Proceedings of the 7th
World Congress on Intelligent Control and Automation
June 25 - 27, 2008, Chongqing, China

Neural Network Ensemble Based on Rough Sets Reduction and Selective Strategy*

Yaonan Wang
College of Electrical and Information Engineering
Hunan University
Changsha, Hunan Province 410082, China
yaonan@hnu.cn

Dongbo Zhang, Huixian Huang
Institute of Information Engineering
Xiangtan University
Xiangtan , Hunan Province, 411105, China
zhadonbo@yahoo.com.cn, huanghx0618@hotmail.com

***Abstract* –Based on rough sets reducts, a new neural network ensemble method is proposed. Reducts with robustness and good generalization ability are achieved by a dynamic reduction technology. Then according to different reducts, multiple BP neural networks are designed as base classifiers. And with the idea of selective ensemble, the best neural network ensemble can be found by some search strategies. Finally, by combining the predictions of component networks with voting rule, classification can be implemented. Compared with conventional ensemble feature selection algorithms, less time and lower computing complexity is needed of the method in this paper.**

***Index Terms* - Rough sets, Reduction, Neural network ensemble, Remote sensing image classification**

I. INTRODUCTION

By designing multiple neural network predictors and integrating their outputs, NNE [1] can significantly improve the generalization ability of learning system. As a kind of engineering neural computation technique with broad application prospect, NNE has been one of current research focuses in the field of machine learning and neural computation [2].

According to theoretical analysis of Krogh [3], to lower the ensemble generalization error, high accuracy neural network components are needed, in addition the error between them should be as much as possible uncorrelated and diverse. There are various approaches to generate diverse neural network components, for example, the components in ensemble can be built through different object functions [4], network structures, number of hidden neurons[5] and initial weights[6]. Except for these, some literatures introduce cross-validation technique [3], genetic algorithm [7][8], negative correlation learning [9][10] and clustering algorithm [11][12] etc., to generate neural network components. Among them, the approaches mostly important and widely used are Boosting [13] and Bagging [14]. Besides, neural network ensemble based on feature selection is also a valid technique, which has been showed in many recent literatures [15][16][17].

Ensemble feature selection will be implemented by search strategies such as mountain climbing [18], genetic algorithm[16], sequent forward and sequent backward[19] etc., when the evaluating function balancing accuracy and diversity is given. Because the search is in the total of feature space, so if the dimension of the feature space increases, the number of feature subsets expands rapidly and time for training neural network increase sharply. To solve this problem, we propose a new feature subset selection method based on rough sets reduction. By knowledge reduction, some amounts of reductive subsets are found as the candidate feature subsets to generate neural network components, which will decrease the number of feature subsets needed to be considered. According to the characteristics of rough sets reducts, which has the same classification ability as total of feature without reduction, the accuracy of neural network member will be guaranteed, in addition the dimension of input feature space and complexity of component networks can be greatly reduced.

II. PRELIMINARY

A. Concept of Rough Sets

An information system can be denoted by 4 tuple $S = (U, A, V, f)$, where U is the universe, A is the set of feature attributes, $V = \bigcup_{a \in A} V_a$, V_a is the value domain of attribute a, information function $f : U \times A \to V$, determine value on attribute A for every object. Generally, information system S can also be denoted by simplified form, i.e. $S = (U, A)$.

Every subset $B \subseteq A$ can define a binary indiscernible relation on U, i.e., $ind(B)$. By it, U can be partitioned with equivalent classes U / B, which satisfies: $\bullet\ U / B = \{X_1, X_2, \cdots, X_n\}$; $\bullet\ X_i \subseteq U$, $X_i \neq \phi$, $X_i \cap X_j = \phi$, $i \neq j\ i, j = 1,2,\cdots,n$; $\bullet \bigcup_{i=1}^{n} X_i = U$.

Every $X_i \in U / B$ ($i = 1,2,\cdots,n$) is called basic equivalent class, among them, the class containing x is denoted as $B(x)$.

X is B definable or B exact, if it can be demonstrated with the union of some equivalent classes in

* This work is partially supported by NSFC Grant #60775047 to Y.N. Wang and Natural Science Foundation of Hunan Province Grant #06JJ5112 to H.X. Huang

U/B, otherwise, X is B undefinable or B rough set. Given a rough set X, $\overline{B}(X)$ and $\underline{B}(X)$ are called upper and lower approximation set of X, respectively:

$$\underline{B}(X) = \bigcup\{B(x) \big| B(x) \subseteq X, x \in U\} \quad (1a)$$

$$\overline{B}(X) = \bigcup\{B(x) \big| B(x) \bigcap X \neq \phi, x \in U\} \quad (1b)$$

Moreover $POS_B(X) = \underline{B}(X)$ is B positive region of X.

