New hybrid method for feature selection and classification using meta-heuristic algorithm in credit risk assessment

Jalil Nourmohammadi-Khiarak¹ · Mohammad-Reza Feizi-Derakhshi² · Fatemeh Razeghi³ · Samaneh Mazaheri⁴ · Yashar Zamani-Harghalani² · Rohollah Moosavi-Tayebi⁴

Received: 28 July 2018 / Accepted: 27 May 2019 / Published online: 11 June 2019
© The Author(s) 2019

Abstract

Credit risk is a factor that arises from the failure of the party to the contract. It is one of the most important factors of risk production in banks and financial companies. Still, there is no standard set of features or indices which have been declared through all credit institutions and according to the classification of customers, they are able to do through terms of credit value. In this paper, a meta-heuristic of imperialist competitive algorithm with modified fuzzy min–max classifier (ICA-MFMCN) is offered to identify an optimal subset of features and increased through accuracy classification and scalability through assessment of credit risk. Performance of proposed ICA-MFMCN classification is approved and recognized using a real credit set that has been selected from a UCI dataset. Classification accuracy is comparable for what has been indicated through resources. The experimental which result, in obtaining new classification by utilizing the proposed are promising for future classification are selection processes in assessment of credit risk through retail, indicating that ICA-MFMCN is one of the ways which can be used to add existing data mining techniques.

Keywords Credit risk · Feature selection · Imperialist competitive algorithm and modified fuzzy min–max classifier

1 Introduction

Credit risk is rooted from the fact that the contract party is not able to perform contractual obligations [1–5]. Impact of the risk is measured by the cost of replacing the cash from default on party to the contract. Losses from credit risk may be occurred before the actual failure party to contract. Generally, the possible loss credit can be explained as a result of a credit event. A credit event can be real when the capabilities of contract party change at the completion of its obligations [1]. The risk arises when the recipients of facilities do not have the ability to repayment of installment's debt to the

bank. Therefore, assessment of credit risk has attracted the attention of many researchers in recent years. Many years ago, focusing on researches of the customer's credit risk assessment was completely incoherent. Quantitative models of customer credit assessment were created after business credit and mostly due to the problem of data availability. Information was limited to financial institutions databases [2]. Currently, certain data are available in several countries publicly and Tax offices and researchers have developed the scoring quantitative techniques to different credit. Because of the crisis, banks are exposed to challenges in finding new ways to precede business that should be safer and to be efficient and helpful completely. They were under the pressure of the crisis, because it showed the level of dangers that were inclusive in the business banking. In retails, risks are considered often without an accurate estimation of degrees and their likely consequences. Many decisions that have been included in the borrowing customer's business have made it necessary to rely on models and algorithms more than human authority and based on such algorithm decisions on hard data [3]. In terms of high economic value, banks have reached high rates of profit for their owners and are not

✉ Jalil Nourmohammadi-Khiarak
Jalil.Nourmohammadi@elka.pw.edu.pl

¹ Faculty of Electronics and Information Technology, Warsaw University of Technology, Warsaw, Poland

² Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran

³ Faculty of Electrical and Computer Engineering, Islamic Azad University, Sari, Iran

⁴ Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM), Seri Kembangan, Malaysia

exposed to the challenge of finding new ways of the business.

Despite the difficulties many banks faced and some have been bankrupted. Investors have been increasingly cautious and are not willing to invest their money in such troubled banks. Moreover, regulators strengthen their capital need of the banks [4] and they resist in the face of such crisis situations. State governments and international organizations are occupied in rescue situations and prevent further consequences of the crisis; it was a catastrophic situation because one of the fundamental causes of the crisis was the way that banks reacted to it. Little detailed methods were used in risk assessment of retail and the collateral for the loan funds was considered as a main successor [5]. The casualties were inevitable when it became clear that this collateral is not worth as much as the original assessed value and credit risk of customers has not properly assessed valuable. All reasons should be worked on for the crisis and even methods which the bank gets busy to business; in terms of the risk taking at the time of work, funds, loan the operations should be faster, safer, more accurate and based on data banks should utilize in better ways. This capital is not only in terms of monetary but also in terms of the data which are collected in the database. Capital in terms of the customer must be better managed by banks. This information should be transformed to knowledge and then money.

