Efficient detection of financial fraud detection by selecting optimal ensemble architecture using optimization approaches

[*1] C. Gayathri, [2] R.Umarani

[*1] *Research Scholar, Dept of Computer Science, Karpagam University, Coimbatore - 641021, Tamilnadu, India.*
[2] *Associate professor of Computer Science, Sri Sarada College for Women, Salem - 638016, Tamilnadu, India.*
[*1] gayathriphd2015@gmail.com, [2] umainweb@gmail.com

Abstract

Objective: To provide the secured and flexible environment which can overcome the current shortcomings and at the same time offer better identification of fraudulent behaviour in the optimized manner.
Method: Two optimization approaches namely Particle Swarm Optimization (PSO) and Fire Fly approach (FFA) are used for optimal selection of ensemble classifier architecture. The different base classifiers used in this work are ADTree, Cart, Prism and Ripper. In our previous research work we attempted to use Optimal Ensemble Classification with PSO (OEC-PSO) for improved detection of financial fraudulent activities. OEC-PSO proceeds with the single ensemble architecture to obtain better classification result by changing combination of classifier and the subset feature in every iteration. This is resolved in this research work by introducing the Optimal Ensemble Architecture Selection using PSO (OEAS-PSO) which would construct different ensemble classifier architecture which would be changes randomly in every iteration along with combination of classifier and subset feature. However, PSO lacks from the performance degradation while selecting the better ensemble classifier in case of presence of more noises such as redundancy. It will lead to more number of iteration, thus the computation overhead would be increased. This is resolved by introducing the Optimal Ensemble Architecture Selection using firefly approach (OEAS-FFA). The firefly approach overcomes the issues of PSO by selecting the optimal ensemble architecture that can provide accurate classification result. Finally, the weighted average fusion method is applied on the selected optimal ensemble classifier to retrieve the final result.
Results: The overall research of this work is evaluated in the Matlab simulation environment to find its performance improvement. This evaluation is conducted between the approaches called the OEC-PSO, OEAS-PSO and OEAS-FFA. The performance evaluation is conducted between these approaches in terms of performance measures called the accuracy, precision and recall.
Conclusion: This analysis work is conducted on the UCI data set from which it is concluded that the OEAS-FFA provides optimal result than the other approaches in terms reduced misclassification cost. The findings of this work demonstrate that the proposed research OEAS-FFA provides better result than the previous approaches.

***Keywords*:** Ensemble architecture construction, multi classification problem, accuracy, misclassification cost

1. Introduction

Online financial transaction, the electronic payment system, eliminates the burden of users from visiting the banks for completing their transaction behaviour. This system attracts more number of users to make use of online transaction services such as credit card transactions which also attracts the fraudulent users. The fraudulent users would attempt to gain access to the transaction information that are happening through the insecure medium to steal the money. The identity information that is hacked by malicious users can make use of it and can act as like genuine users to steal the amount which cannot be predicted by the machines.

This issues needs to be resolved in the proposed research methodology for the efficient handling of the online transactions, such that the users can perform online transaction in the secured manner. Finding of this malicious behaviour would be more challenging task where one cannot know the original identity of users who are attempting online credit card transaction. One of the better approaches for finding the malicious behaviour that are resides in the credit card transaction are finding the divergence that occurs in the continuous data transaction. There are different approaches are used in the previous research methodologies for finding the divergence behaviour which needs to be adapted for well identification of the malicious transaction behaviour.

Classification is one of the approaches which are used frequently by different application for learning the pattern structure from which the divergence of patterns while malicious transaction occurrence can be found. There are different classifications approaches are present in the existing research methodologies that focus on finding the patterns structure and classifying them as whether it is fraudulent pattern or not. Each and every classifier has merits and demerits based on the working procedure.

