Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.250 – 263

RESEARCH ARTICLE

A Research on Ensembles Method for One Class Classification Using Convex Hull Polytope Model

Jay Bhatt[1], Chaita Jani[2]

[1]M.E in C.E, Kalol Institute of Technology & Research Centre, India

[2]Asst. Prof. in Kalol Institute of Technology & Research Centre, India

[1] bhattjay1991@gmail.com; [2] jani.chaita@gmail.com

Abstract— Classification is a data mining task that allocated similar data to categories or classes. One of the most general methods for classification is ensemble method which refers supervised learning. After generating classification rules we can apply those rules on unidentified data and achieve the results. In one-class classification it is supposed that only information of one of the classes, the target class, is available. In an ensemble classification system, different base classifiers are combined in order to obtain a classifier with higher performance. In this work, a new one-class classification ensemble strategy called Approximate Polytope Ensemble is presented. The geometrical theory of convex hull is used to define the boundary of the target class defining the problem.

Keywords— Bagging, Boosting, Classification, Ensembles, One Class Classification, convex hull, polytope

I. INTRODUCTION

The One Class Classification problem is diverse 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. This problem is called 'one class classification' (OCC). For OCC several models have been planned. Most often the methods focus on outlier detection. Ensembles Method has been successfully useful to solve a variety of classification and function approximation problems. Ensembles Method Use a mixture of models to raise accuracy of problems. It merge a series of n learned models M1, M2,…, Mn with the aim of creating an better 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. Here, the geometrical concept of convex hull is used to define the boundary of the target class defining the problem. Expansions and contractions of this geometrical structure are introduced in order to avoid over fitting. Then, the decision whether a point belongs to the convex hull model in high dimensional spaces. Finally, a tiling strategy is proposed in order to model non-convex structures.

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.

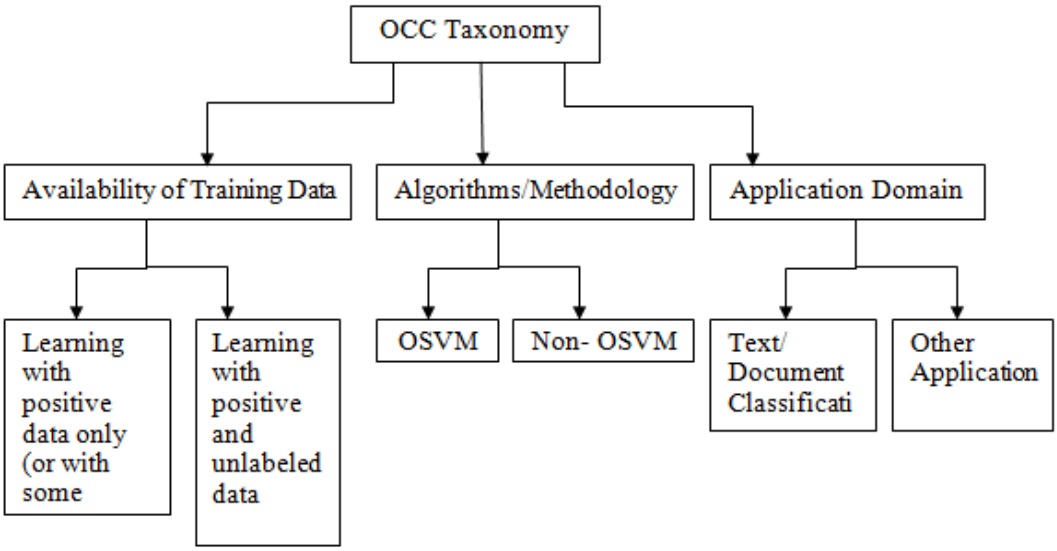

Figure 1: Proposed technique for OCC [7]

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 unlabelled 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 objects 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].

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.

