Received: November 20, 2018

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Feature Selection using Optimized Multiple Rank Score Model for Credit Scoring

Pannir Selvam Ramila Rajaleximi[1]* **Mohammed Saleem Irfan Ahmed[2]** **Ahmed Alenezi[2]**

[1]*Research and Development Centre, Bharathiar University, Coimbatore 641046, India*
[2]*Department of Computer and Information Sciences, College of Science and Arts, Al Ula, Madhina*
* Corresponding author's Email: psleximi@gmail.com

Abstract: Fraud identification and prevention is becoming the decisive challenge for any financial institutions and financial service industries. Majority of the banks use a credit score, a numerical or statistical value to estimate the customer's creditworthiness based on their credit history. A credit score model will be built by employing training dataset and further, an analytical process will be conceded to estimate the credit score of each customer. Thus, for any credit score model, the lenders acquire various data about the customer from various external agencies. The collected data may encompass irrelevant parameters which will not help in making any decisions and also decelerates the global performance of the model. Accordingly, feature selection is an imperative process to eliminate less relevant attributes for any dataset, especially in credit scoring. This paper emphasizes on enlightening the performance of the attribute selection process using multiple rank score model. The proposed method accumulates the results using the optimized threshold algorithm and outperforms well in selecting the quality attribute for the underlying credit score model. The experimental evaluation has been carried out and it is proved that the accuracy and performance of the suggested method are comparatively better than the existing single rank techniques.

Keywords: Credit score, Feature selection, Relevant attributes, Rank score model, Optimized threshold algorithm.

1. Introduction

Financial institutions or banks become the central focus for 90% of the public in India unambiguously after demonetization. Obviously, as the bank possess money transaction, it is one of the attractive places for the fraud to do activities inline with money transaction. The fraud in the financial sector has been in existence from the initial period over a decade. However, over the years, scams in the financial sector have become more sophisticated due to the extended technology based services offered to customers all over the universe. The Indian financial sector is also suffering the pain owed to the escalation of fraud incidents. A survey signifies that 93% of fraud has grown over the last two years [1]. According to the Reserve Bank of India (RBI), in the past five years, over 23,000 cases of fraud encompassing a monstrous sum of rupees 1 lakh crore have reported from several banks. As per the details given by central bank, the highest sum of rupees 28,459 crores was associated in all these cases of fraud informed from April 2017 to March 2018 [2]. Though fraud activities are common in web applications which are exposed to various vulnerabilities [3] for an online financial transaction to happen, the main focus of this paper is only about managing credit risk in the financial sector.

Each financial institutions promotes its own credit risk policy that controls the bank's credit-granting undertakings and adopts essential techniques for manipulating the activities. This approach provides the satisfactory level of a risk-reward trade-off for its accomplishments [4]. Basically, the financial sectors practice credit scoring model that encompasses information about the customers and their transactions which will be helpful in making smart and effective decisions [5]. Based on the customer's credit information, the credit score model examines and calculates the credit score. Credit score models are quantifiable examination

exploited by credit consultants to estimate the customers' value in acquiring credit. The scores are engendered based on the statistical analysis of the customer's behaviours, and characteristics of past loans to predict the liabilities on future loans [6]. The credit risk prediction is the most dominating issues faced by financial institutions. Due to intensified competition, enhancing customer fulfilment with risk control capability and improving transaction proficiency have become the foremost significance in the financial sector [7]. A credit scoring is essential for purchasing various loans, and necessary services such as insurance and acquiring credit cards. However, automated credit-scoring tools raise longstanding concerns of accuracy and unfairness [8].

In India, four credit information companies are licensed by RBI. Each company serves with individual credit scores using different models. The most widely used method is CIBIL credit score which is a three-digit number that signifies a summary of customers' credit history and their credit rating. This score ranges from 300 to 900, with 750 being the average score and 900 being the best score [9]. The two significant measures such as accurateness and unambiguousness have to be taken care while implementing a decision model for credit scoring [10]. Only the scores given by the external agencies are not sufficient for making a decision. As a result, the financial institutions develop their credit risk policy depending on the transactions made and other details regarding the customer. Thus, the bank collects information from various places and implements their model in addition to the scores provided by the external agencies. The collected information is fundamental for making any decisions. However, the underlying dataset comprises of attribute outliers which are not relevant for the credit score model. The attributes that are not relevant for the study and redundant while assimilating the data acquired from two or more places are named as attribute outliers or feature outliers. These attribute outliers not only increase the computation time but also reduce the efficacy and accuracy of the model [11]. Thus feature selection plays an essential part in any decision making model.

This paper concentrates on selecting features that are relevant for classifying the credit risks using optimized multiple rank score. The proposed method delivers better accuracy since, it uses multiple rank score for the features. Finally, the rank scores are accumulated based on which the decision will be made in selecting features that are useful for elevating the underlying decision model to the next level. Further, the method utilizes an optimized algorithm in accumulating the scores.

The paper is systematized as follows. Section 2 provides the literature survey. Section 3 discloses the architecture of the proposed optimized multiple rank score model for feature selection in two phases with algorithm pseudocode. In section 4 the experimental analysis has been carried out with publically available datasets. Section 5 reveals the results acquired from various experiments. The final section concludes the paper with the summary.

