International Journal of Advance Research in Science and Engineering
Volume No.07, Issue No.02, February 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

A CREDIT SCORING PREDICTION MODEL BASED ON HARMONY SEARCH BASED 1-NN CLASSIFIER FEATURE SELECTION APPROACH

Dr.V.Krishnaveni[1], Dr.K.G.Santhiya[2], Mr.P.Ramesh[3] , Mr.S.Jaganathan[4]

[1] Associate Professor , Department of Computer Science ,
Kongu Arts and Science College(Autonomous),Erode, Tamilnadu, India
[2] Associate Professor , Department of Computer Science,
Kongu Arts and Science College,Erode, Tamilnadu, India
3 Assistant Professor and Head, Department of Computer Science,
Kongu Arts and Science College,Erode, Tamilnadu, India
[4] Assistant Professor, Department of Computer Science,
Kongu Arts and Science College,Erode, Tamilnadu, India

ABSTRACT

This article presents a method of feature selection to improve the accuracy and the computation speed of credit scoring models. In this paper, a credit scoring model based on Harmony Search based 1-NN classifier and feature selection method has been proposed to evaluate the credit risks of applicants. The harmony Search optimization method is used to select the relevant features so as to improve the Classification accuracy of the 1-NN classifier. In this article, an algorithm to reduce the dimensionality has been proposed by using Music-Inspired harmony Search optimization algorithm. Consequently, the classification accuracy of the 1-NN classifier has been improved along with a remarkable runtime reduction. Thus the proposed model can perform feature selection and model parameters optimization at the same time to improve its efficiency. The performance of our proposed model was experimentally assessed using two public datasets which are Australian and German datasets. The obtained results showed that an improved accuracy of the proposed model has been achieved compared to other commonly used feature selection methods. In particular, the proposed method could attain the average accuracy of 80.4% with a significantly reduced running time of 57 minutes on German credit dataset and the highest average accuracy of 93.5% with the running time of only 34 minutes on Australian credit dataset. This method can be usefully applied in credit scoring models to improve accuracy with a significantly reduced runtime.

Keywords: Credit Scoring, Machine learning, Feature Selection, Harmony Search Optimization, 1-NN classifier

I. INTRODUCTION

Credit scoring has become one of the main analytical ways for financial institutions to assess credit risk. The purpose of credit scoring is to classify the applicants into two groups: applicants with good credit and applicants with bad credit. Applicants with good credit have great possibility to repay financial obligation while, applicants with bad credit have high possibility of defaulting. The credit risk analysis plays an important role in categorization of customers which allows the customers to be divided into two sets, good and bad. An important goal of the credit risk prediction is to construct the best classification model for a particular data set. There are a lot of irrelevant and redundant features in financial data in general and credit data in particular. When the data is noisy and unreliable by the redundancy and the deficiency in data, the accuracy of classification can be reduced that may lead to bad decisions. In that case, a feature selection strategy is deeply needed in order to filter the redundant features. In order to select a subset of relevant features, feature selection is needed. The subset is sufficient to describe the problem with high precision. Feature selection thus reduces the dimension and the computational complexity of the problem and saves on the cost of measuring non selected features.[1-9]

Credit scoring is a statistical method used to evaluate the credit risk against customers through using customer data and activities. Credit scoring is performed by the bank based on judgmental view of credit experts, credit groups or credit bureaus. Nowadays, some commercial banks began implementation of credit scoring for clients but it has not been widely applied in the test phase and still need to improve gradually. [3][4]

Today credit scoring and internal customer rating is widely used in banking activities to assess the ability to perform financial obligations of a customer against a bank. Beside normal activities, the risk evaluation and identification functions are also very important in the credit activities of the bank. Credit risk level changes to individual clients and is identified through an assessment process. This process was based on financial data and existing non-financial customer's at the time of credit grading and evaluation. [10-13]

There are many methods that have been investigated in the last decade to improve the accuracy in credit scoring. Artificial Neural Networks (ANN) [10-13] and Support Vector Machine (SVM) [14–19] are two commonly soft computing methods used in credit scoring modelling. Recently, other methods like evolutionary algorithms, stochastic optimization technique have shown promising results in terms of prediction accuracy.

