THE REPUBLIC OF TURKEY

BAHCESEHIR UNIVERSITY

FEATURE SUBSET GENERATION FOR ENSEMBLE LEARNING USING FEATURE CLUSTERING AND MUTUAL INFORMATION

Master Thesis

HANA AMAR

ISTANBUL, 2016

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

COMPUTER ENGINEERING

FEATURE SUBSET GENERATION FOR ENSEMBLE LEARNING USING FEATURE CLUSTERING AND MUTUAL INFORMATION

Master Thesis

HANA AMAR

Supervisor: Asst. Prof. C. Okan şakar

ISTANBUL, 2016

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
COMPUTER ENGINEERING

Thesis Name: Feature Subset Generation For Ensemble Learning Using Feature Clustering And Mutual Information

Student Name: Hana Amar

Defense of Thesis Date: 25/5/2016

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assoc. Prof. Dr. Nafiz Arıca
Graduate School Director

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Sciences.

Asst. Prof. Tarkan Aydın
Program Coordinator

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Sciences.

Examining Comitee Members Signature

Thesis Supervisor
Asst. Prof. C. Okan şakar --------------------------------

Member
Asst. Prof. Tarkan Aydın --------------------------------

Member
Assoc. Dr. Fethullhah Karabiber --------------------------------

DEDICATION

I dedicate this achievement to my family, especially my mother who prayed for me to get to where I am. I also dedicate this success to my children and my husband who where inspiring and supportive throughout my studies.

Thanks to you all

ISTANBUL, 2016

HANA AMAR

الاهداء والشكر

أهدي هذا الإنجاز لعائلتي الكبيرة ، وخاصة والدتي التي صلت و دعت لأجلي ،لأصل لما وصلت إليه الآن، و أيضا أهدي هذا النجاح لأولادي وزوجي الذي كان اكثر شخص ملهم و داعم طوال فترة الدراسة و لولاه لما تمكنت من تحقيق نجاحي .

كذلك لا انسى الشكر الجزيل للعم يوسف لكل ما قدمه من اجلي , وجزيل الشكر والعرفان للمهندس أحمد الرفاعي .

ختاما لا يسعني إلا ان اشكر الجميع وكل من ساعدني ودعمني ماديا او معنويا او قدم لي النصيحة من بداية دراستي حتى وصولي هده المرحلة و تحقيق هدفي .

HANA AMAR
هناء عمار

ISTANBUL 2016
اسطنبول 2016

ACKNOWLEDGEMENTS

I can only thank God and all who helped, supported and advised me from the beginning of my studies until I reached my goal. Especially, my father-in-law **“uncle Yousif**” and engineer **Ahmed Alrfaie**.

In addition, I thank those who nominated me for the opportunity to study in Turkey. Also, I would like to thank my coordinator **Dr. Tarkan Aydın**, my supervisor **Dr. C. Okan Şakar** and all teachers at Bahcesehir University who taught me English language and all lecturers who worked to provide comfort for all students.

ISTANBUL, 2016 HANA AMAR

ABSTRACT

FEATURE SUBSET GENERATION FOR ENSEMBLE LEARNING USING FEATURE CLUSTERING AND MUTUAL INFORMATION

Hana Amar

Computer Engineering

Supervisor: Asst. Prof. C. Okan Şakar

May 2016, 32 pages

Ensemble Learning (EL) is considered one of the most effective techniques which is applied to address supervised machine learning problems. In this thesis, we used clustering and feature selection algorithms in order to generate multiple feature subsets from a single feature set to apply EL method. For this purpose, first we clustered the features and obtained many feature subsets. Then, we fed these subsets of features to support vector machine classifier (SVM) to get individual class predictions and combined those predictions using majority voting. After that, we gave the predictions to minimum Redundancy-Maximum Relevance algorithm (mRMr) feature selection algorithm and ranked the feature subsets according to their mRMR scores for generating diverse and accurate subsets which are vital factors for EL. Experimental results on various biomedical datasets show that our method improves the single set accuracies.

