Solving Multi-Class Classification Tasks with Classifier Ensemble based on Clustering

Mohammad Rafiul Haque (ID: 011151326)
Alam Al Saud (ID: 011151346)
Annajiat Yasmin Bipasha (ID: 011143029)
Sabbir Hossain (ID: 011133045)

Department of Computer Science and Engineering

United International University

A thesis submitted for the degree of

BSc in Computer Science & Engineering

September 2019

Abstract

Ensemble learning is very popular for few decades for solving classification problems, because it generates and combines a diversity of classifiers using the same learning algorithm for the base-classifiers. In this paper we propose a method for generating classifier ensembles based on clustering. But with the continuous expansion of data availability in many large-scale, such as surveillance, security, Internet, and finance. It becomes critical to advance the fundamental understanding of knowledge discovery and analysis from raw data to support decision-making processes. Unremarkable computers can't effectuate the demand as they have unsubstantial memory space and delimited speed. As a result of these types of issues, contemporary prediction gets delayed. To eschew these problems, we have done some research and approached a efficacious algorithm so that we can reduce the amount of data by collecting only the informative data from the whole data set by using clustering. So that the mammoth data can be handled quite conveniently. After clustering we get new sub data sets. From each cluster we make fewer chunks of data than the archetype data set with the informative data with our new algorithm. After mingled these sub data sets we have run different ensemble algorithms on this new data set for comparability or exemplification. Comparing the results brandish that the accuracy rate is almost similar or increasing or decreasing. In some case we get more accuracy and in some case we get almost same accuracy. Once in a while we get less accuracy but even if we get less accuracy, the convenience side is that the amount of data is shortened.

Acknowledgements

This work would have not been possible without the input and support of many people over time. We would like to express our gratitude to everyone who contributed to it in some way or another. First, we would like to thank our academic advisor, Dr. Dewan Md. Farid, Associate Professor, Department of Computer Science and Engineering, United International University, Bangladesh. We consider ourselves lucky to have him as our supervisor. We are thankful to him for his constant guidance, advice, encouragement throughout the whole period of time. Our sincere gratitude goes to Dr. Salekul Islam, Head of the Department of Computer Science and Engineering, United International University, Dr. Swakkhar Shatabda, Asst. Professor Undergraduate Coordinator, Computer Science and Engineering, United International University, for their helpful attitude and for the sufficient time and support they gave us whenever we needed, otherwise this thesis could not have been done. We remember, even in the most frustrating period of ours how their continuous guidance kept us on the right track. Last but not least we give huge credit to our family for their unconditional love and immense emotional support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning is concerned about building systems that learns from data. In recent years, there are enormous amount of data being produced and machine learning algorithms are being heavily used in various fields. There are different machine learning algorithms having their own advantages and disadvantages.

1.1 Motivation

We have done this work to make more effective the ensemble methods with clustering and to solve the dilemma we face while working with large amount of data. To handle a humongous size data set is always rough and tough for any data scientist and its causing problem to deals with crowd-pleasing machine learning algorithm. In today's life it is a necessity to pledge with monstrous amount of data regularly but it gets difficult to store these large amounts of data in devices and our work gets slower most of the time. So we thought if somehow we can use fewer amount of data and at the same time the accuracy rate doesn't vary much from the preeminent result then it would have been more convenient for all. So from this motive, we started researching on this topic and approached a method that can reduce the amount of data almost by half in number and sometimes also give almost same accuracy as we get from the main data set. Machine learning can be defined by two things. One is using the data and other is answering the questions. These two parts can be considered as the two sides of machine learning. Both of them have equal importance in machine learning. Using data is what we refer to as training while answering questions is referred to as making predictions or inference.

Machine learning can be defined by two things one is using the data and other is answering the questions. These two parts can be considered as the two sides o machine learning, both of them have equal importance in machine learning. Using data is what we refer to as training while answering questions is referred to as making predictions or inference. There is a lot of data in the world not only generated by people but also by various devices like computers, phones etc. Machine learning promises to derive meaning from all of that data. In our daily life we can see a lot of example of machine learning around us but it's not always apparent that machine learning is behind all these. For example we can say while we tag someone or something in a picture there is clearly machine learning at play for sure but it may not be immediately apparent that recommending next video to watch is also powered by machine learning. Google search engine is the biggest example of this. When we search something in Google search we use so many machine learning systems, as well as text and speech systems too. These powerful capabilities can be used to a wide range of fields, from diabetic retinopathy and skin cancer detection to retail and transportation in the form of self-parking ad self-driving vehicles. Today almost every company use machine learning in their products in some way. It's rapidly becoming an expected feature.

There are many ways in which the machine learns-

- Supervised learning: supervised learning uses labeled data to train the model. Here the machine knew the features of the object and also labels associated to those features.

- Unsupervised learning: the learning with unlabeled data is known as unsupervised learning.

- Reinforcement learning: Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

Now let's do some digging on data mining. Data Mining is the science that uses all the techniques of machine learning in order to extract useful and important patterns from data. Data Mining usually has to do with extracting useful information from massive data sets, that is, Big Data. By data mining we mean to find required knowledge from a large amount of data set. For example we can say when someone search something in Google there are indefinite numbers of data there, but among them which data is required that is extracted by the help of data mining. When companies need to take a good decision the do it through data mining. Collection, extraction, warehousing, analysis & statistics are the things which are involved in data mining. Microsoft excel, Microsoft access, Microsoft visual studio all these software are used in data mining. Through data mining we can do many things like anomaly detection, association rule learning, clustering, classification, regression, summarization etc.

Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model. That is why ensemble methods placed first in many prestigious machine learning competitions. In ensemble learning we have multiple learners. If a test example is given then these multiple learners will give multiple learners will give multiple outputs. These outputs may be all same or all different or some of them will be same and some of them will be different. Multiple learners may give multiple decisions. We have to combine these decisions, for that we need to generate base learners. Base learners has to be different and there are many ways in which these learners can be different, they may be using different algorithms, different hyper parameters, different representations, different training sets. There are many methods for ensemble learning. Here we all discuss about bagging, boosting and random forest algorithms.

- **Random forest:** Random forest is a process or method which operators by constructing multiple decision trees during training phase. The decision of the majority of the trees is chosen by the random forest as the final decision.
- **Bagging:** Bagging is a method for ensemble learning. Bagging stands for "bootstrap aggregation". In bagging the mining models receive the same weight but in boosting algorithm weighting is used to give more influence to the more successful ones. Bagging combines different classifiers into a single prediction model. In bagging voting is used for classifying a new instance.
- **Boosting:** There are many boosting algorithms among which AdaBoost is the renowned one. A series of classifiers are considered in AdaBoost and it combines each individual classifiers votes to classify an unknown or known instance. In boosting each training instance has a weight.