2. Literature survey

2.1 Credit scoring

A credit score model makes loaning process faster. However, it's almost unrealistic to evaluate the enormous amount of data due to the collection of details from various sources, for which feature selection techniques have been employed to predict the financial distress [12]. The financial institutions collect information and generate credit reports regarding the credit applicant and his credit history such as the aggregate number of accounts and its types, an age of his accounts, the frequency of transactions, whether he pay his debts on time. Based on these details, the decision will be made for the new customers [13]. When assessing the risk related to credit, different aspects have to be taken into account. According to the research, various types of scoring are encapsulated as follows [14]:

Application scoring: The scoring which evaluates the creditworthiness for fresh aspirants. It is estimated using various parameters such as social, demographic, financial, and other data during the scrutiny of the application.

Behavioural scoring: The scoring is calculated for the existing candidates based on their behavioural patterns that support credit risk management.

Collection scoring: The scoring clusters the customers for whom the significant actions to be carried out at the first signs of delinquency.

Fraud detection: Fraud / scam scoring models generally rate the customers based on their behaviour and a relative possibility that he may be dishonest.

2.2 Feature selection

However, to increase the proficiency of the model, the feature selection plays a substantial role. The features that are not relevant must be removed which moderates the search space as well as time and speed up the procedure [15, 16]. The modest feature selection algorithm examines each potential subset of attributes until it catches the one which decreases the error rate. Generally, several statistical methods such

as correlation [17] and proximity based approaches [18] have been recommended in finding outliers and in ranking based on scores. The feature selection algorithms can be categorized as filter approach, wrappers approach, and embedded method [19]. Embedded method, a combination of filter and wrapper approaches increases the computational complexity. Wrapper methods consume a prophetic model to score feature subsets whereas filter methods employ a proxy measure as a substitute to score a feature subset [20].

Filter methods compute faster and still catches the effective feature set by allocating the scores for the attributes through distance, information and correlation as a measure. The common methods that are used in filter based approaches are pearson's correlation coefficient, classifier based filters, information gain, gain ratio and relief algorithm. However, the multivariate filters such as correlation based filters are slower, less scalable and may include redundant features whereas the univariate filters such as information gain and gain ratio ignore the attribute dependencies and consider them independently [21]. Another main drawback of the univariate, multivariate and Relief method is that there is no interaction with the classifiers [22]. Classifier based attribute selection includes decision table, JRip, ZeroR classifiers and the attribute selection depends on the classifiers used. Though there are various ranking methods available for attribute selection, all the methods provide different ranks for the same attribute. Another core issue in applying filter method is determining the ending condition of an algorithm which can be solved by applying cross-validation [23, 24].

2.3 Multiple ranks aggregation

Given a set of ranking or ranking scores $R_1, R_2, \ldots, R_n$ for the objects $O_1, O_2, \ldots O_m$, rank aggregation yields a single ranking R in connection with the existing ranks $R_1, R_2, \ldots R_n$. Several methods have been proposed and suggested for integrating the ranks and scores. Assimilating the scores is a statistical technique where each object O_i has n scores and the score of an object O_i is computed using an aggregate scoring function $f(R_{i1}, R_{i2}, \ldots, R_{in})$. Normally the aggregate functions such as min, max, sum and average are used.

$$f(R_{i1}, R_{i2}, \ldots, R_{in}) = MIN\{R_{i1}, R_{i2}, \ldots, R_{in}\} \quad (1)$$

$$f(R_{i1}, R_{i2}, \ldots, R_{in}) = MAX\{R_{i1}, R_{i2}, \ldots, R_{in}\} \quad (2)$$

$$f(R_{i1}, R_{i2}, \ldots, R_{in}) = \{R_{i1} + R_{i2} + \cdots + R_{in}\} \quad (3)$$

$$f(R_{i1}, R_{i2}, \ldots, R_{in}) = \{R_{i1} + R_{i2} + \cdots + R_{in}\} / n \quad (4)$$

In case of combining ranks, several algorithms exist such as Pairwise majority, Borda count, Kemeny optimal aggregation and Spearman's footrule distance [25].

Kemeny optimal aggregation finds the ranking R that minimizes the function as in Eq. (5).

$$K(R, R_1, \ldots, R_n) = \sum_{i=1}^{n} K(R, R_i) \quad (5)$$

Spearman's footrule distance calculates the ranking R that minimizes the function as in Eq. (6).

$$F(R, R_1, \ldots, R_n) = \sum_{i=1}^{n} F(R, R_i) \quad (6)$$

Other algorithms such as Fagin's algorithm [26] and threshold algorithm are extensively used algorithms to aggregate the scores. However, the threshold algorithm is deliberated to be the optimal algorithm since it is correct for every monotone aggregate function f.

3. Optimized multiple rank score model

The architecture of the proposed optimized multiple rank score based feature selection model for credit scoring is depicted in Fig. 1. The proposed framework is divided into two phases. In phase I, the multiple ranks based feature selection has been made using ranker search method with different attribute evaluation strategies. Each ranker provides rank for each attribute which may not be same. And so, for *n* ranker, *n* attribute raking will be the output for the first phase. The second phase is rank score accumulation for which *n* attribute raking acquired from phase I will be fed as an input. The ranking is then transformed as scores and finally, the single ranking will be collected as an output by applying the threshold algorithm. Thus, the final list contains optimal features by eliminating the features that are less relevant. The details about phase I and phase II of the proposed model is explained in this section.

