Algorithm For the Loan Credibility Prediction System

Soni P M, Varghese Paul

Abstract: ***Now a day's people approach or select bank loans to fulfill their needs, which are very common. This practice has been increasing day by day especially for business, education, marriage, agriculture as well. But several people take advantage and misuse the facilities given by the bank. With technology developing at such a peak stage in these days, data mining plays a key role in computer science to solve such issues. Classification is the most suitable predictive modeling technique in data mining to predict the loan repayment capability of a customer in a banking industry. There are various methods to improve the accuracy of a classification algorithm. The accuracy of random forest classification algorithm can be improved using Ensemble methods, Optimization techniques and Feature selection. Various feature selection methods are available. In this research work a novel hybrid feature selection algorithm using wrapper model and fisher score is introduced. The main objective of this paper is to prove that new hybrid model produces better accuracy than the traditional random forest algorithm. This paper also compares the result obtained from other classification methods and feature selection methods to prove that proposed algorithm produces better classification accuracy. The experiments were being done using tools such as weka, R, and python programming. This research aims at introducing a new technique which can increase the progress of banking sector. The accuracy level of this new algorithm in finding the potential of the customer is much higher than the data mining classification algorithm and thus it proves to be very helpful for bank officers.***

Index Terms: ***Classification, feature, LCPS, fisher score, threshold***

I. INTRODUCTION

The computerization of financial operations, connectivity through World Wide Web and the support of automated software's has completely changed the basic concept of business and the way the business operations are being carried out [1]. With the development of computers and network technology, the capacities in the aspects of information production and data collection have been dramatically improved [18]. Data mining place a key role in computer science to solve many of the data related problems. Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set [17].

In this era, a major application of data mining is "Customer Relationship Management". The different areas in which Data mining Tools can be used in the banking industry are customer segmentation, Banking profitability, credit scoring and approval, Predicting payment from Customers, Marketing, Detecting fraud transactions and Cash management and Forecasting operations [2]. A bank loans officer needs analysis of their data in order to learn which loan applicants are safe and which are risky for the bank [3]. The loan credibility prediction system analyzes the applicant's information and classifies him into either applicant with safe credit or applicant with risky credit. Applicant with safe credit has a higher possibility to repay the loan promptly while applicant with risky credit has a high possibility of failing to repay the loan amount. The financial institution's profitability definitely depends on the accuracy of the model. Applying random forest classification algorithm, it is very effective to build a successful predictive model that helps the bankers to take the proper decision regarding safe and risky loan applicants. Random Forest classifier produces high accuracy in both weka and R under the credit data set [4]

To improve the accuracy of Loan Credibility Prediction System (LCPS) several experiments were being done. Classification technique is a commonly used method in data mining which enables the bank to identify the potential of the customer ,i.e. whether the customer is able to repay the loan amount or not. Classification is an important task in areas, like pattern recognition, decision making, and data mining. The classification task can be roughly described as: given a set of objects E ={e1,e2,...,en}, also named examples, cases, or patterns, which are described by m features, assign a class ci from a set of classes C = {c1,c2,...,cj} to an object ep, ep = (ap1,ap2,...,apm) [5]. The accuracy of prediction made during the classification can be improved by applying several other techniques. The major techniques are ensemble, optimization and feature selection. Ensemble method improves the accuracy to a certain extent in multiple data sets. The main idea behind ensemble methodology is to combine a set of multiple learning algorithms in order to obtain better predictive performance that can be obtained from using a single learning algorithm [18]. Optimization has also produced better classification accuracy by changing the values of various selected parameters. The mtry is the main parameter which has been used to improve the accuracy in R programming. The parameter mtry refers to the number of variables selected at each split in making the decision tree of the random forest classification.

Feature selection is one of the most important data preprocessing techniques in data mining and it is closely related to dimensionality reduction. The main idea behind feature selection is ranking the individual features based on some criteria and then searching for an optimal feature

Revised Manuscript Received on June 10, 2019.

Soni P M, Research Scholar, Bharathiyar University, Coimbatore, T.N, India

Varghese Paul, Professor, Department of IT, CUSAT, Cochin, Kerala, India

subset based on evaluation criteria to test the optimality. The accuracy level considerably increased after feature selection methods were applied to the classifier.

