

Ensemble Learning Based Methods For Classification: An Overview

¹Maitri Rathod, ²Fenil Patel, ³Pearl Naik, ⁴Radhika Javiya, ⁵Vivek Joshi

¹Ms., ²Mr., ³ Ms., ⁴Ms.,⁵Prof.

¹Computer Science And Engineering Dept,
¹R.N.G. Patel Institute of Technology, Bardoli, India

Abstract : *Housing is not a homogeneous commodity, but rather a multidimensional bundle of services. If proliferation of urban slums, shanties and squatters' settlements is to be drastically minimized or completely eliminated, it would be more beneficial for government and housing suppliers to provide such basic social amenities and important housing accessories.*

Ensemble approaches in classification is a very popular research area in recent years. An ensemble consists of a set of individually trained classifiers such as neural networks or decision trees whose predictions are combined for classifying new instances. It integrates multiple classifiers to build a classification model, and also used for improving the prediction performance.

Keywords— *Housing, Ensemble-methods, Classification, Boosting, Bagging*

I. INTRODUCTION

Housing has been conceptualized in various ways by different researchers and authors. It has been conceived as the structure, design and built-in-equipment (such as the amount of space, the heating, lighting, sanitary and similar facilities) environmental amenities such as water supply, sewage, solid waste disposal are all part of the package of services designated as housing. It has also been seen to include the layout and equipment such as the open spaces, streets, walkways, utilities, nursery and elementary schools, shops, etc.

Despite the various conceptualizations, housing is highly complex in nature and characteristics simply because it is a multidimensional bundle of services. Housing is not a homogeneous commodity, but rather a label for a collection that are all distinct to some degree. Conceptually, the rent that is paid on residential property is functionally related to three different housing traits like number of kitchens, rooms, baths, etc.

Classification is a data mining (machine learning) technique used to predict group membership for data instances. Combining classifiers is a very popular research area known under different names in the literature such as committees of learners, mixture of experts, classifier ensembles, multiple classifier systems, and consensus theory [3]. The basic idea is to use more than one classifier, hoping that the overall accuracy will be better. Ensemble performances depend on two main properties: the individual success of the base learners of the ensemble, and the independence of the base learners' results from each other[6].

It is possible to build diverse base learners by using same or different type base learners. When the same type base learners are used, the diversity is created by using different training data set for each base learner in the ensemble. There are several methods for creating different training data sets such as Bagging (BG), Boosting, Random Forests (RFs).

The existing ensemble methods create different training data sets by deleting/weighting samples. Based on this observation, adding new features (extended spaces) to the original data set is proposed.

The main idea of an ensemble methodology is to combine a set of models, each of which solves the same original task, in order to obtain a better composite global model, with more accurate and reliable estimates or decisions than can be obtained from using a single model.[12,3,10]

A. Performance

The development of ensemble learning methodologies and techniques has gained a significant attention from the scientific and industrial community. The basic idea behind these methods is the combination of a set of diverse prediction models for obtaining a composite global model which produces reliable and accurate estimates or predictions. Theoretical and experimental evidence proved that ensemble models provide considerably better prediction performance than single models [1].

B. Working

An ensemble consists of a set of models and a method to combine them. Assuming we have a set of models, generated by any of the learning algorithms called base predictors, for classification and regression problems. For learning a set of models given the knowledge that they will be combined, to make an ensemble learning model "how well an ensemble can perform?". Ensemble learning is a two-step process in which first features are extracted using pre-trained base predictor models and these features are predicted using classifiers. The ensemble of the networks yields better results as compared to single network [2].

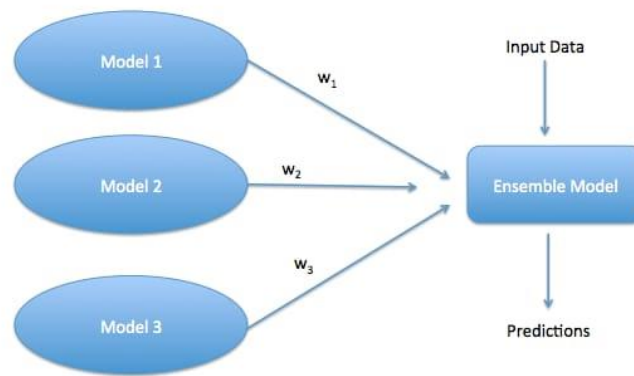


Fig. 1. Simple Ensemble Learning process

As shown in Figure 1, the base predictor models predict an output on their own based on the input data and sends the result to Ensemble model which then predicts the best possible output.