Ensemble classification is the one of the approaches that can utilize the merits of all the base classifiers. In our research work, ensemble classification approach is adapted for the optimal classification in terms of accurate detection of the fraudulent behaviour. The main contribution of our research work is to detect the financial frauds by classifying the financial transaction data set in the accurate manner. This accurate classification is done by using the ensemble classification approach. More optimal output depends on the better ensemble which is obtained by using the optimization approaches like PSO and firefly algorithm. By using these approaches, different ensemble architectures are constructed with different combination of classifiers and the subset of features. This work makes use of four base classifiers namely, ADTree, Cart, Prism and Ripper approaches. Finally the optimal ensemble found by using the optimization algorithm, is used for testing data in which weighted average method is used for getting final fusion result.

In [1], [2] introduced the ensemble classification by using the dual base classifiers to provide convenient way for the users get accurate result for the classification. The ensemble classification is done by adapting the binary classification approach which ends with the accurate classification result. This classification is done with the consideration of the reduced misclassification cost.

In [3], pruning method is integrated with the ensemble classification approach which might end with the more accurate result with reduced classification cost. The pruning method removes the unwanted candidate item sets from the available data sets to obtain the more accurate classification. In [4], ensemble classification is done by using the methodology called the ada boost ensemble classification approach. This increases the accuracy of final ensemble result by adapting Knn classification approach which would find most nearest solution for every base classifier.

In [5], [6], novel classification approach namely concept drift is introduced which optimize the classification solution by considering the time series factors. This is done by using the process called drifting which will find the most important patterns present in the time series data in terms of correlation [7]. Kernel classification approach is utilized to improve the classification accuracy.

In [8], genetic algorithm is used for finding the better ensemble classification approach in terms of improved detection accuracy with reduced misclassification cost. This approach is implemented mainly for the unbalanced data set which can tolerate the class imbalance problem and can produce more accurate result. In [9], micro array data classification approach is introduced which attempt to find the disease presence in the hospital application scenarios, so that ill patients can be treated with more care. The accurate identification is obtained y adapting the ensemble classification approaches.

In [10], tree based ensemble approach is introduced which focus on detecting the fraudulent patterns that are used to find the pedestrian disease. This approach can detect the patterns fastly which would lead to the improved system performance. In [11], difference machine learning approaches are discussed in terms of finding the most qualified data's that can lead to most accurate results. The optimization techniques [12], [13], [14] can be used in the evaluation of the fraud detection.

The above methodologies proves that the many of the research works provides an improved path for predicting the fraudulent behaviour present in the various transaction systems in terms of improved privacy and the security. In the following sub section, detailed description about the proposed research scenario of this research work to detect the fraudulent behaviour present in the insurance database system is discussed.

2. Ensemble classifier architecture construction

In this research work, accurate and efficient detection of financial fraudulent detection behaviours by using the different set of classifiers in the optimized manner is done. To do so, we have conducted various researches previously that focus on finding the financial fraudulent behaviour in the optimized manner with less misclassification cost. In this research work, accuracy of detection of fraudulent behaviour is improved by introducing the ensemble architecture construction process with previous research work. In our previous research work, Optimal Ensemble Classification using PSO is introduced that focus on finding the optimal ensemble by changing the combination of classifiers and the subset feature in every iteration. This work leads to accurate classification result. However, this

work do not concentrates on the ensemble architecture that is the representation of different base classifier in what way their results are going to be combined. In this research work, two approaches are introduced for optimal ensemble architecture selection. Those are Optimal Ensemble Architecture Selection using PSO (OEAS-PSO) and Optimal Ensemble Architecture Selection using firefly approach (OEAS-FFA). Both these mechanism are used to find the accurate classification result by randomly changing the combination classifiers, its ensemble architecture and the subset feature which is given as input in every iteration. This would be continued until the accurate result has been obtained. The overall flow of the work is given in Figure 1.