This paper introduces a new hybrid feature selection algorithm using wrapper method and fisher score method. The new algorithm is termed as wrapper-fisher feature selection algorithm. In this work, LCPS uses a wrapper-fisher feature selection algorithm to select the most significant features which will improve the accuracy of Random Forest (RF) classification. After studying various past data from the bank it is possible to identify several attributes that can influence the customer behavior. The most influencing attribute can be considered while a new customer approaches the bank for loan and thus we can identify the potential of customer. Here by enabling the bank officers to identify fraud applicants by using the final application of this research work.

The contents of this paper are systemized as follows. The next section discusses about the dataset and tools used for conducting the experiment. Section III describes the proposed experiment. Section IV explains about the study and implementation of feature selection and classification using random forest algorithm. The proposed model and its implementation are discussed in section V. Section VI gives a comparative study of accuracies obtained from two models. Section VII gives the conclusion followed by acknowledgement, future scope and references.

II. DATA SET AND TOOLS

I have selected cooperative bank data for this experiment and believe the cooperative bank officers will benefit the most out of it. Moreover standard data set is being used to implement this algorithm in my experiment, because of which any loan application can be processed through this algorithm. Data collection was completed through procedures including on site observation and interview with the concerned bank officer. A detailed study about the loan processing and banking transactions were also made for the same. The data available consists of 10000 records of bank loan transaction data including 25 data fields. Some of the fields were removed directly by manual data preprocessing. The most widely used benchmark data in the feature selection research are several artificial and real-world problems originating from the University of California at Irvine, known as the UCI data repository [12]. The same experiment was conducted using the Standard credit data set also. The data available consists of 1000 records of bank loan transaction data including 21data fields.

The most popular data mining tools such as weka, R programming was used for predicting the classification accuracy before and after feature selection. Optimization was also done in weka and R. To implement wrapper-fisher feature selection algorithm, python programming were considered. Python programming language is the leading software used to implement this algorithm. The feature list of original dataset and standard dataset after manual data preprocessing are displayed in table 1.

Table 1: Features of original and standard data sets

Features of Original dataset	Features of Standard dataset
Loan No.	Age
Loan Date	Sex
Due date	Job
Loan amount	Housing
Opening	Saving accounts
Payment	Checking account
Receipt	Credit amount
int_rcvd	Duration
fine_rcvd	Purpose
Mem No	Risk
action	
secured	
Loan Balance	
interest Rate	
Category	
Purpose	
gender	
Occupation	

III. PROPOSED EXPERIMENT

The frame work of the proposed experiment is shown in the figure 1. In this experiment the data was selected from two domains and performed data preprocessing to build dataset suitable for further processing. To improve the quality of data and consequently the mining results, the collected data is to be pre processed so as to improve the efficiency of data mining process [19]. Data preprocessing is one of the critical step in data mining process which deals with preparation and transformation from the initial data set to the final data set [19]. The four important data pre processing techniques are data cleaning, data integration, data reduction and data transformation. Here feature selection has a major role in preprocessing to select suitable features that affect the accuracy of algorithm and it comes under data reduction. The two data sets were implemented using traditional random forest algorithm as well as new proposed feature selection algorithm. The comparison of accuracies obtained from two models were made and arrived at a conclusion that loan credibility behaviour of a customer can be predicted more accurately using new proposed feature selection algorithm using wrapper model and fisher score concept.