Key words : Ensemble Learning (EL), Feature Clustering, Feature Subset Generation (VG), Minimum Redundancy-Maximum Relevance Algorithm, Support Vector Machine (SVM).

ÖZET

TOPLULUK ÖĞRENME İÇİN ÖZNİTELİK KÜMELEME VE KARŞILIKLI BİLGİ KULLANARAK ÖZNİTELİK ALTKÜMESİ OLUŞTURMA

Hana Amar

Bilgisayar Mühendisliği

Tez Danışmanı: Asst. Prof. C. Okan Şakar

Mayıs 2016, 32 Sayfa

Topluluk öğrenmesi (TÖ) gözetimli makine öğrenmesi problemlerinin çözümünde uygulanan en etkili yöntemlerinden birisidir. Bu tez çalışmasında TÖ yönteminin uygulanabilmesi için bir öznitelik kümesinden çok sayıda öznitelik kümesioluşturulması amacıyla kümeleme ve öznitelik seçimi algoritmalarının kullanıldığı bir yöntem önerilmiştir. Bu amaçla öncelikle öznitelikler kümelenmiş ve çok sayıda öznitelik alt kümesi elde edilmiştir. Daha sonra bu öznitelik alt kümeleri bireysel tahminlerin elde edilebilmesi için destek vektör makineleri (DVM) sınıflandırıcısına beslenmiş ve çoğunluk oylaması tekniğiyle birleştirilmiştir. Bu aşamadan sonra, bu sınıf tahminleri minimum Artıklık-Maksimum İlgililik (mAMİ) öznitelik seçimi yöntemine verilmiş ve öznitelik alt kümeleri TÖ yöntemi için çok önemli faktörler olan farklı ve nitelikli alt kümeler oluşturulması için mAMİ skorlarına göre sıralanmıştır. Biyomedikal veri kümeleri üzerinde yapılan deneysel çalışmalar yöntemimizin tekil küme başarımlarını geliştirdiğini göstermektedir.

Anahtar Kelimeler: Topluluk Öğrenmesi, Öznitelik Kümeleme, Öznitelik Alt Kümesi Oluşturulması, Minimum Artıklık-Maksimum İlgililik Algoritması, Desktek Vektör Makineleri.

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

AL :	Active Learning
AMD :	Adaptive Maximum Disagreement Query Strategy
ATS :	Active Thompson Sampling
EL :	Ensemble Learning
MCC :	Matthews Correlation Coefficient
MRMR :	Minimum Redundancy Maximum Relevance
MVL :	Multi- View Learning
QBC :	Query By Committee
RBF :	Radial Basis Function
RMKMC :	Robust Multi-View K- Means Clustering
SVM :	Support Vector Machine
VC :	Vapnik- Chervonenkis dimension
VG :	View Generation

1. INTRODUCTION

Most experts assure that the advanced technologies and developed programming endeavor to find out a suitable hypothesis or problem solution (class label) from data not completely available. Hence, some modern techniques and algorithms are successful to train a model by using suitable data sets, learn a machine, and get prediction. That was why, classification methods are one of many effective supervised applications in machine learning and data mining. Where all instances in any dataset are represented by using their features with known labels. Compared to unsupervised learning where the samples are unlabeled. Moreover, other proficients thought about integration of some techniques and algorithms or subsets together for optimization of algorithms execution and effectiveness. So, from here came ensemble learning EL technique. That was the reason to consider about more controversial point which is (whether View Generation VG by clustering of features, using Support Vector Machine SVM classifier and applying EL method according MRMR scores could reach a perfect performance or not). And when could the best outcomes be obtained. Also, many subsets were generated should be sufficient, accurate and diverse which is the base condition in EL. Additionally, using clustering and MRMR help to achieve this purpose based on more influential features. MRMR in this study was employed for choosing different views which have high score. Basically, the goal from this thesis was producing multi views and then, applying EL technique on different data sets. Also, we employed many methods together to reach the aim (e.g. k-means algorithm, SVM classifier, MRMR, EL technique). At the end, the aim of this research was achieved where the EL accuracy was improved compared to using single data set with all features. For more clarification look at the third and fourth chapters. Generally, all explanation was divided into four sections and five chapters. the beginning will be with classification introduction.

