

# A Survey on One Class Classification using Ensembles Method

**Jay Bhatt**

*PG Student*

*Department of Computer Engineering  
Kalol Institute of Technology & Research Centre*

**Nikita S Patel**

*Asst. Professor*

*Department of Computer Engineering  
Kalol Institute of Technology & Research Centre*

---

## Abstract

In Data mining Classification is a data mining function that allocated similar data to categories or classes. One of the most common methods for classification is ensemble method which refers supervised learning. After generating classification rules we can apply those rules on unknown data and reach to the results. In one-class classification it is assumed that only information of one of the classes, the target class, is available. This means that just example objects of the target class can be used and that no information about the other class of outlier objects is present. In One Class Classification (occ) problem the negative class is either absent or improperly sampled. There are several classification mechanisms that can be used. In an ensemble classification system, different base classifiers are combined in order to obtain a classifier with higher performance. The most widely used ensemble learning algorithms are AdaBoost and Bagging. The process of ensemble learning method can be divided into three phases: the generation phase, in which a set of candidate models is induced, the pruning phase, to select of a subset of those models and the integration phase, in which the output of the models is combined to generate a prediction.

**Keywords:** Bagging, Boosting, Classification, Ensembles, One Class Classification, Positive and Unlabeled Data.

---

## I. INTRODUCTION

Data classification is the categorization of data for its most effective and efficient use. Data classification is a two-step process: 1.Learning (Model Construction) 2.Classification (Model Usage). In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or “learning from” a training set made up of database tuples and their associated class labels. In the second step, the model is used for classification. A test set is used, made up of test tuples and their associated class labels. These tuples are randomly selected from the general data set. They are independent of the training tuples, meaning that they are not used to construct the classifier. The One Class Classification problem is different from the multi-class classification problem in the sense that in one-class classification it is assumed that only information of one of the classes, the target class, is available. This means that just example objects of the target class can be used and that no information about the other class of outlier objects is present because these data are either difficult or impossible to collect. In practice, it happens quite frequently that the normal state has a good representation, however the abnormal states are rare and the abnormal class is ill-defined, in such a case we have to judge on the abnormality using information from the normal class only. The problem is called ‘one class classification’ (OCC). OCC can be seen as a special type of two-class classification problem, when data from only one class is considered Thus, objects in the target class can be considered as typical, while objects in the negative class can be considered as atypical. For one-class classification several models have been proposed. Most often the methods focus on outlier detection. Conceptually the most simple solution for outlier detection is to generate outlier data around the target set. Ensembles Method has been successfully applied to solve a variety of classification and function approximation problems. Ensembles Method Use a combination of models to increase accuracy problems. It Combine a series of  $k$  learned models  $M_1, M_2, \dots, M_k$  with the aim of creating an improved model  $M^*$ . There are three Popular ensemble methods, Bagging: averaging the prediction over a collection of classifiers, Boosting: weighted vote with a collection of classifiers, Stacking: combining a set of heterogeneous classifiers.

## II. INTRODUCTION TO ONE CLASS CLASSIFICATION IN CLASSIFICATION

In this section, we first introduce the one class classification [7]. Then, we present how to evaluate the performance of the classifier in one class classification. Finally, we recall several techniques to address the one class classification. One class classification is a binary classification task for which only one class of samples is available for learning. The Learning from the available target samples only means that the classifier does not require any hypothesis on the outlier data to estimate the decision boundary. A taxonomy with three broad categories for the study of OCC problems.