So in our proposed method what we do is dividing the data set into some clusters. Then from that cluster we will create a new data set by taking data from nearest center and nearest border. Here we will apply bagging algorithm. There will be no random selection in creating initial subset as we will use clustering. We will give specific data while creating initial subset so there will be no random selection.

1.2 Objectives

- Making ensemble learning more robust with clustering was one of our aim to do this thesis. In our thesis we have emphasized on solving big data problem by collecting informative data.
- Our another objective is to improve data collection algorithms so that informative data can be collected more efficient.
- If somehow classification models can become more robust and faster then it will be better. So in our thesis we have work on how to make these models more robust and faster.

1.3 Organization of the Thesis

The thesis is organised as follows:

Chapter 2 provides related works.

Chapter 3 presents the proposed method.

Chapter 4 discusses the results and experimental analysis.

Chapter 5 presents the conclusions, summaries the thesis contributions, and discusses the future works.

Chapter 2

Related Work

Decisions of multiple classification models can combined into a single final result by ensemble learning (Koziol et al. 2009). The main objective of ensemble methods is not only improving overall classification performance, but also more accurate generalization capability in classifying unseen instances (Yang et al. 2010). The performance of ensemble method mainly affects by the accuracy and the diversity of the base classifiers [1].Some researchers (Nagi et al 2013) conducted an empirical study using nine high-dimensional cancer datasets and three classifiers[2]. The researchers proposed a new ensemble method and compared class-specific accuracy of their method versus each single classifier as well as Bagging and Boosting. Another work by Tan [3]] used seven cancer gene expression datasets along with the decision tree classifier, and two ensemble methods: Bagging and Boosting with decision trees as the classifier. In 2014, one research group introduced SelectBagging ensemble method in their paper (Dittman et al. 2014). In their work they observed how Select-Bagging performed compared to when no ensemble approach is applied. Bagging was first proposed by Leo Breiman [4] in 1994. Bagging was invented to improve classification by combining classifications of randomly generated training sets. Its name was deduced from the phrase "bootstrap aggregating"[4]. Using bagging Kristına Machova, Frantisek Barcak, Peter Bednar describe a set of experiments that can improve results of classification algorithms [5]. Application of bagging to cluster analysis can substantially improve clustering accuracy and yields information on the accuracy of cluster assignments for individual observations [6]. Boosting is one of the ensemble learning method which primarily reduce bias, variance in supervised learning [7], and it converts weak learners to strong ones [8]. Kearns and Valiant(1988, 1989) [9] [10] raised a question that if a set of weak learners can create a single strong learner. An affirmative answer to their question from Robert Schaphire [11] which was published in a 1990's paper has had significant ramifications in machine learning and statistics, most notably leading to the development of boosting [12]. After that Freund and Schapire (1996; 1997) introduces a more effective algorithm that is AdaBoost Algorithm [13]]. Since the invention, it has become very popular with 4 both theoreticians and practitioners of machine learning. With this way many method is proposed about boosting like Logitboost. But a year later, Freund [14] developed a much more efficient boosting algorithm which, although optimal in a certain sense, nevertheless suffered from certain practical drawbacks. The first experiments with these early boosting algorithms were carried out by Drucker, Schapire and Simard on an OCR task. Gunnar Rätsch ,Bernhard Schölkopf ,Alexander Johannes Smola published a paper on knowledge discovery on data stream in conference[15]. They proposed a new boosting algorithm which similarly to v-Support-Vector Classification allows for the possibility of a pre-specified fraction v of points to lie in the margin area or even on the wrong side of the decision boundary. From the data stream the algorithm will filter the collective data and selective data patterns.

Tin Kam Ho [16] invented the very first algorithm for Random Decision Forests. He used random subspace method [17] for this invention, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg [18] [19]]. An extension of this algorithm was developed by Leo Breiman [20] and Adele Cutler [21], and "Random Forests" is their trademark . Random Forests algorithms are efficient, multi class and they are able to handle large attribute space. In face recognition [22], bioinformatics [23] random forests algorithms are being used widely. With the help of Random forest an efficient medical image retrieval method using image classification was proposed in [24]. 3D object segmentation in 3D medical imaging modalities is proposed in [25]. Juan J. Rodríguez, Ludmila Kuncheva, Carlos J. Alonso approached a new ensemble method [26] called rotation forest. Individual accuracy and diversity in Ensemble Learning is encouraged by the idea of Rotation Forest Algorithm. Junshi Xia ,Peijun Du they made a good application on the rotation forest [27]. Rotation Forest, has been applied to hyperspectral remote sensing image classification for the first time. There are other research has been occurred about combining those above ensemble learning methods. And the combination of Bagging , Boosting and Random forest makes the classifier more stro an robust.S. B. Kotsiantis and P. E. Pintelas published a paper [28] on solving the problem of noisy data with the combination of bagging and boosting algorithm. In "adabag: An R Package for Classification with Boosting and Bagging" Esteban Alfaro , Matıas Gamez and Noelia Garcıa approached a method call adabag [29]. In this method they tried to n implements AdaBoost.M1, SAMME and bagging algorithms with classification trees as base classifiers. There is also a notable work that has to be mentioned , Sotiris Kotsiantis approached a method with the ensemble of bagging, boosting, rotation forest . The creates multiple subspaces with those ensemble algorithm. Researchers have given great effort on solving the real world problems with ensemble learning method. Among many examples ther is a good example of real application , which is on person recognition written by Suutala and Roning [30]. The main goal of this application is tracking the identities and locations of different person. Monitoring on those people requires the identification and their unique habitation and behaviour . In this system , the main advantages is people you want to monitor they dont need to wear any kind of sensor. Another notable application occured by Chawla and Bowyer [31] and that solve a very big problem of face recognition [32] [33] [34] under different face expression and difference amount of light. They tested recognition of subjects that they trained with but also subjects that they did not train with. Their data consisted of images of multiple subjects, each with multiple pictures taken under two different lighting conditions and with two different facial expressions (neutral and smiling).