Figure 1. Overall Flow of the Proposed Research

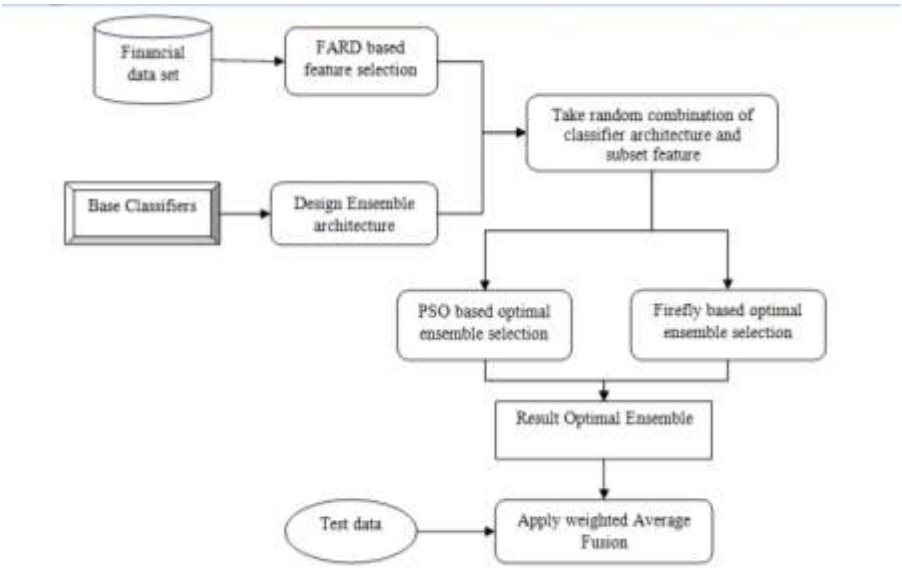


Figure 1 depicts overall flow of the research methodology in terms of different combination of the ensemble classifier and its architecture. This research work provides a better and accurate classification result in terms of selection of optimal combination of ensemble architecture and the subset feature that is given as input. The different ensemble architecture designs that are considered in this work are given as follows:

Figure 2. Different Ensemble Architectures

Figure 2.a. Binary classification

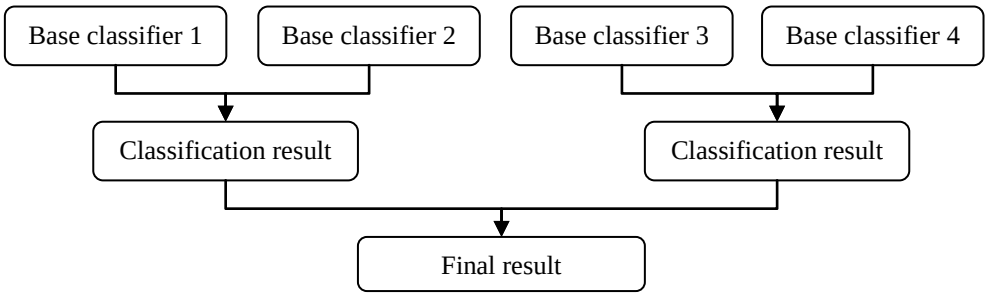
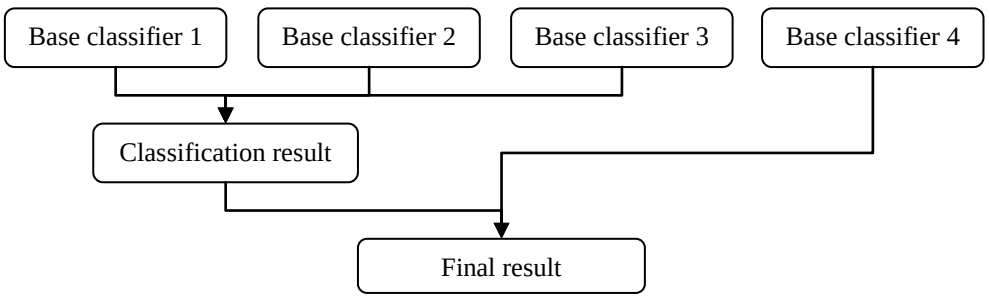