3.1 Phase I: Multiple ranks based feature selection

The phase I evaluates the features based on the ranker search method and it uses several attribute evaluator methods. Each attribute evaluator method supports ranker search method and provides different attribute ranks. Each combination of search method and attribute evaluation scheme produces a set of attribute ranks. Thus, *n* ranker methods produce *n*

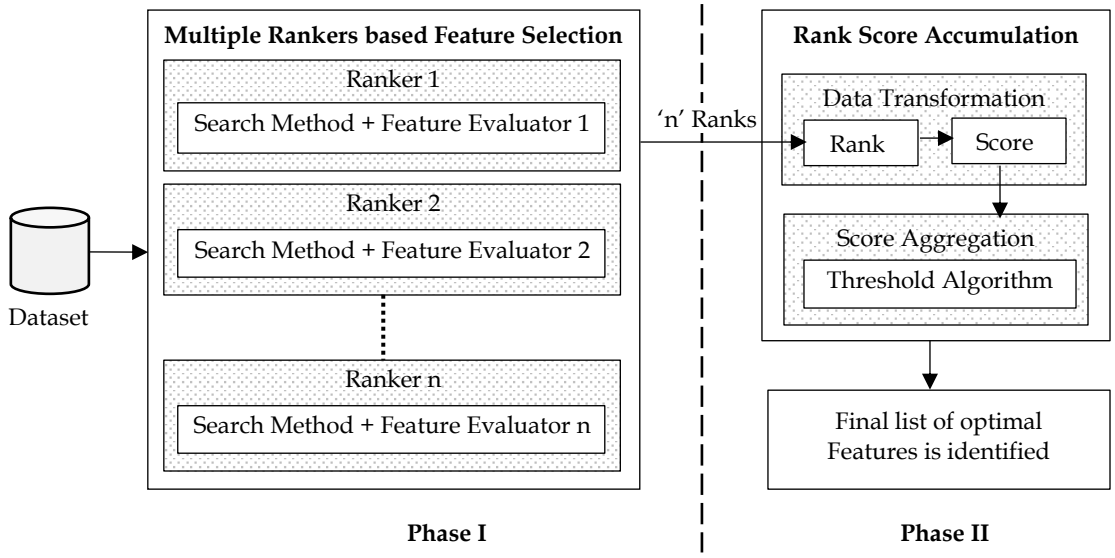

Figure. 1 Architecture of optimized multiple rank score model for feature selection

ranks as an output. It is obvious that each rank produced in phase I are not unique since the feature evaluation algorithm employed for ranking the features are of different types. The input and output of phase I is expressed in the following equation.

Consider the attributes $A_1, A_2, \ldots, A_m$ in a dataset which will be fed as an input for the proposed architecture in phase I where *m* represents the number of attributes. The number of rankers utilized in phase I be *n*. Then the output will be *n* ranking list in which each attribute A_i has $\{R_{i1}, R_{i2}, \ldots, R_{in}\}$ ranks. Thus, the overall output of phase I can be represented in a matrix form with rows representing attributes and columns representing ranking methods, i.e., *m x n* matrix as in Eq. (7), where *m* represents attributes and *n* represents rankings obtained using different methods. Table 1 shows the several attribute evaluation schemes and their description.

Thus, applying *5* ranker search method will obviously produce *5* ranking list which will be the input for the next phase. The detailed architecture of phase I is depicted in Fig. 2. The pseudocode for the proposed multiple ranks based feature selection method is shown in Fig. 3.

3.2 Phase II: Rank score accumulation

The ranks generated for the attributes after applying various attribute evaluation techniques are not same. Each attribute evaluation techniques provides different ranks for the attributes. The next step is to convert the ranks to normalized scores. Though the scores are produced for the attributes using attribute evaluation techniques, they differ drastically in range. Therefore, the ranks are to be converted to scores in a normalized way. The score

$$\begin{pmatrix} A_1 \\ \vdots \\ A_m \end{pmatrix} \rightarrow Phase\ I \rightarrow \begin{pmatrix} R_{11} & \cdots & R_{1n} \\ \vdots & \ddots & \vdots \\ R_{m1} & \cdots & R_{mn} \end{pmatrix} \quad (7)$$

for each attribute in each list is calculated by inverting the ranks.

The scores can be normalized and evaluated using the formula in Eq. (8)

$$Score_i = \frac{1}{Rank_i} \quad (8)$$

Table 1. Description of various attribute evaluators

Attribute Evaluator	**Description**
Classifier	The method assesses the value of a feature by using available classifier such as ZeroR, Decision Table, JRip, M5Rules, OneR, PART
Correlation	The method calculates the value of a feature by determining the Pearson's correlation between the attribute and the class
Gain Ratio	The method determines the value of a feature by computing the gain ratio with respect to the class.
Information Gain	The method examines the significance of a feature by quantifying the information gain with respect to the class.
Relief	The method appraises the value of a feature by recurrently sampling an instance and considering the value of the given feature for the nearest instance of the same and different class.
Symmetrical Uncertainty	The method determines the importance of a feature by evaluating the symmetrical uncertainty with respect to the class.