Using machine learning algorithms is going to discover new way to reduce the existing risk in credit risk assessment. Recently, researcher studied this field [4–6]; however, there is still a huge gap between accuracy and being trustable these algorithms to be used at banks. Therefore, we proposed a new method based on combining a metaheuristic algorithm with a classification method to get better accuracy. So, in this paper we want to make a high accuracy algorithm to be used in industrial and real-world applications. To evaluate the proposed method, a dataset at UCI has been used from German validation datasets from UCI repository of machine learning datasets that is available and has been made from 700 examples of valid applications and 300 applications with bad credit. We have obtained slightly better results in comparison with other methods, especially classification without feature selection.

The paper has been organized as follows; Sect. 2 explains related work, Sect. 3 proposes method completely; Sect. 4 discusses about experimental results with evaluation and comparison of performance and the last section presents the conclusion of the paper and will be introduced in future works.

2 Related work

The present information which is in the database can be used for assessment of credit risk, but these data have large dimensions usually. Unrelated features in the educational database can make lower precision results in the classification analysis. Selecting the most important features is required to increase forecast accuracy, speed and scalability [6–9]. In different fields, for example, to evaluate credit risk, using appropriate features set is critical to the success of the learning process and as a result by itself is a major problem. To evaluate feature selection methods that are connected strongly to scalability algorithms and also it provides selection of incremental feature, a new and efficient hybrid classification has been proposed in the research.

Feature selection problem is one of the challenging problems in machine learning and statistical pattern identification field [10, 11]. This problem is important in many applications (such as classification), because in these applications there are a lot of features, many of which are either unused or do not have much information [12]. Removing these features does not create problems in informational term but raises the computational load for the intended application. In addition, it leads to storing a lot of useful information along with useful data. Various methods of feature selection are trying to select the best subset from 2^N candidate subsets [13]. In all these methods, a subset is selected as the answer based on the type of application and type of definition that can optimize the value of an evaluation function. Any selection method tries to select best features. However, when the amount of N is high, the range of answers will increase. As a result, time complexity of finding optimum result will be grown exponentially.

In general, different methods of feature selection based on the type of the search are divided into different categories. In some techniques, the entire possible space is searched [8]. In other methods that can be heuristics, or random search [12], the search space becomes smaller in exchange for losing some of the efficiency. Aha and Bekhet showed that feature selection improves the classification performance and provides evidence [6]. Dannas et al. applied a feature selection for data collection using feature subset selection algorithm based on relationships along with the search of Taboo Search in features' subsets [14]. It should be noted that the recovery of even a fraction of a percent is enough to attract the attention of many scientific and practical methods. To test the hypothesis as well as to assess the accuracy of the proposed technique, tests are performed on a genuine validate [15].

To handle the feature selection challenge, this study presents analysis and credit risk assessment based on combining a classification method and optimization algorithm. New classification, in the name of ICA-MFMCN, is derived

from imperialist competitive algorithm [16] and a modified fuzzy min–max classifier [17, 18]. Feature selection plays an important role in big data. In the previous works, in the feature selection field and optimizing their performance, imperialist competitive algorithm has shown significant results. In the paper [19], the authors tried to select the best features' set to classify any type of rice, based on the sample image using imperialist competitive algorithm. The outcome has shown that the imperialist competitive algorithm had significant improvement in different issues, especially in feature selection. To this aim, mentioned algorithm is used for feature selection in present paper. There are many classification methods and the reason of using modified fuzzy min–max classifier is that this type of classification shows more accurate results as compared with neural network algorithm and support vector machine. According to the modified fuzzy min–max classifier method in the paper [17], it can be used in new classification methods. Therefore, in this paper, by combining imperialist competition algorithm and modified fuzzy min–max classifier, an appropriate feature selection method is presented for credit risk data.

3 Proposed method

In this section, the proposed method will be explained which consists of Imperialist Competitive Algorithm and a modified fuzzy min–max classifier.

3.1 Imperialist competitive algorithm

According to Fig. 1, the ICA, like other evolutionary optimization methods, starts with initial population. In this algorithm, every element of the population is called a country. Countries are divided into two categories of the colony and imperialists. Every imperialists depending on its power, a number of colonized countries turned to their domination and are controlled by them [16].