If attribute set A is combined with conditional attribute set C and decision attribute set D then information system is also called decision system and can be denoted as $S = (U, C \bigcup D)$. Generally, there are often exists some extent of dependency and correlation between C and D, which can be defined by dependency factor:

$$\gamma_C(D) = \left|POS_C(D)\right| / \left|U\right| \quad (2)$$

Where $\left| * \right|$ computes the cardinality of the set, $POS_C(D) = \bigcup_{X \in U/D} \underline{C}(X)$ means C positive region on D, dependency factor $\gamma_C(D)$ weighs the ratio that the objects can be classified correctly to corresponding class label with the knowledge demonstrated by attribute C.

Because of the correlation of the condition attributes, so not all of them are necessary for decision, from this view, the problem of attribute reduction is derived. Reduction can demonstrate the dependency and correlation between condition attributes and decision attributes with most simplified style, while the classification performance will not decrease.

If the nonempty subset $C' \subseteq C$ satisfies:

a) $\gamma_C(D) = \gamma_{C'}(D)$;

b) there are not exist $C'' \subset C'$ that satisfies $\gamma_{C'}(D) = \gamma_{C''}(D)$.

Then C' is a reduct of C related to D. Generally, reduct is not unique and multiple reducts of an information system describe it from different input feature subset space, thus a large amount of redundant information can be supplied with them.

B. Analysis of Ensemble Generalization Error of NNE

Suppose that by neural network ensemble learning, a function $f: R^m \to R^n$ will be approximated，where the ensemble comprises N component neural networks $f_i (i = 1, \cdots, N)$. And the weight $\omega_i (i = 1, \cdots, N)$ satisfies the conditions that $\omega_i \geq 0$ and $\sum_{i=1}^{N} \omega_i = 1$ is assigned to every component network. Then predictions of component networks are combined with weighted averaging method. For convenience of discussion, here we assume that each component network only contains one output unit, i.e. $n = 1$. However the corresponding conclusions can be easily generalized to situations where each component neural network has more than one output units.

Now suppose the input $x \in R^m$ satisfies probability distribution $p(x)$. And the expected and actual output of *i*th component network on x is $d(x)$, $f_i(x)$ respectively. Then the ensemble output is:

$$\overline{f}(x) = \sum_{i=1}^{N} \omega_i f_i(x) \quad (3)$$

The generalization error E of the ensemble network on x *is*:

$$E = \int p(x)(\overline{f}(x) - d(x))^2 dx \quad (4)$$

E_i is the generalization error of *i*th component neural network on x:

$$E_i = \int p(x)(f_i(x) - d(x))^2 dx \quad (5)$$

And the diverse measure between component network f_i and ensemble network $\overline{f}$ is defined:

$$A_i = \int p(x)(f_i(x) - \overline{f}(x))^2 dx \quad (6)$$

According to theoretical analysis of Krough [3]:

$$E = \overline{E} - \overline{A} \quad (7)$$

Where $\overline{E} = \sum_{i=1}^{N} \omega_i E_i$ is the weighted averaging generalization error of all of component networks. And $\overline{A} = \sum_{i=1}^{N} \omega_i A_i$ is the weighted averaging diverse measure between component network and ensemble network.

It is known from Eq. (7) that the generalization error of the ensemble can be reduced in two aspects: one is improve the accuracy of component network, which is generally a difficult problem in many actual applications. The other is increase the diversity of component network. So to improve the generalization ability of ensemble network, we need to reduce the error correlations of component network as much as possible.

III. PRINCIPAL FOR ROUGH SETS REDUCTS BASED NNE

A. The Model for Rough sets Reducts Based NNE

Based on the theory of rough sets reduction, the method to generate component networks for ensemble is

proposed. Fig.1 is the model of NNE based on multiple rough sets reducts.