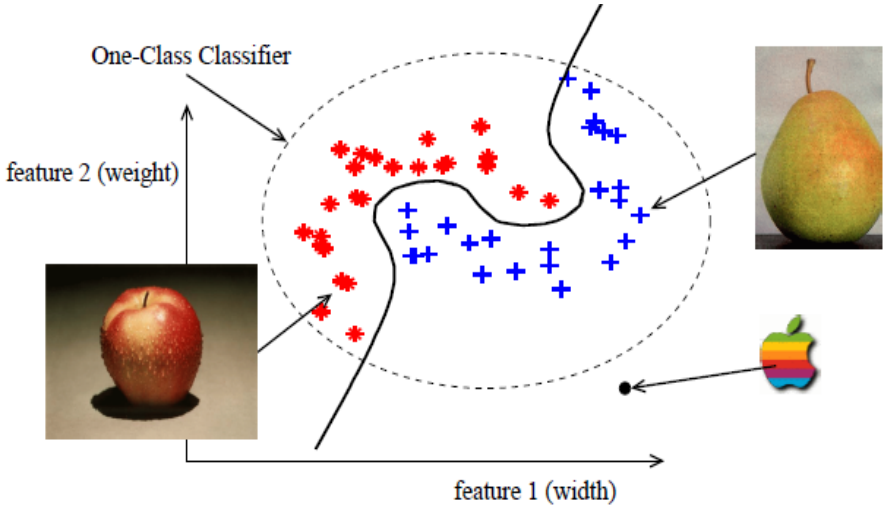

Figure 2 Conventional and a one-class classifier [3]

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 unlabelled 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.

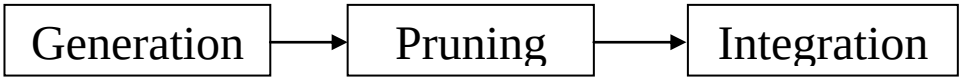

Figure 3 Ensemble learning model

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.