1.1 CLASSIFICATION

As mentioned above, it is an important approach. Also, its applications were used in many areas and domains in real life. So, they can make good assumptions for samples class in the bank or for the customers in the market e.g. because of the past experiences and customer manner. Also, class value can be predicted sometimes if the past data similar to the future data, and compute the accuracy [4] [5]. Probability should be taken into account with classification, because most its algorithms tend to give probabilistic results if the point being one of possible classes i.e. with highest probability [6].

In addition, supervised learning where values of class of training set is available for both correct identified instances and incorrect ones. But in unsupervised, the learning could be obtained by clustering methods. So, data divided into groups and categorized based on some measures of correlations or distances that were taken into account in this thesis which worked cooperatively with supervised algorithms.

Classification has a lot of more popular, effective, and perfect applications, and algorithms which were employed, such as k-Nearest Neighbor Classifier, Decision Tree, Naive Bayes (linear classifier), Logistic Regression Algorithms, Support Vector Machine Classifier, Neural Networks [7][9].

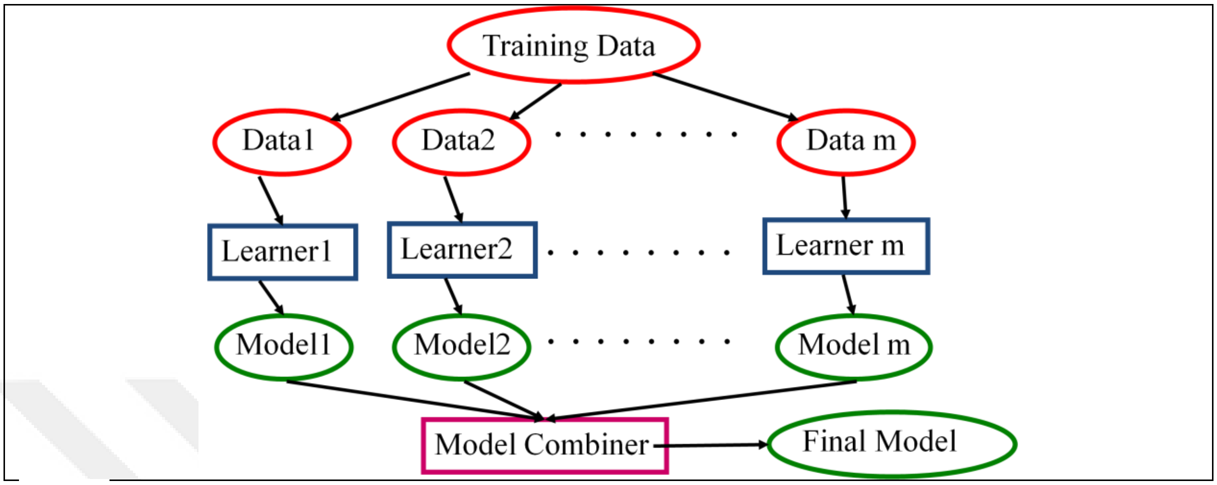
Also, there were some factors which identified any algorithm and affected the results. The experts should be careful about these factors, such as number of training examples, dimensionality of the feature space, number of attributes, data type (maybe linearly separable),and dependency of features [11].

Moreover, classification has two kinds of problems binary and multi class classification problems [12]. In binary classification, there are only two groups (outputs) like all data sets in this research. Whereas in multiple classes classification, it will be more than two groups for assigning a sample to one of them. Although there are some problems that faced in practical life such as: text categorization (e.g. spam filtering), fraud detection, optical character recognition, machine vision (e.g. face detection), natural language processing (e.g. spoken language understanding), market segmentation (e.g. predict how customer will act), and bio informatics data (e.g. classify proteins according to their function and type) [10].