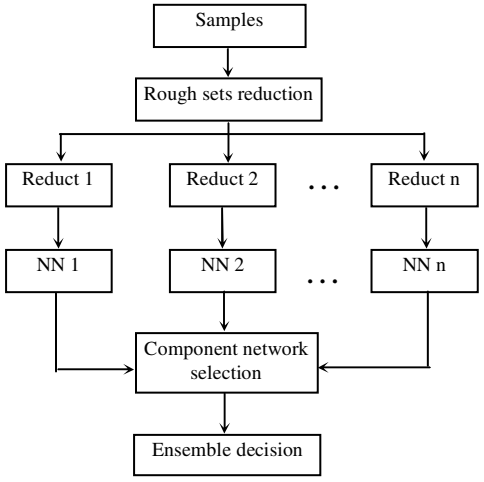

Fig. 1 Model of neural network ensemble based on multiple rough sets reducts

According to characteristics of rough sets reduct, the accuracy of the component network can be guaranteed. And due to different reducts related to different subset feature spaces, so it is favorable to increase the diversities of component networks. The generalization ability can be improved by making full use of the redundant information supplied by rough reducts.

B. Dynamic Reduct Approach Based on GA and Resampling Method

The computing for rough sets reducts is a NP-hard problem. And the computation is exceedingly great for high dimensional system and/or large data sets. Thus in this condition, GA (genetic algorithm) [20] will be a favorable approach to be adopted.

While designing classifier, the data sets are often partitioned into training samples and test samples, and only training samples are used to design classifier. However, rough sets reduct is often sensitive to samples distribution and due to the randomness of sampling, the training samples distribution and corresponding reducts will be different for every random sampling operation. Thus to find reducts with the characteristics of stable and powerful generalization ability, a dynamic reduct approach [21] is adopted here:

1) n times random sampling are implemented according to the pre-specified sampling ratio, and a set of subsystems that compose system S is achieved, i.e., $S = \{S_1, S_2, \cdots S_n\}$. Among them subsystem i is $S_i = (U_i, A)$, where $U_i \subseteq U$ and A is original attribute set;

2) By GA, searching out corresponding attribute reducts $RED_{S_i}(A)$ for every subsystem $S_i \in S$, $i = 1,2,...,n$;

3) Count out the frequencies of all of reducts that appear in step 2).

Higher frequency of a reduct means the reduct is stable and it has powerful generalization ability. Then all of reducts belong to $RED_{S_i}(A)$, $i = 1,2,...,n$ will be output if the frequency of them exceed specified threshold th_f:

$$DRED(A, th\_f, S) = \{B \subseteq A \mid \frac{\left|\{S_i \in S \mid B \in RED_{S_i}(A),\ i = 1,...,n\}\right|}{|S|} \geq th\_f\} \quad (8)$$

C. Component network selection

Though, large amount of reducts can be generated through above dynamic reduct approach. However, the errors of neural networks designed on them may be correlated. And ensemble generalization error may not reduce, in contrary, it even possible goes up[22]. So the ensemble using all the individual networks may not result in best generalization ability, which is both demonstrated in literatures of Zhou[23] and Wu[24] . Based on this point, the researches proposed the idea of selective neural network ensemble.

Remote sensing image classification used in our experiment belongs to the problem of pattern recognition. Due to convenience for discussion, the corresponding analysis is based on two class labels classification problem. But note that the following derivations can also be generalized to situations where more than two class labels may be contained.

Suppose that N neural networks $f_i (i = 1, \cdots, N)$ are combined to approximate a function $f : R^m \to \ell$, where $\ell = \{-1, +1\}$ is the set of class labels. The outputs of individual neural network are combined through majority voting. We assume that there are m instances and expected output of these instances is $[d_1, d_2, \cdots d_m]^T$, where $d_j \in \{-1, +1\} (j = 1,2, \cdots m)$ is the expected output of j th instance. And actual output of i th component network on these instances is $[f_{i1}, f_{i2}, \cdots f_{im}]^T$, where $f_{ij} \in \{-1, +1\} (i = 1,2,..., N$; $j = 1,2, \cdots m)$ is the actual output of i th component network on j th instance. Apparently, if actual output and expected output of i th component network on j th instance is consistent then $f_{ij} d_j = 1$, otherwise $f_{ij} d_j = -1$. Thus the generalization error of the i th component neural network on those m instances is:

$$E_i = \frac{1}{m}\sum_{j=1}^{m} error(f_{ij} d_j) \tag{9}$$

Where $error(x)$ is a function defined as:

$$error(x) = \begin{cases} 1, & x = -1 \\ 0.5, & x = 0 \\ 0, & x = 1 \end{cases} \tag{10}$$