In colonial competitive algorithm when there are good countries that belong to the same attractive region, this trend converges to local optimum, rapidly. Wonderful country becomes the dominant on population quickly and arrives to premature convergence. Thus, among some countries revolution that can be useful and is belonging to different regions. Changing the algorithm parameters can affect the performance of the model. It should be noted that after every step of the algorithm, conditions for the next step of algorithm can be changed dynamically. Furthermore, can change the initial population, the selected layout, and type of intersection, mutation rate or other parameters of ICA are modifiable and the results of these modifications can be controlled. The flowchart of ICA in Fig. 1 explains what is used in proposed

Fig. 1 The flowchart which describes ICA utilized in the proposed method

method. Benefits of the proposed algorithm can be summarized as follows:

- The higher ability to optimize compared with different algorithms of optimization, by different variety of optimization problems.
- Suitable speed finding optimal answer.

ICA has been used for feature selection in the proposed method; so in next section, it will be explained how this algorithm will be utilized for the purpose. In the next section,

M-FMCN will be described which is one of the best existing classifiers.

3.2 Modified fuzzy min–max classifier

The specific kind of neuro-fuzzy is Fuzzy Min–Max which has high efficiency instead of other machine learning methods [18] which can be learned in a very short time. The Fuzzy Min–Max used in this paper has a close similarity version in [17], which is the most frequently applied type in the written works. In the proposed work, a Fuzzy Min–Max classifier is used which uses modified compensatory neurons. This classifier can work online and it is a supervised method based on fuzzy Min–Max. Overlapping division was created in borders on the whole. In this approach for managing overlapping division, a changed compensatory node based on radius-based transition function that can increase the classification accuracy is used. In the algorithm, time and space complexity have been decreased. In addition as it has been shown in the experimental results, the proposed method is less sensitive to external parameters compared with other methods [17].

3.3 Hybrid ICA and M-FMCN

To implement the proposed algorithm by this modification, first using the ICA algorithm selects the optimal features and in the next step using the M-FMCN is classified on the selected features and credit risk creates the appropriate categories. The proposed method flowchart has been shown in Fig. 2.

General idea for the proposed method is combining the ICA and M-FMCN, moreover, we will start with feeding input data, then ICA starts selecting the best features. There is one interesting point that should be noted, not only M-FMCN is used inside ICA as a cost function but also M-FMCM is used as a classifier for final decision. Therefore, this connection makes ICA more accurate to select features. In ICA optimization, there is a cost function like all other optimizations which should be minimized. Considering a suitable cost function for every problem is very crucial. Therefore, according to our goal in this paper we calculate accuracy for each iteration based on each feature till to achieve high accuracy.

In this paper, using random methods, the initial population (countries) as binary numbers 0 and 1 of credit risk data features are produced. In fact, a primary sample's length is equal to the amount of data feature. The value 1 in country shows that the column should be selected, while the 0 value shows that this feature does not have the capability of selection. A sample of production in a country is shown in Fig. 3.

After the production of the initial population amounts, cost function of every country is calculated, as well as pri-

mary imperialists and their empires would be created. Cost function amount of a country is predicted with F as follows:

$$\text{Fit} = F(X_{\text{country}}) = F(f_1, f_2, \dots, f_n). \quad (1)$$

The initial population number is specified with n and the number of imperialists, with m which is powerful countries selected as imperialists. The rest of set is specified with i , gained from $i = n - m$ and this set is colonized according to the split among imperialists. To broadcast countries among imperialists, the following equation will be applied.

$$\text{Fit}_m = \max(\text{cost}_i) - \text{cost}_m. \quad (2)$$

A number of colonies will be given to every imperialist, whose number fits its power i th. The Eq. (2) shows the power of the imperialist (Fit_m) that cost_i is the gained cost of the i th imperialist, $\max\{\text{cost}_i\}$ is the imperialist maximum cost and cost_m is the paid cost for normalizing this imperialist. The colonized countries are divided according to Eq. (3).

$$P_n = \left| \frac{C_n}{\sum_{i=1}^{N_{\text{imp}}} C_i} \right|, \quad (3)$$

where C_n and C_i are the cost of the imperialist and country, respectively. The Imperialist Competitive Algorithm starts by having primary states of every imperialist. Evolution process is in a loop that continues to satisfy a stop condition. The power of an imperialist is defined according to Eq. (4).

$$T \cdot C_n = \cos t(\text{imperialist}_n) + \text{mean}\{\text{Cost}(\text{colonies of empire}_n)\}. \quad (4)$$

By repeating the current process in the Fig. 2 flowchart, weak empires lose their strength and are destroyed by weakening. Imperialist competition causes a state that only one imperialist would exist. This is when colonial competition algorithm stops by achieving the target function's optimum point. In this paper, stop state is a specified repetitive number. When the feature selection process is finished with the imperialist competition, in the next phase of Fig. 2 flowchart, the selected features are sent to M-FMCN classifier. In this classifier which uses the fuzzy logic for classifying, the classification of credit risk data is done and finally the classification accuracy is extracted and set to the output.