Figure 2.b. Triple class consideration



ISSN (online) : 2277-5390
ISSN (Print): 2277 - 5382

Figure 2.c. Multi class consideration

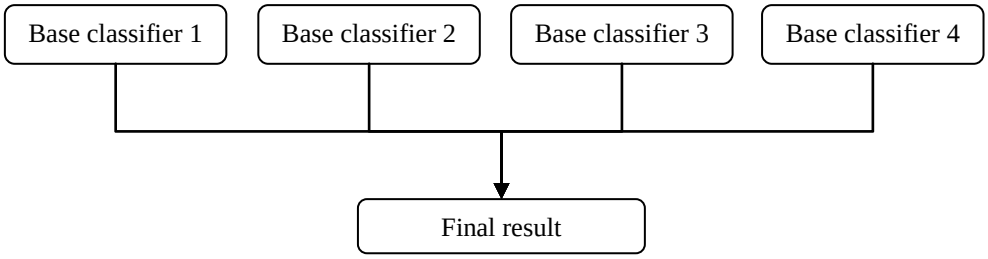


Figure 2a-c depicts the different ensemble architectures that are considered in this work for evaluation. These architectures would be changed randomly in every iteration of ensemble process. In the following sub sections, two research methodologies that are used are discussed in the detailed manner.

2.1. Optimal Ensemble Architecture Selection using PSO (OEAS-PSO)

This is an optimal ensemble approach which is introduced in this work to improve the classification accuracy, thus the efficient detection of the financial fraudulent can be done. This approach optimizes the fraudulent behaviour detection process by adapting the nature of Particle Swarm Optimization (PSO) approach. PSO is a based on the biological behaviour of the birds in terms of their food foraging functions. The particles would attempt to move towards the location in which more food contents are available. This proposed research methodology adapts this nature for finding the more optimal ensemble, thus the more accurate detection of fraudulent behaviour can be done. OEAS-PSO will reach the optimal solution by continually updating the fitness function in every iteration. In this work fitness value is considered as the misclassification cost which should be minimized for more optimal solution. This fitness value is adjusted in every iteration until it reaches the end criterion by randomly changing the particles (combination of classifier and its architecture, subset feature). This method would result with the optimal ensemble which is then used in the testing phase for fraudulent pattern detection. The ensemble result is obtained by using the methodology called the weighted average fusion technique. This will take the weighted average value of every base classifier present in the ensemble which is then retrieved as the final result. The weighted average value of each classifier is calculated as follows:

$$C(\bar{x}) = \underset{i=1}{\overset{k}{\arg\max}} \left(\frac{\sum_{j=1}^{L} W_j Y_{ij}}{L} \right)$$

Where

Y_{ij} = ith output of jth base classifier

k = number of classifiers

L = number of base modules

W = weight value of classifier j

This approach leads to better classification output by selecting the optimal ensemble which is given in the following algorithm.

Input: Insurance data set

Output: Financial fraud detection results

1. Load the Insurance data

2. Select the relevant subset feature using FARD

3. Design the different ensemble architecture randomly using four base classifiers

4. Initialize particles with random selection of ensemble architecture with different subset features

5. Calculate the fitness value of each particle

Here fitness value is misclassification cost $\mathrm{Err}^{*}(y) = \frac{|\{x \epsilon X : c(x) \neq y\ (x)\}|}{|X|}$

6. until the end criterion met

7. Assign the best particle among the set of particles in terms of less misclassification cost

8. Update the particle velocity based on best velocity as like follows:

$$v_{i,d} \leftarrow \omega\, v_{i,d} + \phi_p\, r_p\, (p_{i,d} - x_{i,d}) + \phi_g\, r_g\, (g_d - x_{i,d})$$

9. Update the particles position

$$x_i \leftarrow x_i + v_i$$

10. If fitness (xi) > P

11. Update the particle best position $p_i \leftarrow x_i$

12. If ($f(p_i) < f(g)$)

13. Replace the global best solution as current best solution

14. End if

15. Randomly change the ensemble architecture and the subset feature

16. Repeat

17. Return best ensemble classifier

18. Load the test data on optimal ensemble classifier

19. Apply average weighted scheme

20. Return final result

The above algorithm provides a working procedure of OEAS-PSO approach which will retrieve the most optimal ensemble result.

PSO approach used in this work might lack from the selection of most optimal ensemble in case of presence of the more noises ie redundancy of particles in the environment. This drawback needs to be resolved for getting the most accurate and optimal result in the future. To do so, firefly approach is replaced instead of PSO which can operate well in case of presence of noises too. The working scenario is described detailed in the following sub section.