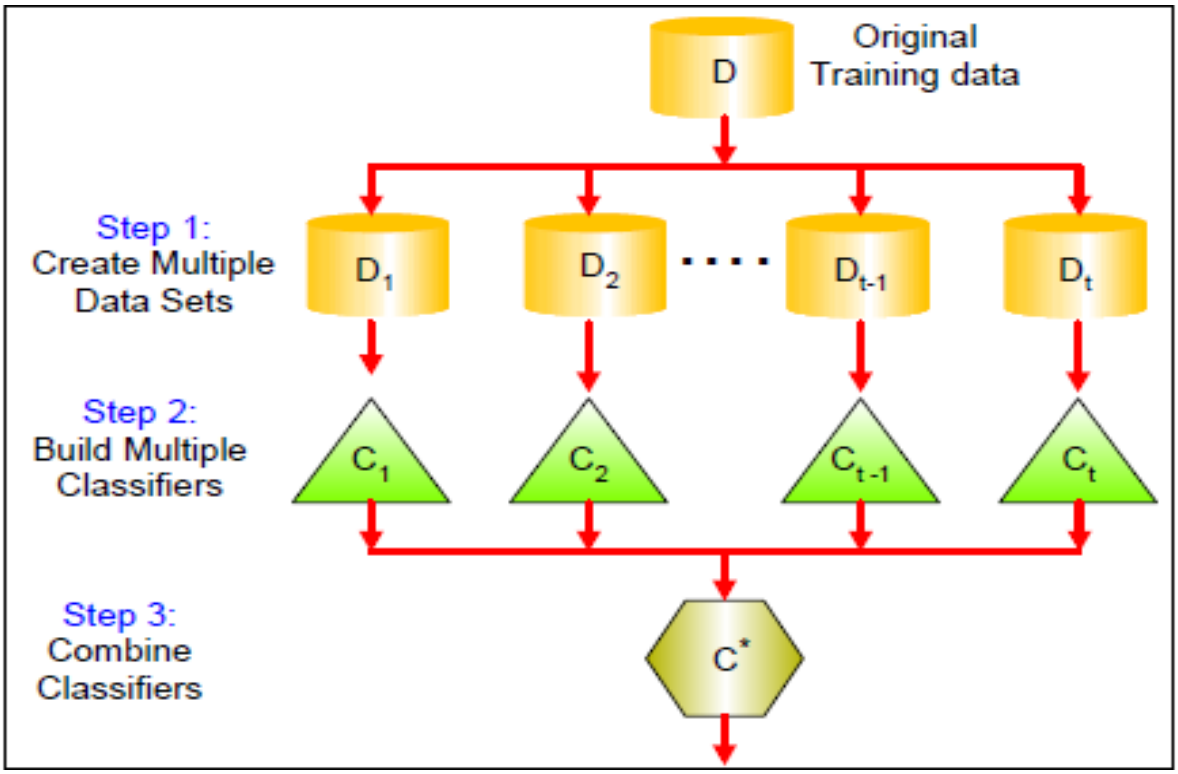

Figure 4 Ensemble Method

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 recomputing any kind of weights therefore, neither is necessary to adapt the weight update formula nor to change computations in the algorithm.

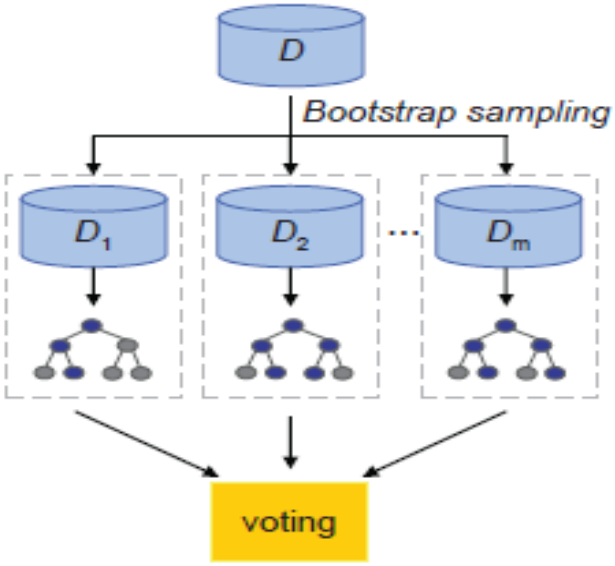

Figure 5 Bagging [9]

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.

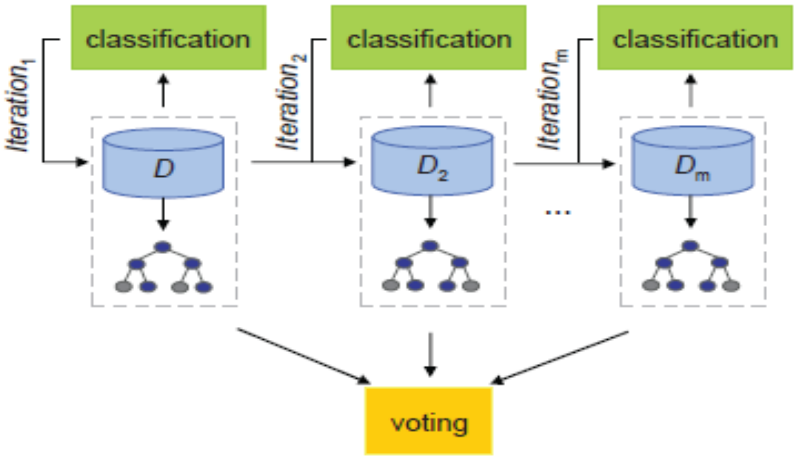

Figure 6 Boosting [9]

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. PROPOSED WORK

In this work, three main contributions are proposed in the context of One-Class classification: 1) The geometric structure of the convex hull is proposed for modelling the boundary of the one-class classification problem. Shrunken or enlarged versions of the baseline convex hull of the training data are used to avoid over-fitting and to find the optimal operating point of the classifier. These versions are called extended convex polytopes and their growth is governed by a parameter n. 2) This limitation is circumvented by approximating the D-dimensional expanded convex polytope decision by an ensemble of decisions in very low-dimensional spaces like d << D.

