ENSEMBLE LEARNING METHODS FOR DECISION MAKING: STATUS AND FUTURE PROSPECTS

SHAHID ALI[1], **SREENIVAS SREMATH TIRUMALA**[2], **ABDOLHOSSEIN SARRAFZADEH**[1]

[1] Department of computing, UNITEC institute of Technology, Auckland, New Zealand
[2]School of Computer and Mathematical Sciences, AUT University, Auckland, New Zealand
E-MAIL: alis12@wairaka.com, ssremath@aut.ac.nz, hsarrafzadeh@unitec.ac.nz

Abstract:

In real world situations every model has some weaknesses and will make errors on training data. Given the fact that each model has certain limitations, the aim of ensemble learning is to supervise their strengths and weaknesses, leading to best possible decision in general. Ensemble based machine learning is a solution of minimizing risk in decision making. Bagging, boosting, stacked generalization and mixture of expert methods are the most popular techniques to construct ensemble systems. For the purpose of combining outputs of class labels, weighted majority voting, behaviour knowledge space and border count methods are used to construct independent classifiers and to achieve diversity among the classifiers which is important in ensemble learning. It was found that an ideal ensemble method should work on the principal of achieving six paramount characteristics of ensemble learning; accuracy, scalability, computational cost, usability, compactness and speed of classification. In addition, the ideal ensemble method would be able to handle large huge image size and long term historical data particularly of spatial and temporal. In this paper we reveal that ensemble models have obtained high acceptability in terms of accuracy than single models. Further, we present an analogy of various ensemble techniques, their applicability, measuring the solution diversity, challenges and proposed methods to overcome these challenges without diverting from the original concepts.

Keywords:

Support vector machine (SVM), ensemble, classifier, diversity.

1 Introduction

Machine Learning (ML) is the paradigm to be adopted for problem solving using self-learning procedures[1]. This principle is applicable to discrete problem solving method where a difficult task is divided into number of sub tasks. Further, a section of ML called meta machine learning is dedicated for identifying efficient methods to segregate a problem into number of sub problems that can be solved using any simple technique. In meta machine learning researchers have adopted the approach of using multiple learners to solve the problem and combined their decisions. This meta machine learning has been in focus since a decade has been named as ensemble or committee machine. There are different ways machines can learn that will make up an ensemble and there are various methods to combine the decisions of these learners.

An ensemble is a supervised learning algorithm which implies that it can be trained and deployed for predictions. The trained ensemble generates single hypothesis. However, this hypothesis is not essentially contained in the hypotheses space of the models from which it is built. Ensembles have shown more flexibility in representation [2]. This flexibility in theory results in over fitting the training data than a single model. But, the practical ensemble techniques tend to reduce over fitting of the training data [3].

The main objective of ensemble learning is to find a optimal performance through the combination of multiple classifiers. In ensemble learning a single or multiple algorithms are deployed to generate different base classifiers. These base classifiers are strategically combined together through a combination method decisions making in classifying new data instances. Since ensemble is a combination of multiple methods, assessing the prediction of an ensemble requires a lot of computation compare to that of a single model. So ensembles may be consider as a technique by working on poor learning algorithms by performing a lot of computations [4].

This paper is organized as follows: Section II introduces related work on ensemble based systems for critical decision making. Section III provides the methods for evaluating ensemble based decision making systems. Conclusion and direction for future work is presented in section IV.

2 Related Work

Perhaps the earliest work on ensemble systems was started in 1979 where researchers proposed to use ensemble systems in divide-and-conquer principle by dividing the feature space into two or more classifiers [5]. Over a decade later, variance reduction property ensemble system was proposed [6]. In this study it was concluded that generalization performance of a neural network can be enhanced by deploying an ensemble of similarly configured networks. In 1990, the ensemble systems are brought to the heart of machine learning research by providing the strong classifier in Probably Approximately Correct (PAC) by combination of weak classifiers through a method called boosting[7]. Method for separating points in multidimensional spaces was proposed with the help of stochastic processes named Stochastic Determination (SD) approaches [8]. The basic philosophy of this method is, that it takes poor solutions as an input and exhibits good solutions. Stochastic processes looked promising that led later on to random subspace method for constructing SVM based ensemble systems.

Stacked generalization method was proposed in 1992[9] with the philosophy of minimizing the generalization error rate of one or two generalizers. A general theoretical framework was proposed for ensemble methods construction In another study, results from multiple neural networks were combined through fuzzy logic, which displayed precise classification [10].