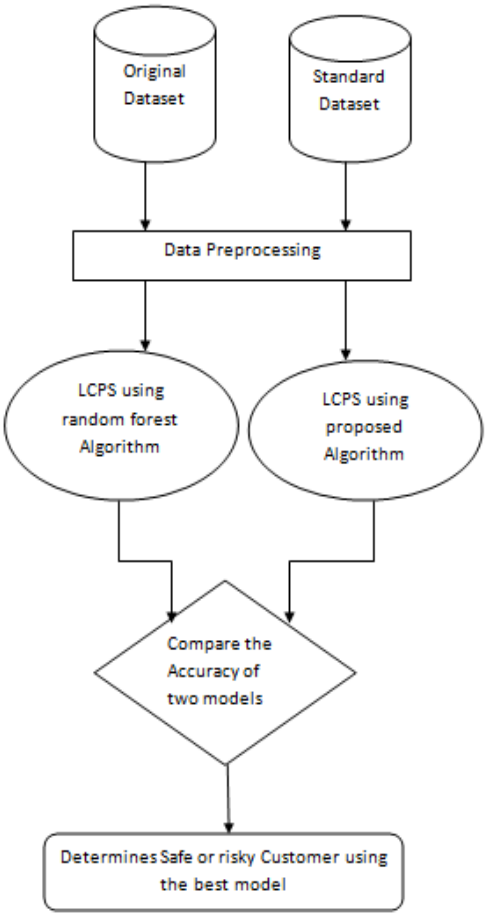

Fig 1: Block diagram of the proposed experiment

IV. LCPS USING RANDOM FOREST

To predict the loan credibility behaviour of a customer, various classification algorithms are used. Each algorithm produces different accuracy. Initially the model was created using random forest algorithm. Feature selection is a data preprocessing technique that has major importance in this research work. This section discusses the process of feature selection and classification process using random forest algorithm.

Feature Selection

Many feature selection techniques are available. Before doing classification feature selection can be applied so as to increase the accuracy level. During the last two decades, feature selection techniques have become an active and fruitful research field in machine learning [6]. Feature selection is an important data mining task which can be effectively utilized to develop knowledge based model for the Loan Credibility Prediction System. Feature selection plays a major role in data preprocessing. Generally dataset consists of relevant, irrelevant and redundant features. Irrelevant and redundant features do not contribute anything to determine the target class and at the same time reduces the accuracy of the model created. The process of eliminating such features from a dataset is termed as feature selection. The search strategy usually employs feature ranking [8, 7] or subset search [9] techniques. In feature ranking, a weight or score is assigned to each feature according to its individual merit.

Ranking methods are not able to remove redundant features within the dataset. Subset search evaluates the quality of subsets of features. The best performance of the selected features can be achieved when both the feature selection and classification stages are optimized together using the same criterion function [10]. Criteria function can be either classifier independent [11] (i.e., filter approach) or classifier specific [11] (i.e., wrapper approach or embedded method). The feature selection methods are depicted in figure 2

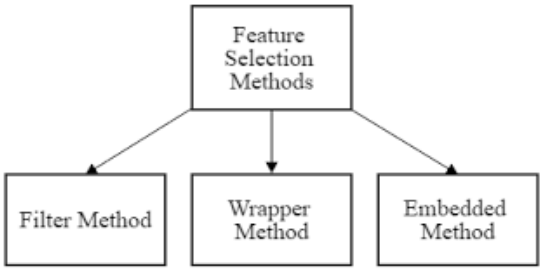

Fig 2: Feature selection methods

The features and their corresponding rank obtained after applying feature selection on the original dataset are listed in the table 2

Table 2: Ranked features of original dataset

Sl.	Features	rank
0	Opening	1
1	int_rcvd	1
2	fine_rcvd	1
3	interest Rate	1
4	Loan amount	2
5	Receipt	3
6	Action	4
7	Purpose	5
8	Occupation	6
9	days	7
10	Payment	8
11	gender	9

The code segment for feature selection related with random forest classification for the loan credibility prediction system implemented in python is explained below:

```
fromsklearn.model_selection import train_test_split
fromsklearn.ensembleimportRandomForestClassifier
feature_names = X.columns
X_train, X_test, y_train, y_test =
train_test_split(X, y, stratify=y, test_size=0.20, random_state=42)
clf = RandomForestClassifier ( random_state=13, class_weight="balanced", max_depth=10, n_estimators=150)
```

clf.fit(X_train, y_train)

feature_imp = pd.Series (clf.feature_importances_, index=feature_names).sort_values(ascending=**False**)

feature_imp

Table 3 depicts the list of features indexed on their feature importance of standard data set

Table 3: Ranked Features of standard data set

Sl.	Features	rank
0	Age	1
1	Credit amount	1
2	Duration	1
3	Saving accounts	2
4	Purpose	3
5	Checking account	4
6	Housing	5
7	Job	6
8	Sex	7