1.2 ENSEMBLE LEARNING

It is a broad issue of machine learning styles where multiple subsets were produced and many hypothesis were trained to solve the same problem. In contrast to ordinary way of machine learning approaches which tried to learn one hypothesis from a single data set. Ensemble methods were aimed to construct a set of diverse hypotheses and combine them to improve performance of classification. By the way, it was the main idea of this thesis. Empirically, generating multiple subsets, combining different features and various models was suitable to get better predictive performance. that means, higher accuracy was got with integration of some learners. Some ensemble techniques (especially bagging Sampling with replacement) were reduced problems related to over fitting and variance of the training data. The common algorithm of EL were Bayes Optimal Classifier BOC. It works by an ensemble of all the learners in the hypothesis space and bootstraps them. Then, it build classifier on each. Bootstrapping aggregation abbreviated as bagging, involved having each model in the ensemble vote with equal weight. In order to promote model variance. Also with bagging, each model randomly drawn subset and trained one learner with each set. In the other hand, the ensembles were not always Successful with different types of data and some models were not divers according their MRMR score. Additionally, most ensemble approaches were promoted diversity and independency among the models were combined (it is a main key in EL technique). Furthermore, there were other approaches of ensembles such as stacking, boosting [14] [15] (where resulted a larger increasing in accuracy than bagging). So, choosing the right integration was manner more than it was science. Additionally, voting method in ensembles used for classification techniques whereas averaging for regression approaches. Besides, decision trees algorithms were commonly applied with ensembles (e.g. Random Forest) [17] [16]. Also, slower algorithms can benefit from ensemble techniques too. Next figure explain bagging way which was used in this study and its resource .

Figure 1.1: Ensemble learning using bagging technique



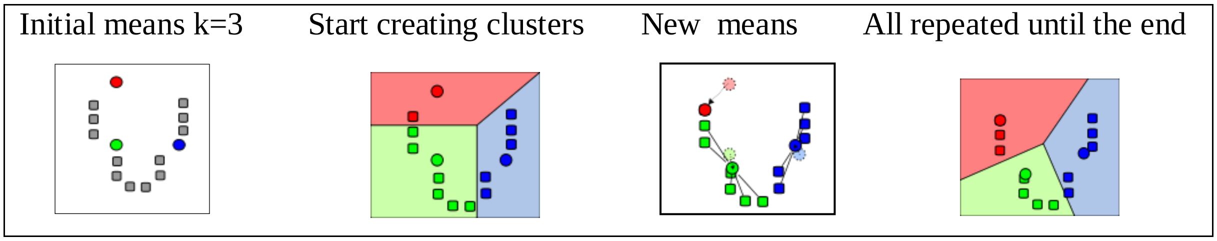
Source: www.cs.cornell.edu/Courses/cs4700/2008fa/PPT/CS4700-EL.ppt

1.3. CLUSTERING:

Clustering deals with unlabeled data. This method is based on grouping each similar samples together into different clusters which is considered as a goal of clustering technique. Also, it is a normal job of k-means**.** Although clustering is an unsupervised learning technique, it is also employed as a preprocessing step with supervised algorithms in some machine learning applications. In this study, K-means clustering algorithm was applied to cluster features in order to generate multi views that is a main aim of this thesis. Although some problems like time complexity were existed with a very big data [18][19][20].

Clustering technique can be used in many applications. Also, there are some characteristics were required such as (scalability, different features, dealing with noise, insensitivity, interpretability, and usability) [21][22]. K-means algorithm was a heuristic and exclusive approach. Also, it was called a centroid model as figure (1.2) show.