The M-FMCN classifier is a hybrid and homogeneous method of classification that uses a multi-layer tree structure. This method, against its appearance, does not work like decision tree and mostly behaves like cascade classification hybrid methods. In this method, to increase performance and the accuracy of the network performance, hyper boxes with different sizes and in different levels of the network

Fig. 2 The hybrid proposed method flowchart

Fig. 3 Sample of center the initial population (country) of solutions set

are used, so that the classification in the uniform and the non-boundary area is conducted using large hyper boxes and in the boundaries which need higher accuracy using hyper boxes which are smaller and have higher accuracy.

Here, unlike the earlier methods of FMM (fuzzy min—max), for managing overlapping areas, hyper-box contraction is not used. Instead, data within each overlapping area are classified in the next layers and with higher accuracy. This task is done such a way that by moving from the first layer to the last one, hyper-box size and the space covering are reduced and decision accuracy is increased. In Fig. 4, the hyper-box formation of the first and second layer and their classification method in the proposed method is shown.

In this network, every node of a tree is a distinct classifier which is responsible for classification in a special area of the sample space and has distinct outputs. The existing node in the first layer (root) mostly covers all the sample space and is mostly responsible for large uniform classification areas. The second layer nodes cover the homogeneous areas of the first layer and classify the samples which are in these areas. Respectively, every node of the i th layer covers one of the homogeneous areas of the i th layer and is actually responsible for classification in the mentioned area. After the formation of the general tree, the best output of these nodes is selected and considered as the output of the entire network. In this paper, every node of this tree is introduced as sub-network. Similar to other methods of FMM, all the hyper boxes are

Fig. 4 Proposed classification method in first and second layer

Table 1 Summary of ICA-MFMCN parameters

Parameter	Settings
Population size	50
Maximum number of features	12
Minimum Specifications	5
Fitting criterion	Accuracy
Fitting function	Min–Max fuzzy neural network
Type of neural network	Forward multilayer network
Size of the hidden layer	$2 + 1/(\text{number of properties} + \text{number of classes})$
Training cycles	[300; 600]
Quickly learn	[0.3; 1.0]
Momentum	[0.2; 0.7]
Choice plan	Roulette wheel
Dynamic selection pressure	Yes
Retain the best people	Yes
Likely to attract	Number of features/1
The possibility of revolution	0.9
Maximum fitting	Unlimited
Maximum number of generations	50
Using initial stop	Good

created and set in the training phase and used in the test phase. Moreover, like classification methods, the management of the homogeneous areas is conducted after the creation and setting of the available hyper boxes in a network. This task, as mentioned previously, reduces the time and space complexity of the algorithm. Respectively, classification is conducted using M-FMCN method. For get more details, refer to [17]. Table 1 shows the parameters that have been used in this paper.

4 Empirical results

ICA-MFMCN proposed classification functions have been validated using genuine validation data set collected in the UCI database. The test has performed on Germany dataset to validate the results presented from ICA-MFMCN algorithm that are available to public. These results are comparable with other outcomes mentioned in this paper.

4.1 Describing data collection

German validation datasets from UCI repository of machine learning datasets are available and have been made from 700 examples of valid applications and 300 applications with bad credit [20]. The main dataset, as provided by Professor Hoffmann for every application, contains 20 inputs features that describe credit history, account balance, loan purpose, loan amount, employment status, personal information, age, job titles and housing. The main data set has 13-Stratified features that some of them have been converted to binary feature sets in a way that they can be managed by the MFMCN properly. German credit datasets converted 30 regular features of the correct data type and 2 special features (ID and Labels) that are available at the following website (<http://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/download-course-materials>). All features of descriptive statistics are shown in Table 2.