By voting, the sum output on j th instance of N networks ensemble can be counted:

$$sum_j = \sum_{i=1}^{N} f_{ij} \tag{11}$$

And the ensemble output on j th instance is:

$$\bar{f}_j = sgn(sum_j) \tag{12}$$

where $sgn(x)$ is a sign function:

$$sgn(x) = \begin{cases} 1, & x > 0 \\ 0 & x = 0 \\ -1, & x < 0 \end{cases} \tag{13}$$

Obviously, $\bar{f}_j \in \{-1,\ 0,\ +1\},\ (j = 1,\ 2,..., m)$. If the ensemble output is consistent to expected output on j th instance then $\bar{f}_j d_j = 1$; else if it is not consistent then $\bar{f}_j d_j = -1$, otherwise $\bar{f}_j d_j = 0$, which denotes the correct and wrong decisions are identical. Thus the ensemble generalization error is:

$$E = \frac{1}{m}\sum_{j=1}^{m} error(\bar{f}_j d_j) \tag{14}$$

If k th component network is removed from ensemble then the rest of networks ensemble output on j th instance is:

$$\bar{f}_j^{-k} = sgn(sum_j - f_{kj}) \tag{15}$$

And the new ensemble generalization error is:

$$E^{-k} = \frac{1}{m}\sum_{j=1}^{m} error(\bar{f}_j^{-k} d_j) \tag{16}$$

From Eq.(14) and Eq. (15), we know that if k th component network satisfing Eq. (17) is excluded from ensemble then $E^{-k} \le E$, which means that the ensmble is better than the one including k th component network.

$$\sum_{j=1}^{m} \{error(sgn(sum_j) d_j) - error(sgn(sum_j - f_{kj}) d_j)\} \ge 0 \tag{17}$$

It is indicated from above derivation that if the individual networks satisfying Eq. (17) are excluded from the ensemble then the performance is better than that of the ensemble including all of them.

Though for two class labels classification, we can select component networks by Eq. (17), but in actual multiple class labels classification applications, the computational cost required for excluding those neural networks that should not join the ensemble is still too extensive to be met. Thus in real world application, we adopt independent validation samples to evaluate the generalization error of ensemble by Eq. (14), and find the best ensemble neural network based on some searching strategies.

If N neural network classifiers have been designed based on rough sets reducts then the methods finding best or close to best ensemble network are numerous, for example, exhaustive approach, by it all of the possible combinations (excluding empty, there are $2^N - 1$ possibilities) are computed, and the combination with minimum generalization error is found. The exhaustive approach can find best neural network ensemble easily when N is small. But if N is large (for example $N > 30$) then the computational cost is too extensive to be met. In this condition, GA is a better selection for global optimal searching.

Finally, majority voting based ensemble strategy is used to combine the predictions of Component networks.

IV. Experiment Results And Analysis

In experiment, a 1024×1026 Landsat TM 7 bands remote sensing image is adopted. By observation, land cover can be categorized into five classes, C_1-plant, C_2-river, C_3-higher density built area, C_4-lower density built area, C_5-bare. Total of 7217 samples are extracted, where 30% （2165）and 20% （1443） of them are selected as training samples and valid samples respectively, and the rest 50% （3609）are test samples. Expected output of each class is encoded as $d_1(1,0,0,0,0)$, $d_2(0,1,0,0,0)$, $d_3(0,0,1,0,0)$, $d_4(0,0,0,1,0)$, $d_5(0,0,0,0,1)$.

Because intensity of each band is already discrete value varies from 0～255, so they can be used directly to construct decision table. To achieve stable and powerful generalization ability rough sets reducts by dynamic reduct technique, 40 random sampling operations are conducted. Each time only 30% samples are extracted from training set. The total of rough sets reducts we find in all of these 40 operations is 56, however only 12 reducts whose frequency is over 15. Given the threshold of frequency $th\_f = 15/40 = 0.375$ ，then the reducts whose frequency are more than 15 are selected as final rough reducts (TABLE. I).