The performance of bagging and boosting methods were compared [11].This was found that bagging performs better than boosting in low noise system. Additionally, it was stated that bagging outperforms single classifier. Generally bagging method was considered to be more suitable for constructing ensembles. Robust driven ensemble approach for classification was invented for classification [12]. Bagging, boosting and random subspace methods were used. This was concluded that bagging is effective for weak and unstable classifiers and boosting is beneficial for weak and simple classifiers. Moreover, it was evident from the study, that random subspace method is helpful for weak and unstable classifiers, which has decreasing learning curve.

The relationship of various classifiers combination methods were investigated [13] and found that double fault measure of diversity and the measure of difficulty both showed correlation with majority of vote and Naive Bayes combinations, which was not expected. Constructing different classifiers through stacked generalization was investigated [7]. This study proved that stacked generalization performs better in choosing best classifier in ensemble compare to cross validation. Two extended versions of stacking: probability distributions and multi-response linear regression were proposed. The extended versions of stacked generalization performed well compared to existing stacking and cross validation methods versions.

Boosting margin in SVM based ensemble creation was investigated [13]. It was derived from the research that boosting the margin results in boosting classifier complication and maximizing the margins is attractive but not necessarily at the expense of other factors. A new approach to classifier ensemble design named combined fusion selection was proposed in which each classifier was substituted with mini-ensemble of a pair of sub-classifiers with a random linear form [14]. Till today all the ensemble learning methods benefited from this approach. A new local boosting algorithm for dealing with classification was proposed, based on boosting via resampling version of Adaboost [15]. The research results of this study were more accurate and robust than independent Adaboost approach.

The performance of four SVM ensemble constructing techniques namely Bagging, Adaboost, Arc-x4 and stacking was evaluated [13] studied. The results of this study demonstrated that bagging is considered to be an effective technique for various problems because of its better performance and higher generality. A new ensemble method was proposed based on manipulating the class labels [16]. This method generated different new class labels with the help of Cartesian product of the class attribute and built a component classifier. Extensive experiments and bias variance results in this showed that their method significantly reduce the bias of base learner, which is considered important factor in constructing ensemble based systems.

Least Square SVM (LS-SVM) and Proximal SVM (PSVM) was deployed for multiclass classification [17]. LS-SVM and PSVM were used for binary classification applications and cannot be applied directly to regression or multiclass applications. The authors unified and simplified the framework of LSVM and PSVM into extreme learning machine. Active Support vector machines are very popular in the field of relevance feedback [18]. They perform better when the size of training data is small, but it results in some unsatisfactory relevance results quite frequently. To overcome this problem, bagging algorithm was used to construct ensemble and outputs of classifiers were combined via majority of voting. The results of this study showed, that bagging algorithm is more effective method in constructing ensemble than the state of the art approaches.

Ensembles are well known methods and are applied across various research disciplines, which are obtained by combining highly accurate classifiers with less ones [19]. However, questions like, what is the best way of constructing ensemble? And issues like how best to understand the decisions made by ensembles? Are still unanswered [19]. When a new learning

problem appears, first instance the question arises, what is the best approach that needs to be applied for construction of an ensemble of classifiers? [20, 19]. In reality, there is no best ensemble method [20]. Similarly, there is not even a single best learning algorithm. However, some methods work principally better than others and some methods might do well than others in certain circumstances.

3 Evaluation of Methods for Constructing Ensemble based Decision Making

In this section four methods, i.e., bagging, boosting, stacked generalization and mixture of experts will be evaluated against six characteristics of ensemble learning for decision making, to help practitioners selecting the most suitable ensemble method for their specific research needs.

3.1 Predictive Performance - Accuracy

Predictive performance to considered to be main feature for selecting the algorithm [21]. Moreover, predictive performance measures accuracy, which can be used to benchmark algorithms. In this regard, bagging method is considered to be high in accuracy, because of its easy implementation and its functionality on limited data size. Boosting method has low accuracy, because of its suffering from over-fitting problems and its failure to understand complex composite classifier. Stacked generalization method has low accuracy as combining lower level models to higher level models is a complex task. Mixture of experts results in low accuracy considering the fact that assigning weights to the classifiers from the output of $Tier1$ classifier to $Tier2$ classifier is a complex task too.

3.2 Scalability

Scalability refers to the method ability to function on large datasets [22]. Bagging method has low scalability as it operates on limited data size. Boosting method operates on unlimited data size, hence having high scalability. Stacked generalization method operates on medium size of training data resulting in medium scalability. Whereas mixture of expert method functions on low data size, hence having low scalability.