II. TRADITIONAL ENSEMBLE ALGORITHMS

This section describes the ensemble algorithms.

Bagging, was proposed by Breiman. Bagging creates a new training data set for each base learner by resampling the original training data set with replacement.

Decision Tree, is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems

Boosting, is an ensemble modeling technique which attempts to build a strong classifier from the number of weak classifiers.

Random Forest, was proposed by Breiman. It can be formulated as bagging plus the second form of random subspaces.

III. THEORETICAL FOUNDATIONS

3.1 Ensemble Classifiers

An ensemble consists of a set of individually trained classifiers (such as neural networks or decision trees) whose predictions are combined for classifying new instances.

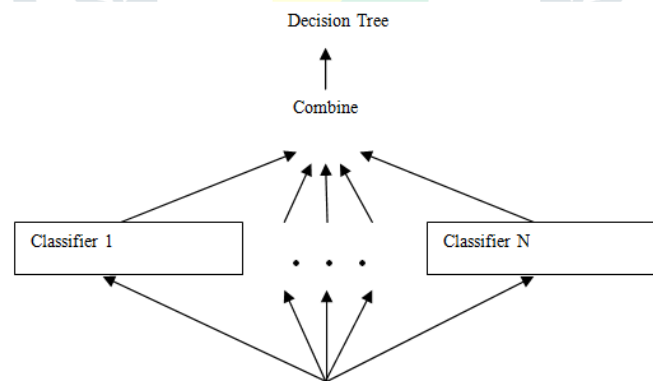


Fig 2. Ensemble of Decision Tree.

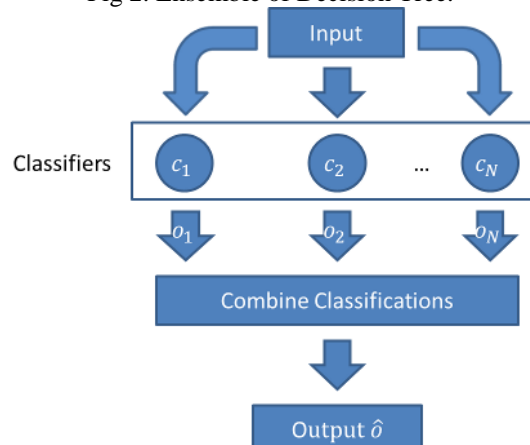


Fig 3. General concept of Ensemble Classifier.

In Fig 2 the input is taken from different classifiers such as C_1, C_2, \dots, C_N and it is combined together such as O_1, O_2, \dots, O_N to combine the classifications to form the output \hat{O} .

An ensemble Classifier is often more accurate than any of the single classifiers in the ensemble [8]. Bagging [7] and Boosting are two popular methods for producing ensembles. These methods use re-sampling techniques to obtain different training sets for each of the classifiers. Bagging stands for bootstrap aggregating which works on the concept of bootstrap samples. If original training dataset is of size N and m individual classifiers are to be generated as part of ensemble then m different training sets- each of size N , are generated from original dataset by sampling with replacement.

The multiple classifiers generated in bagging are independent to each other. In case of boosting, weights are assigned to each sample from the training dataset. If m classifiers are to be generated, they are generated sequentially such that one classifier is generated in a single iteration. For generating classifier C_i , weights of training samples are updated based on classification results of classifier C_{i-1} . The classifiers generated by boosting are dependent on each other.

The theoretical and empirical research related to ensemble has shown that an ideal ensemble consists of highly correct classifiers that disagree as much as possible. Opitz and Shavlik [11] empirically verified that such ensembles generalize well. Breiman [7] showed that bagging is effective on unstable learning algorithms. The four approaches for building ensembles of diverse classifiers are presented as:

- (1) Combination level: Design different combiners.
- (2) Classifier level: Use different base classifiers.
- (3) Feature level: Use different feature subsets.
- (4) Data level: Use different data subsets.

Random Forest is a classifier consisting of a collection of tree structured classifiers. It generates an ensemble of decision trees. To achieve diversity among base decision trees, Breiman selected the randomization approach which works well with bagging or random subspace methods [7], [6].

