A Novel Selective Ensemble Learning Based on K-means and Negative Correlation

Liu Liu(✉), Baosheng Wang(✉), Bo Yu(✉), and Qiuxi Zhong(✉)

College of Computer, National University of Defense Technology,
Changsha 410073, China
{hotmailliuliu,hotmailwbs,
BoYUnudt,Qiuxizhong}@163.com

Abstract. Selective ensemble learning has drawn high attention for improving the diversity of the ensemble learning. However, the performance is limited by the conflicts and redundancies among its child classifiers. In order to solve these problems, we put forward a novel method called KNIA. The method mainly makes use of K-means algorithm, which is used in the integration algorithm as an effective measure to choose the representative classifiers. Then, negative correlation theory is used to select the diversity of classifiers derived from the representative classifiers. Compared with the classical selective learning, our algorithm which is inverse growth process can improve the generalization ability in the condition of ensuring the accuracy. The extensive experiments demonstrate that the robustness and precision of the proposed method outperforms four classical algorithms from multiple UCI data sets.

Keywords: Ensemble learning · K-means · Negative correlation · Neural network

1 Introduction

With the development of machine learning, various methods are put forward continually [1–3]. Since Hansen and Salamon proposed neural network ensemble learning [4], ensemble learning has become an important branch of machine learning. It is also considered as the first of the four directions in machine learning by Dietterich [5]. Although many researchers have proposed different methods [6–8] to improve the conventional algorithms on good generalization ability and the diversity, but these can still lead to huge computational overhead for most users.

Usually ensemble learning is defined as, using different algorithms to solve the same problem, and using a strategy to integrate the separate solutions to achieve a joint decision [9]. The ensemble learning can be divided into two categories. One is only a method as the basic algorithm [10], such as Random-Forest algorithm [11], EENCL [12], GA-SVMs [13]. And the other is a variety of methods as the base, such as ensemble diversity measures. Ensemble learning is to diversity advantage into better learning accuracy, but the diversity maybe lead to significant time and space complexity. Selective ensemble learning method [14] was presented to give a solution to the problem. “Selective ensemble” concept was first introduced by Zhou [15], and had

X. Sun et al. (Eds.): ICCCS 2016, Part II, LNCS 10040, pp. 578–588, 2016.
DOI: 10.1007/978-3-319-48674-1_51

theoretical and experimental demonstration. Due to selective ensemble learning makes the weak learning algorithm combined with a strong learning algorithm, it has been a strong focus on the international academic community. But the difficulty lies in how to choose the algorithm by eliminating redundancy algorithm to reduce the computational complexity, and how to enhance the integration of diversity through the selection algorithm.

In the remainder of this article, the related knowledge to selective ensemble learning is introduced in Sect. 2. In Sect. 3, a novel selective ensemble learning (KNIA) is described, which has dynamic adaptability. This algorithm uses k-means to generate the representative classifiers, and then applies a negative correlation to select the derived algorithms. Different from traditional methods, this new method trains and chooses new algorithms by incremental learning, rather than training and testing all the algorithms. KNIA algorithm has more accuracy than other classical algorithms. This novel approach will significantly reduce the number of training and enhance the algorithm's adaptability and diversity. In Sect. 4, there are three experiments to show our algorithm has higher performance than other classical algorithms. To the impartiality of the experimental results, our experiments adopted the five experimental data from UCI. Finally, we summarize this paper and outline some future researches.

2 Related Work

Compared with other ensemble learning algorithms, the goal of the selective ensemble algorithm is to choose a subset from a large library of classifiers. The selective algorithm excludes the redundant classifiers to construct a streamlined ensemble learning. Thereby, it is able to reduce the storage space, and the computational complexity of ensemble learning approach. In the other words, the diversity and the quality among basal classifiers are important factors for improving the selective ensemble learning performance [16]. According to the characteristics of the current mainstream selective ensemble learning methods, it can be divided into three categories: according to the correlation about the subset of algorithms, according to the data set processing method, and according to the classification algorithm optimization. Correlation between the underlying algorithms generally can be divided into clustering algorithm based on distance metric and negative correlation algorithm utilizing the output of difference. Clustering is one of the unsupervised learning for analyzing data relationship. Clustering has the ability to divide similar data into the same class. This classification ability is based on some features, such as Euclidean distance. Negative correlation algorithm will be described in the following sections.