3.3 Computational Cost

It is important to know about the efficiency of method,i.e., does it produce results in reasonable amount of time [23]. In terms of computational cost, bagging and booting methods are considered to highly efficient. Both methods obtain ensemble of classifiers efficiently through robust training of data, resulting in low computational cost. In stacked generalization method training of data requires reasonable amount of time and rectifying improper training by $Tier2$ is also time consuming, hence resulting in low efficiency and high computational cost. Mixture of experts method requires reasonable time for training data for classifiers, hence resulting in high computational cost too.

3.4 Usability

Machine learning is considered to be iterative process. To improve the performance of an ensemble system practitioners change parameters to generate better classifiers. Bagging and boosting methods are considered to be highly usable. Parameters of both these algorithms are flexible for generating better classifiers. Stacked generalization method has low usability, as once the weights to $Tier1$ classifier are assigned they are not flexible, resulting in low usability. Parameters of assigning weights to classifier in mixture of expert method are partially flexible to generate better classifiers, hence resulting in medium usability.

3.5 Compactness

Compactness can be measured by ensemble size and complexity of classifiers in ensemble methods [22]. In this regard, bagging method results are highly compact because it only works on limited training data size and results are easy to understand. Boosting method on the other side has low compactness, due to its functionality on unlimited data size, whereas boosting of decision trees could result in thousands (or millions) of nodes, hence difficult to visualize them. Both stacked generalization and mixture of expert methods has medium compactness, as they operate on low to medium size of training data.

3.6 Speed of Classification

Speed of classification indicates the ability of method to perform the classification in a certain time frame [24]. Bagging method results in robust classification because it operates on limited data size. Speed of classification for boosting method is moderate compare to bagging, because it operates on unlimited data size and fails to resolve over-fitting issue. Stacked generalization and mixture of expert methods are slow in classification because in these methods each classifier is trained on

separate training data and at the end output of these classifiers are ensemble to make decision. Table 1 provides the comparison of learning algorithms methods.

Attributes	Bagging	Boosting	Stacked Generalization	Mixture of Experts
Accuracy	High	Low	Low	Low
Scalability	Low	High	Medium	Low
Computational Cost	Low	Low	High	High
Usability	High	High	Low	Medium
Compactness	High	Medium	Low	Low
Speed of Classification	High	Medium	Low	Low

Table 1: Comparison of Methods

3.7 Importance of Diversity Measures in a Decision Making

Diversity is considered to be cornerstone of ensemble systems. Diversity plays important role in creating an ensemble, where each classifier is as different as possible and still considered to be consistent with the training set considered to be important feature for obtaining better ensemble performance. In this chapter we will briefly discuss diversity and various approaches towards its creation and measures. If a classifier makes a perfect generalization performance then there would not be a need to deploy ensemble techniques. Over lapping of data, noise and outliers makes it impossible for a classifier to propose [25]. The success of an ensemble system relies on its ability to correct the errors of its classifiers. The strategy in ensemble system is to create various classifiers and combining their outputs to improve the performance of a single classifier. If all the classifiers result in same output, then correcting the error will not be possible. Therefore, individual classifiers are required in ensemble system to make different errors on different instances [26]. The intuition is that if each classifier makes different errors, then combining these errors strategically will reduce total error. Therefore, the overall strategy in ensemble system is to make classifier as unique as possible. Classifiers are created whose decision boundaries are different from others, such classifiers are said to be diverse.

There are several quantitative measures for diversity assessment. The most common one is pair wise measures, addressed between two classifiers. For example, an ensemble of having n classifiers, then its total pairwise diversity measure can be calculated as, the mean pairwise measure of overall $n.(n-1)/2$ pairs of classifiers is represented from (1)[27].

$$F_{Total} = \frac{2}{n(n-1)} \sum_{\forall_i \neq j} f_{i,j} \tag{1}$$

Where $f_{i,j}$ is diversity or similarity measure of two classifiers outputs i and j. The following two pairwise measures are considered to be useful in diversity.

The disagreement is the probability that the two classifiers will disagree, whereas, double fault measure is considered to be the probability that both classifiers are not correct. Increase in the value of disagreement and double fault value increases diversity results. This rule makes the assumption that diversity is considered to be highest, when half of the classifiers are correct and half of them are incorrect. Based on this assumption entropy measure can be represented from (2)[28].

$$E = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T - [T/2]} min\{\zeta_i, (T - \zeta_i)\} \tag{2}$$

Where ξ_i is the number of classifier out of T, and N is the dataset cardinality [28]. It varies between 0 and 1, where 0 indicates all classifiers are same, and 1 represents weight diversity.