5 Experimental results

After initial production solutions, a subset of reduced features has been created as a complex of features in the initial solution. According to Table 2, it can be seen that original set

Table 2 Credit data sets converted to descriptive statistics

Attr.	Code	Description	Statistics	Range
<i>att1</i>	ID	Observation no.	$\text{avg} = 500.500 \pm 88.819$	[1, 1000]
<i>att2</i>	CHK_ACCT	Checking account status	$\text{avg} = 1.577 \pm 1.258$	[0, 3]
<i>att3</i>	DURATION	Duration of credit in months	$\text{avg} = 20.903 \pm 12.059$	[4, 72]
<i>att4</i>	HISTORY	Credit history	$\text{avg} = 2.545 \pm 1.083$	[0, 4]
<i>att5</i>	NEW_CAR	Purpose of credit	$\text{avg} = 0.234 \pm 0.424$	[0, 1]
<i>att6</i>	USED_CAR	Purpose of credit	$\text{avg} = 0.103 \pm 0.304$	[0, 1]
<i>att7</i>	FURNITURE	Purpose of credit	$\text{avg} = 0.181 \pm 0.385$	[0, 1]
<i>att8</i>	RADIO/TV	Purpose of credit	$\text{avg} = 0.280 \pm 0.449$	[0, 1]
<i>att9</i>	EDUCATION	Purpose of credit	$\text{avg} = 0.050 \pm 0.218$	[0, 1]
<i>att10</i>	RETRAINING	Purpose of credit	$\text{avg} = 0.097 \pm 0.296$	[0, 1]
<i>att11</i>	AMOUNT	Credit amount	$\text{avg} = 3271.258 \pm 2822.737$	[250, 18, 424]
<i>att12</i>	SAV_ACCT	Average balance in savings account	$\text{avg} = 1.105 \pm 1.580$	[0, 4]
<i>att13</i>	EMPLOYMENT	Present employment since	$\text{avg} = 2.384 \pm 1.208$	[0, 4]
<i>att14</i>	INSTALL_RATE	Instalment rate as % of disposable income	$\text{avg} = 2.973 \pm 1.119$	[1, 4]
<i>att15</i>	MALE_DIV	Applicant is male and divorced	$\text{avg} = 0.050 \pm 0.218$	[0, 1]
<i>att16</i>	MALE_SINGLE	Applicant is male and single	$\text{avg} = 0.548 \pm 0.498$	[0, 1]
<i>att17</i>	MALE_MAR_WID	Applicant is male and married or a widower	$\text{avg} = 0.092 \pm 0.289$	[0;1]
<i>att18</i>	CO-APPLICANT	Application has a co-applicant	$\text{avg} = 0.041 \pm 0.198$	[0;1]
<i>att19</i>	GUARANTOR	Applicant has a guarantor	$\text{avg} = 0.052 \pm 0.222$	[0, 1]
<i>att20</i>	PRESENT_RESIDENT	Present resident since – years	$\text{avg} = 2.845 \pm 1.104$	[1, 4]
<i>att21</i>	Applicant	owns real estate	$\text{avg} = 0.282 \pm 0.450$	[0, 1]
<i>att22</i>	PROP_UNKN	Applicant owns no property (or unknown)	$\text{avg} = 0.154 \pm 0.361$	[0, 1]
<i>att23</i>	AGE	Age in years	$\text{avg} = 35.546 \pm 11.375$	[19, 75]
<i>att24</i>	OTHER_INSTALL	Applicant has other installment plan credit	$\text{avg} = 0.186 \pm 0.389$	[0, 1]
<i>att25</i>	RENT	Applicant rents	$\text{avg} = 0.179 \pm 0.384$	[0, 1]
<i>att26</i>	OWN_RES	Applicant owns residence	$\text{avg} = 0.713 \pm 0.453$	[0, 1]
<i>att27</i>	NUM_CREDITS	Number of existing credits at this bank	$\text{avg} = 1.407 \pm 0.578$	[1, 4]
<i>att28</i>	JOB	Nature of job	$\text{avg} = 1.904 \pm 0.654$	[0, 3]
<i>att29</i>	NUM_DEPENDENTS	Number of people for whom liable to provide maintenance	$\text{avg} = 1.155 \pm 0.362$	[1, 2]
<i>att30</i>	TELEPHONE	Applicant has phone in his or her name	$\text{avg} = 0.404 \pm 0.491$	[0, 1]
<i>att31</i>	FOREIGN	Foreign worker	$\text{avg} = 0.037 \pm 0.189$	[0, 1]
<i>att32</i>	LABEL	Credit rating	mode = 1 (700), least = 0 (300)	1 (700), 0 (300)