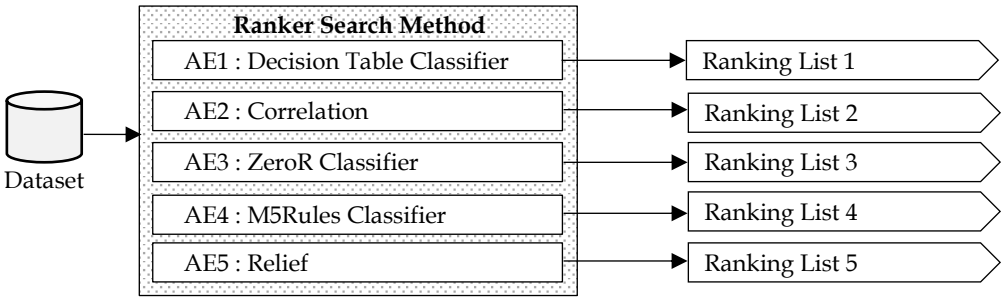

Figure. 2 Detailed architecture of phase I

Algorithm 1 : Multiple ranks based feature selection

Input : Set of training data with *m* attributes
n ranker search method
AE_i – Set of attribute evaluators
Output : R_i - Set of attribute ranks

1. Begin algorithm
2. $R_i = \{\}$
3. Apply the training data with *m* attributes
4. For *i* from *1* to *n* do
5. Choose the ranker search method
6. Choose the attribute evaluator AE_i
7. If AE_i is a classifier
8. Choose the classification algorithm
9. End if
10. Set the cross validation fold as 10
11. Generate the ranks R for *m* attributes
12. $R_i = R_i \cup R$
13. End For
14. End algorithm

Figure. 3 Multiple ranks based feature selection algorithm

Lemma 1 : *Let $A_1, A_2, \ldots, A_m$ be the set of attributes and let R be the corresponding ranks where R_i is the rank of A_i for i=1, 2, … , m. Then score of the attribute can be calculated as*

$$S_i = {}^{1}/_{R_i} \quad \forall\, i = 1, 2, \ldots, m \tag{9}$$

Inversion is a correct choice since the score is indirectly proportional to ranks. When the rank value is increased, obviously the score must be decreased. Once the ranks are converted to scores, the next step is to apply optimal threshold algorithm proposed by Fagin et al. [27] for integrating the ranking scores. The algorithm works on the sorted scores in such a way that at each sequential access, it sets the threshold *t* as the aggregation of scores, do random access to compute the score of each element and maintain a list of top-k elements. The process stops when the scores of the top list are greater or equal to the threshold and finally returns the top *k* elements. The idea behind the threshold algorithm is absolutely true since, any element that it has not seen in the sequential access of top-k elements cannot have a value higher than the stopping value [28].

Algorithm 2 : Rank score accumulation

Input : *Ri* - Set of *n* Ranks with *m* Attributes
Output : *R* - Single Rankset

1. Begin algorithm
2. //Converting ranks to scores
3. For *i* from *1* to *n* do
4. For *j* from *1* to *m* do
5. $S_{ij} = 1/R_{ij}$ //Convert ranks to scores
6. End For
7. End For
8. //Combining scores using threshold algorithm
9. Sort all the *n* rank score list S_i
10. For each sequential access
11. Set the threshold *t* as an aggregate of scores
12. Randomly access and compute the scores of all attributes
13. Maintain a list of top-m elements
14. If (scores of top-m list ≤ threshold *t*)
15 Exit
16. End If
17. End For
18. Return top-*m* attributes

Figure. 4 Rank score accumulation algorithm

After identifying the final scores, the threshold can be set in such a way that the attributes having the scores less than the given threshold can be removed as they are less relevant. Always the final scores after integrating multiple scores will be $n \leq S_i \leq n/m$, where *n* is the number of ranks generated in phase I and *m* is the number of attributes used in the dataset for feature selection.

4. Experimental Analysis

The experimental analysis has been made for the proposed feature selection approach with the dataset available at UCI repository and other publically available datasets. The dataset descriptions such as the names of the datasets, number of instances, number of attributes and their types are presented in Table 2. The German, Australian and Japanese credit

dataset are downloaded from UCI repository and HMEQ is available at www.kaggle.com [29].

The proposed method has been analyzed using the above 4 datasets. The attribute ranks are collected using various attribute evaluator techniques such as Classifier attribute evaluator, Correlation, Gain ratio, Info gain and Relief algorithm. The sample results with a set of ranks after processing phase I for Japanese credit approval dataset is shown in Table 3. From the table, it is clear that each method provides different ranks for each attribute. The evaluation has been made using Waikato Environment for Knowledge Analysis (Weka) tool of version 3.9.3. Once the ranks are identified, the next step is to transform ranks into corresponding scores. Finally, the scores are integrated and single rank set for the attributes are identified using optimal threshold algorithm. The transformation of ranks to rank scores and integration of results after finding the ranks using attribute evaluators [30] such as ZeroR Classifier, Decision Table, M5Rules, Correlation [22] and Relief algorithm [24] available in Weka tool for Australian Credit dataset is shown in Table 4.

Table 4 shows the final rank of the attributes after applying optimal threshold algorithm. In this analysis using Australian Credit Dataset, the threshold value is set as 0.5. Thus the two attributes A1 and A11 having scores less than 0.5 are removed.