Using of processing data sets to design ensemble learning generally has two ways, one is the data set using different sampling methods, such as Bagging [17], Boosting [18, 19], AdaBoost [20]; the other is a set of data to extract its own characteristics, for example wrap, Attribute Bagging [21], FS-PP-EROS method [22]. The former tries to extract the dataset from different aspects as different training sets. So we can train up the different characteristics of methods with the different training sets. Difference between the two is to use different strategies to choose the training sets. The latter focuses on the choice of properties. However, the former produces different training sets that are obtained by random sampling.

The selective ensemble algorithm is optimized by coordinating the relationship between sub-algorithms. At present the Optimization method mainly makes use of the genetic algorithm, such as GASEN [23], PSO optimization algorithm [24]. GASEN assigns initial weights to all classifiers and employs genetic algorithm to evolve the weights. Through the evolved weights, it chooses some classifiers combined into the ensemble.

Selective ensemble algorithm apart from the three categories: Kappa Pruning [25], Stacked Generalization [26], mixtures-of-experts [27] and so on. However, these algorithms have a common drawback that will end once the learning algorithm to form a fixed pattern, not suitable for incremental learning, thus reducing the generalization ability of the model. To solve this problem, a novel method will be introduced which has adaptive and incremental capacity in the next section.

3 Algorithm Description

3.1 Selective Ensemble Learning

Construction of selective ensemble learning is divided into three parts: the first part is to determine the sampling methods of data collection, the second part is to choose an appropriate set of algorithms, and the third part is how to integrate independent algorithms to generate the final decision. To simplify the description of selective ensemble learning, M1 represents the sampling method, M2 represents the selection strategy which is used to choose the algorithms with diversity, and M3 stands for the final joint algorithm. M2 generally chooses the representative algorithms which have the minimum generalization error to constitute the ensemble learning, $\mathrm{E} = Agrmin_{S_{i \in S}}(e(S_i))$. Where S is the space-based classifier, s_i represents an optional sub-base algorithm space, e denotes an error subspace. Assume that dataset space for training set D and validation set $\underline{D}$, M1 uses the i-th sampling subset $D_i = \{(x_i, y_i) | M1(D) \in D\}$. If M2 uses a classic clustering algorithm, it can be described as selective integration algorithm:

Step 1: The data set in the proportion is divided into a training set D and validation set $\underline{D}$. D_i is the subset of D, which is sampled by Bootstrap method;

Step 2: S uses the training set D_i to train each algorithm, then the validation set D tests the performance of classifier s_i and their output matrix set $\mathrm{O} = \{o_1, o_2, \cdots, o_n\}$ are recorded;

Step 3: M2 is a clustering process, which is as follows:

Input: It output $\mathrm{O} = \{o_1, o_2, \cdots, o_n\}$ of each individual classifier is input of clustering algorithm;

Output: The plurality of $\{s_1, s_2, s_3, \cdots\}$ are the output of M2;

Processing: Let o_{ih}^{l} and o_{ih}^{j} is the output of the algorithm l and j on the h-th sample of the training set D_i. d denotes the differences between $o_{ih}^{l} = \{y_1, y_2, \cdots, y_n\}$ and

$o_{ih}^{j} = \{z_1, z_2, \cdots, z_n\}$, the distance function is $\mathrm{d}\left(o_{ih}^{l}, o_{ih}^{j}\right) = \sqrt{(y_1 - z_1)^2 + (y_2 - z_2)^2 + \cdots + (y_n - z_n)^2}$. If $\mathrm{d}(o_{ih}^{l}, o_{ih}^{j})$, l and j will be classified as same category. After completion of the first classification, the various types of center will be updated, and then continue the previous procedure. Until the final point of the center of the sample does not change, then the algorithm ends;

Step 4: Each subclass will adopt the algorithm with the smallest generalization error for a representative algorithms.