Figure 1.2: K-means algorithm



Source: https://en.wikipedia.org/wiki/K-means_clustering

1.4 SUPPORT VECTOR MACHINE SVM

Support vector machine was the most popular successful algorithm in classification machine learning [24]. It was a base classifier of this thesis. It worked well with our data. Also, Its outputs were often affected by the data which may be separable, non separable, or non linear data [23]. Also, the support vectors could define the discriminant. Therefore, the optimization for kernel function parameters could improve the results which exactly happened here. The main idea of support vector machine was the two classes were separated by hyper plane (the discriminant) with maximum margin between classes [25][28]. The aim of SVM in this research produced a model on training set and predicted class label on the test set. If there was a big number of features the processes in SVM will be very slow. But, with using kernel trick might be more faster. Thus, basic kernels are: linear, polynomial, radial basis function RBF, sigmoid and their parameters. Besides, the large margin separators have lower VC dimension which affect hypothesis H. Then, it will be a small difference between the training and test error scores. So, the maximum number of points that can be split by H. VC meant the Vapnik Chervonenkis dimension of H and measures the size of the model class (i.e. how many dimension was there) [27][26].

2. METHODS

As mentioned in introduction chapter. The combination of more than one approach was very useful. It was clear in the results. All steps will be clarified in next sentences. So, the original data was transposed as a first step of an algorithm. After that, we have clustered data attributes by using k-means algorithm. Thus, we tried some distances and many clusters numbers to find appropriate ones. The five clusters was appropriate where grouping every similar attributes together. That was run many times with using bootstrapping (shuffled by replacement). We have supposed B=10. As a result, fifty subsets (Views) were generated. And then, all features subsets have been sent to the SVM classifier once to get predictions.

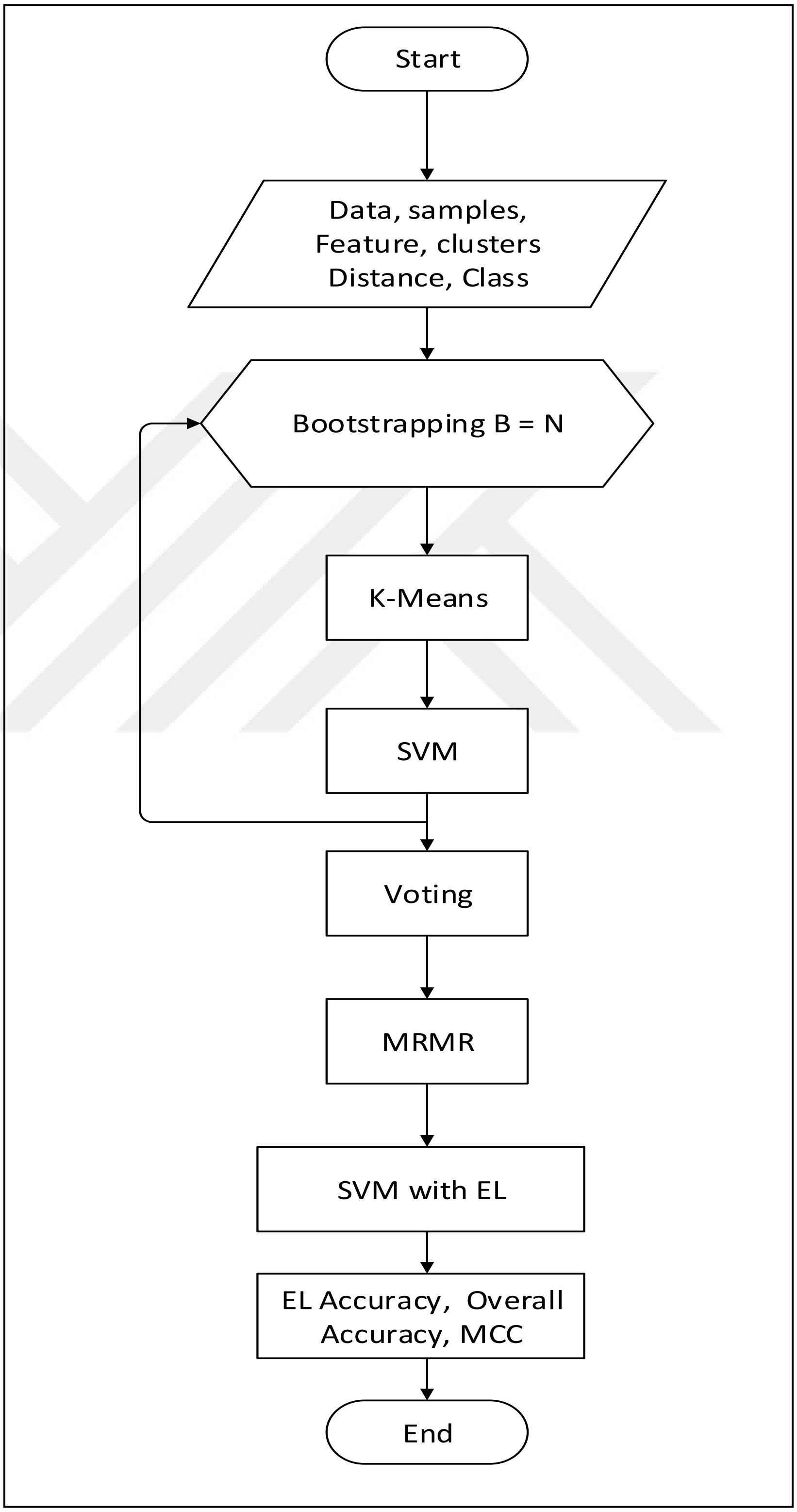
The data was divided into two groups train and test set (before sending to SVM). So, the features were placed into their views according to the features indexes (within k-means part). The run was inside for loop where B and clusters numbers were mentioned above. Therefore, five subsets (models) were produced initially. Those were sent to the classifier 10 times so, fifty subsets (hypothesis) with their target values were generated at the end. Then, computing number of correct classified samples of each subset and save it into one matrix of all results were done. In addition, number of misclassified instances was computed.

