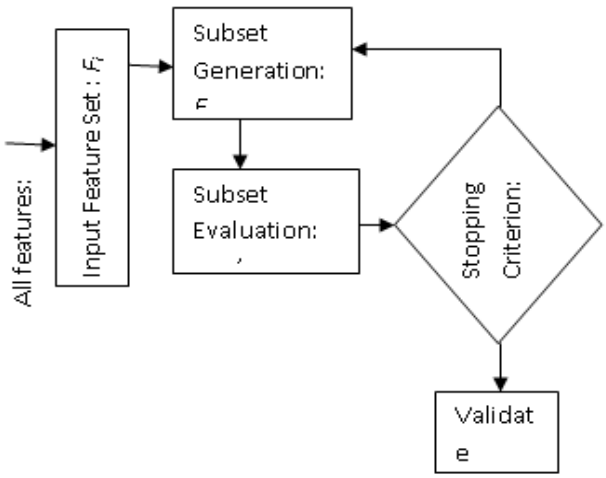

Fig.1. Feature Selection Process

The first step involves searching the feature space and generates the subset that predicts the class most accurately. There are various algorithms that perform the generation of optimum feature subset [11]. While searching the feature space the cost search need to be minimized and generate the optimal subset. Two most common methods used in traversing the feature space are sequential forward selection and backward elimination. In Sequential Forward Selection the search begins with an empty set, adding attributes one at a time. The Backward Elimination method uses the entire set of attributes to begin the search and starts eliminating one at a time until a stopping criterion have been met. Other variations such as random method are also employed to add or delete variables to generate an optimal subset. More complicated search like best first are also used to generate optimal subset [5]. The second step in the process of optimal feature set selection is the evaluation of the features selected. It uses a predetermined evaluation function to measure the goodness of the generated subset [10]. Based on the measurement the evaluated sets are ranked. The ranking of the evaluated sets are used further to select the subsets. Information gain, Correlation Analysis, Gini Index are some of the functions used in the evaluation step. The induction algorithm itself forms the evaluation function by predicting the classification accuracy in the case of wrapper method. The feature search is stopped by a stopping criterion in the third step of the process. If the new feature added to the set does not improve the classification accuracy, the process may be stopped. There are other options to determine the stopping criteria. This includes performing the process for a predetermined number of iterations; perform until a predefined number of features are selected or until top n features with highest rank are selected. Validating the result forms the final step in the process. Methods of feature selection are filter, wrapper and hybrid techniques.

III. FEATURE SELECTION METHODS

Based on the training set labeling, feature selection methods are categorized into supervised, unsupervised and semi-supervised feature selection. Filter, wrapper, and hybrid models are the different categories in supervised feature selection [12].

A. Filter Method

The filter method is independent of the learning algorithm used in the classification step. It uses the properties of the data itself to reduce the number of features used [6], [13]. This approach not only makes possible a reduction in the number of features used in the induction algorithm but also improves the performance of classification algorithm. While it enjoys the benefit of using the same feature in different learning algorithms, it suffers the major drawback of not interacting with the classifier algorithm [6], [14]. Filter based methods are in general faster. Various methods like Pearson correlation, Information gain, Correlation-based feature selection, Gain ratio, etc are used as independent evaluation criteria in filter approach. Various search strategies such as bi-directional search, forward selection, backward elimination, and best first are also used in this approach. The features are ranked based on some evaluation measure selected in the filter approach. A relevant score is generated for each feature in the feature set [15].

B. Wrapper Method

The wrapper algorithms use a predetermined induction algorithm in the feature selection process. Wrapper methods show greater tendency for better results compared to filter methods due to the usage of induction algorithm in the performance evaluation [6]. Given a predefined classifier, a search strategy is applied by the wrapper model on a subset of features, evaluates the selected subset by the classifier performance and repeats the process of selection and evaluation until the desired quality is reached. However the method in general has a greater computational cost and it increases with increase in the number of attributes [13].