In this paper, a new method for feature selection based on Harmony Search integrated with 1-NN classifier for predicting credit score has been proposed. The competitors' average accuracy has been compared with that of the proposed approach. The two real world datasets, the Australian and German credit datasets have been used as benchmark datasets. It has been observed that the proposed method outperforms the competitors.

This paper is organized as follows: Section 2 describes the background of Harmony Search, feature selection and 1-NN classifier. The details of the proposed model are described in Section 3. Section 4 presents the experiments and the obtained results which show an accuracy improvement of the proposed model. Finally concluding remarks and future works are presented in Section 5.

II. BACKGROUND

This section describes about Feature Selection, harmony Search and 1-NN Classifier.

2.1 Feature Selection

The high-dimensional feature vectors often impose a high computational cost when classification is performed. Feature selection plays a major role as a pre-processing technique in reducing the dimensionality of the datasets in the fields of data analysis and data mining applications. Feature selection is an act of identifying a small subset of features to be employed for classification when a classification problem has involved a large set of potential features. The data without feature selection may be redundant or noisy, and may degrade the accuracy rate of classification.

Finding an optimal subset of features in feature selection is inherently combinatory, since the usefulness of each feature needs to be determined [8]. Hence, feature selection is an optimization problem. An optimal approach is necessary to measure all possible subsets.[10][11]

2.2 Harmony Search

Harmony Search (HS) is a music-based meta-heuristic optimization algorithm. It was inspired by the observation that the aim of music is to search for a perfect state of harmony. A HS algorithm repeatedly constructs harmonies and performs pitch adjustment for each of them until a satisfactory harmony is identified. This harmony in music is analogous to find the optimality in an optimization process. During the recent years, the HS has been applied in the fields of function optimization, mechanical structure design and pipe network optimization. [20][21]

In [21] two novel methods are presented by applying harmony search to feature selection. In particular, it demonstrates the potential of utilising this search mechanism in combination with fuzzy-rough feature evaluation. The role of Harmony search in selecting quality subset by introducing additional parameter control schemes to reduce the effort and impact of parameter configuration is depicted in [21].

In [21], the speed of the Harmony Search algorithm and the 1-NN classifier has been exploited and a new fast and accurate approach has been proposed for feature selection.

As the Harmony Search algorithm is known to be generally quite effective for rapid global search of large search space in difficult optimization problems, in this work, this algorithm is combined as a wrapper with the classifier 1-NN in a bid to find the optimal feature subset that yields the best performance.

2.3 1-NN Classifier

This work improves the classification performance of the classifier, 1-nearest Neighbour. The nearest neighbour (1-NN) classifier is commonly used due to its simplicity and effectiveness. It is good to the extent that it is conceptually simple, can be used even with few examples and a wonderful performer in low dimensions for complex decision surfaces. It is bad to the extent that it suffers a lot from the curse-of-dimensionality and the classification is slow. This study proposes a novel Music-inspired Harmony search optimization algorithm.

International Journal of Advance Research in Science and Engineering
Volume No.07, Issue No.02, February 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

III. THE PROPOSED METHOD

Harmony search (HS) is a meta-heuristic that simulates the improvisation process of musicians. Solutions of the optimization process correspond to musicians and the harmony of the notes generated by a musician corresponds to the fitness of the solution.

Main steps of the algorithm are given below:

1: Initialize the problem and algorithm parameters.

2: Initialize the harmony memory.

3: Improvise a new harmony.

4: Update the harmony memory.

5: Repeat steps 3-4 until the stopping criterion is met

In HS algorithm, a new Harmony vector, $x' = x_1', x_2', \ldots x_{N-1}', x_N'$, is generated based on three rules: (1) memory consideration (2) pitch adjustment and (3) random selection. Generating a new harmony is called 'improvisation'. The HMCR, Harmony Memory Consideration rate, which varies between 0 and 1, is the rate of choosing one value from the historical values stored in the HM, while (1 – HMCR) is the rate of randomly selecting one value from the possible range of values.