5. Results

The performance metrics play a vital role in any sort of research. The performance of the proposed optimized multiple rank score (OMRS) model is compared with traditional information gain ratio

Table 2. Dataset description

Dataset ID	Dataset Name	Instances	Attribute Count	Attribute Types
D1	Australian Credit Approval	690	14	6 numerical, 8 categorical
D2	German Credit	1000	20	7 numerical, 13 categorical
D3	Japanese Credit Approval	690	15	6 numerical, 9 categorical
D4	HMEQ	5960	12	10 numerical, 2 categorical

Table 3. Attribute ranks using ranker search method and using various attribute evaluation techniques for D3

Attributes	Classifier Attribute Evaluators					Correlation	Gain Ratio	Info Gain	Relief	Symmetrical Uncertainty
	ZeroR	Decision Table	JRip	OneR	PART					
A1	15	11	13	13	15	15	15	15	6	15
A2	6	13	10	15	10	9	12	13	10	13
A3	4	7	7	14	7	5	6	8	9	7
A4	2	14	14	11	13	7	8	10	7	11
A5	3	15	15	12	12	6	9	11	8	10
A6	7	6	6	6	6	13	10	6	2	6
A7	5	8	9	7	8	12	11	7	3	9
A8	8	5	4	5	5	4	5	4	11	5
A9	13	1	1	1	1	1	1	1	1	1
A10	14	3	2	3	3	2	3	3	4	3
A11	12	2	3	2	2	3	2	2	13	2
A12	10	10	11	10	11	14	14	14	5	14
A13	11	12	12	9	14	11	13	12	12	12
A14	9	9	8	8	9	10	7	9	14	8
A15	1	4	5	4	4	8	4	5	15	4

Table 4. Calculation of scores based on ranks and integration of scores to find final ranks for Australian credit dataset.

Attributes	Classifier Attribute Evaluators						Correlation		Relief		Final Result	
	ZeroR		Decision Table		M5Rules							
	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Score	Rank
A1	14	0.07	14	0.07	14	0.07	13	0.08	13	0.08	**0.37**	14
A2	6	0.17	11	0.09	11	0.09	10	0.1	10	0.1	0.55	12
A3	4	0.25	7	0.14	8	0.13	7	0.14	8	0.13	0.79	9
A4	2	0.5	8	0.13	10	0.1	8	0.13	4	0.25	1.11	7
A5	3	0.33	4	0.25	6	0.17	4	0.25	2	0.5	1.5	4
A6	7	0.14	6	0.17	7	0.14	6	0.17	7	0.14	0.76	10
A7	8	0.13	5	0.2	4	0.25	5	0.2	5	0.2	0.98	8
A8	9	0.11	1	1	1	1	1	1	3	0.33	3.44	1
A9	10	0.1	2	0.5	3	0.33	2	0.5	12	0.08	1.51	3
A10	13	0.08	3	0.33	2	0.5	3	0.33	14	0.07	1.31	6
A11	12	0.08	13	0.08	12	0.08	12	0.08	11	0.09	**0.41**	13
A12	11	0.09	12	0.08	13	0.08	11	0.09	1	1	1.34	5
A13	5	0.2	10	0.1	9	0.11	14	0.07	6	0.17	0.65	11
A14	1	1	9	0.11	5	0.2	9	0.11	9	0.11	1.53	2

Table 5. Summarization of Test Options

Test options	Test options name	No. of instances used			
		D1	D2	D3	D4
TO1	10-fold cross validation	690	1000	690	5960
TO2	Percentage split (66% training set)	235	340	235	2026
TO3	Full Training Set	690	1000	690	5960

(IGR) [31] for attribute selection and classification using the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm for various test options. Basically, the classification model mainly used for fraud detection or intrusion detection uses precision, recall, F-measure, accuracy, true positive rate, and false positive rate [32]. The performance of the proposed system is analyzed based on the classification rate for the four datasets using various test options such as 10-fold cross validation, percentage split and, training set. The test options are summarized in Table 5.

The performance of the proposed system is measured through the correct and incorrect classification rates along with few statistical measures such as kappa statistic, mean absolute error, root mean squared error, relative absolute error, and root relative squared error which is given in Table 6. In Table 6, the bold style depicts the highest classification accuracy and kappa statistics. The proposed algorithm provides better accuracy rate and kappa statistics for 8 out of 12 tests. For Australian dataset, the suggested algorithm provides higher classification rate for 10-fold cross validation and, percentage split test options. The number of attributes selected using phase I for dataset D1 is 12, D2 is 15, D3 is 13 and, for D4 is 10. In case of German credit and, Japanese credit datasets, the proposed method

Table 6. Performance evaluation through statistical metrics

Test Options	Existing (E) / Proposed (P) Method	Correctly Classified Instances	Correct Classification %	Incorrectly Classified Instances	Incorrect Classification %	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error %	Root relative squared error %
Australian Credit Approval Dataset										
TO1	IGR (E)	586	84.93	104	15.07	**0.70**	0.22	0.35	44.80	70.97
	OMRS (P)	588	**85.22**	102	14.78	**0.70**	0.22	0.35	44.44	70.22
TO2	IGR (E)	202	85.96	33	14.04	0.72	0.24	0.34	48.78	67.89
	OMRS (P)	207	**88.09**	28	11.91	**0.76**	0.20	0.32	40.48	63.69
TO3	IGR (E)	629	**91.16**	61	8.84	**0.82**	0.16	0.28	31.89	56.48
	OMRS (P)	611	88.55	79	11.45	0.77	0.19	0.30	37.64	61.35
German Credit Dataset										
TO1	IGR (E)	710	71	290	29	0.23	0.38	0.45	89.47	98.24
	OMRS (P)	716	**71.60**	284	28.40	**0.24**	0.38	0.45	89.52	97.75
TO2	IGR (E)	259	**76.18**	81	23.82	**0.31**	0.35	0.43	83.65	96.63
	OMRS (P)	255	75	85	25	0.27	0.36	0.42	86.48	95.28
TO3	IGR (E)	729	72.90	271	27.10	0.26	0.39	0.44	92.17	96.02
	OMRS (P)	752	**75.20**	248	24.80	**0.36**	0.36	0.42	85.95	92.73
Japanese Credit Approval Dataset										
TO1	IGR (E)	592	**85.80**	98	14.20	0.71	0.21	0.35	43.42	69.42
	OMRS (P)	592	**85.80**	98	14.20	**0.72**	0.22	0.34	44.62	68.20
TO2	IGR (E)	201	**85.53**	34	14.47	**0.71**	0.24	0.35	47.60	69.94
	OMRS (P)	194	82.55	41	17.45	0.65	0.22	0.39	45.37	77.43
TO3	IGR (E)	610	88.41	80	11.59	0.76	0.20	0.32	40.73	63.82
	OMRS (P)	620	**89.86**	70	10.14	**0.79**	0.18	0.30	36.37	60.31
HMEQ Dataset										
TO1	IGR (E)	5190	**87.08**	770	12.92	**0.51**	0.20	0.33	63.58	83.65
	OMRS (P)	5179	86.90	781	13.10	0.49	0.21	0.34	65.57	84.11
TO2	IGR (E)	1740	85.88	286	14.12	**0.47**	0.22	0.34	68.38	86.83
	OMRS (P)	1745	**86.13**	281	13.87	0.45	0.22	0.35	68.51	87.61
TO3	IGR (E)	5310	89.09	650	10.91	0.58	0.19	0.31	60.46	77.76
	OMRS (P)	5313	**89.14**	647	10.86	**0.60**	0.19	0.31	59.23	76.97