Step 5: Using the weights of the algorithm joint. Weight method is to give a certain weight $w_i(\sum_{i=k} w_i = 1)$ to each individual algorithm, and joint decision-making $\sum_{i=k} w_i d_i$.

From the above process, we can know that the traditional methods only choose the child algorithms which have the minimum classification error, without taking into account the relationship between algorithms. But some child algorithms have the positive correlation is that weaken the advantage of the ensemble learning. In the following section, we will introduce a novel algorithm, which reflects the process of selecting the child algorithms with complementary relationship.

3.2 KNIA Algorithm

The purpose of selective ensemble algorithm is to choose the appropriate classifiers instead of the whole algorithm library. ZHOU [15] has proved that many could be better than all. Based on the previous theory, the selective adaptive algorithm can be used instead of the traditional method. Our method does not need training all classifiers, compared with conventional selective ensemble algorithm. The novel algorithm uses an incremental strategy to build an adaptive architecture. K-means produces farther algorithms. And the negative correlation theory is used to select child-classifiers which are derived. If the result cannot achieve the intended effect, a number of new algorithms are constructed, which screened through a negative correlation theory. If the updating achieves the desired effect, the end of the algorithm; if the effect is increased, but the algorithm is not suitable for the expectations, then the algorithm returns to the previous step; if the effect is the same or deteriorated after the addition, then the newly added methods are deleted, and new classifiers are selected. Before the novel algorithm is described, some relevant affirmed must be explained. Suppose there are k representative classifiers, which are to derive the sub-algorithms. Assume that the set $S_{init} = \{c_1, c_2, c_3, \cdots, c_q\}$ represents the initial set of classifiers. $\mathrm{Train}(\mathrm{x}, \mathrm{y}) = \{(x_i, y_i), \mathrm{i} \in 1, 2, \cdots, \mathrm{N}\}$ and $\mathrm{D} = \{(x_i, y_i), \mathrm{i} \in \mathrm{m}\}$ represent the training data set and the validation data set, $T_1, T_2, T_3, \cdots \in Train(x, y)$ indicates the sample training set. Bagging sampling method is used [24], when the sampling frequency is enough large, there are approximately 36.8% of the data will not be drawn. Then the data is not drawn as the verification data.

Negative correlation [28, 29]: the relationship between child interaction classifiers is by a penalty term θ_i to the performance. If $e_i(N)$ represents the algorithm i at the N-th training error function, you can get the following error expression:

$$E_i(d) = \frac{1}{N}\sum\nolimits_{d=1}^{N} e_i(d) = \frac{1}{N}\sum\nolimits_{d=1}^{N} (F_i(d) - y(d))^2 + \frac{1}{N}\sum\nolimits_{d=1}^{N} \gamma p_i(d)$$

Wherein $p_i(d) = (F_i(d) - F(d))\sum_{i \neq j}(F_j(d) - F(d))$. When $\gamma = 0$ the penalty term is zero, which represents the sub algorithms are an independent. When $\gamma = 1$ calculating $\frac{\partial E_i(d)}{\partial F_i(d)} = F(d) - y(d)$. While similar definition, error function integrated algorithm can be expressed as $E_{en}(n) = \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N} F_i(d) - y(d)\right)$, then $\frac{\partial E_{en}(n)}{\partial F_i(d)} = \frac{1}{M}(F(d) - y(d))$, so as to get the following formula $\frac{\partial E_{en}(n)}{\partial F_i(d)} \propto \frac{\partial E_i(n)}{\partial F_i(d)}$. This indicates that ensemble learning error is passed ultimately by each individual sub algorithm, in order to obtain better accuracy. The ensemble learning uses a negative correlation to select the diversity of sub-classifiers [30].