Chapter 3

Ensemble Learning

3.1 Random Forest

Random forest or random decision forest is a method that operates by constructing multiple decision trees during training phase. The decision of the majority of the trees is chosen by the random forest as the final decision. It is a supervised algorithm used for classification and regression.We can visualize it from its name. As the name conceive, this algorithm generates the forest with a number of trees. Ordinarily, the more tress in the forest the more herculean the looks like. In the similar process in the random forest classifier, there is a straight relationship among the number of trees in the forest and the outcomes it can get: the higher the number of trees gives the higher accuracy outcomes. But one thing to note is that generating the forest is not the similar as building the decision with information gain or gain index method. If you learn the decision tree algorithm. You might be guessing are we generating more number of decision trees and how we can make more number of decision trees. As all the calculation of nodes choosing will be similar for the similar dataset. In random forests create multiple trees as converse to an individual tree in court model to classify a new object based on attributes every tree gives a classification and same the tree votes for that class. The forest pick the classification having the highest votes all the several trees in the forest and in the case of regression takes the average of the outputs by different trees. The decision of the majority of the trees is chosen by the random forest as the final decision. You can see how a random forest work step by step is showed below-

Advantages:

- It is very accurate learning algorithm on Ensemble Learning. It produces highly accurate model classifier.
- It is very efficient on big dataset.
- It can run operation on very big data within short time.

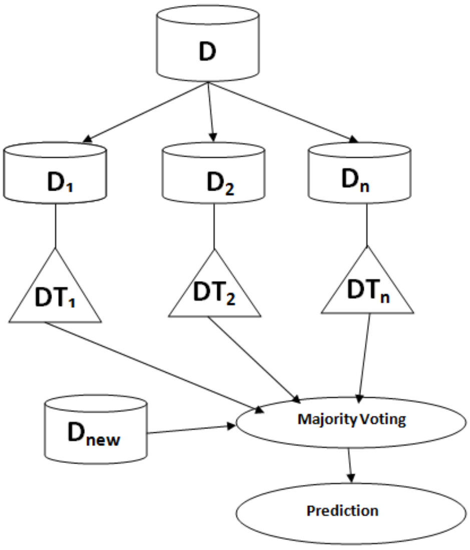

Figure 3.1: Random Forest - Random Forest algorithm process

- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- It has methods for balancing error in class population unbalanced data sets.
- It saves the decision forest for using in future on different data.
- The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.

Disadvantages:

- The main disadvantage of Random forests is their complexity. They are much harder and time-consuming to construct than decision trees.
- Overfitting can easily occur.
- In real time prediction a large number of trees may make the algorithm slower.
- Random forests have been observed to over fit for some datasets with noisy classification/regression tasks.
- For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable for this type of data.

3.2 Bagging

Bagging is a classic technique for generating lots of predictors and combining it together in a simple way to do a better job. Ensemble methods try to use a collection of predictors to do a better job than any single predictors would have done alone. Bagging stands for bootstrap aggregation. Bagging is a technique for learning many classifiers each using only portions of the data and then combining them through a model averaging technique. The idea behind this is to reduce over fitting of a classic model. To avoid over fitting we would memorize the data set and we would get a far lower training error on that training data set.The idea behind bagging is to do a similar kind of data splitting or resampling technique but instead of using them to check to see whether we over fit instead we try to combine them so that we can produce a better classifier or predictor.

Bagging is a classic ensemble technique or producing better predictors than any single predictor. It's a technique that tries to reduce the complexity of a model class. So if a model class is chosen that's very prone to over fit in it apply bagging to provide a collection of learners in that class that are less complex and less prone to over fit and it's quite simple to implement which is resampling the data once for each learner. Each learner is trained on an individual resampling and we create a predictor that might over fit on that sample but then b averaging them together produces something that's robust to small variation to the data. It essentially plays on the bias variance straight on choosing something that's prone to over fit in them thus has low bias but reducing its variance in model averaging. The price is the computational cost which if we learn k bagged predictors our prediction time computation becomes k times larger than it would have been before.

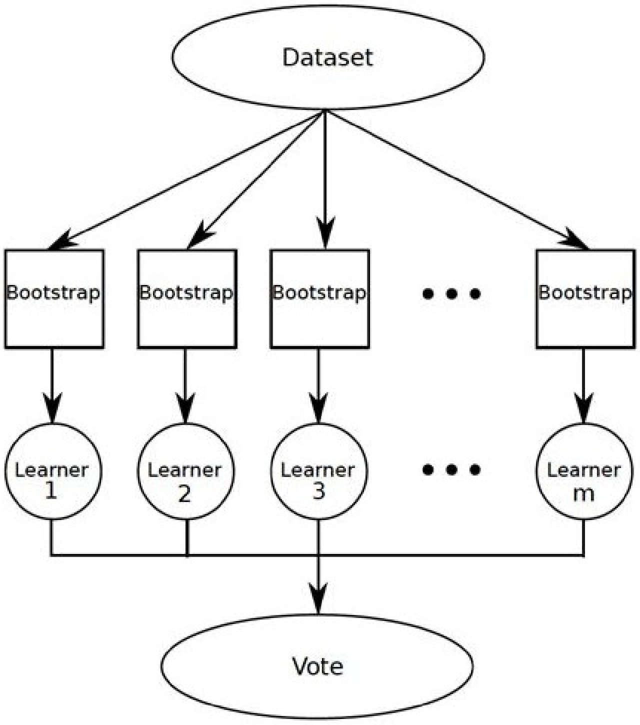

Figure 3.2: Bagging - Bagging algorithm process

Advantages:

- Easy to implement.
- Works well with many classifiers.
- Provides an unbiased estimate of the test error.
- Reduces variance and helps to avoid over fitting.
- Improves ability to ignore irrelevant features.

Disadvantages:

- With lots of data, we usually lean the same data.
- Averaging over these doesn't help.

3.3 Boosting

Adaboost, concise for Adaptive Boosting, is a machine learning technique. Adaboost is another ensemble learning algorithm in machine learning where these are made of multiple classification and regression algorithm like random forest, bagging and boosting. Adaboost is the most similar algorithm with bagging or Booststrap Aggregation. This machine learning algorithm is conceptually simple to understand. Adaboost is a one of the best Meta formulated algorithm by Yoav Freund and Robert Schapire, who won the 2003 Global Prize for their excellent job. Adaboost classification algorithm assembles infirm classifier algorithm to form rigid classifier. A separate algorithm can classify the objects weakly. Even if we assemble abundant classifiers with election of training set at each iteration and imposing appropriate amount of weight in final voting for prediction, after all we can have good accuaracy for in total classifier.Adaboost is more impressible to noisy data. In some problems it may be less capable to the overfitting proble than other classifier algorithm. The separate classifier can be weak, however as long as the representation of every one is slightly better than random conception, the ultimate model can be proven to converge to a dynamic classifier. The actual difference between adaboost and bagging methods. While the training stage is sequential for adaboost, other thing is adaboost allocated weights to every resulting model by average weights, and adaboost contineously train and evaluate until achieving a better learner than a random guessing.