Classification Model

There are several classification algorithms that are used to create software models to predict the accuracy of determining the loan credibility behaviour of a customer. Some examples of classification algorithms are random forest, naïve bayes, k-nearest neighbour, J48, JRip, SMO, and Adaboost etc. In this research work, random forest classifier is selected to build a model for the loan repayment credibility of a customer. Random forests form a family of methods that consist in building an ensemble (or forest) of decision trees grown from a randomized variant of the tree induction algorithm [15]. In this algorithm actual learning is performed by the fit method. This method is called with training data. For example as two arrays X_train and y_train in supervised learning estimators. Its task is to run a learning algorithm and to determine model-specific parameters from the training data and set these as attributes on the estimator object [15]. The code segment for the classification process of loan credibility data set using random forest algorithm is explained as below:

fromsklearn.model_selectionimporttrain_test_split

fromsklearn.ensembleimport RandomForestClassifier

feature_names = X.columns

X_train, X_test, y_train,

y_test = train_test_split(X, y, stratify=y, test_size=0.20, random_state=42)

clf = RandomForestClassifier(random_state=13, class_weight="balanced", max_depth=10, n_estimators=150)

clf.fit(X_train, y_train)

y_pred=clf.predict(X_test)

print("Accuracy:", metrics.accuracy_score(y_pred, y_test))

print("matrix", confusion_matrix(y_test, y_pred))

print("auc", roc_auc_score(y_test, y_pred))

print("f1", f1_score(y_test, y_pred))

fromsklearn.metricsimportclassification_report

print('report', classification_report(y_test, y_pred))

Classification report generated by the random forest prediction model of the original dataset is as shown in figure 3.

report	precision	recall	f1-score	support
0	0.60	0.33	0.43	9
1	0.99	1.00	0.99	673
micro avg	0.99	0.99	0.99	682
macro avg	0.80	0.67	0.71	682
weighted avg	0.99	0.99	0.99	682

Fig 3: RF Classification report of original dataset

The accuracy of the model was obtained as 0.9882697947214076.

V. LCPS USING PROPOSED MODEL

The main aim of my research is to develop a new feature selection algorithm. This feature will enable the banks to predict accurately if the customer can repay the loan on time or not. Kudo and Sklansky applied a two-stage feature selection, where a filter method preceded a wrapper algorithm. This combination was shown to select better feature subsets than if the wrapper model is applied directly [13]. Filter-based methods rank the features as a pre-processing step prior to the learning algorithm, and select those features with high ranking scores [14]. Wrapper-based methods score the features using the learning algorithm that will ultimately be employed [13]. Wrapper-Fisher algorithm is a proposed hybrid feature selection algorithm that comprises the features of both wrapper model and fisher score concept. LCPS produces better accuracy while it was executed by proposed algorithm. The proposed model experimented on two different data sets such as original dataset and standard data set with improved accuracy than traditional random forest algorithm.

Wrapper-Fisher Algorithm

Wrapper methods are based on greedy search algorithms as they evaluate all possible combinations of the features and select the combination that produces the best result for a specific machine learning algorithm Fisher score is one of the ranking methods to select the important features that predict the credibility behavior of a customer. The proposed algorithm selects each feature independently according to their scores under the Fisher criterion, which leads to a suboptimal subset of features. The score of the i-th feature *Si* will be calculated by Fisher Score as in equation (1)

$$S_i = \frac{\sum n_j(\mu_{ij} - \mu_i)^2}{\sum n_j \times \rho^2_{ij}} \quad (1)$$

Where μij and ρij are the mean and the variance of the i-th feature in the j-th class, respectively, nj is the number of

instances in the j-th class and $\mu i$is the mean of the i-th feature. Wrapper fisher algorithm is implemented in python programming. As per the algorithm first step to set the threshold and check for the value of fisher score for each attribute and then decide on the importance of each features/attributes in the Dataset. So for defining threshold make a copy of dataset then shuffle values in each column and score is calculated. Now algorithm use real features and compares its fisher score with the fisher score of the shuffled features which is there as a threshold. So if the fisher score of real features is greater than shuffled (threshold) it will take it as an important features otherwise discard it. Boruta uses the above concept and eliminates features with least importance for further processing. The lists of parameters used in Boruta are [16] :

1. maxRuns: maximal number of random forest runs. Default is 100.
2. DoTrace: It refers to verbosity level. 0 means no tracing. 1 means reporting attribute decision as soon as it is cleared. 2 means all of 1 plus reporting each iteration. Default is 0.
3. getImp: function used to obtain attribute importance. The default is getImpRfZ, which runs random forest from the ranger package and gathers Z-scores of mean decrease accuracy measure.
4. holdHistory: The full history of importance runs is stored if set to TRUE (Default).