The description of the novel algorithm is as follows:

Step 1: Initializing the set has q algorithms (S_{init}), wherein d << n

Step 2: Using k- means K, different algorithms ($S_{\mathrm{en}} = \{c_1, c_2, \cdots, c_k\}, \mathrm{K} = |S_{en}|$) are obtained from S_{init}. The process refers to the third step of the last section;

Step 3: Each of the classifier derives M child classifiers, so there are K * M child classifiers. The child classifiers are trained by the samples which are collected by Bagging method

Step 4: K * M classifiers are processed by a negative correlation method. It excludes the algorithms which does not meet the requirements and holds h child classifiers. $S_{en=}S_{en} + S^{'} = \{c_1, c_2, \cdots, c_K, c_{K+1}, \cdots, c_{K+h}\}$, $\mathrm{K} = |S_{en}| + h$, $\mathrm{h} \leq |S_{en}|$;

Step 5: If the updating S_{en} reaches the maximum precision, then stopping. Otherwise, it goes on to step 2;

Step 6: Finally, the classifiers are connected by the majority voting method

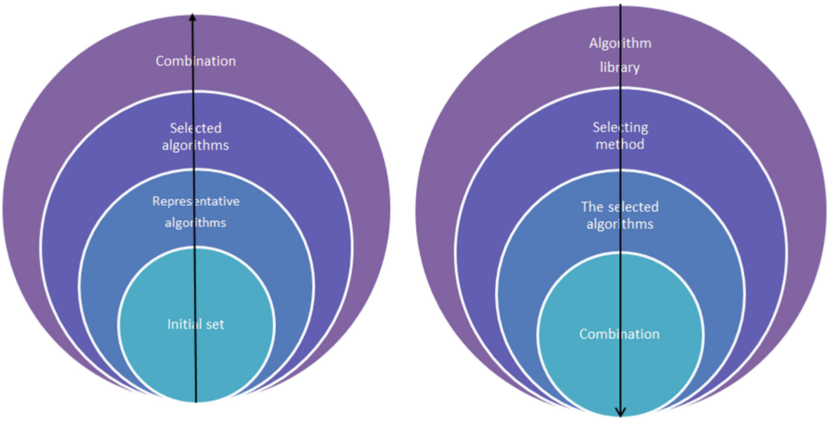

Fig. 1. Comparison chart of the process of the algorithms

We can easily get the novel algorithm's time complexity is O (n1 * m + NktT), wherein the n1 refers to the total number of application of the algorithm, T denotes iteration about our method and O(Nkt) stands for the complexity of k-mean. And the previous algorithm time complexity is O(n * m + nkt), where n represents the size of library, k denotes the numbers of clusters and t is iteration about k-mean. n1 and N are far less than n. And T is a fixed number. Therefore, the complexity of our method is less than the previous algorithms, when the n is very large.

Can be seen from the Fig. 1, our algorithm with adaptive is inverse growth process, which compares with the original algorithm. Algorithms growth process is the direction of arrow along a straight line. The original algorithm is to reach equilibrium by training an algorithm library. However, KNIA is along the opposite view. It incrementally chooses new classifiers with negative correlation to achieve the result of optimization. KNIA uses the K-means algorithm to generate the father algorithms, rather than generating the candidate classifiers directly. This has the advantage that the different fathers have a greater probability of producing the diversity of children. By calculating the negative correlation, the algorithms can be selected with the appropriate subset of difference. KNIA is a derived method can improve the adaptive ability, and strengthen generalization ability. And only a limited number of training algorithms can achieve the desired effect.