Every component obtained by the memory consideration is examined to determine whether it should be pitch-adjusted. This operation uses the PAR parameter, Pitch Adjustment Rate, which is the rate of pitch adjustment as follows:

Pitch adjusting Decision for x_i' ← { Yes with probability PAR; No with probability (1- PAR)

The value of (1 - PAR) sets the rate of doing nothing. If the pitch adjustment decision for x_i' is YES, x_i' is replaced as follows:

$x_i' = x_i' \pm \text{rand}(\,) * BW,$

where

BW is an arbitrary distance bandwidth

rand ( ) is a random number between 0 and 1

In this work, the fitness of the feature subset has been evaluated by one nearest neighbour (1-NN).

The main steps involved in 1-NN classification method are as follows:

Step 1: Calculate distances of all training vectors to test vector using Euclidian Distance

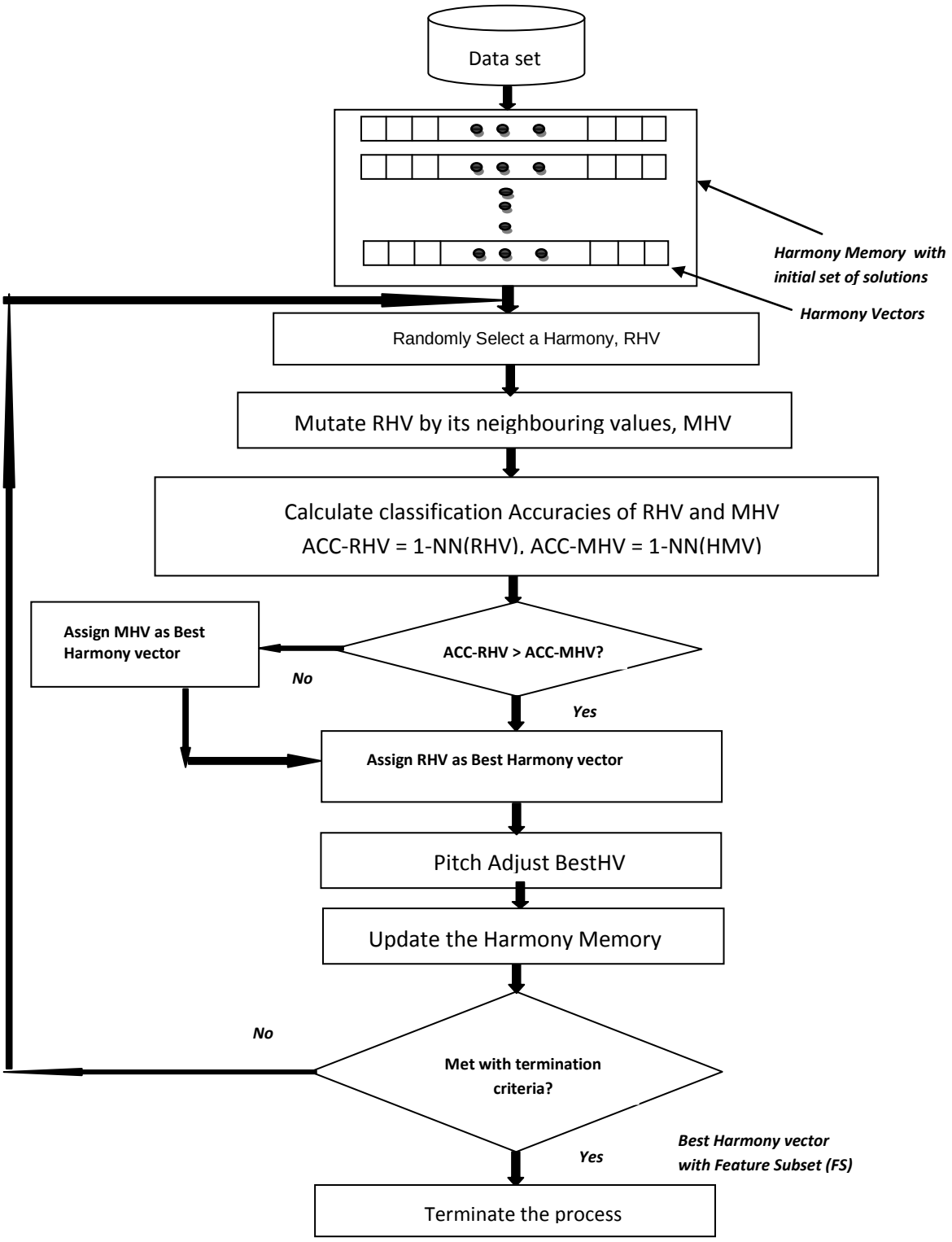
Step 2: Pick the closest vector to the test vector

Step 3: Assign the class of the closest training vector a the class of the test vector same value/class as the nearest instance in the training set is predicted for the test example

While the Harmony Search (HS) algorithm is used as a wrapper to select the best harmony vector with selected features, the 1-NN classifier is used to find the classification accuracy.