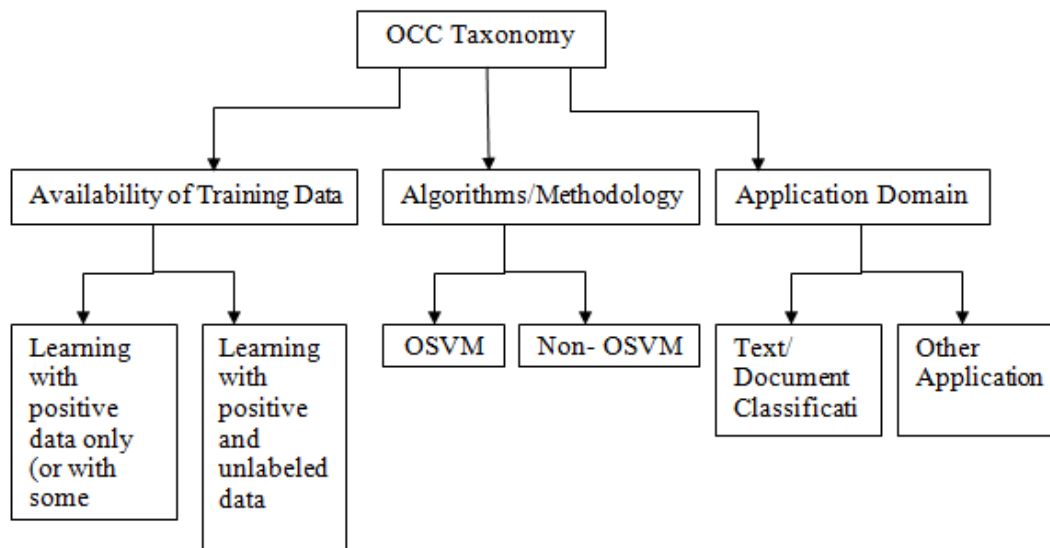


Fig. 1: Proposed Technique for OCC [7]

A Proposed technique for OCC is shown in Figure 1. It represents the three broad categories for the study of OCC problems, Availability of Training Data: Learning with positive data only (or with a limited amount of negative samples) or learning with positive and unlabeled data, Methodology Used: Algorithms based on One Class Support Vector Machines (OSVMs) or methodologies based on algorithms other than OSVMs, Application Domain Applied: OCC applied in the field of text/document classification or in other application domains. In figure 2 an example of a training dataset is given for the apple-pear problem. Each object has two feature values (for instance the width and the height of the object; the exact features are not important for this discussion). Each training object  $x$  can therefore be represented as a point in a 2-dimensional feature space. Here the apples are indicated by stars, the pears by pluses. In principle, objects can be scattered all around the (2-dimensional) feature space, but due to the continuity assumption, apples are near apples and pears near pears. Furthermore, there are physical constraints on the measurement values (weights and sizes are positive, and are bounded by some large number) [3].

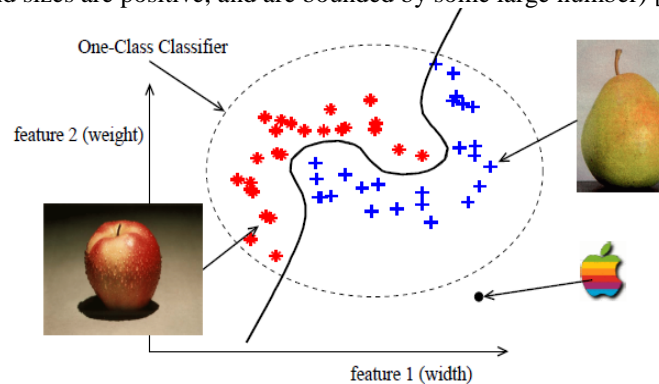


Fig. 2: Conventional and A One-Class Classifier [3]

In the apple-pear example the two classes can be separated without errors by the solid line in figure 2. Unfortunately, when the outlier apple in the right lower corner is introduced, it cannot be distinguished from the pears. To identify the outlier, a one-class classifier should be trained. An example of a one-class classification is given by the dashed line.

#### A. Availability of Training Data:

OCC problems have been considered widely under three broad frameworks:

- Learning with positive class only.
- Learning with positive class and some quantity of weakly distributed negative class.
- Learning with positive and unlabeled data.

The last type has received greatly research interest among the text classification [7]. The main plan behind these strategies is to build a decision boundary around the positive data so as to differentiate the outliers from the positive data.

### III. OCC VS. MULTI-CLASS CLASSIFICATION

In a usual multi-class classification problem, data from two or more classes are accessible and the decision boundary is supported by the existence of example samples from all class. Different researchers have used other terms to define one class classification such as Outlier Detection [7], Novelty Detection or Concept Learning. As defined before, in OCC tasks, the negative class is

either missing or limited in its sharing, so only one side of the classification boundary can be made definitively by using the data. This makes problem of one-class classification harder than the problem of usual multi-class classification. The task in OCC is to describe a classification boundary about the positive or target class, such that it accepts as many objects as possible from the positive class, while it minimize the possibility of accepting non-positive or outlier objects. As only one side of the boundary can be described, in OCC, it is tough to make a decision, on the base of just one class how closely the boundary should fit in each of the information around the data. It is also harder to make a decision which attributes should be used to discover the best division of the positive and negative class objects. that's why it is to be accepted that occ algorithms will require a huge number training instances comparative to usual multi-class classification algorithms.