Furthermore, saving all subsets for each model independently in its matrix. After that, the voting between real class and each model prediction was gotten. Putting the outcomes in different matrix was done. So, there were fifty results to send them with real class to the MRMR function and rank them which gave the diversity score. Since it was the main condition in this work.

Based on diversity score the best **N** views were selected and trained by SVM with ensemble learning once to get target values (i.e. sending first sub set, then the next two views, 3 views, 4, 5,...., until **N** subsets together).

And then, the accuracy was obtained again each time which called ensemble learning accuracy. Also, Matthews Correlation Coefficient MCC was employed to compute the balance measurement between positive and negative instances, all steps will be more evident in the next figure.

Figure 2.1: Flowchart of an algorithm steps



2.1 ACTIVE LEARNING TECHNIQUE AL

Ensemble learning approach could be applied with active learning (AL) specially if the decorate strategy was implemented [1] [2] according previous literature. AL is a supervised machine learning technique where labels of the data is known. The learner is in control where the user or expert could be asked to choose suitable training instances. The major idea behind active learning is that a machine learning algorithm can achieve better performance with small number of training samples.

Also, we can say that AL is a special case of semi-supervised machine learning because they have same goal that is achieving a good learning performance (classification accuracy). But, they have different approach that Semi-supervised learning employs unlabeled data whereas an active learning handles with labelled instances. AL faced several problems in many applications such as unlabelled data that an active algorithm is provided. It must pay to reach any value. So, that will increase the cost in contrast to labeled samples [35].

Another strong assessment in active learning work that there was a good benefit from known labels for training part. To illustrate, the labels come from an empirical experiment (e.g. biological, chemical and clinical studies). Then the experts always expected the results because of the usage of experimental setting.

Active learning strategies tried to limit the number of labelled samples which was needed to train an effective classifier as natural resumption in spam filtering applications.

Batch mode active learning is another type of active learning, its algorithms select multiple examples at one time [3]. In contrast other active learning algorithms that select only one example to ask its label at one time. This learning has been successfully implemented in many applications.

All the previous points depends on which strategy was used such as:

I. **Uncertainty sampling US:** Labeling the examples which the current model was least certain as the correct output should be (decision tree classifier by an active naive Bayes e.g.). Also, it is a simplest and most popular used strategy of query.

II. **Query by committee QBC:** It is a group of models where the training on the current samples which are known class values. And voting will be on the unlabelled data. Then, labeling those instances which the learner disagrees by a maximum entropy. The QBC approach tend to maintain a committee of models, which are all trained on the current set of labelled instances.