2.2. Optimal Ensemble Architecture Selection using firefly approach (OEAS-FFA)

OEAS-FFA approach is introduced in this work for overcoming the issues that arises in the OEAS-PSO because of noise intolerant behaviour of the particles. Firefly algorithm is a Meta heuristic behaviour which is based on the signalling nature of fire flies to attract other fire flies. The firefly would emit the more lightening signal, if they want to attract the other flies. The firefly with more lightening would be attracted by other fire flies, thus they will reach them. Here the lightening is act as a fitness values which should be better for win in the environment. This nature of fire flies is adapted in this work to select the most optimal ensemble classifier with good architecture and subset features. Here the ensemble classifier that can generate more accurate result with reduced misclassification cost would be selected as most optimal ensemble. Firefly approach can work well in case of presence of the noises too. It wont converge to the performance degradation in the noisy environment, and the final solution would be retrieved

more accurate than the OEAS-PSO. This is proved in the experimental scenario part. The algorithm for OEAS-FFA is given as follows:

Input: Insurance data set

Output: Financial fraud detection results

1. Load the Insurance data

2. Select the relevant subset feature using FARD

3. Design the different ensemble architecture randomly using four base classifiers

4. Initialize the objective function f(x) misclassification cost

misclassification cost $\mathrm{Err}^{*}(\mathrm{y}) = \frac{|\{x \epsilon X : c(x) \neq y\ (x)\}|}{|X|}$

5. Generate an initial population of fireflies X_i

6. Formulate light intensity I so that it is associated with f (x)

I = f (x)

7. Define absorption coefficient γ

8. While (t < MaxGeneration)

9. for i = 1 : n (all n fireflies)

10. for j = 1 : n (n fireflies)

11. if ($I_j > I_i$),

12. Move firefly i towards j;

13. Vary attractiveness with distance r via exp (-γr)

Randomly change the ensemble architecture and subset feature

14. Evaluate new solutions and update light intensity;

15. end if

16. end for j

17. end for i

18. Rank fireflies and find the current best;

19. end while

20. Post-processing the results and visualization;

21. Apply weighted average fusion technique

22. End with classification result

The above algorithm provides a better and accurate Ensembling classification result than the OEAS-PSO approach. This approach will vary the lightening parameter value in every iteration of the firefly approach to reach the final optimal result. The evaluation results are proved in the matlab simulation environment which is discussed detailed in the following section.

3. Experimental results

The experimental tests have been conducted in the MATLAB simulation environment to find whether the financial fraud detection occurs or not. The insurance data set would consist of different transaction behaviour of the users in the timely manner which is utilized for the accurate detection of the classification result. The performance evaluation is done between the existing methodology named as Optimal Ensemble Classification using PSO (OEC-PSO) and the proposed methodologies named as Optimal Ensemble architecture selection using PSO (OEAC-PSO) and Optimal Ensemble Architecture Selection using Fire Fly Approach (OEAC-FFA). The comparison is made against the performance measures called the accuracy and precision. The performance comparison is given in the detailed manner as like follows:

3.1. Accuracy comparison

Accuracy is defined as the degrees of reduced misclassification error rate in terms of classifying different number of features present in the environment. Accuracy of the proposed research work should be more in the proposed research work in terms of reduced misclassification error rate than the existing approach. The accuracy comparisons of the proposed and existing methodologies are given in Figure 3.

$$\text{Accuracy} = \frac{\text{Number of correctly classified data}}{\text{Total number of data}} \times 100$$