Advantages:

- Very simple to implement.
- Fairly good generalization.
- The prior error need not be known ahead of time.

Disadvantages:

- Can over fit in presence of noise.
- Suboptimal solution.
- Hard to implement in real time platform.
- Time and computation expensive.
- Not very speedy to train or score.
- Compared to linear classifier it has lack o interpretability.

3.4 Proposed Method

Our main topic of thesis is Ensemble learning. With ensemble learning we use this to make a weak learner to strong learner. Popular approaches of EL are bagging,boosting and randomforest. Both are robust method of Ensemble learning. But we when we deal with big data things get more difficult to get perfect inference and calculation. It require more power and time. So in this paper we proposed a different approach. Before go with EL we cluster the data set in severel part. from those data set we took 30% of center data and 30% of data from border line. Then we apply EL on those clustered filter data. With this method we get 40% less data. Thus we can solve the big data problem. For collecting the informative data form dataset we have applied our algorithm. In which agorithm we use distatnce to maintain distance to measure the points.

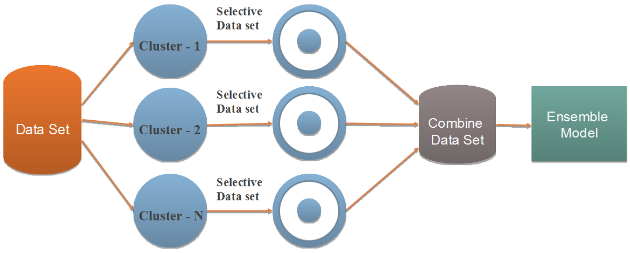

Figure 3.3: Methodology - Proposed Method

3.4.1 Methodology

Step 1: Preprocess the data set.
Step 2: cluster the data set with k-mean clustering algorithm into 3 cluster.
Step 3: Take 30% of center data and 30% of border side data.
Step 4: apply on different models based on different EL methods and collect results.
Step 5: observe the accuracy of final clustered data from different models.

3.4.2 Description

Step 1: At first we select dataset. And preprocess the dataset whether there is any missing value or character value. If there is missing value we put some perfect precise value instead . After that we find whether there is any non-nominal value. Like character. If there is we set a value for each character.
Step 2: Then we apply k-mean clustering algorithm to cluster the preprocess data set. We create 3 clustered data set from main dataset. Here we faces some difficulties that some dataset turn into overfitted in clustered data set. So we had to solve the ovefitting problem.
Step 3: After getting perfect clustered dataset we take 30% of center data and 30% of border side data. To get those data, first we sort all instances based on Euclidean distance from centeroid. For doing this we applyed our algorithm . the psudo code given below

Then we took the first 30% data and last 30% of data from the sorted dataset. With this filter instances we create different clustered files. Here we also faces some difficulties of overfitting problem. Sometime the filtered dataset has only one class. When we apply the dataset on different model we were getting over fit alarm.

Step 4: Then we Apply those filtered data set into different models. Those model was based on adaboost, bagging, random forest algorithm. Then we collect results of different model.

Step 5: In this step we observed the results and calculate the difference of final result on filtered data from main dataset result. Sometime the final rest remains same and some time it gets more accurate result and sometime we got less accurate result.

Chapter 4

Experimental Analysis

In this section we describe how data is used for the evaluation of the ensemble method; we provide details of the experiments and then we present the results . In order to empirically test the proposed ensemble learning methods , we ran a number of experiments on several UCI data sets (Blake and Merz 1998) and compared with previous bagging boosting methods. Some data set shows very good result with our proposed approach and some data set remains in same results. Very few datasets has shown to negative result.So the evolution of our proposed approach performance was conducted as follow : (A) We collected some dataset and take some experimental analysis and observe the results . (B) then we made experiment with same dataset with our proposed approach. For experimenting we have used weka software for observing the results and that makes more convenient to compare the performance.

4.1 Data sets

There are some data set details descriptions we are used in this thesis work.

Table 4.1: Dataset descriptions.

Data sets	Attribute Types	No.of Instances	Attribute	Classes
Nursery	Nominal	12960	8	5
Contraceptive Method Choice	Nominal, Integer	1473	9	3
PRIMA	Real	768	9	2
Glass	Real	214	9	7
Page Blocks Classification	Real	5473	10	5
Wine Quality	Real	4898	12	7
Cleveland	Real	303	14	5
Bach Choral Harmony	Nominal	5665	17	102
Bank Marketing	Real	45211	17	2
Statlog (Image	Real	2310	19	7
Vehicle Segmentation	Nominal	846	19	4
Diabetic Retinopathy Debrecen	Real	1151	20	2
Anuran Calls (MFCCs)	Real	7195	22	60
Wall-Following Robot Navigation	Real	5456	24	4
Phishing Websites	Integer	2456	30	2
Turkiye Student Evaluation	Integer	5820	33	5
Dermatology	Nominal	366	33	6
Annealing	Real	798	38	5

4.2 Experiment

This table show the classification accuracy and precision, recall and F-score values for a AdaBoost (C4.5) classifier with 10-fold cross validation.

Table 4.2: AdaBoost (C4.5) classifier with 10-fold cross validation.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	99.4907	0.995	0.995	0.995
Contraceptive Method Choice	50.0339	0.498	0.5	0.498
PRIMA	71.7448	0.717	0.717	0.717
Glass	74.291	0.738	0.743	0.739
Page Blocks Classification	97.0217	0.97	0.97	0.97
Wine Quality	66.129	0.655	0.661	0.656
Cleveland	53.7954	0.511	0.538	0.524
Bach Choral Harmony	74.6514	0.985	0.747	0.850
Bank Marketing	89.6198	0.889	0.896	0.892
Statlog (Image Segmentation	98.2684	0.983	0.983	0.983
Vehicle	76.0047	0.757	0.76	0.758
Diabetic Retinopathy Debrecen	65.4214	0.655	0.654	0.655
Anuran Calls (MFCCs)	86.3238	0.993	0.863	0.923
Wall-Following Robot Navigation	99.835	0.998	0.998	0.998
Phishing Websites	97.1144	0.971	0.971	0.971
Turkiye Student Evaluation	84.2955	0.843	0.843	0.843
Dermatology	95.9016	0.959	0.959	0.959
Annealing	94.4862	0.945	0.945	0.945

This table show the classification accuracy and precision, recall and F-score values for a AdaBoost (CART) classifier with 10-fold cross validation.