This new geometric structure is called approximate convex polytope decision ensemble. As a result, a very efficient and powerful one-class classifier is obtained. 3) However, many real world problems are not well modelled using a convex polytope. Thus, an ensemble of convex polytopes is finally proposed in order to approximate the non-convex boundary of the one-class classification problem. The algorithm is based on a tiling strategy and each convex polytope is approximated by the approximate convex polytope decision ensemble.

- First, I used Approximate Polytope Ensemble (APE), to create a model for OCC.

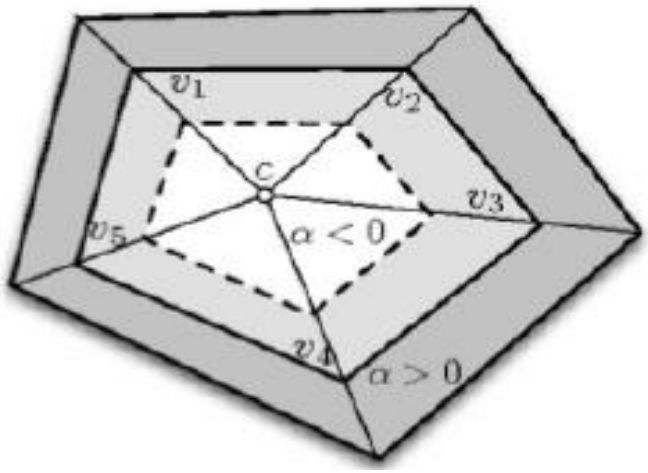

Figure 7 Illustration of the expanded convex polytope in the 2D space. [24]

- Then, the geometrical concept of convex hull is used to define the boundary of the target class defining the problem.
- Second, the decision whether a point belongs to the convex hull model in high dimensional spaces is approximated by means of random projections and an ensemble decision process. This algorithm describes the test procedure. It is possible to check if the test point lies inside the projected polytope.
- Finally, a tiling strategy is proposed in order to model non-convex structures because, the main drawback of APE is that the boundary of training data may not be well modelled by a convex polytope. Hence, an extension of the algorithm to cope with non-convex boundaries is also proposed.