Figure 3. Accuracy comparison

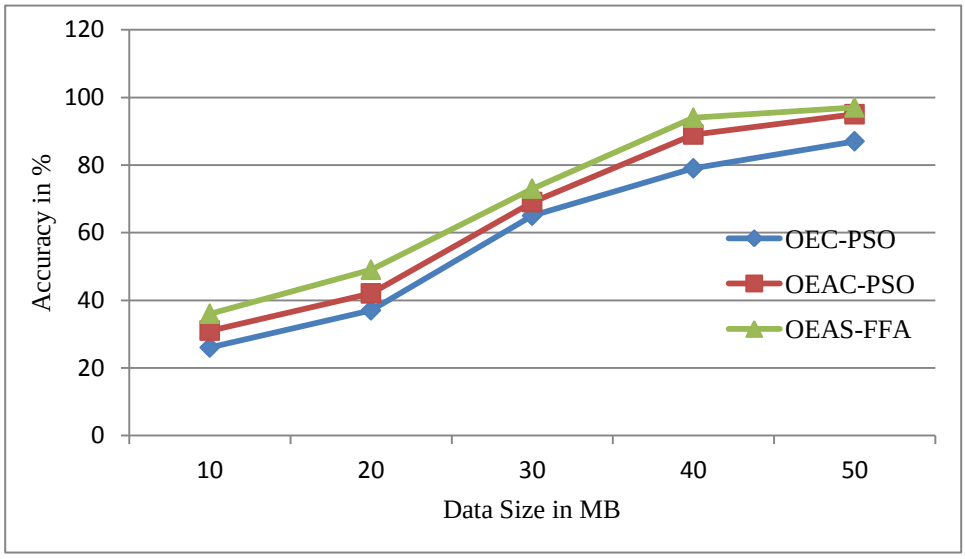


Figure 3 depicts the performance comparison of the accuracy parameter value of both existing and proposed research scenario in terms of different data sizes. In the x axis, different data sizes are taken and in y axis accuracy value obtained while predicting the fraudulent behaviour is taken. From this graph, it is proved that the proposed research work have more accuracy than the existing approach in terms of efficient detection of fraudulent behaviour.

4.2. Precision comparison

Precision value is defined as the amount of correctly predicted result over a total number of predicted results. This parameter is used to indicate the overall performance improvement of the proposed methodology in terms of predicting accurate results. The precision value is calculated as like follows:

$$\text{Precision} = \frac{\text{Number of correctly classified data}}{\text{Total number of classified data's}} \times 100$$

Figure 4. The performance comparison of the precision parameter value of both existing and proposed research scenario.

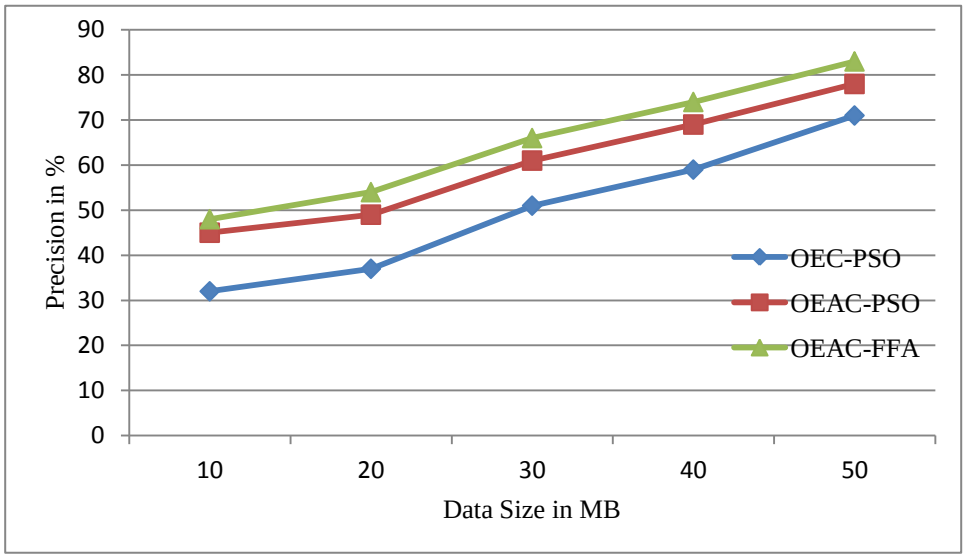


Figure 4 depicts the performance comparison of the precision parameter value of both existing and proposed research scenario in terms of different data sizes. In the x axis, different data sizes are taken and in y axis precision value obtained while predicting the fraudulent behaviour is taken. From this graph, it is proved that the proposed research work have more precision rate than the existing approach in terms of efficient detection of fraudulent behaviour.