Table 4.3: AdaBoost (CART) classifier with 10-fold cross validation.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	99.5216	0.995	0.995	0.995
Contraceptive Method Choice	55.3293	0.552	0.553	0.552
PRIMA	73.4375	0.733	0.734	0.734
Glass	71.0280	0.701	0.710	0.701
Page Blocks Classification	96.9852	0.968	0.970	0.969
Wine Quality	63.5770	0.747	0.636	0.687
Cleveland	56.1056	0.470	0.561	0.504
Bach Choral Harmony	56.7167	0.961	0.567	0.713
Bank Marketing	89.4318	0.886	0.894	0.890
Statlog (Image Segmentation	97.9221	0.979	0.979	0.979
Vehicle	76.1229	0.756	0.761	0.758
Diabetic Retinopathy Debrecen	65.8558	0.659	0.659	0.659
Anuran Calls (MFCCs)	84.3502	0.993	0.844	0.912
Wall-Following Robot Navigation	99.7434	0.997	0.997	0.997
Phishing Websites	96.9064	0.969	0.969	0.969
Turkiye Student Evaluation	85.1546	0.852	0.852	0.851
Dermatology	41.8033	0.400	0.418	0.399
Annealing	93.8596	0.939	0.939	0.938

This table show the classification accuracy and precision, recall and F-score values for a AdaBoost (Naive Bays) classifier with 10-fold cross validation.

Table 4.4: AdaBoost (Naive Bays) classifier with 10-fold cross validation.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	91.8210	0.962	0.918	0.939
Contraceptive Method Choice	49.3551	0.520	0.494	0.498
PRIMA	75.6510	0.752	0.757	0.753
Glass	49.0654	0.498	0.491	0.456
Page Blocks Classification	90.8460	0.938	0.908	0.919
Wine Quality	44.2630	0.462	0.443	0.432
Cleveland	54.7855	0.530	0.548	0.538
Bach Choral Harmony	74.5102	0.985	0.745	0.848
Bank Marketing	89.1177	0.881	0.891	0.885
Statlog (Image Segmentation	80.1299	0.819	0.801	0.779
Vehicle	44.7910	0.510	0.448	0.413
Diabetic Retinopathy Debrecen	56.8202	0.695	0.568	0.507
Anuran Calls (MFCCs)	73.9541	0.986	0.740	0.845
Wall-Following Robot Navigation	52.4560	0.625	0.525	0.528
Phishing Websites	90.7915	0.908	0.908	0.908
Turkiye Student Evaluation	82.7835	0.841	0.828	0.832
Dermatology	95.9016	0.600	0.959	0.959
Annealing	54.6366	0.839	0.546	0.569

This table show the classification accuracy and precision, recall and F-score values for a AdaBoost (SVM) classifier with 10-fold cross validation.

Table 4.5: AdaBoost (SVM) classifier with 10-fold cross validation.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	93.1327	0.967	0.931	0.949
Contraceptive Method Choice	50.9844	0.513	0.510	0.510
PRIMA	77.3438	0.769	0.773	0.763
Glass	57.0093	0.719	0.570	0.636
Page Blocks Classification	92.9289	0.930	0.929	0.909
Wine Quality	52.0621	0.602	0.521	0.559
Cleveland	60.0660	0.736	0.601	0.662
Bach Choral Harmony	71.3806	0.872	0.654	0.747
Bank Marketing	87.1251	0.726	0.872	0.823
Statlog (Image Segmentation	93.1602	0.932	0.932	0.931
Vehicle	74.4681	0.733	0.745	0.733
Diabetic Retinopathy Debrecen	67.6803	0.686	0.677	0.676
Anuran Calls (MFCCs)	80.5976	0.991	0.806	0.889
Wall-Following Robot Navigation	71.4260	0.718	0.714	0.709
Phishing Websites	92.6911	0.927	0.927	0.927
Turkiye Student Evaluation	85.6357	0.858	0.856	0.856
Dermatology	96.1749	0.962	0.962	0.962
Annealing	83.5840	0.821	0.836	0.823

This table show the classification accuracy and precision, recall and F-score values for a Bagging (C4.5) classifier with 10-fold cross validation.

Table 4.6: Bagging (C4.5) classifier with 10-fold cross validation.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	97.3302	0.988	0.973	0.980
Contraceptive Method Choice	52.9532	0.528	0.530	0.529
PRIMA	76.9531	0.766	0.770	0.767
Glass	74.2991	0.738	0.743	0.739
Page Blocks Classification	97.2045	0.971	0.972	0.971
Wine Quality	65.8024	0.770	0.658	0.710
Cleveland	55.7756	0.505	0.558	0.529
Bach Choral Harmony	74.5278	0.985	0.745	0.848
Bank Marketing	90.4448	0.897	0.904	0.900
Statlog (Image Segmentation	97.4026	0.974	0.974	0.974
Vehicle	74.4681	0.737	0.745	0.740
Diabetic Retinopathy Debrecen	66.4639	0.667	0.665	0.665
Anuran Calls (MFCCs)	84.4892	0.993	0.845	0.913
Wall-Following Robot Navigation	99.5784	0.996	0.996	0.996
Phishing Websites	97.8833	0.979	0.979	0.979
Turkiye Student Evaluation	85.4639	0.855	0.855	0.855
Dermatology	95.0160	0.959	0.959	0.959
Annealing	94.3609	0.945	0.944	0.944

This table show the classification accuracy and precision, recall and F-score values for a Bagging (CART) classifier with 10-fold cross validation.