Table 3 Primary solutions and reduced feature set for dataset of German credit

	Current solution	IG	Gain ratio	Correlation	Voting	
att2	☑	☑	☑	☑	☑	☑
att3	☑	☑	☑	☑	☑	☑
att4	☑	☑	☑	☑	☑	☑
att5	☑					☑
att6			☑	☑		☑
att7						
att8		☑		☑		☑
att9						
att10						
att11		☑	☑	☑	☑	☑
att12	☑	☑	☑	☑	☑	☑
att13		☑		☑		☑
att14	☑					☑
att15	☑					☑
att16	☑					☑
att17						
att18						
att19	☑					☑
att20						
att21	☑	☑	☑	☑	☑	☑
att22	☑	☑	☑	☑	☑	☑
att23		☑	☑			☑
att24		☑	☑	☑	☑	☑
att25						
att26		☑	☑	☑	☑	☑
att27						
att28						
att29	☑					☑
att30						
att31	☑					☑

Table 4 Compare the average prediction of two techniques for dataset of German credit

Methods	Iteration							
	1	2	3	5	10	20	30	50
GA–NN [15]	75.65	76.88	77.36	77.81	78.03	78.18	78.37	78.50
Hybrid GA–NN [15]	77.75	78.16	78.25	78.37	78.55	78.74	78.83	78.90
ICA- MFCN	78.5	78.89	78.95	79.05	79.13	79.36	79.01	80.02

of features has been reduced from 30 to 20 features in stages to reduce the search space. Due to the number of trials, the initial parameters of colonial competitive algorithm are set on the experimental values to 1, except for the maximum number of features that is set on 16. Furthermore, the initial parameters of M-FMCN are set on the same values of test 1, except number of test cycles that are set in the interval [30, 100].

All basic parameters run up to another, did not change except the parameters that are shown in Table 3. According to these, for German validation dataset 99 measurements have been performed that are expressed in terms of percentage accuracy for every technique. ICA-MFMCN technique in the eighth run reached the best average prediction accuracy of 80.4%, with chosen unique plan described and type of one_point intersection. ICA-MFMCN technique reached

Fig. 5 Comparison of dataset of German credit results with a distance on the coordinates of the point agree a scale measure

Table 5 Statistics tests of the differences among the ICA-MFMCN and the GA-NN for the German credit

Test	Level of significance	<i>p</i> value
Paired <i>t</i> test (two-tailed)	0.009	0.000
D'Agostino and Pearson omnibus normality test	0.049	0.714
Shapiro–Wilks normality test	0.048	0.390
Wilcoxon matched-pairs signed-rank test (two-tailed)	0.016	0.009

the best result that is 81.0% in the seventh run with selected unique plan and uniform intersection type. The best solution for ICA-MFMCN and genetic algorithm–neural network (GA-NN) techniques is obtained with 12 and 16 selected features, respectively.

To validate the findings of Test 1, summary comprehensive results of Test 2 are presented in Table 4 and Figs. 3 and 4. Table 4 and Fig. 3 show that the difference between the average predictions two techniques in 8 first-generation ICA-MFMCN is narrow, while the differences of accurately predicted between generations 8th and 50th ICA-MFMCN remain the same size (see Fig. 5).

According to Fig. 4, it is clear that the objective function value enters to the stage of saturation after the eighth generation that is similar to results of Test1. Objective function value in the last 40 generations grows less than only first 4 generations from 10 generations to 50. The results presented in Tables 4 and 5 show a clear connection between ICA-MFMCN and GA-NN techniques and their performance. According to the finding, it is concluded that there is an impact of reducing the search space, and incremental step, with the effects of strategies to create an initial population of an ICA-MFMCN in the general classification, and it is considered a positive relationship. Moreover, this was supported by statistical tests (see Fig. 6).

Table 6 compares the results of the proposed ICA-MFMCN method with other methods such as GA-NN method in German credit dataset. The average rate of accuracy shown referring to the results of dataset is valid. Results obtained that use from the training data set, have not been shown. Adaptive review shows that ICA-based algorithm

suggests a credible alternative to select optimum feature for credit risk assessment that provides the best performance.

5.1 Statistical analysis

The proposed ICA-MFMCN method examined with GA-NN method which is used the paired *t* test. The null hypothesis is rejected because the calculated *t* value is greater than the critical two-tail tabulated value 3.3554 with the chosen level of significance 0.009, in comparing the paired *t* test with a *t* value of 6.8136 and 8 *df*. The paired *t* test presumes that sets of values from population of pairs in which the alteration among sets follows a Gaussian distribution [21] are tested. Normality test has been applied to test the assumptions. There is no sufficient indication to support the claim that we have experimented our pairs of values from a population of sets that is different from a Gaussian distribution (see Table 5), because of the D'Agostino and Pearson omnibus normality test *p* value = 0.714 is greater than $\alpha = 0.049$ and the Shapiro–Wilk normality test *p* value = 0.39 is greater than $\alpha = 0.048$.