4 Experiment

4.1 Comparison of Various Models

One main objective of the experiment is to compare the performance difference between BPNN, Random-Forest method, Decision-Tree method, BPEnsemble and our algorithm. Five data sets from UCI are used in the experiment, which are Wine data set, Breast_Cancer dataset, Vote dataset, Chronic_Kidney dataset, and Phishing_ Web_sites_Features dataset. Wine dataset records three kinds of Italian wine ingredients, and it contains 178 samples. And each sample contains 13 different ingredients and a given sample label. Breast-cancer dataset contains 569 various health indicators, including 30 findings parameters and a result label. Vote dataset is the devoting information of the U.S. congress, and it contains 435 examples and 16 properties. Chronic_Kidney_disease dataset is about the chronic kidney disease, which was collected from the hospital nearly 2 months of period. This dataset includes 400 instances and 25 attributes. Phishing_Websites_Features dataset is the latest data set about phishing websites, which contains11055 examples. There are 30 features and a result label in every example (Fig. 2).

BPNN refers to the feed-forward neural network, which has 10 neurons. BPEnsemble is an ensemble learning method using BPNN as a sub algorithm. It contains 100 BPNN, and each BPNN adopts eight neurons. Random-Forest uses the Boostrap sampling method, meanwhile it contains 500 randomly Decision-Tree. KNIA initial state integrates 5 BPNN, 10 Decision-Tree, 5 support vector machine (SVM) and 1 extreme learning machine. To evaluate the performance of all the training methods

Table 1. Experiment results

Data set	BPNN	Decision-Tree	Random-Forest	BPEnsemble	KNIA
Wine	0.9438	0.9326	0.9888	0.9663	1
Bcancer	0.9420	0.9420	0.9560	0.9855	0.9888
Vote	0.9471	0.9632	0.9632	0.9524	0.9643
Cdisease	0.9524	1	1	0.9762	1
PWsites	0.8751	0.8878	0.9367	0.9647	0.9763

described above, we use the five data sets to train the methods. The results are shown in Table 1. Observing the table of results, we notice that:

(1) When the ensemble algorithms only are considering, KNIA outperforms BPEnsemble and Random_Forest in four of all five data sets. And it can find that the Random-Forest and BPEnsemble algorithms with the integrated nature are better than the other two algorithms. This shows that ensemble learning method has the natural advantages, which can integrate the advantages of large classifiers.

(2) The results show that the accuracy of our new method is improved by an average of 5.38%, 5.38%, 1.694% and 1.686% compared respectively with other methods. Although the performance of the new algorithms and the best algorithms is similar or equal in some data set, our algorithm is always better than the worst algorithm. This shows that our algorithm has very good generalization ability.

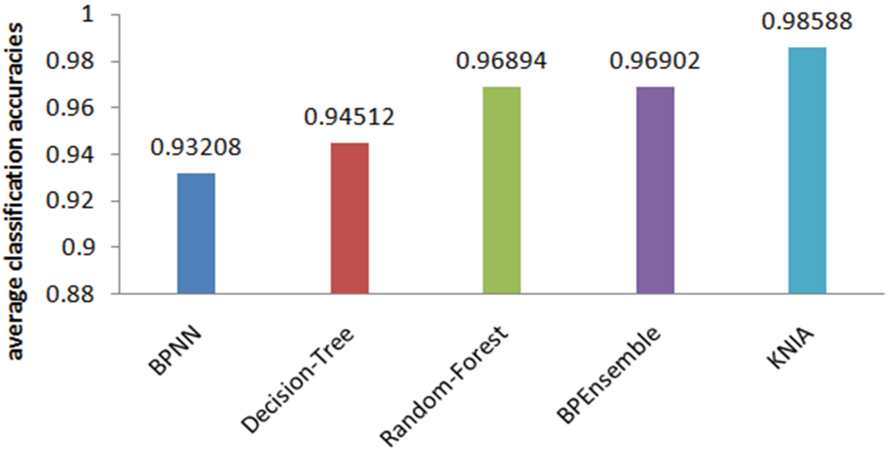

Fig. 2. Average performance of five algorithms

(3) Compared with other algorithm integration algorithm, we found that the result of the integration algorithm is significantly affected by selection. And the experimental results show that the new algorithm has a negative correlation is better than all integration or random selection, because the negative correlation property is able to eliminate and mitigate the mutual exclusion.