Table 4.7: Bagging (CART) classifier with 10-fold cross validation.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	97.3688	0.988	0.974	0.981
Contraceptive Method Choice	54.0394	0.538	0.540	0.538
PRIMA	77.3438	0.769	0.773	0.769
Glass	72.4299	0.712	0.724	0.712
Page Blocks Classification	97.2045	0.971	0.972	0.971
Wine Quality	63.3116	0.737	0.633	0.681
Cleveland	57.4257	0.488	0.574	0.520
Bach Choral Harmony	59.0115	0.966	0.590	0.733
Bank Marketing	90.4138	0.896	0.904	0.899
Statlog (Image Segmentation	96.5801	0.966	0.966	0.966
Vehicle	72.2222	0.709	0.722	0.714
Diabetic Retinopathy Debrecen	66.1164	0.662	0.662	0.662
Anuran Calls (MFCCs)	83.3912	0.992	0.834	0.906
Wall-Following Robot Navigation	99.4685	0.995	0.995	0.995
Phishing Websites	96.3998	0.961	0.961	0.961
Turkiye Student Evaluation	86.0137	0.861	0.860	0.860
Dermatology	36.8852	0.352	0.369	0.358
Annealing	94.2356	0.942	0.942	0.942

This table show the classification accuracy and precision, recall and F-score values for a Bagging (Naïve Bays) classifier with 10-fold cross validation.

Table 4.8: Bagging (Naïve Bays) classifier with 10-fold cross validation.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	90.2778	0.952	0.903	0.927
Contraceptive Method Choice	48.4725	0.512	0.485	0.489
PRIMA	75.2604	0.748	0.753	0.749
Glass	50	0.504	0.500	0.469
Page Blocks Classification	90.4988	0.937	0.905	0.917
Wine Quality	44.4671	0.659	0.445	0.531
Cleveland	55.1155	0.532	0.551	0.541
Bach Choral Harmony	74.0159	0.985	0.740	0.845
Bank Marketing	87.8990	0.884	0.879	0.881
Statlog (Image Segmentation	80.3463	0.821	0.803	0.783
Vehicle	45.6265	0.526	0.456	0.422
Diabetic Retinopathy Debrecen	56.8202	0.690	0.568	0.508
Anuran Calls (MFCCs)	74.5379	0.987	0.745	0.849
Wall-Following Robot Navigation	52.8226	0.627	0.528	0.532
Phishing Websites	90.7372	0.907	0.907	0.907
Turkiye Student Evaluation	82.7663	0.841	0.828	0.832
Dermatology	97.5410	0.976	0.975	0.975
Annealing	57.5188	0.841	0.575	0.598

This table show the classification accuracy and precision, recall and F-score values for a Bagging (SVM) classifier with 10-fold cross validation.

Table 4.9: Bagging (SVM) classifier with 10-fold cross validation.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	93.0401	0.967	0.930	0.948
Contraceptive Method Choice	50.8486	0.508	0.508	0.506
PRIMA	77.7344	0.774	0.777	0.767
Glass	57.4766	0.722	0.575	0.640
Page Blocks Classification	93.6050	0.932	0.936	0.922
Wine Quality	52.0008	0.601	0.520	0.558
Cleveland	57.7558	0.514	0.578	0.541
Bach Choral Harmony	77.0874	0.987	0.771	0.866
Bank Marketing	87.2665	0.872	0.723	0.791
Statlog (Image Segmentation	92.8571	0.929	0.929	0.928
Vehicle	72.2222	0.709	0.722	0.714
Diabetic Retinopathy Debrecen	66.5508	0.683	0.666	0.663
Anuran Calls (MFCCs)	80.3614	0.988	0.804	0.887
Wall-Following Robot Navigation	71.1510	0.715	0.712	0.706
Phishing Websites	92.7815	0.928	0.928	0.928
Turkiye Student Evaluation	85.5842	0.857	0.856	0.856
Dermatology	96.9945	0.970	0.970	0.970
Annealing	83.4586	0.836	0.835	0.801

This table show the classification accuracy and precision, recall and F-score values for a Random Forest classifier with 10-fold cross validation.

Table 4.10: Random Forest classifier with 10-fold cross validation.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	99.0664	0.995	0.991	0.993
Contraceptive Method Choice	51.5275	0.511	0.515	0.512
PRIMA	77.0833	0.766	0.771	0.767
Glass	79.9065	0.794	0.799	0.793
Page Blocks Classification	97.5333	0.975	0.975	0.975
Wine Quality	70.2532	0.791	0.703	0.744
Cleveland	55.7756	0.476	0.558	0.510
Bach Choral Harmony	75.3928	0.985	0.754	0.854
Bank Marketing	90.3895	0.893	0.904	0.896
Statlog (Image Segmentation	98.0087	0.980	0.980	0.980
Vehicle	76.0047	0.752	0.760	0.755
Diabetic Retinopathy Debrecen	69.1573	0.696	0.692	0.692
Anuran Calls (MFCCs)	88.3113	0.994	0.883	0.935
Wall-Following Robot Navigation	99.5418	0.995	0.995	0.995
Phishing Websites	97.3044	0.973	0.973	0.973
Turkiye Student Evaluation	86.6838	0.868	0.867	0.867
Dermatology	94.2623	0.943	0.943	0.942
Annealing	94.7368	0.948	0.947	0.948

4.3 Experimental Results

This table show the classification accuracy and precision, recall and F-score values for a Bagging (C4.5) classifier with 10-fold cross validation.

Table 4.11: Bagging (C4.5) classifier with 10-fold cross validation on cluster data set.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	96.4497	0.985	0.964	0.487
Contraceptive Method Choice	52.7273	0.522	0.527	0.523
PRIMA	74.4541	0.743	0.745	0.743
Glass	70.6349	0.691	0.706	0.697
Page Blocks Classification	96.2287	0.962	0.962	0.962
Wine Quality	64.3052	0.756	0.643	0.347
Cleveland	51.6854	0.464	0.517	0.487
Bach Choral Harmony	75.3239	0.987	0.753	0.425
Bank Marketing	89.1061	0.884	0.891	0.886
Statlog (Image Segmentation	94.0666	0.943	0.941	0.941
Vehicle	73.5178	0.722	0.735	0.726
Diabetic Retinopathy Debrecen	66.1337	0.661	0.661	0.661
Anuran Calls (MFCCs)	88.7344	0.994	0.887	0.469
Wall-Following Robot Navigation	99.5719	0.996	0.996	0.996
Phishing Websites	97.0890	0.971	0.971	0.971
Turkiye Student Evaluation	86.0745	0.861	0.861	0.861
Dermatology	88.5321	0.890	0.885	0.884
Annealing	91.5612	0.918	0.916	0.916

This table show the classification accuracy and precision, recall and F-score values for a Bagging (CART) classifier with 10-fold cross validation.