Fig 3.1 depicts the proposed HS-1-NN wrapper approach for feature selection

.



IV. EXPERIMENT AND RESULTS

Our experiment has been implemented to test the proposed algorithm with some datasets including two UCI public datasets, German credit and Australian credit.

In this paper, 1-NN with the original dataset is used as the base-line method. Two methods, the proposed method and the base-line method, were performed on the same training and testing datasets to compare their efficiency. In order to test the consistency of obtained results, those implementations were repeatedly done 20 times.

For HS algorithm, the following parameters are set to the values recommended in [20] as these values lead to the optimal solution with significant convergence rate: Size of the harmony memory = 10, HMCR = 0.9, PAR = 0.3, BW = 0.01. The maximum improvisation number was set as 700 for all test problems. For the proposed algorithm, the same parameter setting has been maintained.

4.1 German credit approval dataset

The German credit dataset consists of 1000 loan applications, with 700 instances of creditworthy applicants and 300 instances of rejected applicants. For each applicant, 20 attributes describe the credit history, account balances, loan information and personal information. The final results were obtained by averaging these 20 independent trials.

Different classifiers over the German credit datasets were compared and their performances are shown in Table 4.1. Baseline is the classifier without feature selection. Classifiers used in our investigation include: Linear SVM, CART, k- NN, Naïve Bayes, MLP. Various feature selection methods are used for comparison including filter approach and wrapper approach. The wrapper approach includes two methods: Genetic algorithms (GA) and Particle Swarm Optimization (PSO).

TABLE 4.1. PERFORMANCE COMPARISON OF DIFFERENT CLASSIFIERS OVER THE GERMAN CREDIT DATASET

Classifier	**Wrapper Methods**		**Baseline**
	GA	PSO	
Linear SVM	78.62	74.23	77.18
CART	75.28	74.86	74.30

k-NN	70.84	68.68	70.86
Naïve Bayes	73.46	72.19	70.52
MLP	74.59	71.50	71.76
1-NN			76.77
HS-1-NN			80.4

As shown in Table 1 for comparing the performances of various methods, we saw that the accuracy of HS-1NN on the subset of newly selected features has been obviously improved, The average accuracy is 76.77% on the original data. After applying the feature selection, the average accuracy increases to 80.4%.
Furthermore, our method relying on a parallel processing strategy allows the time to run 20 trails with 5-fold cross validate taking only 57 minutes while other methods must run several hours. This result emphasizes the efficiency of our method in terms of running time due to efficiently filtering the redundant features.

4.2. Australian credit approval dataset

The credit data of Australia consists of 690 applicants, with 383 instances of credit worthy and 307 default examples. Each instance contains both numerical features, categorical features, and discriminant feature.

TABLE 4.2. PERFORMANCE COMPARISON OF DIFFERENT CLASSIFIERS OVER THE AUSTRALIAN CREDIT DATASET

Classifier	**Wrapper Methods**		**Baseline**
	GA	PSO	
Linear SVM	85.52	85.52	85.52
CART	85.25	85.46	85.20
k-NN	86.06	85.31	84.58

Naïve Bayes	86.52	87.09	68.55
MLP	85.60	86.00	84.15
1-NN			87.82
HS-1-NN			93.5

Table 4.2 shows the performances of different classifiers and selection methods over the Australian credit datasets for comparison. The obtained results indicate that the accuracy of HS-1-NN on a subset of 9 selected features has been obviously improved. The average accuracy is 87.82% on the original data, while the average accuracy increases to 93.5% after applying the feature selection in our method. Based on parallel processing, time to run 20 trails with 5-fold cross validate taken by our method can be reduced to only 34 minutes

V. CONCLUSION

In this paper, a new Harmony Search based wrapper method has been proposed for feature selection for 1NN Classifier and the same has been applied as a prediction model for credit scoring system. Feature selection provides an effective method in determining the highest classifier accuracy of a subset or obtaining the acceptable accuracy of the smallest subset of features. The accuracy of classifier using the selected features is improved compared with other methods. Fewer features allow a credit department to focus on collecting relevant and essential variables. The runtime has also been significantly reduced and as a consequence the workload of credit evaluation personnel can be reduced because the proposed approach does not have to take into account a large number of features in the assessment process, which requires much less effort in computation. This paper has investigated and compared different methods over two real world credit datasets. Experimental results show that the proposed method is effective in credit risk investigation. The method offers a quick assessment with improved accuracy of the classification.