III. **Expected model change:** The most influential samples could be labeled to change the generated hypothesis. If we knew the labels Since discriminative probabilistic models are usually trained with gradient based improvement. Thus, the change can be measured by the magnitude of the gradient.

IV. **Expected error reduction:** Saving the points which were the most decreased for generalization error of learners.

V. **Variance reduction:** Selection for instances which reduced output variance that was one of the many effective factors of error.

VI. **Balance exploration and exploitation:** Obtaining the classes of samples is defined as a difficult choice between the exploration and the exploitation over the instances representation. This strategy worked by handling the active learning problem (Bouneffouf algorithm e.g.). To give another example of AL which was a sequential algorithm such Active Thompson Sampling ATS. It took one instance from the pool and asked the user to label this point in each iteration.

2.2 VIEW GENERATION AND MAXIMUM DISAGREEMENT FOR CLASSIFICATION

As we know the goal of active learning is to obtain best classification performance with fewer labeled samples. In contrast to passive learning where the training set is often chosen randomly without interaction with the classifier.

Furthermore, view generation that means, the single feature set was broken down to several sub sets. Also, in AL from few instances with using views generation could result many sub sets [33]. Producing diverse subsets was a main key of AL. Sometimes MVL Multi View Learning is not required to do with the algorithm, because of large number of features. But there are some limited researches on sub sets generation due to the performance will improve even small or large number of attributes.

The generation of different subsets corresponds to splitting of attributes which generalizes the task of feature choosing [32]. There are many ways and strategies to produce subsets that is similar to taking data from more than one resource. Also, there are two conditions for subsets generation as sufficiency and diversity. A partition of features is to several disjoint subsets with using suitable manner like Adaptive Maximum Disagreement AMD query strategy [31]. This strategy was applied on hyper spectral data sets and gave perfect outcomes according to this literature [30].

2.2.1 AL, EL, Simi Learning Based On Maximum Disagreement

Disagreement is important point of many techniques in machine learning to reach the best model among set of learners that worked on the same data set. Also, the handling only with labeled data in AL whereas simi supervised learning works on unlabelled samples. In addition, other types of learning need a big number of training instances. Thus, some experts conducted reasonable techniques such creation multiple subsets. And then, fitting them to the base function (e.g. classifier) and getting outputs. After that, they decided which one is divers (different from each other) and sufficient model according some characteristics or attributes. (i.e. measure the disagreement between sub sets). Generally, disagreement between subsets should be small if they are very close to the target output.

So, this manner can be viewed as an indication of how far every subset is from the selected model. The disagreement was between group of hypothesis. At the end, the set which was maximum disagreement should be selected [34]. Moreover, other researchers consider about multiview as Adaptive Maximum Disagreement AMD [36] [37]. The purpose of classification is to learn a hypothesis h: X → Y to predict value of the class correctly. A learner could access the whole data with all variables with a single subset, while the available features were chattered into disjoint subsets to generate multiple subsets. So, it is assumed that each subset is sufficient to learn the target model. That means, the hypothesis from any subset agree with this aim. Individually, the disagreement among all potential models could be computed with various features. If at least two subsets were disagree that may not decrease redundancy of features. It focus on the most informative samples (i.e. relevant features) with the maximum disagreement at each iteration. AMD actually handles the variety of different subsets. So, samples that have higher contradiction indicate to large difficulty in defining the decision. However, bootstrapping all the views to learn from the training set help to enhance their agreement. The agreement of different hypotheses (learners from different views) represents the intersection of those hypothesis with the target. Thus, any model from those generated will not be too far from the desired model. Also, a sampling within CAMD which is a smaller candidate subset that computational load is certainly reduced. Briefly, the subsets generation idea by this strategy could be applied with EL, AL, Simi supervised techniques.