Table 4.12: Bagging (CART) classifier with 10-fold cross validation on cluster data set.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	96.2310	0.983	0.962	0.486
Contraceptive Method Choice	56.8182	0.560	0.568	0.562
PRIMA	75.3275	0.750	0.753	0.750
Glass	70.6349	0.691	0.706	0.697
Page Blocks Classification	96.3504	0.963	0.964	0.963
Wine Quality	62.0572	0.731	0.621	0.336
Cleveland	56.1798	0.709	0.562	0.313
Bach Choral Harmony	74.2049	0.975	0.742	0.412
Bank Marketing	89.1282	0.882	0.891	0.885
Statlog (Image Segmentation	93.6324	0.940	0.936	0.936
Vehicle	71.5415	0.697	0.715	0.701
Diabetic Retinopathy Debrecen	66.1337	0.663	0.661	0.662
Anuran Calls (MFCCs)	87.0885	0.993	0.871	0.464
Wall-Following Robot Navigation	99.5107	0.995	0.995	0.995
Phishing Websites	96.4857	0.965	0.965	0.965
Turkiye Student Evaluation	86.1605	0.862	0.862	0.861
Dermatology	89.9083	0.907	0.899	0.897
Annealing	89.0295	0.826	0.890	0.429

This table show the classification accuracy and precision, recall and F-score values for a Bagging (Naive Bays) classifier with 10-fold cross validation.

Table 4.13: Bagging (Naive Bays) classifier with 10-fold cross validation on cluster data set.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	90.1595	0.950	0.902	0.463
Contraceptive Method Choice	54.2045	0.547	0.542	0.540
PRIMA	71.6157	0.711	0.716	0.712
Glass	58.7302	0.622	0.587	0.580
Page Blocks Classification	88.2299	0.925	0.882	0.896
Wine Quality	44.1417	0.474	0.441	0.434
Cleveland	54.4944	0.511	0.545	0.525
Bach Choral Harmony	65.8127	0.978	0.658	0.393
Bank Marketing	85.7264	0.862	0.857	0.859
Statlog (Image Segmentation	79.3054	0.792	0.793	0.780
Vehicle	42.2925	0.515	0.423	0.400
Diabetic Retinopathy Debrecen	57.4128	0.706	0.574	0.525
Anuran Calls (MFCCs)	84.191	0.991	0.842	0.455
Wall-Following Robot Navigation	56.7584	0.655	0.568	0.568
Phishing Websites	92.1116	0.921	0.921	0.921
Turkiye Student Evaluation	83.3524	0.850	0.834	0.839
Dermatology	91.7431	0.923	0.917	0.915
Annealing	66.4557	0.838	0.665	0.688

This table show the classification accuracy and precision, recall and F-score values for a Bagging (SVM) classifier with 10-fold cross validation.

Table 4.14: Bagging (SVM) classifier with 10-fold cross validation on cluster data set.

Data sets	Classification accuracy (%)	Precision (weighted avg.)	Recall (weighted avg.)	F-score (weighted avg.)
Nursery	91.8575	0.961	0.919	0.470
Contraceptive Method Choice	54.5455	0.536	0.545	0.540
PRIMA	73.5808	0.731	0.736	0.726
Glass	42.0635	0.714	0.579	0.320
Page Blocks Classification	91.6058	0.907	0.916	0.900
Wine Quality	52.7248	0.601	0.527	0.281
Cleveland	52.2472	0.455	0.522	0.486
Bach Choral Harmony	72.4382	0.972	0.724	0.415
Bank Marketing	85.0500	0.826	0.850	0.812
Statlog (Image Segmentation	89.7974	0.899	0.898	0.897
Vehicle	67.9842	0.652	0.680	0.654
Diabetic Retinopathy Debrecen	65.5523	0.664	0.656	0.656
Anuran Calls (MFCCs)	84.4228	0.989	0.844	0.455
Wall-Following Robot Navigation	73.8838	0.740	0.739	0.737
Phishing Websites	93.8311	0.939	0.938	0.938
Turkiye Student Evaluation	85.9026	0.861	0.859	0.859
Dermatology	91.2844	0.916	0.913	0.911
Annealing	77.4262	0.550	0.774	0.321

Chapter 5

Conclusions & Future Work

5.1 Conclusions

After all the researching and studying and doing all the experiments we have tried to make an approach to solve the problem we face while dealing with large amount of data at some extent. We have used clustering and different ensemble learning methods to reduce the data and get the informative data with data border and center data colection algorithm. After collecting those informative data the dataset shrinks. This solve the bigdata problem effectively . On this informative dataset we applied ensemble learning algorithm . In some cases we have also got higher accuracy rate. In some cses we got lower accuracy. And also in some cases the result remain same. So here we can get the result with the smallest version of dataset. It depends on dataset and models. For now we haven't used large amount of data in real due to the lack of proper devices. So we have used small datasets for the experiment but in future we will improve ourselves and work with big datasets and we will also try to solve data imbalance problems as well.

5.2 Future Work

Many different analysis , experiment have been left for the future due to lack of time (i.e. the experiments with real data are usually very time consuming, requiring even days to finish a single run). Future work concerns deeper analysis of particular mechanisms, new proposals to try different methods, or simply curiosity. There are some ideas that I would have liked to try . This thesis has been mainly focused on collecting informative data from big data set. We have imposed our algorithm to perform our purpose. But in the future the following ideas could be implemented and tested: 1. Improve the round data collection algorithm . Make it more robust and accurate with analysis and experiment. Thus it can work with any shape of data and can collect more accurate informative data. 2. Experiment with more big data . For making more reliable our proposed approach we will experiment on more big data.