Beside the paired *t* test, we have examined the ICA-MFMCN which is a new technique and the GA-NN technique applying Wilcoxon matched-pairs signed-rank test. This test offers a strong substitute to the paired *t* test, while the population of difference scores is symmetrically distributed. While the test is only slightly less powerful than the *t* test, the data are normally distributed. In addition, it is considerably more powerful when the different scores are regularly distributed with heavy tails.

The null hypothesis is tested by the Wilcoxon matched-pairs signed-rank test. As depicted in table of exact critical values for the Wilcoxon's matched-pairs signed-rank test [21], is the smaller of the sums less than 2, there will be a significant discrepancy among the classifications. The sum of the ranks for positive differences is $R_{\text{pos}} = 45$, and the sum of the ranks for negative differences is $R_{\text{neg}} = 0$. The smaller sum of the ranks is $R_{\text{neg}} = 0$ leading to rejecting the null hypothesis at the 99.0% confidence level and refusing the idea which the difference is due to chance, and also we can find out that the populations contain different medians in favor of the proposed novel ICA-MFMCN technique. This

Fig. 6 Comparison of the results of dataset of German credit with equal spacing on the coordinates of the point

Table 6 Comparison of the results with other techniques (German credit dataset)

Methods	Selected features		Accuracy	
	Average	Std.	Average (%)	Std. (%)
ICA-MFMCN	12.02	1.36	80.02	0.281
Hybrid GA-NN [15]	12.36	1.57	78.90	0.297
GA-NN [15]	13.91	1.81	78.50	0.402
SVM + GA [22]	13.30	1.41	77.92	3.970
GP [22]	13.20	2.10	78.10	4.120
NN [23]	All	–	74.67	–
SVM [24]	All	–	76.59	0.328

inference is in accordance with the t test and the hypothesis H1 of the current investigation.

6 Conclusion and future work

By competition growth and risk, also a reduced profit margin in retail credit industry, options should be faster, safer and more accurate and based on data. To meet these needs, credit institutions should consider more sophisticated techniques to assess credit risks. Feature selection is one of the challenges in creating a complex method. In this paper, an exploration of new advanced hybrid feature selection has been proposed to deal with these problems. New technique

uses the early experience of experts and fast algorithms performance to future ranking and also optimization capabilities of an ICA. The proposed algorithm is used from the UCI repository of machine learning credit dataset to estimate the performance. The average rate of accuracy obtained is compared by parametric paired t test and signed ranking test with Wilcoxon matched pairs on the values obtained from the GA-NN technique. Statistical test results show that the available data support the hypothesis of searching for reliability level of 1%. Since the feature selection in data mining activities is epidemic, a subject of research can be more existing algorithm application on datasets in other regions. Therefore, according to the specification of the data, rapid filtering techniques can be used to reduce the searching space.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- de Sousa, M.M., Figueiredo, R.S., Instituto Federal Goiano Campus Ceres, GO, Brazil, and Federal University of Goiás, GO, Brazil: Credit analysis using data mining: application in the case of a credit union. *J. Inf. Syst. Technol. Manag.* **11**(2), 379–396 (2014). <https://doi.org/10.4301/S1807-17752014000200009>
- Ala'raj, M., Abbod, M.F.: A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert*