To generate each single tree in Random Forest Breiman followed following steps: If the number of records in the training set is N , then N records are sampled at random but with replacement, from the original data, this is bootstrap sample. This sample will be the training set for growing the tree. If there are M input variables, a number $m \ll M$ is selected such that at each node, m variables are selected at random out of M and the best split on these m attributes is used to split the node.

The value of m is held constant during forest growing. Each tree is grown to the largest extent possible. There is no pruning. In this way, multiple trees are induced in the forest; the number of trees is pre-decided by the parameter N tree. The number of variables (m) selected at each node is also referred to as m_{try} or k in the literature. The depth of the tree can be controlled by a parameter node size (i.e. number of instances in the leaf node) which is usually set to one [7]. Once the forest is trained or built as explained above, to classify a new instance, it is run across all the trees grown in the forest.

Each tree gives classification for the new instance which is recorded as a vote. The votes from all trees are combined and the class for which maximum votes are counted (majority voting) is declared as classification of the new instance. This process is referred to as Forest RI in the literature [8].

It is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems

It is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems

IV. EXISTING SURVEYS ON ENSEMBLE OF CLASSIFIERS

Given the potential usefulness of ensemble methods, it is not surprising that a vast number of methods are now available to researchers and practitioners. The variety of ensemble techniques have arisen several taxonomies in the literature which aim to categorize ensemble methods from the algorithm designer point of view. Sharkey [13] proposed taxonomy for ensemble of neural networks. This taxonomy suggests three dimensions:

- (1) The first dimension indicates if the ensemble's members are competitive or cooperative. In the competitive mode, a single member is selected to provide the classification. In cooperative mode the classifications of all members are combined.
- (2) The second dimension indicates if the ensemble is created top-down or bottom-up. In top-down mode the combination mechanism is based on something other than the classifiers outputs. Bottom-up techniques take the outputs of the members into account in their combination. Bottom up methods are subdivided into fixed methods (such as voting), and dynamic methods (such as stacking).
- (3) The third dimension indicates if we combine ensemble, modular, or hybrid components; it is distinguish between modular systems and pure ensemble systems. The main idea of pure ensemble systems is to combine a set of classifiers, each of which solves the same original task and to obtain a more accurate and reliable performance than using a single classifier.

On the other hand, the purpose of modular systems is to break down a complex problem into several manageable problems.

Although several surveys on ensemble for classification tasks are available in the literature [5] and there are several papers which suggest taxonomy for ensemble methods [6], in this paper the four main contributions are introduced:

- 1) A new unified taxonomy is suggested to categorize all significant ensemble methods developed in the field. As indicated in [5]: "Still there is no agreed upon structure or taxonomy of the whole field, although a structure is slowly crystallizing among the numerous attempts." On the one hand because existing taxonomies usually concentrate on some aspects (for example [6] concentrates on diversity), the new proposed taxonomy tries to organize existing taxonomies into a coherent and unified taxonomy.
- 2) Due to the fact that ensemble learning is an active research field, this paper proposes an updated survey which refers to new researches from the last three years that have not been previously covered by existing surveys.
- 3) This paper covers efficient and mature ensemble methods that do not belong to the mainstream, and therefore are usually not mentioned in existing surveys.
- 4) It also proposes several selection criteria, presented from the practitioner's point of view, for choosing the most suitable ensemble method.

V. TAXANOMY OF ENSEMBLE CLASSIFIER

A typical ensemble method for classification tasks contains the following building blocks:

Training set

A labeled dataset used for training the ensemble. Most frequently the training set is a collection of instances (also known as samples or observations). Each instance is described by attribute-value vectors. The input space is spanned by the attributes used to describe the instances. In semi-supervised methods of ensemble generation, such as ASSEMBLE [9], unlabeled instances can be also used for the creation of the ensemble.

Inducer

The inducer is an induction algorithm that obtains a training set and forms a classifier that represents the generalized relationship between the input attributes and the target attribute.

Ensemble generator

This component is responsible for generating the diverse classifiers.

Combiner

The combiner is responsible for combining the classifications of the various classifiers.