4.2 The Impact on the Number of Neurons

In order to prove the superiority of the new method from several aspects, the KNIA integrates the BPNN algorithm in this part. Neuron was a very important parameter for most neural networks, so there were two experiments to be carried out for different neurons. One was to compare the use of different neurons in the new methods, the other was the comparison of BPEnsemble and KNIA using different neurons. The first part of the experiment was aimed at KNIA, and it was assumed that the number of neurons was the same in the neural network sub algorithm, and the second was to compare the KNIA with different neurons (Fig. 3):

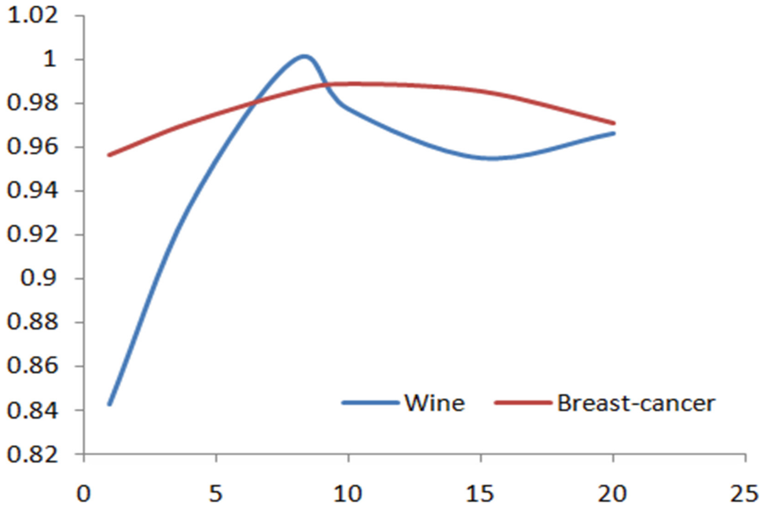

Fig. 3. Different neurons comparison

We used the method with 1, 4, 8, 10, 15 and 20 neurons. The result shows that the test result has a close relationship with the neurons. Not the more neurons the better, nor is it good for different data to use the same. For Wine and Breast-cancer, KNIA method can get the best effect by using 8 and 10 neurons respectively (Fig. 4).

As can be seen from the above, KNIA and BPEnsemble have similar curves, but the KNIA method is better than the BPEnsemble method. This shows that even though the two methods use the BP as the integrated sub algorithm, but the different integration methods have different effects. Secondly, the performance of the algorithm is related to the number of neurons, and it is very important to select the appropriate number of neurons. The reason of this phenomenon is that too many neurons can lead to over fitting of training, but too few of neurons can lead to the network's ability to classify.