The concept of wrapper-fisher algorithm is explained below. It consistes of mainly nine steps.

Step 1: Initialize fisher score threshold value as T.

Step 2: Select the first attribute and calculate its fisher score afs1

Step 3: If afs1 > T, add first attribute into the set TS otherwise add to NTS

Step 4: Repeat step 2 and Step 3 for all attributes in the dataset.

Step 5: Rank the set TS and NTS based on the fisher score

Step 6: Apply classification using TS and compute accuracy

Step 7: Add the first element of NTS to set TS and new set is termed as NTS

Step 8: Apply classification using NTS and compute accuracy

Step 9: if current accuracy > previous accuracy then repeat steps 7 to step 9 until the accuracy remain unchanged otherwise stop

The block diagram portraited in figure 4 explains the logic of the proposed algorithm more clearly. Let 'n' be the total number of features in the dataset and T be the initial fisherscore threshold. The variable ' i ' is used as count variable, First step is to calculate fisher score of each feature and if it is greater than threshold value T then add it to a set called TS otherwise add it to a set called NTS. TS (Threshold Set) is the set of features that satisfies the threshold value and NTS(Non Threshold Set) is the set of features that fails to satisfy the threshold value. Here both the concept of greedy based wrapper method and fisher score were implemented. The next step is to rank the set TS and NTS based on the threshold value. Random Forest Classification starts with the features of TS and measures the accuracy as ACC1. After setting ACC1 , take the first feature of NTS and measure the classification accuracy as ACC2. If it produces a better accuracy , then continue the process until the two features of NTS produces the same accuracy. Otherwise discard the feature of NTS and ACC1 remains as the highest

The corresponding ranked features after feature selection of the original dataset is displayed in table 2. Similarly table3 displays the ranked features of the standard data set. The code segment for obtaining the ranked features using boruta algorithm is explained below:

```
BorutaPy(alpha=0.05,
estimator=RandomForestClassifier(bootstrap=True, cl
ass_weight='balanced',
criterion='gini', max_depth=6, max_features='auto',
max_leaf_nodes=None, min_impurity_decrease=0.0,
min_impurity_split=None, min_samples_leaf=1,
min_samples_split=2, min_weight_fraction_leaf=0.0,
n_estimators=52, n_jobs=4, oob_score=False,
random_state=<mtrand.RandomState object at 0x000
001CF203354C8>, verbose=0, warm_start=False),
max_iter=100, n_estimators='auto', perc=100
random_state=<mtrand.RandomState object at 0x0000
01CF203354C8>,
two_step=True, verbose=2)
feature_df=pd.DataFrame(loan_df.drop(['secured'],axis=
1).columns.tolist(), columns =['features'])
Feature_df['rank']=boruta_selectorranking_feature_df=fe
ature_df.sort_values('rank',ascending =
True).reset_index(drop=True)
Feature_df
```

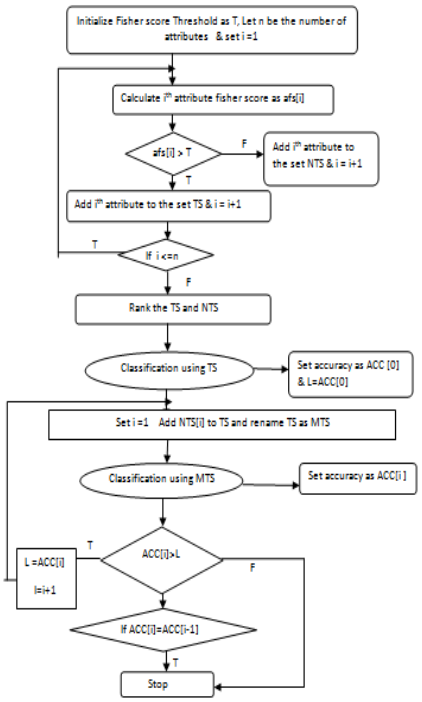

Fig 4: Block diagram Wrapper-Fisher feature selection algorithm.