#### IV. STATE OF THE ART ON ENSEMBLES TECHNIQUES

Data classification plays important role in the field of data mining. The increasing rate of data diversity and size decrease the performance and efficiency of classifier. The decreasing performance of classifier compromised with unvoted data of classifier. Now the merging of two or more classifier for better prediction and voting of data are used, such techniques are called Ensemble classifier. Now the merging of two or more classifier for better prediction and voting of data are used, such techniques are called Ensemble classifier. Good ensemble methods are that in which each individual classifiers are accurate and diverse But ensemble methods are combination of predictions made by a set of individual classifiers. Accurate classifier is meant to be produce accurate prediction than the random classifier and diverse classifier is meant to be produce prediction independently. Ensembles of classifiers, where a variety of classifiers are pooled before a final classification decision is made. Ensemble learning consists on the solution of two problems: (1) how to generate the ensemble of models? (Ensemble generation); and (2) how to integrate the predictions of the models from the ensemble in order to obtain the final ensemble prediction? (Ensemble integration). Ensemble pruning has been reported, at least in some cases, to reduce the size of the ensembles obtained without degrading the accuracy. Pruning has also been added to direct methods successfully increasing the accuracy.

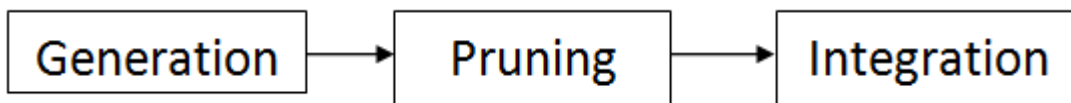


Fig. 3: Ensemble Learning Model

Ensembles are sets of learning machines that combine in some way their decisions, or their learning algorithms, or different views of data, or other specific characteristics to obtain more reliable and more accurate predictions in supervised and unsupervised learning problems. Ensembles method Use a combination of models to increase accuracy. The basic Ensemble Method model is as shown in figure 4.

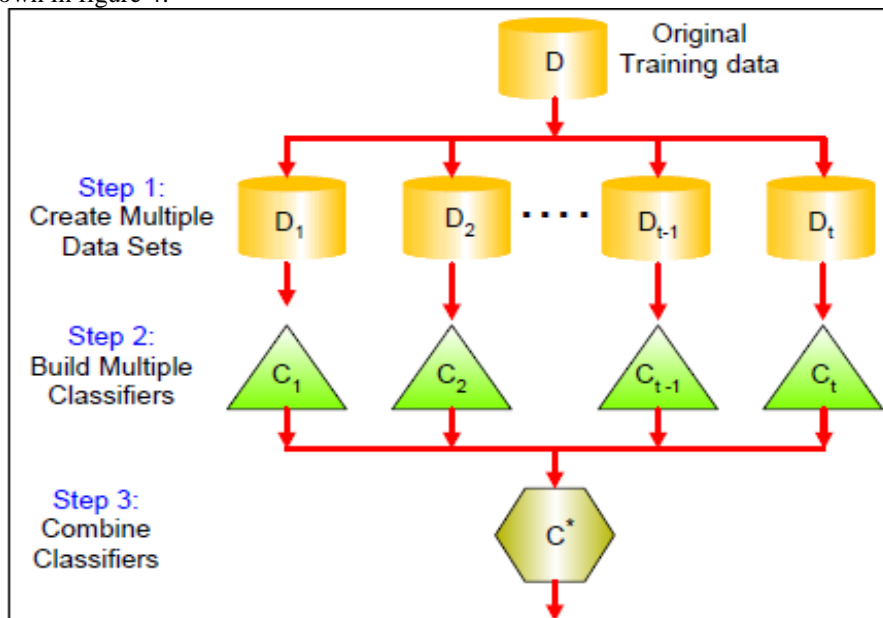


Fig. 4: Ensemble Method

The Popular ensemble methods are:

- (1) **Bagging:** Each member of the ensemble is generated by a different data-set. It is good for unstable models. Where small differences in the input data-set yield big differences in output. Many approaches have been developed using bagging ensembles to deal with class imbalance problems due to its simplicity and good generalization ability. The hybridization of bagging and data pre-processing techniques is usually simpler than their integration in boosting. A bagging algorithm does not require to recompute any kind of weights therefore, neither is necessary to adapt the weight update formula nor to change computations in the algorithm.