A. **FLOW OF WORK:**

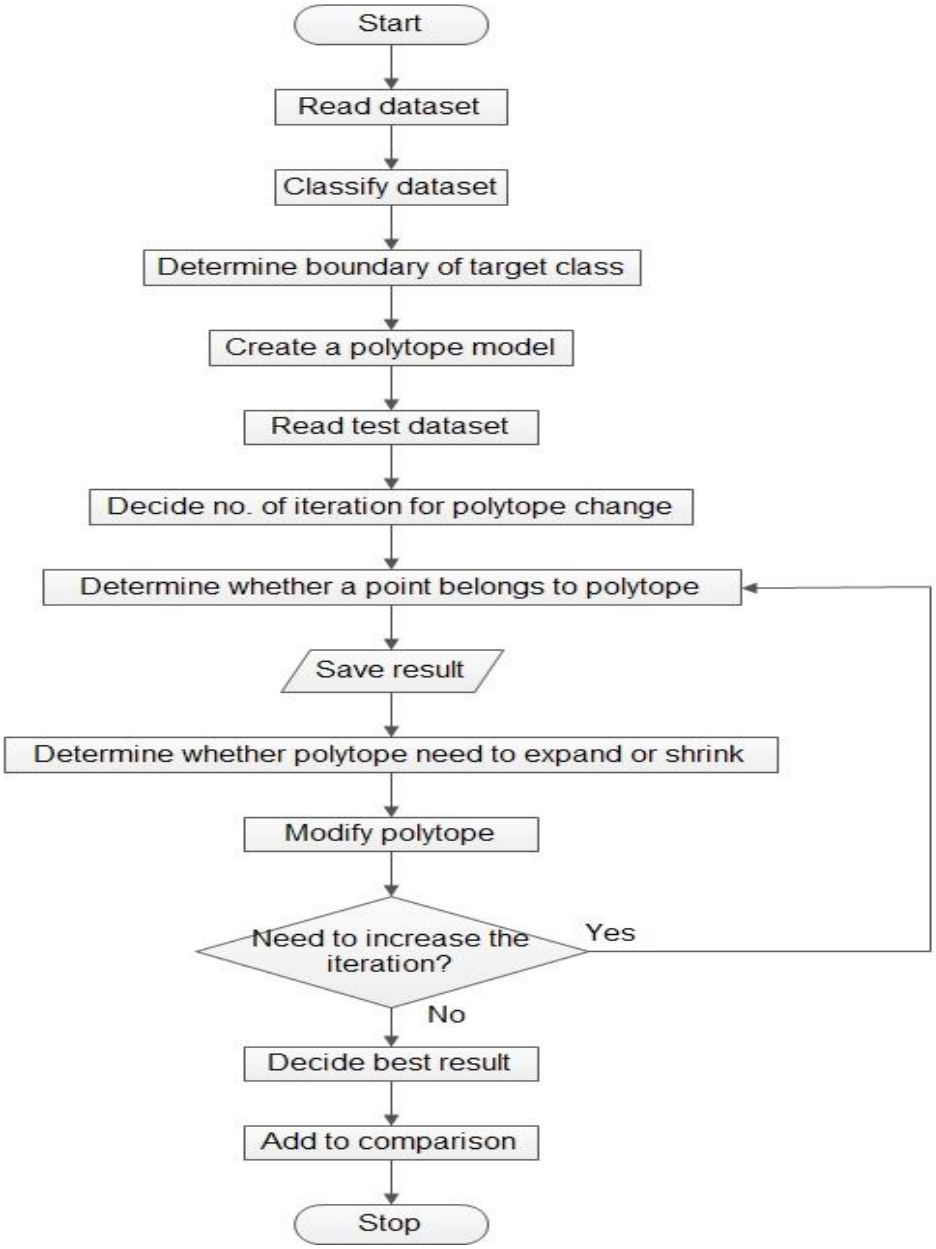

Figure 8 Flow of Proposed Work.

B. **THE PROPOSED ALGORITHM:**

1. First, read the dataset.
2. Classify the dataset based on different classification methods.
3. Create an initial convex hull. The initial convex hull has not any fixed boundary points.
4. The geometrical concept of convex hull is used to define the boundary of the target class (polytope) defining the problem.
5. Now, read the test dataset for finding whether points are inside or outside.
6. Decide the no. of iteration for polytope change.
7. The decision whether a point belongs to the polytope in high dimensional spaces is approximated by means of random projections and an ensemble decision process.
8. Generate Rules that map Inside-Outside test point for convex hull.

   RULES:

   XT = PT X (PROJECT DATA)

   VTN = [VT + VI { (VI-C) / ||VI – C|| }] WHERE VI IS IN {V}T .

   IF XT IS NOT IN CONV VTN THEN

   RESULT = OUTSIDE

   BREAK.
9. Save the result for inside-outside points.
10. Decide whether structure need to shrink or expand the convex hull governed by parameter n.
11. Check whether we need to increase the no. of iteration for best result?
12. Add the output result for comparison.

C. **PHASE 1: CONSTRUCTIVE ALGORITHM**

1. Create an initial Convex hull. The initial convex hull has not any fixed boundary points.
2. Sort the points by x-coordinate, resulting in p1 ,…..,pn .
3. Compute the upper bound and lower bound of convex hull to decide the boundary.
4. Add all the test points to this convex hull to check whether that point inside or outside the convex hull.

Algorithm CONVEXHULL (P)

INPUT: A SET P OF POINTS IN THE PLANE.