Bibliography

[1] T. G. Dietterich, "International workshop on multiple classifier systemsmcs 2000: Multiple classifier systems," *Technical report 98-17, James P. Anderson Co., Fort Washington, Pennsylvania, USA*, April 1980. 5

[2] D. K. B. Sajid Nagi, "Network modeling analysis in health informatics and bioinformatics," *Volume 2, Issue 3, pp 159–173 Classification of microarray cancer data using ensemble approach 5*, September 2013. 5

[3] G. D. Tan AC, "Ensemble machine learning on gene expression data for cancer classification," *Published in Applied bioinformatics 2(3 Suppl):S75-83*, February 2003. 5

[4] L. Breiman, "Bagging predictors," *Machine Learning , Volume 24, Issue 2, pp 123–140*, August 1996. 5

[5] P. B. K Machova, F Barcak, "A bagging method using decision trees in the role of base classifiers," *Acta Polytechnica Hungarica vol.3 No.2*, June 2006. 5

[6] J. F. S Dudoit, "Bagging to improve the accuracy of a clustering procedure," *Bioinformatics, Volume 19, Issue 9, 12 June 2003, Pages 1090–1099, https://doi.org/10.1093/bioinformatics/btg038*, June 2003. 5

[7] L. Breiman, "Bias, variance, and arcing classifiers," *TECHNICAL REPORT. Archived from the original (PDF) on 2015-01-19. Retrieved 19. Arcing [Boosting] is more successful than bagging in variance reduction 5*, January 1996. 5

[8] Z. Zhi-Hua, "Ensemble methods: Foundations and algorithms," *Chapman and Hall/CRC. p. 233-310. ISBN 978-1439830031*, 2012. 5

[9] M. Kearns, "Thoughts on hypothesis boosting," *Unpublished manuscript*, December 1988. 5

[10] L. V. Michael Kearns, "Crytographic limitations on learning boolean formulae and finite automata," *Symposium on Theory of computing. ACM. 21: 433–444. doi:10.1145/73007.73049*, 1989. 5

[11] R. E. Schapire, “The strength of weak learnability,” *Learning. Boston, MA: Kluwer Academic Publishers. 5 (2): 197–227. CiteSeerX 10.1.1.20.723. doi:10.1007/bf00116037*, june 1990. 5

[12] L. Breiman, “Arcing classifier (with discussion and a rejoinder by the author),” *Ann. Stat. volume.26 (3): pp 801–849. doi:10.1214/aos/1024691079*, November 3 1998. 5

[13] Y. Preund and R. E. Schapire, “A decisiontheoretic generalization of online learning and an application to boosting,” *Journal of Computer and System Sciences, volume.55(1): pp 119-139*, August 1997. 5

[14] Y. Freund, “A more robust boosting algorithm,” *arXiv preprint arXiv:0905.2138, 2009 www.arxiv.org*, 2009. 6

[15] A. J. S. Gunnar Rätsch, Bernhard Schölkopf, “Robust ensemble learning for data mining,” *PAKDD 2000: Knowledge Discovery and Data Mining. Current Issues and New Applications pp 341-344*, March 2003. 6

[16] T. K. Ho, “Random decision forests,” *International Conference on Document Analysis and Recognition, Montreal, QC, 14–16*, August 1995. 6

[17] ——, “The random subspace method for constructing decision forests,” *IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844. doi:10.1109/34.709601*, August 1998. 6

[18] E. M. Kleinberg, “Stochastic discrimination,” *Annals of Mathematics and Artificial Intelligence. volum 1 , issue(1–4), pp 207–239. CiteSeerX 10.1.1.25.6750. doi:10.1007/BF01531079*, September 1990. 6

[19] E. Kleinberg, “An overtraining-resistant stochastic modeling method for pattern recognition,” *Annals of Statistics. volume 24 pp 2319–2349. doi:10.1214/aos/1032181157. MR 1425956*, November 1996. 6

[20] E. M. Kleinberg, “On the algorithmic implementation of stochastic discrimination,” *IEEE Transactions on PAMI. volum.22 issue:5 DOI: 10.1109/34.857004*, May 2000. 6

[21] B. L, “Random forests,” *Machine Learning. volum.45 (1): pp 5–32. doi:10.1023/A:1010933404324*, 2001. 6

[22] Y. Tang, “Real-time automatic face tracking using adaptive random forests,” *Master’s thesis, Department of Electrical and Computer Engineering McGill University, Montreal, Canada*, June 2010. 6

[23] G. BA, “An application of random forests to a genome-wide association dataset: Methodological considerations and new findings,” *BMC Geneticsvolume 11, Article number: 49 http:// www.biomedcentral.com/ 1471-2156/ 11/ 49 DOI: 10.1186/1471-2156-11-49*, June 2010. 6

[24] D.-Y. K. Ji-Hyeon Lee, “Keyword annotation of medical image with random forest classifier and confidence assigning,” *In: International Conference on Computer Graphics, Imaging and Visualization, pp. 156–159*, 2011. 6

[25] M. Yaqub, “Improving the classification accuracy of the classic rf method by intelligent feature selection and weighted voting of trees with application to medical image segmentation.” *MLMI 2011. LNCS, vol. 7009, pp. 184–192. Springer, Heidelberg (2011)*, 2011. 6

[26] R. F. Ludmila I. Kuncheva, Juan j. Rodriguez, “A new classifier ensemble method,” *IEEE Transactions on Pattern Analysis and Machine Intelligence 28(10):1619-30 DOI: 10.1109/TPAMI.2006.211*, November 2006. 6

[27] P. D. Junshi Xia, “Hyperspectral remote sensing image classification based on rotation forest,” *IEEE Geoscience and Remote Sensing Letters Volume: 11 , Issue: 1 , page:239 – 243, DOI: 10.1109/LGRS.2013.2254108*, January 2014. 6

[28] P. E. P. S. B. Kotsiantis, “Combining bagging and boosting,” *World Academy of Science, Engineering and Technology International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering Vol:1, No:8*, 2007. 6

[29] N. G. Esteban Alfaro, Matıas Gamez, “adabag: An r package for classification with boosting and bagging,” *Journal of Statistical Software , Volume 54, Issue 2*, August 2013. 6

[30] J. Suutala and J. Roning, “Methods for person identification on a pressuresensitive floor,” *Experiments with multiple classifiers and reject option. (to appear) Information Fusion, Special Issue on Applications of Ensemble Methods*, 2008. 6

[31] N. Chawla and K. Bowyer, “Designing multiple classifier systems for face recognition,” *In N. Oza, R. Polikar, J. Kittler, and F. Roli, editors, Proceedings of the Sixth International Workshop on Multiple Classifier Systems, pages 407– 416. Springer, Berlin*, 2005. 7

[32] K. W. B. P. J. Flynn and P. J. Phillips, “Assessment of time dependency in face recognition,” *In International Conference on Audio and Video Based Biometric Person Authentication, pages 44–51*, 2003. 7

[33] S. S. R. Chellappa, C. Wilson, “Human and machine recognition of faces,” *A survey. In Proceedings of the IEEE, volume 83(5), page 705=740. Institute for Electrical and Electronics Engineers*, 1995. 7

[34] A. Samal and P. Iyengar, “Automatic recognition and analysis of human faces and facial expressions,” *Pattern Recognition, 25(1):65–77*, 1992. 7