provides better accuracy for 10-fold cross validation and full training set test options whereas, with HMEQ dataset, suggested algorithm achieves better results for percentage split and, full training set test options.

Similarly, the performance of the proposed optimized multiple rank score (OMRS) model is compared with existing information gain ratio (IGR) for the 4 credit score datasets using the Information retrieval metrics such as precision, recall and, F-measure by varying the test options. The values are depicted in Table 7. The bold values represent the best performance of the models with true positive (TP) rate, Precision, Recall and, F-measure. Out of 12 evaluations made, the proposed method provides higher TP rate for 8 cases, higher precision for 7 cases and, better F-measure for 8 experiments.

From the evaluation, it is clear that the proposed optimized multiple rank score method offers better classification in majority of the test cases when compared with the existing information gain ratio based feature selection. Fig. 5 represents the percentage of classification rate for existing and, proposed methods.

Another analysis has been made with other single ranker methods such as rule based classifiers [30] that include ZeroR, Decision Table, M5Rules, Correlation [22] and, Relief algorithm [24] with the aggregation of these methods in proposed optimized multiple rank score model (OMRS) with 10-fold cross validation. The performance metrics such as precision, recall and, F-measure are computed for all the 4 datasets and, the details are given in Table 8 where, E represents the existing single filters and P indicates the proposed method with the aggregation of all the existing methods E. From Table 8, the recall and, F-measure values of the proposed method are higher for the 3 datasets (except HMEQ dataset), whereas, precision value for the proposed method is greater for all the 4 datasets.

From the assessment, the average precision for the existing single filters such as ZeroR, Decision table, M5Rules, Correlation and Relief for the 4 datasets are 80.45%, 80.88%, 79.2%, 80.93% and, 80.85% respectively whereas, for the proposed method the average precision is 91.95%. Thus it is proved that the proposed model outperforms the existing

Table 7. Performance evaluation through information retrieval measures

Test Options	Existing (E) / Proposed (P) Method	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	POC Area
Australian Credit Approval Dataset									
TO1	IGR (E)	0.849	0.152	0.850	0.849	0.849	0.696	0.873	0.851
	OMRS (P)	**0.852**	0.142	**0.856**	**0.852**	**0.853**	0.706	0.873	0.851
TO2	IGR (E)	0.860	0.131	0.873	0.860	0.859	0.733	0.864	0.817
	OMRS (P)	**0.881**	0.117	**0.882**	**0.881**	**0.881**	0.763	0.898	0.862
TO3	IGR (E)	**0.912**	0.104	**0.917**	**0.912**	**0.911**	0.825	0.907	0.880
	OMRS (P)	0.886	0.107	0.890	0.886	0.886	0.774	0.915	0.885
German Credit Dataset									
TO1	IGR (E)	0.710	0.501	0.686	0.710	0.689	0.242	0.612	0.657
	OMRS (P)	**0.716**	0.499	**0.692**	**0.716**	**0.694**	0.255	0.615	0.655
TO2	IGR (E)	**0.762**	0.484	**0.741**	**0.762**	**0.744**	0.324	0.637	0.684
	OMRS (P)	0.750	0.510	0.726	0.750	0.730	0.284	0.625	0.683
TO3	IGR (E)	0.729	0.503	0.706	0.729	0.702	0.279	0.613	0.644
	OMRS (P)	**0.752**	0.413	**0.739**	**0.752**	**0.741**	0.371	0.672	0.685
Japanese Credit Approval Dataset									
TO1	IGR (E)	**0.858**	0.142	0.859	**0.858**	**0.858**	0.714	0.874	0.852
	OMRS (P)	**0.858**	0.135	**0.863**	**0.858**	**0.858**	0.719	0.884	0.867
TO2	IGR (E)	**0.855**	0.138	**0.862**	**0.855**	**0.855**	0.717	0.859	0.809
	OMRS (P)	0.826	0.178	0.825	0.826	0.825	0.649	0.844	0.809
TO3	IGR (E)	0.884	0.125	0.885	0.884	0.884	0.765	0.891	0.858
	OMRS (P)	**0.899**	0.114	**0.901**	**0.899**	**0.898**	0.796	0.899	0.869
HMEQ Dataset									
TO1	IGR (E)	**0.871**	0.457	**0.868**	**0.871**	**0.854**	0.544	0.723	0.820
	OMRS (P)	0.869	0.470	0.867	0.869	0.851	0.536	0.709	0.812
TO2	IGR (E)	0.859	0.468	0.848	0.859	0.843	0.489	0.699	0.797
	OMRS (P)	**0.861**	0.496	**0.854**	**0.861**	**0.842**	0.492	0.684	0.786
TO3	IGR (E)	**0.891**	0.411	**0.895**	**0.891**	0.877	0.627	0.741	0.825
	OMRS (P)	**0.891**	0.374	0.889	**0.891**	**0.881**	0.628	0.762	0.837