5. Conclusion

Financial fraud detection is the most concerned task in the real world environment which needs to be avoided for ensuring the secured environment for the users. In this work, two approaches are used for obtaining better classification result namely OEAS-PSO and OEAS-FFA. These approaches are found to be better in selection of the optimal ensemble classifiers than the previous methodology OEC-PSO. This approach proved to better in terms of changing the ensemble architecture and the subset feature in every iteration. The experimental tests were conducted in the matlab simulation environment which is proved that the proposed research provides better result in terms of improved accuracy, precision and recall.

6. References

1. M. Paz Sesmero, Juan M. Alonso-Weber, German Gutierrez, Agapito Ledezma, Araceli Sanchis. An ensemble approach of dual base learners for multi-class classification problems *Information Fusion*. 2015; 24(1), 122-136.
2. Seokho Kang, SungzoonCho, PilsungKang. Multi-class classification via heterogeneous ensemble of one-class classifiers. *Engineering Applications of Artificial Intelligence*. 2015; 43(1), 35-43.
3. Fotini Markato poulou, Grigorios Tsoumakas, Ioannis Vlahavas. Dynamic ensemble pruning based on multi-label classification. *Neuro computing*. 2015; 150(1), 501-512.
4. Guo Haixiang, LiYijing, LiYanan, LiuXiao, LiJinling. BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. *Engineering Applications of Artificial Intelligence*. 2015; 12-11.
5. Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani Thuraisingham. Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints. *IEEE Transactions On Knowledge And Data Engineering*. 2011; 23(6), 859-874.
6. Yuhang Zhang, Hsiuhan Lexie Yang, Saurabh Prasad, Edoardo Pasolli, Jinha Jung, Melba Crawford. Ensemble multiple kernel active learning for classification of multisource remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2015; 8(2), 845-858.
7. Luca Canzian, Yu Zhang, Mihaela van der Schaar. Ensemble of distributed learners for online classification of dynamic data streams. *IEEE Transactions on Signal and Information Processing over Networks*. 2015; 1(3), 180-194.
8. Urvesh Bhowan, Mark Johnston, Mengjie Zhang, Xin Yao. Reusing genetic programming for ensemble selection in classification of unbalanced data. *IEEE Transactions on Evolutionary Computation*. 2014; 18(6), 893-908

9. Zhan-Li Sun, HanWang, Wai-Shing Lau, Gerald Seet, Danwei Wang, Kin-Man Lam. Microarray data classification using the spectral-feature-based tls ensemble algorithm. *IEEE Transactions On Nanobioscience*. 2014; 13(3), 289-299.
10. Yanwu Xu, Xianbin Cao, Hong Qiao. An efficient tree classifier ensemble-based approach for pedestrian detection. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics. 2011; 41(1), 107-117.
11. Ashfaqur Rahman, Daniel V. Smith, Greg Timms. A novel machine learning approach toward quality assessment of sensor data. *IEEE Sensors Journal*. 2014; 14(4), 1035-1047.
12. A. Prakash, C. Chandrasekar. An optimized multiple semi-hidden markov model. *Indian Journal of Science and Technology*. 2015; 8(2), 165-171.
13. Zahra Asheghi Dizaji, Farhad Soleimanian Gharehchopogh. A hybrid of ant colony optimization and chaos optimization algorithms approach for software cost estimation. *Indian Journal of Science and Technology*; 2015; 8(2), 128-133.
14. Reza Effatnejad, Fazlollah Rouhi. Unit commitment in power system t by combination of dynamic programming (DP), genetic algorithm (GA) and particle swarm optimization (PSO). *Indian Journal of Science and Technology*. 2015; 8(2), 134-141.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

***Citation**:*

C. Gayathri, R.Umarani. Efficient detection of financial fraud detection by selecting optimal ensemble architecture using optimization approaches. *Indian Journal of Innovations and Developments.* 2015; 4 (8), December.