OUTPUT: A LIST CONTAINING THE VERTICES OF CH (P) IN CLOCKWISE ORDER.

1. Sort the points by x-coordinate, resulting in a sequence p1. . . pn.
2. Put the points p1 and p2 in a list Lupper, with p1 as the first point.
3. for i 3 to n
4. Do append pi to Lupper.

5. while Lupper contains more than two points and the last three Points in Lupper do not make a right turn
6. Do delete the middle of the last three points from Lupper.
7. Put the points pn and pn−1 in a list Llower, with pn as the first point.
8. for i n−2 down to 1
9. Do append pi to Llower.
10. While Llower contains more than 2 points and the last three points in Llower do not make a right turn.
11. Do delete the middle of the last three points from Llower.
12. Remove the first and the last point from Llower to avoid duplication of the points where the upper and lower hull meets.
13. Append Llower to Lupper, and call the resulting list L.
14. Return L.

D. PHASE 2: CONSTRUCTIVE MODEL ALGORITHM

The steps needed for learning and testing the proposed approach are described in below two Algorithms, respectively. Both algorithms require defining the number of projections. In the learning phase, at each iteration, a random matrix is created. Then, the training set is projected into the space spanned by the random projection matrix. Finally, the vertices of the convex hull of the projected data set are found.

Input: Training set C = {xt} is in RD, i = 1…N , with D the dimensionality of each data example xi;
Number of Projections n

Output: The model M composed of n projection matrices and their respective convex hull vertices.

M= null;

$c = \frac{1}{N} \sum_i x_i, \forall x_i \in C;$

for t = 1..n do
Pt ~ N(0,1) % Create a normal random projection matrix;
Ct : {ptx|x is in C} % Project data onto the low dimensional random space;
{vi}t= conv Ct % Find the convex hull and return the set of vertices;
M=M+(Pt; {vi}t) % Store the set of vertices associated to the convex hull in the projected space and the projection matrix;
end

E. PHASE 3: INSIDE OUTSIDE TEST ALGORITHM

This algorithm describes the test procedure. It is possible to check if the test point lies inside the projected polytope. A point approximately belongs to the model if it lies inside all the t projected polytopes.

Input: A test point x is in RD; The model M; The parameter n
Output: Results is in {INSIDE, OUTSIDE}
Results = INSIDE;
for t = 1..n do
xt = Ptx % Project data.;
vtn = [vt + vi { (vi-c) / ||vi – c|| }] where vi is in {v}t % Find the expanded convex polytope in the low dimensional space;
if xt is not in conv vtn then
Results = OUTSIDE
Break
end
end

F. **PHASE 4: NON-CONVEX DECOMPOSITION ALGORITHM**

The main drawback of APE is that the boundary of training data may not be well modelled by a convex polytope. Hence, an extension of the algorithm to cope with non-convex boundaries is also proposed. The underlying idea of this extension is to divide the non-convex boundary into a set of convex problems. Then, each of the convex problems is solved by means of the approximate convex polytope decision ensemble. The result of this process is another ensemble algorithm called Non-convex APE (NAPE).

Input: Training set C is in RD, with D the number of features;
Number of Projections n
Output: Model M composed of several convex models defined by above algo.
L = null;
Pick a random training point cstart;
L = L + {cstart}% Initialize the list of possible centers with the first random element;
Set all data points x is in C to the value not visited;
While each x with value not visited do
if L = null then
Pick a random a training point with attribute not visited, p is in C.
L = L + {p}
end
p = first(L) % Remove the first element of the list;
Ci : {x is in C | ||x-p||<= r} % Find the set of points to be modelled with a Convex polytope in this iteration;
Set all data points x is in Ci to the value visited;
Mi = TrainAPE(Ci; T) % Find the approximate model associated to the selected set using above algo;
M=M+Mi % Add the new convex model to the final model set.;
L = L + {vi is in C | vi is in Mi} % Add the points of C corresponding to vertices of the projected convex hulls of the current model Mi ;
end

VI.EXPERIMENTAL WORK

This section describes the current status of implementation along with appropriate screen shots.