From the table 2 feature set table 4 lists the features whose fisher score value is greater than threshold and features whose fisher score value is less than threshold of the original dataset. Both sets are termed as TS and NTS respectively. From the table 3 feature set table 5 displays the features of TS and NTS of the standard data set. Sometimes the addition of the feature in the NTS to the optimal feature list results in a better accuracy than the accuracy produced by the features of TS only . But in the original dataset , the addition of the feature "loan amount" from NTS to the optimal feature list produced low accuracy and finally the features selected only from the threshold set. In case of standard data set, the first feature "housing" from NTS produced better accuracy than considering only features in the TS.

Table 4: Features in TS and NTS of original dataset.

Sl	Features in TS	Features in NTS
1	Opening	Loan amount
2	int_rcvd	Receipt
3	fine_rcvd	action
4	interest Rate	Purpose

Table 5: Features in TS and NTS of standard dataset

Sl	Features in TS	Features in NTS
1	Age	Housing
2	Credit amount	Job
3	Duration	Sex
4	Saving accounts	
5	Purpose	
6	Checking account	

Figure 5 portraits the accuracy of classification using the features of TS and features of NTS in the original data set. From the figure it can be proved that adding one feature say loan amount decreases the accuracy to .9873. So the final classification accuracy using new approach is considered as .9912 and clearly it is greater than the traditional classification process. The comparison of TS and NTS using standard data set is depicted in figure 6. The algorithm produced initial accuracy .65 using TS features .

While executing the new algorithm with standard data set, it was found that addition of first feature "housing" from NTS into the optimal feature set increased the accuracy from .65 to .73333. Again the addition of second feature "job" from NTS leads to decrease in the accuracy to .6933 and so we stopped the iteration. So the analysis proved that sometimes the addition of a feature below threshold value also results in a better accuracy. The advantage of wrapper-fisher feature selection algorithm is that it helps to perform feature selection in a successful manner and leads to improve the accuracy of the classification process

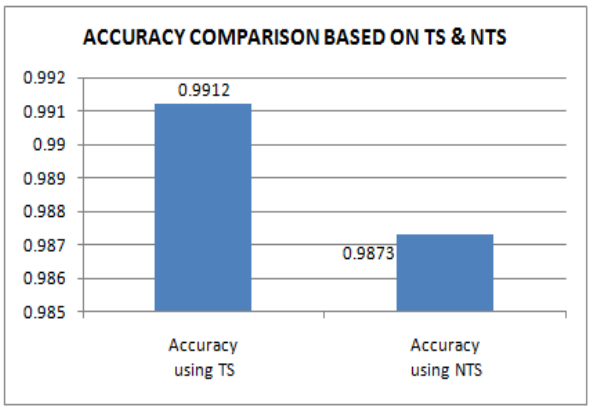

Fig 5: Accuracy comparison of original data set with TS and NTS

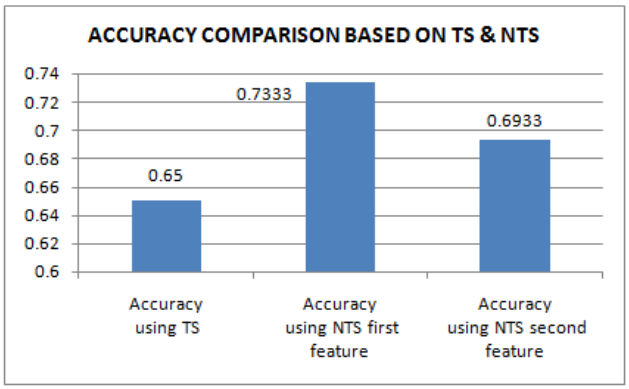

Fig 6: Accuracy comparison of standard data set with TS and NTS

VI. RESULT ANALYSIS

The result analysis and comparison of accuracy performance obtained using the two different models in the original data set and standard data set are expressed in Table 6 and the corresponding graphical representation in figure 7. Wrapper-Fisher algorithm produces better accuracy in both datasets

Table 4: Accuracy comparison of two models

Dataset	Random Forest	Wrapper-Fisher Feature selection
Original	0.98827	0.9912
Standard	0.7222	0.7333

Table 7 explains about the accuracy obtained from various feature selection methods on different data sets. The performance evaluated using the tools weka and python programming. Various feature selection methods experimented were Chisquared, Filtered, InfoGain, OneR, Relief and wrapper-fisher. From the table it is clear that the proposed wrapper-fisher feature selection algorithm helps to improve the accuracy of classification algorithm . Figure 8 also portraits the same.