C. Hybrid Method

A hybrid method uses a filter method in the first pass and a wrapper method in the second pass. The method incorporates the benefits of the above said two methods. The irrelative features are removed by the first pass and a classifier specific wrapper method is applied further to reduce the feature set [13]. Reduction of feature set from n features to a lower number l reduces the computation space from $2n$ to $2l$. The benefit of this hybrid technique is that it decreases the computational costs for wrapper method by retaining its benefits in feature selection strategies. Hybrid filter wrapper approach is used in various studies with different feature selection strategies and classification algorithms. One such approach uses genetic algorithm to optimize the features of SVM classifier [16]. In another approach decision tree method is used with a combination of variable selection techniques [17].

IV. STRATEGIC FRAMEWORK FOR FEATURE SELECTION

The feature selection strategies aim at generating a subset of the original features with better classifier interpretability, lower computational cost and above all better accuracy for the classifiers.

The feature subset selections strategies are shown in the strategic framework where the process of both filter and wrapper approaches are embedded. As discussed the filter approach uses an independent criteria for feature evaluation whereas the wrapper approach evaluates the classifier performance. The strategic framework for the filter wrapper hybrid approach in feature selection method proposed in our study is shown in Fig.2.

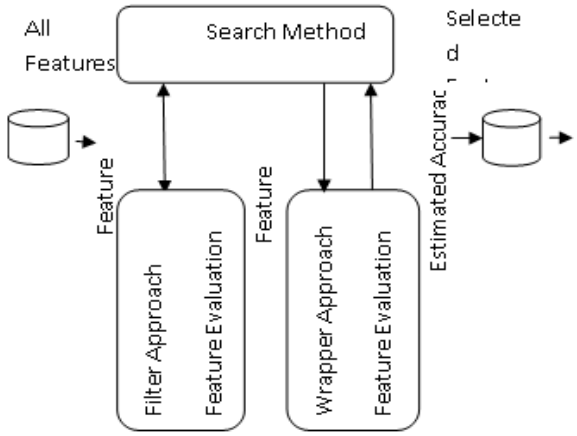

Fig.2. Strategic framework for feature selection method

As discussed earlier in the paper a filter approach with a suitable search technique uses independent criteria for feature evaluation and select the optimal subset. This subset is then used in the wrapper approach in which the classifier performance measure is the criteria to further optimize the feature subset. An estimated accuracy in performance or a best accurate performance is measured as criteria in selecting the optimal set of features. The selected features are used further in the classification study and analysis. A feature selection algorithm is proposed here to illustrate the steps in the feature selection framework and is given in Algorithm 1.

Algorithm 1. Feature Selection Algorithm: Filter Wrapper Hybrid Approach

Input: A: All features.

$\boldsymbol{F_i}$: Subset of features from which the search is started, $\boldsymbol{F_i \subset A}$.

$\boldsymbol{S}$: Stopping Criterion.

Output: $\boldsymbol{F_o}$ *:* Subset of selected features

1: initialize: $F_o=F_i$
2: Ss = *eval* (F_0;A;X); evaluate F_0 for 2 different cases of X : X1, X2
Case X1 : X be an independent criteria .
Case X2 : X be a classification algorithm.
3: while S == Ss do
4: F = *generate* (**A**);
a subset for evaluation is generated.
5: S = *eval* (F; **A**; X);
evaluate the current subset F by X.
6: if S is better than Ss then
7: $Ss = S$
8: $F_o = F$
9: end if
10: end while
11: return

V. EXPERIMENTAL STUDY AND RESULT ANALYSIS

The study was conducted on a dataset collected from a Rural District Bank (RDB). Out of 20404 records the experimental study was conducted on a total of 11113 records after the data pre-processing steps. Manual feature selection was performed in the initial stage and 15 attributes were selected for study. Based on the filter approach a list of 9 attributes was selected in the order of their ranks obtained after the evaluation criteria. The attributes are mentioned in Table I. A wrapper approach with Naive Bayes Classifier algorithm as evaluation criteria was performed and the accuracy of the classifier was taken as the evaluation measure for the optimal subset selection. Both filter and wrapper approaches were used in the hybrid technique used in our study for performing feature selection in our dataset. WEKA toolkit was used for experimental analysis [18]. The evaluation measures used to select feature subsets in our experiment are information gain, gain ratio and Pearson correlation. Information gain with respect to the class was measured to evaluate the worth of each attribute. For a given dataset D and attribute A the information gain was computed for each attribute A as follows:

$$IG(D,A) = E(D) - \sum_{v \in A} \frac{|DA,v|}{|D|} E(DA,v) \quad (1)$$

Where v denoted a value of A and DA, v denoted the set of instances where A has value v and E denoted the overall entropy of the dataset. In filter approaches using information gain and gain ratio evaluation function, the worth of an attribute was evaluated by measuring the information gain and gain ratio with respect to the class respectively. In correlation evaluation function the worth of an attribute was evaluated by measuring the correlation (Pearson's) between it and the class. A ranker search method was used to rank the attributes and an attribute ranking score was generated by individual evaluation. The score of the attributes obtained revealed the worth of the attribute. The ranking score of each attribute and the assigned ranks for each of the evaluation function using filter approach is given below. Table I. shows the rank obtained for attributes using evaluation function Information gain.

Attributes	**Ranker+Infogain**	
	Rank	**Score**
Sex	9	0
Loan Type	8	.0002
Loan Rate	7	.0057
Loan Period	6	.0195

Loan Amount	5	.0266
Age	4	.0301
Income	3	.1135
Asset Value	2	.3861
Loan Balance	1	.5514

Table I. Attribute ranking with Information gain.

Table II. shows the ranks obtained with Information Gain Ratio. It gives the ratio between the information gain and the intrinsic value. For a test intrinsic value can be computed as follows:

$$IV(D,A) = -\sum_{v \in A} \frac{|DA,v|}{|D|} \log_2 \frac{|DA,v|}{|D|} \quad (2)$$

The attributes are rearranged in the order of ranking as shown in Table II.

Attributes	**Ranker+Gain Ratio**	
	Rank	**Score**
Loan Rate	9	0
Sex	8	.0003
Loan Type	7	.0086
Age	6	.0126
Income	5	.0169
Loan Period	4	.0213
Asset Value	3	.0418
Loan Amount	2	.0775
Loan Balance	1	.2483

Table II. Attribute ranking with Gain ratio.

Among two features X and Y Correlation of feature X with the class variable is considered to be more predictive if correlation of feature X is more superior to Y with the class variable. Pearson correlation measure is a very simple and effective feature selection measure for continuous features [7]. Table III shows attribute ranking with Pearson Correlation evaluation function.

Attributes	**Correlation Ranking Filter**	
	Rank	**Score**
Loan Type	9	.0153
Sex	8	.0159
Loan Rate	7	.0167
Loan Period	6	.0202
Asset Value	5	.0508
Age	4	.0535
Loan Amount	3	.0788
Income	2	.1567
Loan Balance	1	.402

Table III. Attribute ranking with Pearson correlation

The performance of the wrapper model was measured using the accuracy of the classifier used in the model study. It also uses the measures of recall and precision values. Experimental set up uses both training set and testing set out of the data samples used for the study. The Naive Bayes algorithm was selected to run the training set, and the learning result was applied on to the testing set to measure the prediction accuracy. WEKA tool was used to run the classifier and the accuracy was noted for different subsets of selected features. The values obtained with each subset of attributes and the accuracy of Naïve Bayes Classifier with different combination of feature subsets are shown in Table IV. F1, F2, F3 denotes feature subsets with selected attributes. A set of 9 attributes gave an accuracy of 75.39% where as a set of 6 attributes gave 77.43%.

Wrapper Approach: Naïve Bayes Classifier	**Attributes Selected**	**Prediction Accuracy**		
		F1	**F2**	**F3**
	9	75.39		
	8	74.3	76.2	75.1
	7	75.19	77.2	76.9
	6	77.43	75.01	76.1
	5	76.5	77.21	77.8

Table IV. Wrapper Approach: Naïve Bays Classifier accuracy with selected feature subsets.

The least significant attributes were identified using the wrapper approach and by comparison of the accuracy of classifier performance obtained with each subset of features. A set of five attributes gave an accuracy of 77.8% which was the best performance obtained with the selected set of features. Combining the results of both filter and wrapper approaches a list of 5 attributes were selected as the major features contributing to the prediction accuracy using Naive Bayes Classifier algorithm. The five deciding attributes selected for our study were the age, income, loan amount, asset, balance amount. These attributes contribute to the five C's relevant to the borrower in any credit analysis.