First, the geometrical concept of convex hull is used to define the boundary of the target class defining the problem.

Second, the decision whether a point belongs to the convex hull model in high dimensional spaces is approximated by means of random projections and an ensemble decision process. This algorithm describes the test procedure. It is possible to check if the test point lies inside the projected Polytope. Decide the no. Iteration for Expand or Shrink the Polytope for covers the points.

A. **Task :2 Convex Hull - Polytope Model:**

1. CONVEX HULL FOR IONOSPHERE DATASET:

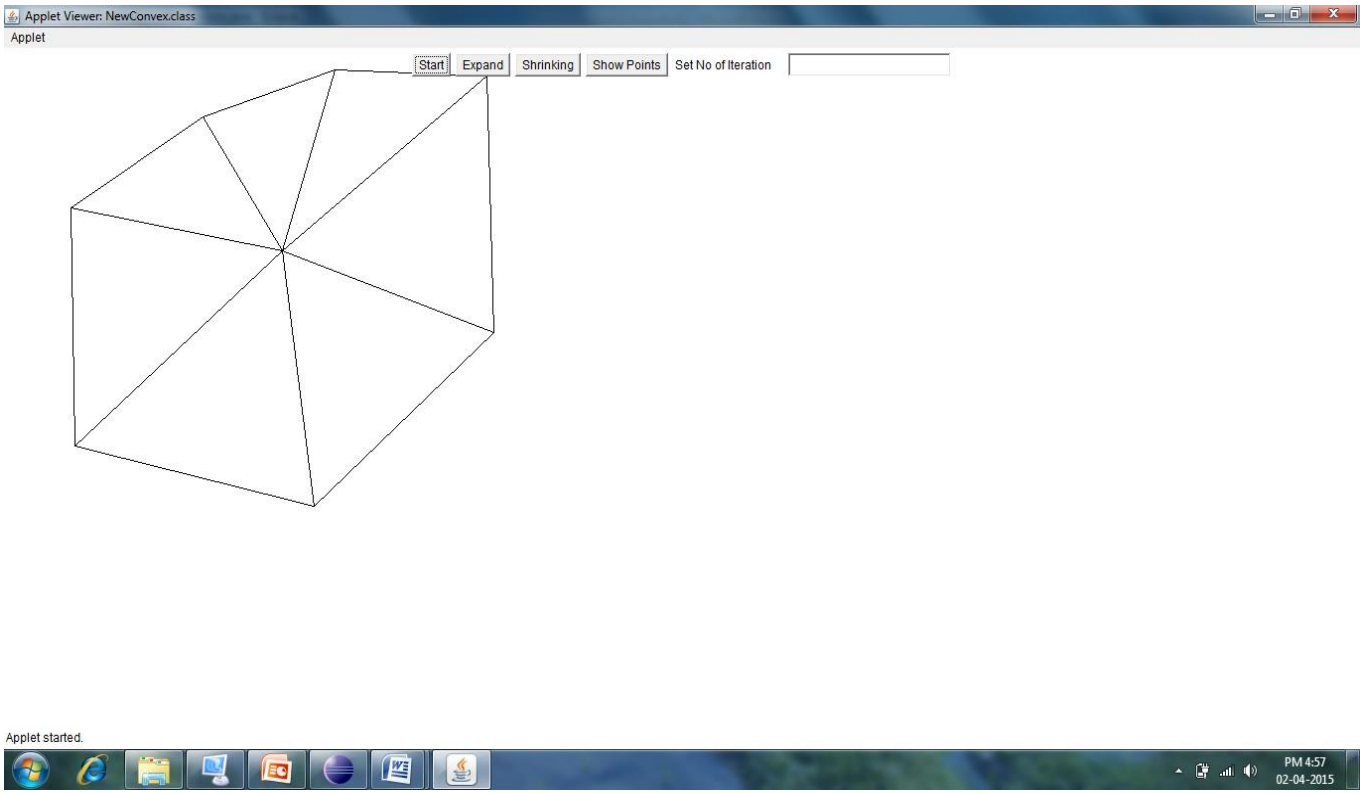

Figure 9 Polytope of Glass Dataset

B. **After Insert Data in the Polytope Model:**

2. INSERT DATA INTO CONVEX HULL OF IONOSPHERE DATASET:

Now, Insert points of Dataset Into created convex hull – Polytope model.