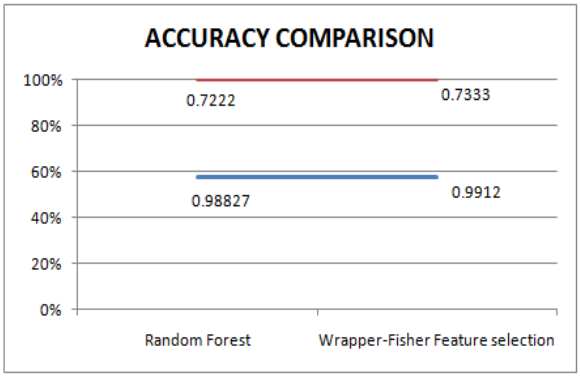

Fig 7: Accuracy Comparison of two models

Table 7: Accuracy obtained from different feature selection methods.

Feature Selection	Tool	Dataset	Accuracy
Chisquared	Weka	Standard	78.4
Filtered	Weka	Standard	74.7
InfoGain	Weka	Standard	74.7
OneR	Weka	Standard	96.1
Relief	Weka	Standard	60.4
Wrapper-Fisher	Python	Original	99.1
Wrapper-Fisher	Python	Standard	73.33

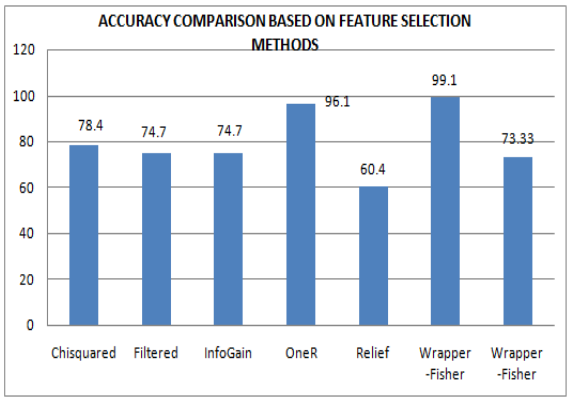

Fig 8: Accuracy comparison based on different feature selection methods

VII. CONCLUSION

In this paper, a novel hybrid feature selection approach is proposed to predict the loan repayment capability behavior of a customer in a cost effective way. Complex set of decision making are need to be taken by bank officers to determine whether to approve loan applicants or not. Normally classification technique solved the problem up to an extent. Now the experiment proved that a model that use feature selection before classification can help the bank officers to take proper decision more accurately. This proposed methodology will protect the bank from further misuse, fraud applications etc by identifying the customers whose repayment capability status is risky especially in the co-operative banking sector. The experiment proved that the classification accuracy have considerably increased after feature selection. The proposed algorithm had produced better accuracy than existing methods. Experiments on standard data sets proved that the proposed algorithm for loan credibility prediction system outperforms many other feature selection methods

VIII. ACKNOWLEDGMENT

I would like to thank my research guide Dr. Varghese Paul, Professor in IT, Cochin University of Science and Technology, who assisted me to do the research work in a successful manner.

IX. FUTURE SCOPE

In future I wish to develop a Data mining application using wrapper-fisher feature selection algorithm and surely it helps the bank officers to take proper decision when a new customer approaches the bank for taking the loan. The proposed hybrid feature selection algorithm for classifying the loan credibility behavior of a customer in a banking industry can also be used for several other applications in the future especially binary classification problems such as prediction of various diseases, prediction of various examination results etc.