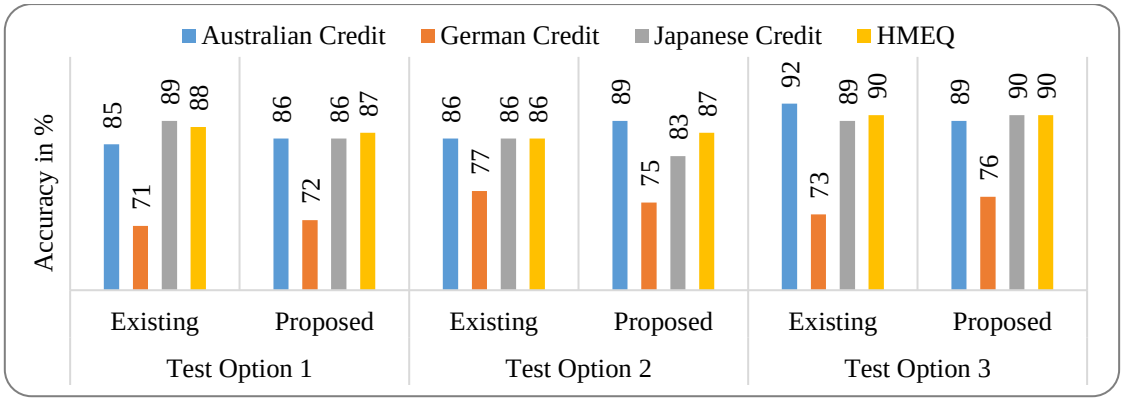

Figure. 5 Comparison on classification percentage of all datasets

Table 8. Performance evaluation of proposed multiple rank score model with individual rank filters

Ranker Methods	Precision	Recall	F-Measure
Australian Dataset			
ZeroR (E)	0.846	0.846	0.846
Decision Table(E)	0.848	0.848	0.847
M5Rules (E)	0.840	0.841	0.840
Correlation (E)	0.850	0.848	0.848
Relief (E)	0.846	0.846	0.846
OMRS (P)	**0.856**	**0.852**	**0.853**
German Credit Dataset			
ZeroR (E)	0.683	0.713	0.677
Decision Table(E)	0.680	0.710	0.679
M5Rules (E)	0.680	0.710	0.679
Correlation (E)	0.673	0.696	0.681
Relief (E)	0.662	0.688	0.668
OMRS (P)	**0.692**	**0.716**	**0.694**
Japanese Credit Dataset			
ZeroR (E)	0.862	**0.858**	**0.858**
Decision Table(E)	0.850	0.849	0.849
M5Rules (E)	0.858	0.854	0.854
Correlation (E)	0.854	0.854	0.854
Relief (E)	0.861	0.854	0.854
OMRS (P)	**0.863**	**0.858**	**0.858**
HMEQ Dataset			
ZeroR (E)	0.827	0.839	0.808
Decision Table(E)	0.857	0.863	0.844
M5Rules (E)	0.790	0.812	0.796
Correlation (E)	0.860	0.863	0.841
Relief (E)	0.865	**0.870**	**0.855**
OMRS (P)	**0.867**	0.869	0.851

individual rank filter model. Thus, the optimized multiple rank score model works better and, provides better classification accuracy for the credit score datasets after removing the outlier attributes.

6. Conclusion

This paper introduces an optimized multiple rank score model for feature selection. Several ranking methods are used and are optimized using threshold algorithm to calculate the score for each attribute. Experiments have been performed for the proposed algorithm with commonly used existing single filter attribute selection methods. Based on the analysis made using statistical measures and, information retrieval metrics, it is proved that the proposed method produces good results for 8 out of 12 cases by finding the attribute relevancy through multiple ranks. The future work focuses on other attribute selection approaches and in improving the classification accuracy of imbalanced datasets specifically for credit scoring datasets.

References

[1] S.V. Ramana and S.G. Krishna, “A study on impact of fraud in Indian banking sector (with special reference on retail banking products)”, *International Journal of Academic Research and Development*, Vol.2, No. 6, pp.544-547, 2017.

[2] The Times of India Business, PTI, May 2, 2018. https://timesofindia.indiatimes.com/business/india-business/over-23000-bank-frauds-worth-rs-1-lakh-crore-reported-in-5-years-rbi/articleshow/63998429.cms

[3] A. Saravanan, M. S. Irfan Ahmed, and S. Sathya Bama, “A Survey on Exposed Vulnerabilities in Web Applications”, *Asia Pacific Journal of Research*, Vol.1, No. 35, pp.84-89, 2016.