REFERENCES

1. Altman, E. I. & Saunders, A. Credit risk measurement: Developments over the last 20 years. J. Bank. Financ. 21, 1721–1742 (1997).
2. Wu, X. et al. Top 10 algorithms in data mining. (2008). doi:10.1007/s10115-007-0114-2
3. Angelini, E., di Tollo, G. & Roli, A. A neural network approach for credit risk evaluation. Q. Rev. Econ. Financ. 48, 733–755 (2008).
4. Bellotti, T. & Crook, J. Support vector machines for credit scoring and discovery of significant features. Expert Syst. Appl. 36, 3302–3308 (2009).

5. Wen, F. & Yang, X. Skewness of return distribution and coefficient of risk premium. J. Syst. Sci. Complex. 22, 360–371 (2009).
6. Zhou, X., Jiang, W., Shi, Y. & Tian, Y. Credit risk evaluation with kernel-based affine subspace nearest points learning method. Expert Syst. Appl. 38, 4272–4279 (2011).
7. Kim, G., Wu, C. H., Lim, S. & Kim, J. Modified matrix splitting method for the support vector machine and its application to the credit classification of companies in Korea. Expert Syst. Appl. 39, 8824–8834 (2012).
8. Liu, H. & Motoda, H. Feature Selection for Knowledge Discovery and Data Mining. (1998).
9. Guyon, I. & Elisseeff, A. An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3, 1157–1182 (2003).
10. Oreski, S., Oreski, D. & Oreski, G. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert Syst. Appl. 39, 12605–12617 (2012).
11. Saberi, M. et al. A granular computing-based approach to credit scoring modeling. Neurocomputing 122, 100–115 (2013).
12. Lee, S. & Choi, W. S. A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. Expert Syst. Appl. 40, 2941–2946 (2013).
13. Ghatge, A. R. & Halkarnikar, P. P. Ensemble Neural Network Strategy for Predicting Credit Default Evaluation. 2, 223–225 (2013).
14. Chaudhuri, A. & De, K. Fuzzy Support Vector Machine for bankruptcy prediction. Appl. Soft Comput. J. 11, 2472–2486 (2011).
15. Ghodselahi, A. A Hybrid Support Vector Machine Ensemble Model for Credit Scoring. Int. J. Comput. Appl. 17, 1–5 (2011).
16. Huang, C.-L., Chen, M.-C. & Wang, C.-J. Credit scoring with a data mining approach based on support vector machines. Expert Syst. Appl. 33, 847–856 (2007).
17. Li, S. T., Shiue, W. & Huang, M. H. The evaluation of consumer loans using support vector machines. Expert Syst. Appl. 30, 772–782 (2006).
18. Martens, D., Baesens, B., Van Gestel, T. & Vanthienen, J. Comprehensible credit scoring models using rule extraction from support vector machines. Eur. J. Oper. Res. 183, 1466–1476 (2007).
19. Wang, Y., Wang, S. & Lai, K. K. A new fuzzy support vector machine to evaluate credit risk. IEEE Trans. Fuzzy Syst. 13, 820–831 (2005).
20. V. Krishnaveni and G. Arumugam ,"The Performance Analysis of a Novel Enhanced Artificial Bee Colony Inspired Global Best Harmony Search Algorithm for Clustering", Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) January 2012, Advances in Intelligent and Soft Computing, 2012, Volume 132/2012, 21-28
21. S Varadhaganapathy, V Krishnaveni, G Arumugam, RR Rajalaxmi "Harmony and bio inspired harmony search optimization algorithms for feature selection in classification", Computer Systems Science And Engineering, Volume 30, Issue 4, Pages 257-272, 2015 CRL Publishing Ltd