REFERENCES

1. VivekBhambri "Application of Data Mining in Banking Sector", IJCSt Vol. 2, ISSue 2, June 2011, ISSN : 2229-4333(Print) | ISSN : 0976-8491(Online)
2. L. Torgo, Functions and data for "data mining with r" R package version 0.2.3, 2012
3. Jiawei Han, MichelineKamber," Data Mining: Concepts and Techniques", Morgan Kaufmann, Elsevier, 2006
4. Soni P M, Varghese paul , "A Novel Hybrid Classification Model For the Loan Repayment Capability Prediction System", International Journal on Future Revolution in Computer Science & Communication Engineering ISSN: 2454-4248 Volume: 4 Issue: 1
5. Marcos EvandroCintra∗, Trevor P. Martin†, Maria Carolina Monard‡, and Heloisa de Arruda Camargo§ , "Feature Subset Selection Using a Fuzzy Method", 2009 International Conference on Intelligent Human-Machine Systems and Cybernetics.
6. Challita N., Khalil M., Beauseroy P. 2016 IEEE IntMultidiscipConfEng Technol. 2016. New feature selection method based on neural network and machine learning; pp. 81–84.
7. I. Guyon and A. Elisseeff. An introduction to variable and feature selection.Journalof Machine Learning Research, 3:1157–1182, March 2003.
8. K. Kirra and L. A. Rendell. The feature selection problem: Traditional methodsand a new algorithm. InProceedings of the 10th National Conference on ArtificialIntelligence, pages 129–134, San Jose, CA, September 1992. MIT Press, Cambridge,MA.
9. P. Pudil, J. Novoviˇcov´a, and J. Kittler. Floating search methods in feature selection.Pattern Recognition Letters, 15:1119–1125, November 1994.
10. C.M.Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1996.
11. A. Jain and D. Zongker. Feature selection: Evaluation, application and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(2):153–158, February 1997.

12. UCI Benchmark Repository – a huge collection of artificial and real-world dataset. University of California, Irvine, Dept. of Information and Computer Sciences. http://www.ics.uci.edu/˜mlearn, June 2004.
13. M. Kudo and J. Sklansky. Classifier-independent feature selection for two-stage feature Selection. In SSPR '98/SPR '98: Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition, LNCS, pages 548–554, London, UK,
14. August 1998. Springer-Verlag.
15. QuanquanGu, Zhenhui Li, Jiawei Han "Generalized Fisher Score for Feature Selection"
16. Gilles Louppe "UNDERS TANDING RANDOM FORE S T S from theory to practice " arXiv:1407-7502v3 [stat.ML] 3 June 2015
17. https://medium.com/@sagar.rawale3/feature-selection-methods-in-machine-learning-eaeef12019cc
18. EsraMahsereciKarabuluta, Selma Ayşe Özelb, TurgayİbrikçibA , "A comparative study on the effect of feature selection on classification accuracy" , Procedia Technology 1 (2012) 323 – 327
19. Kavitha C.R, Soni P.M," A STACKING FRAMEWORK FOR THE PREDICTION OF BINARY CLASSIFICATION PROBLEMS ", International Journal of Civil Engineering and Technology (IJCIET) Volume 8, Issue 5, May 2017, pp. 692–702.
20. Soni P M, Varghese Paul, " AN EFFICIENT DATA PRE PROCESSING FRAME WORK FOR LOAN CREDIBILITY PREDICTION SYSTEM ",IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

AUTHORS PROFILE

Ms. Soni P M obtained BCA degree from Mahathma Gandhi University and MCA from IGNOU. She is pursuing PhD at Bharathiyar University, Coimbatore. Area of research is Data Mining. She has published articles in International and National journals and has also attended Conferences, Seminars and Workshops. She is an academic counselor for IGNOU Computer Application programmes. She has 20 years of teaching experience in handling Computer Science subjects.

Dr. Varghese Paul obtained B.Sc (Engg) degree in Electrical Engineering from Kerala University, M.Tech in Electronics and Ph.D in Computer Science from Cochin University of Science and Technology.

His research areas are Data security using Cryptography, Data Compression, Data Mining, Image Processing and E_Governance. He is the developer of TDMRC Coding System for character representation and encryption system using this coding system. He has got many research publications in international as well as national journals. He has published a text book also.

He had been Professor and Head of Information Technology Department in Cochin University of Science and Technology and currently acts as Research Coordinator there. Also he isResearch Supervisor of Kerala Technological University, MG University Kottayam, Anna Technical University Chennai, Bharathiar University Coimbatore and Bharathidasan University Trichy. 34 research scholars have already completed research studies under his guidance.

Earlier he has worked as Industrial Engineer with O/E/N India Ltd Cochin, Communication Engineer with KSE Board, SCADA Engineer in Saudi Electricity Department and Dean (CS, IT and Research) in Toc H Institute of Science and Technology.

He is a Certified Software Test Manager, Ministry of Information Technology, Government of India. Also, member of Information System Audit and Control Association, USA and Indian Society for Technical Education, India.