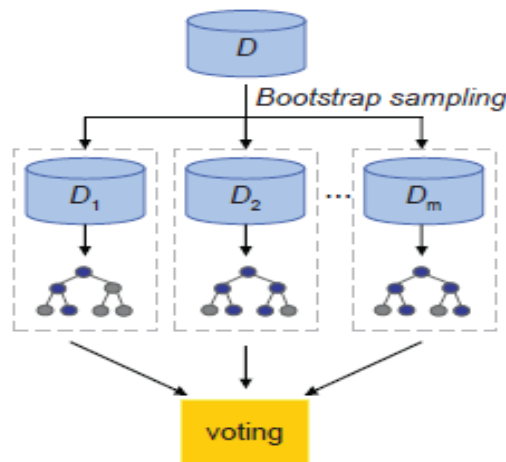


Fig. 5: Bagging [9]

- (2) Boosting: It is a family of ensemble learners. Its Basic idea is Weight the individual instances of the data-set. It iteratively learns models and records their errors and Distribute the effort of the next round on the miss-classified examples. The quantity of focus is measured by a weight, which initially is equal for all instances. After each iteration, the weights of misclassified instances are increased; on the contrary, the weights of correctly classified instances are decreased.

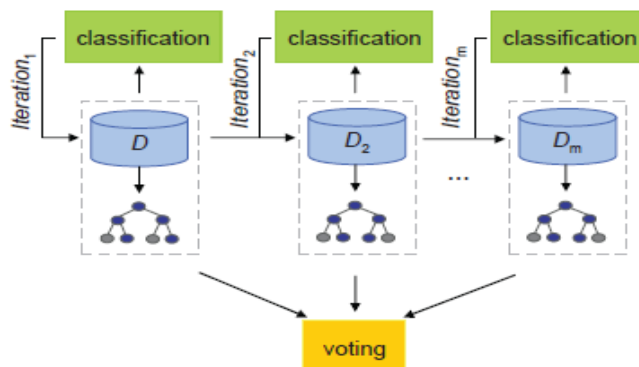


Fig. 6: Boosting<sup>[9]</sup>

- (3) Stacking: Its Basic idea is to have the output of a layer of classifiers as input to another layer. Stacking (sometimes called stacked generalization) involves training a learning algorithm to combine the predictions of several other learning algorithms. First, all of the other algorithms are trained using the available data, then a combiner algorithm is trained to make a final prediction using all the predictions of the other algorithms as additional inputs.

#### A. Learning Ensembles of Classifiers: Description and Representative Techniques

The main purpose of ensemble methodology is to try to increase the performance of single classifiers by inducing several classifiers and combining them to gain a new classifier. so, the basic idea is to build several classifiers from the original data and then summative their predictions when unknown instances are presented. This plan follows the human natural activity that tends to get several opinions before building any significant decision. Ensemble based classifiers generally refer to the mixture of classifiers that are negligible variants of the same base classifier, which can be considered in the broader concept of multiple classifier systems.

### V. CONCLUSION AND FUTURE WORK

In this paper, the goal of One Class Classification is to bring classifiers when only one class the positive class is well categorized by the training data. This survey provides a broad insight into the study of the discipline of OCC. Depending upon the data availability, algorithm use and application, appropriate OCC techniques can be applied and improved upon. It would be fruitful to investigate some more innovative forms of kernel, that have shown greater potential in standard SVM classification. The OCC field is becoming mature, still there are several fundamental problems that are open for research, not only in describing and training classifiers, but also in scaling, controlling errors, handling outliers, using non-representative sets of negative examples, combining classifiers and reducing dimensionality. Another point to note here is that in OSVMs, the kernels that have been used mostly are Linear, Polynomial, and Gaussian. This paper provide that ensemble-based algorithms are worthwhile, improving the results that are obtained by the usage of data pre-processing techniques and training a single classifier. The use of more classifiers makes them more complex, but this growth is justified by the better results that can be assessed. Also In this paper, the

state of the art on ensemble methodologies. Furthermore, we have exposed the positive synergy between sampling techniques and Bagging ensemble learning algorithm.

## ACKNOWLEDGMENT

The authors would like to thank the reviewers for their precious comments and suggestions that contributed to the expansion of this work.