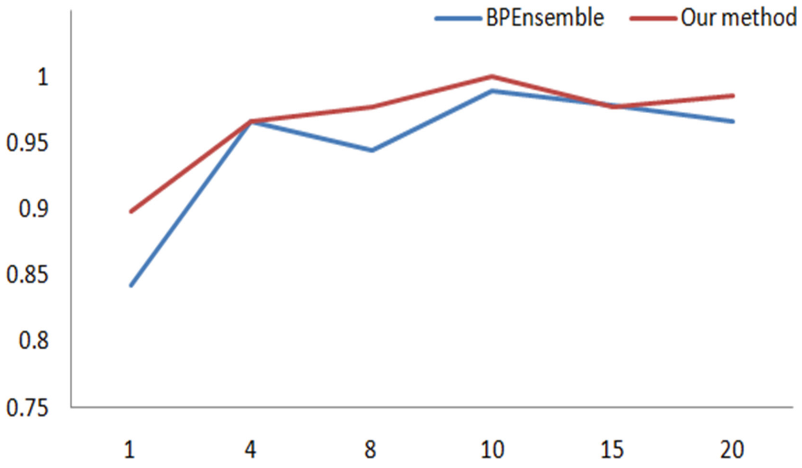

Fig. 4. KNIA and BPEnsemble comparison

4.3 Effect of Classification Algorithm K-means

In the KNIA method, K-means method is used to divide a subset of data, so different K maybe have an important impact on the results. To confirm this, this experiment constructed six different KNIA methods which corresponded to six different k on the data set Wine. The results are as follows:

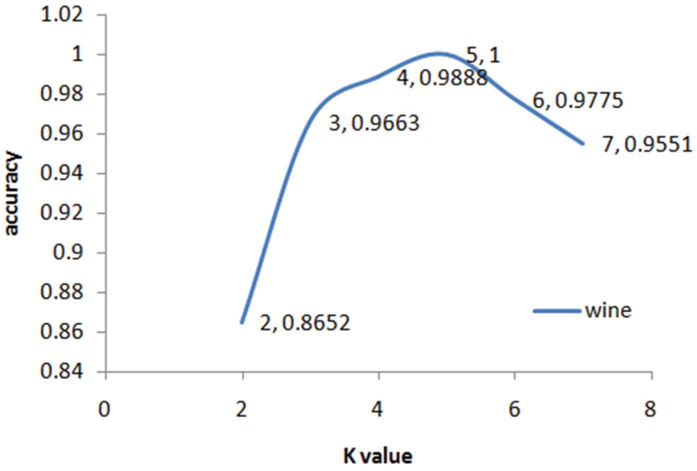

Fig. 5. The effect of K value

Figure 5 shows that the number of K-means clustering has a significant effect on performance. When K is five, the algorithm is close to maximum performance. When K is 2, it shows the worst performance. The reason is that too few of the clusters cannot achieve the purpose of training sufficient number of algorithms, so that the negative correlation theory is not enough to choose a different algorithm. And too many of the different data sets training a large number of method sets, may have a greater probability to obtain a conflict with each other or have a repeat of the method sets, which is not conducive to improve the performance of the integrated algorithm.

5 Conclusion

In this paper, KNIA method is employed to create the heterogeneous ensemble learning. It mainly uses K-means method to construct derived classifiers. And negative correlation theory is used to select the weights with different properties. KNIA method effectively raises the differences between the selected weights. Meanwhile the new algorithm with adaptive algorithm significantly reduces the time complexity. The experiments show that KNIA algorithm compared with other classical algorithms improves 3.54% on average. And we prove that its superiority has universality from different angles. Although there are more than ten thousand records in Phishing_Websites data set, it cannot achieve the level of big data. This may be not sufficient to show the advantages and disadvantages of the algorithm. There is a disadvantage of the novel method is that the advantages of sub-algorithms are not reasonably utilized in the joint phase. So KNIA algorithm still needs us to continue to in-depth study.

Acknowledgment. The author is grateful to Baosheng Wang and Bo Yu for the guidance and advice, and thanks to the support of the project. The work was supported by Science Foundation of China under (NSFC) Grant No. 61472437 and No. 61303264.