- Syst. Appl. **64**, 36–55 (2015). <https://doi.org/10.1016/j.eswa.2016.07.017>
3. Louzada, F., Ara, A., Fernandes, G.B.: Classification methods applied to credit scoring: a systematic review and overall comparison. *ArXiv160202137 Stat.* (2016). <https://doi.org/10.1016/j.sorms.2016.10.001>
 4. Sousa, M.R., Gama, J., Brandão, E.: A new dynamic modeling framework for credit risk assessment. *Expert Syst. Appl.* **45**, 341–351 (2013). <https://doi.org/10.1016/j.eswa.2015.09.055>
 5. Abellán, J., Castellano, J.G.: A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Syst. Appl.* **73**, 1–10 (2017). <https://doi.org/10.1016/j.eswa.2016.12.020>
 6. Bekhet, H.A., Eletter, S.F.K.: Credit risk assessment model for Jordanian commercial banks: neural scoring approach. *Rev. Dev. Finance* **4**(1), 20–28 (2014). <https://doi.org/10.1016/j.rdf.2014.03.002>
 7. Bonsall, S.B., Holzman, E.R., Miller, B.P.: Managerial ability and credit risk assessment. *Manag. Sci.* **63**(5), 1425–1449 (2016). <https://doi.org/10.1287/mnsc.2015.2403>
 8. Khandani, A.E., Kim, A.J., Lo, A.W.: Consumer credit-risk models via machine-learning algorithms. *J. Bank. Finance* **34**(11), 2767–2787 (2010). <https://doi.org/10.1016/j.jbankfin.2010.06.001>
 9. Basel III: A global regulatory framework for more resilient banks and banking systems—revised version June 2011. 01-Jun-2011. [Online]. <http://www.bis.org/publ/bcbs189.htm>. Accessed 03 Jun 2017.
 10. David, A.W., Richard, B.L.: A comparative evaluation of sequential feature selection algorithms. *Comp. Eval. Seq. Feature Sel. Algorithms* (2016). https://doi.org/10.1007/978-1-4612-2404-4_19
 11. Nourmohammadi-Khiarak, J., Zamani-Harghalani, Y., Feizi-Derakhshi, M.-R.: Combined multi-agent method to control inter-department common events collision for university courses timetabling. *J. Intell. Syst.* (2017)
 12. Liu, H., Motoda, H.: *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell (1998)
 13. Bolón-Canedo, V., Sánchez-Marño, N., Alonso-Betanzos, A.: A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.* **34**(3), 483–519 (2013). <https://doi.org/10.1007/s10115-012-0487-8>
 14. Danenas, P., Garsva, G., Gudas, S.: Credit risk evaluation model development using support vector based classifiers. *Proc. Comput. Sci.* **4**, 1699–1707 (2011). <https://doi.org/10.1016/j.procs.2011.04.184>
 15. Oreski, S., Oreski, G.: Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Syst. Appl.* **41**(4 part 2), 2052–2064 (2014). <https://doi.org/10.1016/j.eswa.2013.09.004>
 16. Atashpaz-Gargari, E., Lucas, C.: Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition. In: 2007 IEEE Congress on Evolutionary Computation, pp 4661–4667, 2007. <https://doi.org/10.1109/cec.2007.4425083>
 17. Davtalab, R., Parchami, M., Dezfoulan, M. H., Mansourizadeh, M., Akhtar, B.: M-FMCN: modified fuzzy min-max classifier using compensatory neurons. In: Proceedings of the 11th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, Stevens Point, Wisconsin, USA, 2012, pp 77–82
 18. Davtalab, R., Dezfoulan, M.H., Mansoorizadeh, M.: Multi-level fuzzy min-max neural network classifier. *IEEE Trans. Neural Netw. Learn. Syst.* **25**(3), 470–482 (2014). <https://doi.org/10.1109/TNNLS.2013.2275937>
 19. MousaviRad, S.J., Tab, F.A., Mollazade, K.: Application of imperialist competitive algorithm for feature selection: a case study on bulk rice classification. *Int. J. Comput. Appl.* **40**(16), 41–48 (2012). <https://doi.org/10.5120/5068-7485>
 20. M. Lichman {UCI} Machine Learning Repository. 2013. [Online]. https://www.researchgate.net/publication/243457624_Application_of_Imperialist_Competitive_Algorithm_for_Feature_Selection_A_Case_Study_on_Bulk_Rice_Classification. Accessed Feb 2012
 21. Myers, J.L., Well, A., Lorch, R.F.: *Research Design and Statistical Analysis*. Routledge, Abington (2010)
 22. 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**(4), 847–856 (2007). <https://doi.org/10.1016/j.eswa.2006.07.007>
 23. Khashman, A.: Neural networks for credit risk evaluation: investigation of different neural models and learning schemes. *Expert Syst. Appl.* **37**(9), 6233–6239 (2010). <https://doi.org/10.1016/j.eswa.2010.02.101>
 24. Wang, S., Mathew, A., Chen, Y., Xi, L., Ma, L., Lee, J.: Empirical analysis of support vector machine ensemble classifiers. *Expert Syst. Appl.* **36**(3), 6466–6476 (2009). <https://doi.org/10.1016/j.eswa.2008.07.041>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.