3.8 Correlation

Diversity is measured as a correlation between the output of two classifiers and can be represented from (3)[13].

$$\rho_{i,j} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, 0 \leq \rho \leq 1 \tag{3}$$

when $P = 0$, maximum diversity is achieved, stating classifiers are uncorrelated.

Q-Statistic can be represented from (4) [29].

$$Q_{i,j} = (ad - bc)/(ad + bc) \tag{4}$$

If the same instances are correctly classified by both classifiers, then Q assumes positive values, otherwise negative values. Maximum diversity is obtained when $Q = 0$. However, measuring diversity is not straightforward because there is no formal definition of diversity and research in 2002 raised some doubts about the measures of diversity usefulness in creating ensemble of classifiers [13].

There are various approaches to achieve diversity. The most popular approach to achieve diversity is by using different datasets to train individual classifiers [30]. The datasets are usually obtained through resampling techniques, such as bagging or boosting. In these techniques data subsets are selected randomly from the whole training data. Three classifiers are trained on random and resampling data subsets, resulting in formation of three different decision boundaries. At the end these boundaries are ensemble to obtain accurate classification.

Another comprehensive approach to achieve diversity is by deploying different classifiers [31]. The instability of classifiers can be controlled by changing such parameters, hence resulting in diversity. Similarly, changing the parameters allow the classifiers to be suitable candidates in ensemble setting.

4 Conclusion and Future Work

Ensemble learning is a procedure that is employed to train numerous learning machines and combining their outputs to obtain a better composite global model with more accurate and reliable decisions than can be accomplished through a single model. The ensemble methodology has been used across various disciplines to improve the predictive performance of single models such as: bioinformatics, medicine, finance, manufacturing, information security, information retrieval and image retrieval etc.

As explained in previous sections, though ensemble methods are popular and applied across various disciplines, questions like, what is the best method for constructing problem based ensemble? Is still unanswered. With the above analysis it may be concluded that a single method cannot be used to achieve all the characteristics which are required for optimal performance of ensemble learning for reducing the chances of making a poor decision. It can also be concluded that bagging algorithm achieved high accuracy compare to other methods because of its easy implementation and its functionality on limited data size. Bagging was considered to be highly accurate, usable, compact and robust for constructing ensembles. Boosting algorithm has high usability. In terms of computational cost stacked generalization and mixture of experts methods were considered highly expensive to run and to manage their data. However, the success of ensemble methods depend on other factors as well such as the choice of a base classifier, the procedure in which training set is modified, the selection of combination method and the ability of selected base classifier to solve the problem.

Similarly, questions like how to understand the decisions made by ensembles? Is still unanswered. In our paper we reviewed various methods for combining outputs of ensemble classifiers. After reviewing those methods, it can be concluded that these methods lack comprehensibility, i.e., the knowledge learned by ensemble method is not understandable to the user. Moreover, these combination methods have flexible parameters for training and non-trainable data, therefore, their results were not agreed upon in vast research community. Diversity creation in ensemble learning is very important as accurate results can be obtained only when the classifiers are more diverse. Lack of formal definition to diversity in ensemble learning posed one of the limitations to its functionality. The relation between ensemble diversity and ensemble performance is designed for regression problems; however, it is not yet formalized for classification problem. From our review for methods combining outputs of ensemble based classifiers, we can conclude that there is not even a single best method for creating an ensemble, even not for combining classifiers' outputs. Some methods work mainly better than others and few methods might do better than other in certain circumstances.

Further research is required to develop an ideal method for construction ensemble based should achieve accuracy, scalability, usability, flexibility and should be able to handle large huge image size and long term historical data particularly of spatial and temporal data of environmental analysis. Similarly, a method for combining outputs of classifiers should be developed as well, where the focus of that method should not only to achieve optimal performance but should also focus on providing more information of its insight in decision making.

References

[1] K. Li, Z. Liu, and Y. Han, "Study of selective ensemble learning methods based on support vector machine," *Physics Procedia*, vol. 33, no. 0, pp. 1518 – 1525, 2012.

[2] E. Stamatatos and G. Widmer, "Automatic identification of music performers with learning ensembles," *Artificial Intelligence*, vol. 165, no. 1, pp. 37 – 56, 2005.