[4] RBI Notification on Credit Risks. https://www.rbi.org.in/scripts/NotificationUser.aspx?Mode=0&Id=906

[5] W. Bouaguel, G. B. Mufti, and M. Limam, “Similarity Aggregation a New Version of Rank Aggregation Applied to Credit Scoring Case”, *Mining Intelligence and Knowledge Exploration, Lecture Notes in Computer Science, Springer, Cham*, Vol.8284, pp.618-628, 2013.

[6] S. Arya, C. Eckel, and C. Wichman, "Anatomy of the Credit Score", *Journal of Economic Behavior & Organization*, Vol.95, pp.175-185, 2013.

[7] X. Chen, C. Zhou, X. Wang, and Y. Li, "The Credit Scoring Model Based on Logistic-BP-AdaBoost Algorithm and its Application in P2P Credit Platform", In: *Proc. of International Forum on Decision Sciences*, pp. 119-130, 2017.

[8] M. Hurley and J. Adebayo, "Credit Scoring in the Era of Big Data", *Yale Journal of Law and Technology*, Vol.18, 148, 2016.

[9] RBI Notification. RBI/2009-10/354 RPCD.CO RRB.BC No.62 /03.05.33/2009-10

[10] B. Baesens, C. Mues, M. De Backer, J. Vanthienen, and R. Setiono, "Building Intelligent Credit Scoring Systems using Decision Tables", *Enterprise Information Systems V*, pp.131-137, 2004.

[11] T. Howley, M.G. Madden, ML. O'Connell, and A.G. Ryder, "The Effect of Principal Component Analysis on Machine Learning Accuracy with High-Dimensional Spectral Data", *Applications and Innovations in Intelligent Systems XIII*, pp.209-222, 2006.

[12] D. Liang, C. F. Tsai, and H. T. Wu, "The Effect of Feature Selection on Financial Distress Prediction", *Knowledge-Based Systems*, Vol.73, pp.289-297, 2015.

[13] C.F. Tsai and J.W. Wu, "Using Neural Network Ensembles for Bankruptcy Prediction and Credit Scoring", *Expert Systems with Applications*, Vol.34, No.4, pp.2639-2649, 2008.

[14] M. Siami and Z. Hajimohammadi, "Credit Scoring in Banks and Financial Institutions via Data Mining Techniques: A Literature Review", *Journal of AI and Data Mining*, Vol.1, No.2, pp.119-129, 2013.

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, Vol.112, New York, Springer, 2013.

[16] Y. Liu and M. Schumann, "Data Mining Feature Selection for Credit Scoring Models", *Journal of the Operational Research Society*, Vol.56, No.9, pp.1099-1108, 2005.

[17] S. Sathya Bama, M. S. Irfan Ahmed, and A. Saravanan, "A Mathematical Approach for Improving the Performance of the Search Engine through Web Content Mining", *Journal of Theoretical & Applied Information Technology*, Vol.60, No.2, 2014.

[18] S. Sathya Bama, M. S. Irfan Ahmed, and A. Saravanan, "Relevance Re-ranking Through Proximity Based Term Frequency Model", *International Conference on ICT Innovations*, pp. 219-229, 2016.

[19] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, pp.1157-1182, 2003.

[20] G. Forman, "BNS Feature Scaling: An Improved Representation Over TF-IDF for SVM Text Classification", In: *Proc. of International ACM Conf. on Information and Knowledge Mining*, pp.263–270, 2008.

[21] R. Aziz, C. K. Verma, and N. Srivastava. "Dimension Reduction Methods for Microarray Data: A Review", *AIMS. Bioengineering*, Vol.4, No.1, pp.179-197, 2017.

[22] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A Review of Feature Selection Methods on Synthetic Data", *Knowledge and Information Systems*, Vol.34, No.3, pp.483-519, 2013.

[23] A. Figueroa, "Exploring Effective Features for Recognizing the User Intent Behind Web Queries", *Computers in Industry*, Vol.68, pp.162-169, 2015.

[24] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based Feature Selection: Introduction and Review", *Journal of Biomedical Informatics*, 2018.

[25] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank Aggregation Methods for the Web", In: *Proc. of International Conf. on World Wide Web*, pp.613-622, 2001.

[26] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee, "Comparing and Aggregating Rankings with Ties", In: *Proc. of International ACM Symposium on Principles of Database Systems*, pp.47-58, 2004.

[27] R. Fagin, A. Lotem, and M. Naor, "Optimal Aggregation Algorithms for Middleware", *Journal of Computer and System Sciences*, Vol.66, No.4, pp.614-656, 2003.

[28] P. Cao, and Z. Wang, "Efficient Top-k Query Calculation in Distributed Networks", In: *Proc. of the Annual ACM symposium on Principles of Distributed Computing*, pp.206-215, 2004.

[29] Kaggle Criteo Labs. Display Advertising Challenge. 2014. https: //www.kaggle.com/c/criteo-display-ad-challenge. Accessed 05.09.2018.

[30] E. Kalapanidas, N. Avouris, M. Craciun, and D. Neagu, "Machine Learning Algorithms: A Study on Noise Sensitivity", In: *Proc. of Balcan Conference in Informatics*, pp. 356-365, 2003.

[31] J. Dai and Q. Xu, "Attribute Selection based on Information Gain Ratio in Fuzzy Rough Set Theory with Application to Tumor Classification", *Applied Soft Computing*, No. 1, pp. 211-221, 2013.

[32] S. Sathya Bama, M. S. Irfan Ahmed, and A. Saravanan, "A Survey on Performance Evaluation Measures for information Retrieval System", *International Research Journal of Engineering and Technology*, Vol.2, No.2, pp.1015-1020, 2015.