There are other attributes as well which can contribute to the five C's. But for our experimental analysis we restrict our study to these five attributes for further classifier testing and analysis.

VI. CONCLUSION

Feature Selection plays a key role in the performance of any classification algorithm. A large subset of irrelevant features in the training set will adversely affect the learning performance of the classifiers. An optimal feature set will always improve the learning performance and help build better classifier models. It also lowers the computational complexity with less storage requirement. Use of suitable filter rank aggregation strategy can be applied to select the optimal features in the filter approach method.

An improvement in the accuracy of the classifier can also be achieved by discretizing the selected quantitative attributes. Various discretization techniques can be experimented with naive bayes classifier to improve the performance. We also plan to experiment a fuzzy discretization technique using segmented approach and member function definition with the naive bayes classifier model for further improvement in the classifier performance.

REFERENCES

1. Tsai, C. F. and J. W. Wu, 'Using neural network ensembles for bankruptcy prediction and credit scoring", Expert Systems with Applications, 2008, 34 (4), 2639-2649.
2. Thomas, L, "Consumer credit models: pricing, profit, and portfolios". Oxford University Press.2009.
3. Fernandez, G, "Statistical Data Mining Using SAS Applications" Chapman & Hall/Crc: Data Mining and Knowledge Discovery. Taylor and Francis, 2010.
4. Sadatrasoul, S. M., M. Gholamian, M. Siami, and H. Z. "Credit scoring in banks and financial institutions via data mining techniques: A literature review", Journal of artificial Intelligence and Data Mining. 2013, 1 (2), 119-129.
5. Guyon, I. and A. Elisseeff, "An introduction to variable and feature selection. Journal of Machine Learning Research 2003, 3 (9), 1157-1182.
6. Kohavi, R. and G. H. John, "Wrappers for feature subset selection", Artificial Intelligence, 1997, 97 (1).
7. Rodriguez, I., R. Huerta, C. Elkan, and C. S. Cruz, "Quadratic Programming Feature Selection" Journal of Machine Learning Research, 2010, 11 (4), 1491{1516.
8. Konstantinos Tsiptis and Antonios Chorianopoulo, Data Mining Techniques in CRM: Inside Customer Segmentation, Wiley & Sons LTD, 2009.
9. Segmentation for Credit Based Delinquency Models White Paper Vantage Score ,May 2006
10. H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering", IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4):491-502.
11. Dash, M. and H. Liu, "Consistency-based search in feature selection", Artificial Intelligence, 2003, 151 (1-2), 155-176.
12. P. Mitra, C. A. Murthy, and S. Pal, "Unsupervised feature selection using feature similarity", IEEE Transactions on Pattern Analysis and Machine Intelligence,2002, 24:301– 312.
13. H. Liu and H. Motoda. Computational Methods of Feature Selection. Chapman and Hall/CRC Press, 2007.
14. M.A. Hall and L.A. Smith, "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper", In Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, 1999,volume 235,page 239.
15. Chrysostomou, K., S. Y. Chen, and X. Liu, "Combining multiple classifiers for wrapper feature selection. International Journal of Data Mining, Modeling and Management, 2008, 1 (1), 91-102.
16. Huang, C. L., M. C. Chen, and C. J. Wang, "Credit scoring with a data mining approach based on support vector machines", Expert Systems with Applications,2007, 33 (4), 847-856.
17. Cho, S., H. Hong, and B. C. Ha, "A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the mahalanobis distance: For bankruptcy prediction", Expert Systems with Applications, 2010, 37 (4), 3482-3488.
18. Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2nd edition, 2005.

AUTHORS PROFILE

Ms Femina T Bahari. Research Scholar at Cochin University of science & Technology (CUSAT). Completed M Tech from NIT Calicut. Research interests include CRM and data mining.

Dr Sudheep Elayidom M Professor at CUSAT. Received PhD from CUSAT. First rank holder for BTech and Mtech from M G University, Kerala. Authored book named "Data Mining and warehousing', Cengage Learning. Presented papers in various international and national conferences. Organized, chaired and delivered key note addresses in number of international conferences. Research interests include Big Data and Data Mining.