[3] I. G. Webb and Z. Zheng, "Multistrategy ensemble learning: reducing error by combining ensemble learning techniques," *IEEE Transactions on Knowledge and Data Engineering,*, vol. 16, no. 8, pp. 980 – 991, 2004.

[4] F. Bellal, H. Elghazel, and A. Aussem, "A semi-supervised feature ranking method with ensemble learning," *Pattern Recognition Letters*, vol. 33, no. 10, 2012.

[5] G. Brown, "Ensemble learning," 2009.

[6] Y.-Z. Zhang, C.-M. Liu, L.-K. Zhu, and Q.-L. Hu, "Constructing multiple support vector machines ensemble based on fuzzy integral and rough reducts," in *Industrial Electronics and Applications*, 2007, pp. 1256 –1259.

[7] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, pp. 197–227, 1990.

[8] E. M. Kleinberg, "Stochastic discrimination," *Annals of Mathematics and Artificial Intelligence*, vol. 1, 1990.

[9] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241 – 259, 1992.

[10] S.-B. Cho and J. Kim, "Multiple network fusion using fuzzy logic," *Neural Networks, IEEE Transactions on*, vol. 6, no. 2, pp. 497 –501, 1995.

[11] R. Maclin and D. W. Opitz, "Popular ensemble methods: An empirical study," *CoRR*, vol. abs/1106.0257, 2011.

[12] D. Miller and L. Yan, "Critic-driven ensemble classification," *Signal Processing, IEEE Transactions on*, vol. 47, no. 10, pp. 2833 –2844, 1999.

[13] C. A. Shipp and L. I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Information Fusion*, vol. 3, no. 2, pp. 135 – 148, 2002.

[14] L. Kuncheva and J. Rodriguez, "Classifier ensembles with a random linear oracle," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 4, pp. 500–, 2007.

[15] C. X. Zhang and J. S. Zhang, "A local boosting algorithm for solving classification problems," *Computational Statistics and Data Analysis*, vol. 52, no. 4, pp. 1928–1941, 2008.

[16] Q. Wang and L. Zhang, "Ensemble learning based on multi-task class labels," in *Advances in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science, M. Zaki, J. Yu, B. Ravindran, and V. Pudi, Eds., 2010, vol. 6119, pp. 464–475.

[17] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 2, 2012.

[18] X.-Y. Wang, B.-B. Zhang, and H.-Y. Yang, "Active svm-based relevance feedback using multiple classifiers ensemble and features reweighting," *Engineering Applications of Artificial Intelligence*, 2012.

[19] G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," *Expert Systems with Applications*, vol. 38, no. 1, pp. 223 – 230, 2011.

[20] M. W. Craven and S. J. W, "Extracting tree-structured representations from trained networks," *Advances in Neural Information Processing*, vol. 8, 1996.

[21] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[22] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, pp. 1–39, 2010.

[23] P. Granitto, P. Verdes, and H. Ceccatto, "Neural network ensembles: evaluation of aggregation algorithms," *Artificial Intelligence*, vol. 163, no. 2, pp. 139 – 162, 2005.

[24] B. Pfahringer, G. Holmes, and R. Kirkby, "Optimizing the induction of alternating decision trees," in *Advances in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science, 2001, vol. 2035.

[25] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: a survey and categorisation," *Information Fusion*, vol. 6, no. 1, pp. 5–20, 2005.

[26] P. Melville and R. J. Mooney, "Creating diversity in ensembles using artificial data," *Information Fusion*, vol. 6, no. 1, pp. 99 – 111, 2005.

[27] R. Polikar, "Ensemble based systems in decision making," *Circuits and Systems Magazine, IEEE*, vol. 6, no. 3, pp. 21–45, 2006.

[28] P. Melville and R. J. Mooney, "Diverse ensembles for active learning," in *Proceedings of the twenty-first international conference on Machine learning*, ser. ICML '04. New York, NY, USA: ACM, 2004, pp. 74–.

[29] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. Kegelmeyer, "Ensemble diversity measures and their application to thinning," *Information Fusion*, vol. 6, no. 1, pp. 49 – 62, 2005.

[30] Z.-H. Zhou, "When semi-supervised learning meets ensemble learning," *Frontiers of Electrical and Electronic Engineering in China*, vol. 6, pp. 6–16, 2011.

[31] K. Sirlantzis, S. Hoque, and M. Fairhurst, "Diversity in multiple classifier ensembles based on binary feature quantisation with application to face recognition," *Applied Soft Computing*, vol. 8, no. 1, pp. 437–445, 2008.