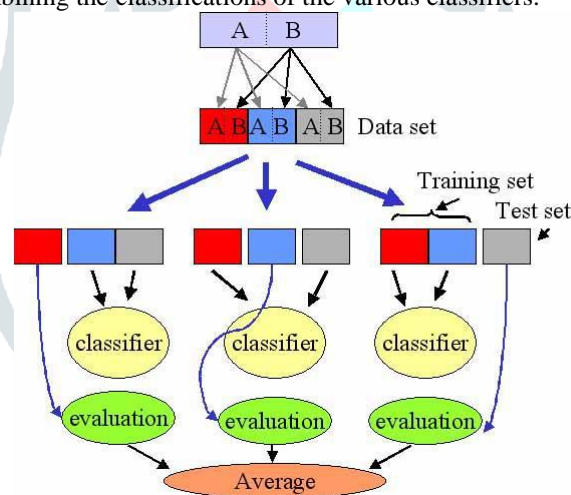


Fig 3: Taxonomy of classifiers

VI. ENSEMBLE DIVERSITY

In an ensemble, the combination of the output of several classifiers is only useful if some of the inputs are different [9]. Creating an ensemble in which each classifier is as different as possible while still being consistent with the training set is theoretically known to be an important feature for obtaining improved ensemble performance. A diversified classifiers lead to uncorrelated errors, which in turn improve classification accuracy.

In the regression context, the bias-variance-covariance decomposition has been suggested to explain why and how diversity between individual models contributes toward overall ensemble accuracy. In the classification context, there is no complete and agreed upon theory [6]. More specifically, there is no simple analogue of variance-covariance decomposition for the zero- one loss function.

Instead, there are several ways to define this decomposition. Each way has its own assumptions. Sharkey [13] suggested taxonomy of methods for creating diversity in ensembles of neural networks. More specifically, Sharkey's taxonomy refers to four different aspects: the initial weights, the training data used, the architecture of the networks, and the training algorithm used.

VII. CONCLUSION

This paper presents an updated taxonomy survey of ensemble methods for classification problems. Ensemble algorithms are mostly used in machine learning because of its well versed performances than other single algorithms. A description of five

dimensions of ensemble taxonomy and traditional algorithms are also included in this paper. The ensemble algorithms using all the features have more accurate base learners.

REFERENCES

- [1] [1] P. E. Pintelas and Ioannis Livieris, Special issue on Ensemble Learning and its applications, Department of Mathematics, University of Patras, GR Patras, Greece, 10 June 2020. Available : <https://www.researchgate.net/publication/342083251>
- [2] Aarti Kumthekar, Ramachandra Reddy G., Ensemble Learning Techniques for Cloud Classification, International Journal of Engineering and Advanced Technology (IJEAT), Available: <https://www.ijeat.org/wp-content/uploads/papers/v9i2/B3957129219.pdf>
- [3] Bauer, E. and Kohavi, R., "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants". Machine Learning, 35: 1-38, 1999.
- [4] C.E. Brodley, Recursive automatic bias selection for classifier construction, Machine Learning 20 (1995) 63-94.
- [5] Dietterich T., Ensemble methods in machine learning. In J. Kittler and F. Roll, editors, First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science, pages 1-15. Springer-Verlag, 2000.
- [6] G. Brown, J.L. Wyatt, and P. Tino, "Managing Diversity in Regression Ensembles," The J. Machine Learning Research, vol. 6, pp. 1621-1650, 2005.
- [7] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, pp. 123-140, 1996.
- [8] L.I. Kuncheva and C.J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their relationship with the Ensemble Accuracy," Machine Learning, vol. 51, no. 2, pp. 181-207, 2003.
- [9] Opitz, D., Feature Selection for Ensembles, In: Proc. 16th National Conf. on Artificial Intelligence, AAAI, 1999, pp. 379-384.
- [10] Opitz, D. and Maclin, R., Popular Ensemble Methods: An Empirical Study, Journal of Artificial Research, 11: 169-198, 1999
- [11] Opitz D. and Shavlik J., Generating accurate and diverse members of a Neural network ensemble. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, Advances in Neural Information Processing Systems, volume 8.
- [12] Quinlan, J. R., Bagging, Boosting, and C4.5. In Proceedings of the Thirteenth National Conference on Artificial Intelligence, pages 725-730, 1996.
- [13] Sharkey, A., Multi-Net systems, In Sharkey A. (Ed.) Combining Artificial Neural Networks: Ensemble and Modular Multi-Net Systems. pp. 1-30, Springer-Verlag, 1999.