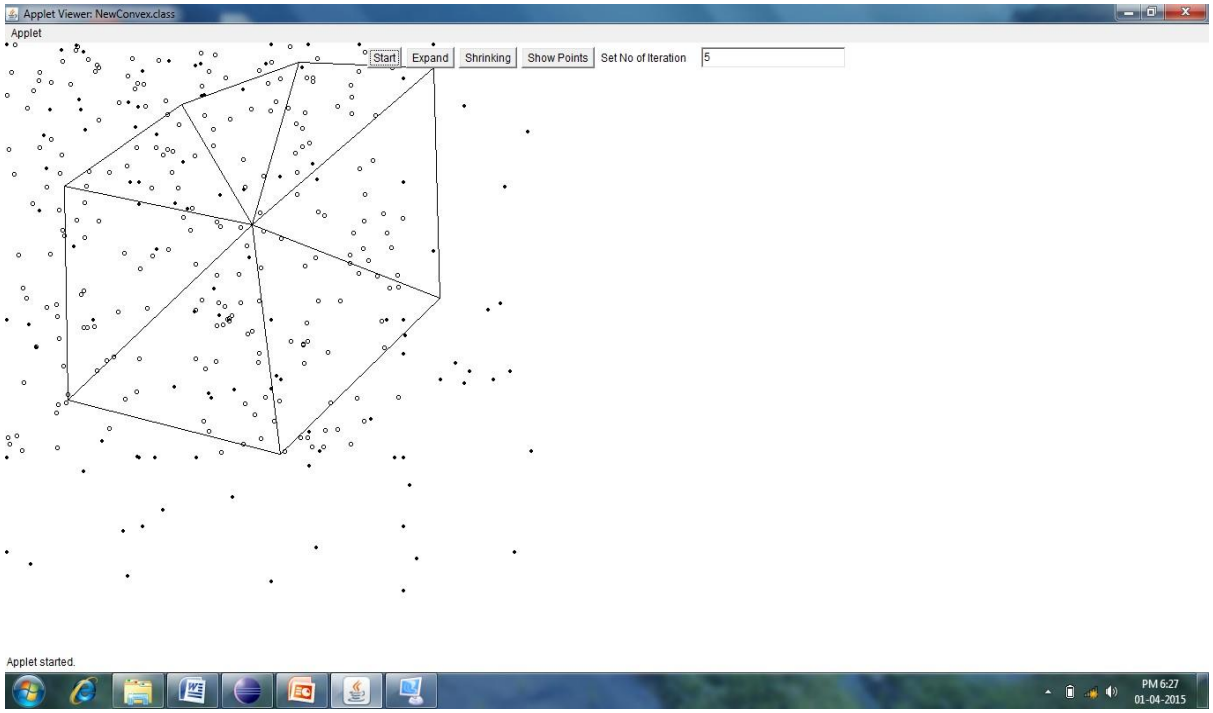

Figure 10 Insert Data into Polytope of Ionosphere Dataset

We discover improved % Accuracy or Reduced % Error Rate from the use of Ensembles of Naïve Bayes and IBK (KNN) classification methods for classify the data into one class classification using convex hull polytope model.

VI.B.1.1 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.

ACKNOWLEDGEMENT

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 Knowledg 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.
[24] Pierluigi Casale, Oriol Pujol, Petia Radeva "Approximate Polytope Ensemble for One-Class Classification," Signal Processing Systems Group Eindhoven University of Technology, Spain.