TABLE. I Rough Set Reductions That Appears More Than 15

index	Reduct bands	index	Reduct bands
1	B4，B6，B7	7	B3，B5，B6，B7
2	B3，B4，B6	8	B2，B4，B5，B6
3	B4，B5，B6	9	B1，B3，B4，B5

4	B1，B3，B5，B6	10	B1，B2，B5，B7
5	B1，B5，B6，B7	11	B1，B2，B4，B5
6	B2，B5，B6，B7	12	B1，B2，B3，B6，B7

Based on above 12 optional reducts, 12 individual neural networks can be constructed as base classifier. Here BP networks with single hidden layer are adopted as candidate component networks. And all of them are designed with the same 15 neurons in hidden layer, and 5 neurons in output layer, which corresponding to 5 class land covers. While the number of neurons of input layer may be different, what is determined by bands number in reductive attribute set. Then the component networks are all trained by back propagation algorithm.

According to the idea of selective ensemble, appropriate component networks that has least ensemble generalization error must be found from 12 optional individual networks. As the total of possible combinations are $2^{12}-1=4095$, and the search space is not very large, so simple exhaustive searching method is used. while searching, ensemble generalization error is evaluated on valid samples set, and best ensemble network with least generalization error is found, which consists of 5 individual networks indexed by 2, 4, 9, 11, 12 (TABLE. I).

The performances of ensemble networks are compared and the results are showed in TABLE II. It is apparent that the best ensemble network achieved by our method has the topmost accuracy (97.82%), which is better than most of other situations, such as best individual network, ensemble all of individual networks and network based on all band attributes. Also it is indicated that ensemble network, such as ensemble all individual network and best ensemble network, both have higher accuracy than that of average of all individual network and any individual network. Moreover the accuracy of best ensemble network increases 1.2% than that of average of all individual networks. In addition, the accuracies of former $1\sim12$ network ensemble are presented in TABLE III, from it, we know that the ensemble classification accuracy is not simply improve as the increase of the size of ensemble. The former 5 neural networks has ensemble accuracy (97.74%), which is even higher than all of 12 neural networks ensemble accuracy (97.54%). The experiment results indicate that it is necessary to ensemble with selective strategy.

TABLE II THE CLASSIFICATION PERFORMANCE COMPARISON OF NEURAL NETWORK ENSEMBLE

	Best individual network	Ensemble all of individual network	Network based on all bands	Best ensemble network	Conventional ensemble feature selection
accuracy（%）	97.41	97.54	97.35	97.82	97.91

TABLE III THE CLASSIFICATION ACCURACY OF FORMER 1～12 INDIVIDUAL NEURAL NETWORK ENSEMBLE

	The size of ensemble											
	1	2	3	4	5	6	7	8	9	10	11	12
accuracy（%）	97.27	97.21	97.47	97.47	97.74	97.57	97.35	97.49	97.59	97.49	97.66	97.54

It is obvious that, compared with conventional ensemble feature selection approaches, ensemble network based on rough sets reducts has the merits of smaller searching space, less time cost and lower computational complex. When conventional ensemble feature selection approaches based on total feature space is used, the possible number of subset feature subspaces is $2^7-1=127$. Then 127 individual neural networks needed to be constructed as base classifier. Thus in process of selective ensemble, possible ensemble combination is $2^{127}-1$. Because the search space is too large and exhaustive searching is not applicable again, so GA (genetic algorithm) is used and the best ensemble accuracy is 97.91% (TABLE II), which is slightly better than that of our method, but it is not cost-efficient from the view of time cost and computational complex. While because large quantity of feature subspaces with lower classification performance t have been excluded by rough sets reduction, our method achieves satisfied performance at the cost of less time and lower computational complex. So it is an excellent ensemble feature selection approach. Limited by the space, the classification images are ignored.

V. CONCLUSION

Conventional ensemble based on feature selection have the defects of large search space, much time cost and high computational complex. In this paper, based on rough sets reducts, the accuracy of component network is guaranteed and most of feature subsets with lower performance are excluded. Thus time cost and computational complex reduce remarkably. Despite of that, the method still has deficiency. If dimension of input feature space is lower, then diverse reducts we can find is fewer. In this condition, the method is difficult to exert effectiveness.

REFERENCES

[1] L.K. Hansen, P. Salamon. “Neural network ensembles,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.12, no. 10, pp. 993-1001, 1990,

[2] Z.H. Zhou, S.F. Chen. “Neural network ensemble,” *Chinese Journal of Computers*, vol. 25, no. 1, pp. 1-8, 2002.