## REFERENCES

- [1] Han Jiawei and Kamber Micheline, *Data Mining: Concepts and Techniques*, second edition, pp. 285-296.
- [2] Han Jiawei, Department of computer science, University of Illinois at Urbana-Champaign, 2006.
- [3] David Martinus Johannes TAX, *One class classification Concept-learning in the absence of counter-examples*, Proefschrift.
- [4] S. B. Kotsiantis, "Supervised Machine Learning- A Review of Classification Techniques," *Informatica* 31 (2007) 249-26.
- [5] Mikel Galar, Alberto Fern'andez, Edurne Barrenechea, Humberto Bustince, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews*, Vol. 42, No. 4, July 2012.
- [6] Amir Ahmad and Gavin Brown, "Random Projection Random Discretization Ensembles—Ensembles of Linear Multivariate Decision Trees," *IEEE Transactions On Knowledge And Data Engineering*, Vol. 26, No. 5, May 2014 5, May 2014.
- [7] Shehroz S. Khan and Michael G. Madden, "A Survey of Recent Trends in One Class Classification," National University of Ireland Galway, Ireland.
- [8] Youngmi Yoon, Sangjay Bien, and Sanghyun Park, "Microarray Data Classifier Consisting of k-Top-Scoring Rank-Comparison Decision Rules With a Variable Number of Genes," *IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews*, Vol. 40, No. 2, March 2010.
- [9] Pengyi Yang, Yee Hwa Yang, Bing B. Zhou and Albert Y. Zomaya, "A review of ensemble methods in bioinformatics," University of Sydney, NSW 2006, Australia.
- [10] Rashedur M. Rahman, Farhana Afroz, "Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis," *Journal of Software Engineering and Applications*, 2013, 6, 85-97.
- [11] Arthur Zimek, Fabian Buchwald, Eibe Frank, and Stefan Kramer, "A Study of Hierarchical and Flat Classification of Proteins," *IEEE/ACM Transactions On Computational Biology And Bioinformatics*, Vol. 7, No. 3, July-September 2010.
- [12] D.Gopika1, B.Azhagusundari, "An Analysis on Ensemble Methods In Classification Tasks," *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 3, Issue 7, July 2014.
- [13] Huimin Zhao, Member, IEEE, and Sudha Ram, Member, IEEE, "Constrained Cascade Generalization of Decision Trees," *IEEE Transactions On Knowledge And Data Engineering*, Vol. 16, No. 6, June 2004.
- [14] Shuo Wang, Student Member, IEEE, and Xin Yao, Fellow, IEEE, "Relationships Between Diversity of Classification Ensembles and Single-Class Performance Measures," *IEEE Transactions On Knowledge And Data Engineering*.
- [15] Sarwesh Site, Dr. Sadhna K. Mishra, "A Review of Ensemble Technique for Improving Majority Voting for Classifier," *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 1, January 2013.
- [16] Lior Rokach, "Taxonomy for Characterizing Ensemble Methods in Classification Tasks: a review and annotated bibliography," Ben-Gurion University of the Negev.
- [17] Joao M. Moreira, Carlos Soares, Alpio M. Jorge and Jorge Freire de Sousa, "Ensemble Approaches for Regression: a Survey," *LIAAD, INESC Porto L.A., R. De Ceuta*, 118, 6, 4050-190, Porto PORTUGAL.
- [18] Matteo Re, *Ensemble methods: a review*.
- [19] Thomas G Dietterich, "Ensemble Methods in Machine Learning," Oregon State University Corvallis Oregon USA.
- [20] Ms. Aparna Raj, Mrs. Bincy G, Mrs. T.Mathu, "Survey on Common Data Mining Classification Techniques," *International Journal Of Wisdom Based Computing*, Vol. 2(1), April 2012.
- [21] Tulips Angel Thankachan, Dr. Kumudha Raimond, "A Survey on Classification and Rule Extraction Techniques for Datamining," *IOSR Journal of Computer Engineering (IOSR-JCE)*, Volume 8, Issue 5 (Jan. - Feb. 2013), PP 75-78.
- [22] Chesner Desir, Simon Bernard, Caroline Petitjean, Heutte Laurent, *One class random forests*, HAL Id: hal-00862706.
- [23] Rashedur M. Rahman, Farhana Afroz, "Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis," *Journal of Software Engineering and Applications*, 2013, 6, 85-97.