References

1. Gu, B., Sheng, V.S., Tay, K.Y., Romano, X., Li, S.: Incremental support vector learning for ordinal regression. IEEE Trans. Neural Netw. Learn. Syst. **26**(7), 1403–1416 (2015)
2. Gu, B., Sheng, V.S.: A robust regularization path algorithm for -support vector classification. IEEE Trans. Neural Netw. Learn. Syst. (2016). doi:10.1109/TNNLS.2016.2527796
3. Gu, B., Sun, X., Sheng, V.S.: Structural minimax probability machine. IEEE Trans. Neural Netw. Learn. Syst. (2016). doi:10.1109/TNNLS.2016.2544779
4. Hansen, L.K., Salamon, P.: Neural network ensemble. IEEE Trans. Pattern Anal. Mach. Intell. **12**(10), 993–1001 (1990)
5. Dietterich, T.G.: Machine learning research: four current directions. AI Mag. **18**(4), 97–136 (1977)
6. Liu, X., Wang, L., Huang, G.B., et al.: Multiple kernel extreme learning machine. Neurocomputing **149**, 253–264 (2015)
7. Yao, W., Chen, X.Q., Zhao, Y.: Efficient resources provisioning based on load forecasting in cloud. IEEE Trans. Neural Netw. Learn. Syst. **23**(2), 247–259 (2012)
8. Tang, J., Cao, Y., Xiao, J., et al.: Predication of plasma concentration of remifentanil based on Elman neural network. J. Central South Univ. **20**, 3187–3192 (2013)
9. Krogh, P.S.A.: Learning with ensembles: how over-fitting can be useful. In: Proceedings of the 1995 Conference, vol. 8, p. 190 (1996)

10. Liu, J., Chen, H., Cai, B., et al.: State estimation of connected vehicles using a nonlinear ensemble filter. J. Central South Univ. **22**, 2406–2415 (2015)
11. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
12. Liu, Y., Yao, X., Higuchi, T.: Evolutionary ensembles with negative correlation learning. IEEE Trans. Evol. Comput. **4**(4), 380–387 (2000)
13. Tao, H., Ma, X., Qiao, M.: Subspace selective ensemble algorithm based on feature clustering. J. Comput. **8**(2), 509–516 (2013)
14. Cheng, X., Guo, H.: The technology of selective multiple classifiers ensemble based on kernel clustering. In: Second International Symposium on Intelligent Information Technology Application, IITA 2008, vol. 2, pp. 146–150. IEEE (2008)
15. Zhou, Z.H., Wu, J., Tang, W.: Ensembling neural networks: many could be better than all. Artif. Intell. **137**(1), 239–263 (2002)
16. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
17. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. CRC Press, Boca Raton (1994)
18. Freund, Y.: Boosting a weak learning algorithm by majority. Inf. Comput. **121**(2), 256–285 (1995)
19. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
20. Bryll, R., Gutierrez-Osuna, R., Quek, F.: Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern Recogn. **36**(6), 1291–1302 (2003)
21. Hu, Q., Yu, D., Xie, Z., et al.: EROS: ensemble rough subspaces. Pattern Recogn. **40**(12), 3728–3739 (2007)
22. Thompson, S.: Pruning boosted classifiers with a real valued genetic algorithm. Knowl.-Based Syst. **12**(5), 277–284 (1999)
23. Fu, Q., Hu, S.X., Zhao, S.Y.: A PSO-based approach for neural network ensemble. J. Zhejiang Univ. (Eng. Sci.) **38**(12), 1596–1600 (2004)
24. Margineantu, D.D., Dietterich, T.G.: Pruning adaptive boosting. ICML **97**, 211–218 (1997)
25. Ting, K.M., Witten, I.H.: Issues in stacked generalization. J. Artif. Intell. Res. (JAIR) **10**, 271–289 (1999)
26. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the EM algorithm. Neural Comput. **6**(2), 181–214 (1994)
27. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
28. Liu, Y., Yao, X.: Ensemble learning via negative correlation. Neural Netw. **12**(10), 1399–1404 (1999)
29. Minku, F.L., Inoue, H., Yao, X.: Negative correlation in incremental learning. Nat. Comput. **8**(2), 289–320 (2009)
30. Liu, Y., Yao, X.: Simultaneous training of negatively correlated neural networks in an ensemble. IEEE Trans. Syst. Man Cybern. B Cybern. **29**(6), 716–725 (1999)