[3] A. Krogh, J. Vedelsby. “Neural network ensembles, cross validation, and active learning,” in Advances in Neural Information Processing Systems 7, G. Tesauro, D. Touretzky, T. Leen eds, Cambridge: MA: MIT Press, 1995, pp. 231-238.

[4] J. Hampshire, A. Waibel. “A novel objective function for improved phoneme recognition using time delay neural networks,” *IEEE Trans Neural Networks*, vol. 1, no. 2, pp. 216-228, 1990.

[5] K. J. Cherkauer. “Human expert level performance on a scientific image analysis task by a system using combined artificial neural networks,” in: Proc the 13th AAAI Work shop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms, Portland, OR, 1996, pp. 15-21.

[6] R. Maclin, J. W. Shavlik. “Combining the predictions of multiple classifiers: using competitive learning to initialize neural networks,” in: Proc the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995, pp. 524- 530.

[7] X. Yao, Y. Liu. “Making use of population information in evolutionary artificial neural networks,” *IEEE Trans Systems, Man and Cybernetics——Part B: Cybernetics*, vol. 28, no. 3, pp. 417-425, 1998.

[8] Z. H. Zhou, J. X. Wu, Y. Jiang, S. F. Chen. “Genetic algorithm based selective neural network ensemble,” in: P roc the 17th International Joint Conference on Artificial Intelligence, Seattle, vol. 2, WA , 2001, pp. 797- 802.

[9] S. Zeke, H. Chan, N. Kasabov. “Fast neural network ensemble learning via negative-correlation data correction,” *IEEE Transactions on Neural Networks*, vol. 16, no. 6, pp. 1707-1710, 2005.

[10] X. H. Fu, B.Q. Feng, Z. F. Ma, et al. “Method of incremental construction of heterogenous neural network ensemble with negeative correlation,” *Journal of Xi’an Jiangtong University*, vol. 38, no. 8, pp. 796 –799, 2004.

[11] K. Li, H. K. Huang. “A selective approach to neural network ensemble based on clustering technology,” *Journal of Computer Research & Devlopment*, vol. 42, no. 4, pp. 594-598, 2005.

[12] Q. Fu, S. X. Hu, S. Y. Zhao. “Clustering-based selective neural network ensemble,” *Journal of Zhejiang University*, vol. 6A, no. 5, pp. 387-392, 2005.

[13] Y. Freund, R. E. Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, 1997.

[14] L. Breiman. “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[15] T. K. Ho. “The random subspace method for constructing decision forests,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, 1998.

[16] D. Opitz. “Feature selection for ensembles,” in Proc. 16th National Conf. on Artificial Intelligence, AAAI Press, 1999, pp. 379-384.

[17] J. J. Ling, Z. Q. Chen, Z. H. Zhou. “Feature selection based neural network ensemble method,” *Journal of Fudan University (Natural Science)*, vol. 43, no. 5, pp. 685-688, 2004.

[18] P. Cunningham, J. Carney. “Diversity versus quality in classification ensembles based on feature selection,”, in 11th European Conf. On Machine Learning, R. L. DeMántaras, E. Plaza eds. Barcelona, Spain: , Springer, LNCS 1810 , 2000, pp. 109-116.

[19] A. Tsymbal, M. Pechenizkiy, P. Cunningham. *Diversity in ensemble feature selection*. Technical Report, Trinity. College Dublin, 2003, pp. 1-38.

[20] S. Vinterbo, A. Ohrn, “Minimal approximate hitting sets and rule templates,” *International Journal of Approximate Reasoning*, vol. 25, no. 2, pp. 123-143, 2000.

[21] J. Bazan, A. Skowron, P. Synak. “Dynamic reducts as a tool for extracting laws from decision tables,” in Proceedings of Symposium on Methodologies for Intelligent Systems, Charlotte, NC, Berlin: Springer-Verlag, LNAI 869, 1994, pp. 346-355.

[22] Y. Liu, X. Yao. “Simultaneous training of negatively correlated neural networks in an ensemble,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 29, np. 6, pp. 716-725, 1999.

[23] Z. H. Zhou, J. X. Wu, W. Tang. “Ensembling neural networks: many could be better than all,” *Artificial Intelligence*, vol. 137, no. 1-2, pp. 239-263, 2002.

[24] J. X. Wu, Z.H. Zhou, X. H. Shen, et al. “A selective constructing approach to neural network ensemble,” *Journal of Computer Research & Devlopment*, vol. 37, no. 9, pp. 1039-1044, 2000.