2.3 MRMR ALGORITHM

Maximum relevance minimum redundancy algorithm (MRMR) is a very helpful method to select features based on mutual information (its ordinary job) [38]. The max relevance between features and class means that most attributes which have big impact on class. Also, min redundancy means that decreasing number of repeated variables. That is why, feature selection was believed as an important issue for classification applications. Thus, many experts applied the algorithm and conducted many studies for the good features how to be selected based on the maximal dependency and mutual information. However, because of the difficulty in implementing the maximum dependency case, the researchers achieved to another approach with same job which was called MRMR [39]. Thus, One of the most popular methods to understand max dependency was maximal relevance by selecting the features with the highest relevance to the target class c. Relevance is recognized as relation between features or mutual information where the selected features have the largest score of mutual information [40]. The optimal recognition phase that means, it minimized classification error. Also integration MRMR with other advanced feature selectors (e.g. wrappers) could give meaningful features to use in classification algorithms. At the end of many surveys, their results show that implementation of MRMR with other approaches to choose attributes were very effective in different applications. Also, here in this research MRMR was used for decreasing repeated attributes and choosing diverse views. Also, MRMR score should be high between each feature and class label in the same subset. But should be low between divers subsets). To give an example, if we have two different learners S1, S2 and a classifier, the features of S1 could be more characteristic. since classification error on S1 was smaller than error on S2 using the same classifier according to the outcomes actually.

2.4 MULTI VIEWS K- MEANS CLUSTERING ON BIG DATA (heuristic algorithm)

In the past, a lot of data are obtained from multiple sources. Although each view could be individually used for finding models by clustering as unsupervised learning and used for solving problems of single subset. The clustering performance could be more optimized by improvement its parameters especially if it used for features clustering. K-means clustering is a method of cluster analysis which aims to partition all points into specific clusters, where each point belongs to different cluster. However, this technique most time could not give high performance with big data. Also, traditional data processing applications are not enough. For that reason, some methods were invented such as (robust multi view K-means clustering RMKMC method) to cluster data of large size. This approach used to describe complex, and difficult data sets including capture, storage, search, visualization, analysis, and sharing. In that study, the performance was evaluated in group of data sets to prove the aim of RMKMC [42] [44].

2.5 ENSEMBLE LEARNING (EL) AND ITS SCHEMES

Ensemble learning (EL) is a supervised learning approach. Where a set of learners were combined. That might be such combination of some algorithms with same data, or generation of multiple subsets from a single data set. And then, using voting measure [50]. EL technique is utilized to optimize the performance of several applications. Most time those outputs are better than results which obtained from any single algorithm or method. Especially, when there is diversity among the views. Bagging is the most type of EL that decreases problems related to over fitting in the training part. It is often possible to build perfect ensembles because of three essential reasons as statistical, computational, and representational problems. Several types of EL were existed like Bayes Optimal Classifier BOC that is one of more popular EL schemes. BOC was an ensemble of all the subsets predictions on average measurement. Bagging word came from bootstrap aggregation scheme. It was the best approach and one of the simplest ways to reach good accuracy. It was primarily constructed for classification especially decision tree algorithms. Moreover, (voting) could be employed with classification models and (combining the predictions by averaging) for regression algorithms. Bootstrapping approach is sampling with replacement to all subsets [46] whereas boosting technique has another idea that the integration for three weak models to build a strong one. It could be called as averaging approach. Also, boosting is a powerful method in the last ten years. It is the most applied ensemble technique. Additionally, it is originally constructed for classification applications (with voting). However, it could be used to regression (with averaging) after some extensions and additions. AdaBoost is a short form of adaptive boosting. It is the most common boosting approach. Moreover, stacked generalization (stacking) is a various method of integration multiple hypothesis. Stacking is less widely utilized than boosting and bagging where it was employed to build models of different kinds (various algorithms). The random subspace method (RSM) is other type of EL to combine learners. It is trained at random selection and the outcomes were integrated by a straightforward majority voting [47].

2.5.1 Summary Of Some EL Studies

Some experts specified four approaches of multiple algorithms combination. Firstly, setting for weak classifiers. Then, the outputs of bagging, boosting, and the random subspace method were compared. So, they reached to bagging was useful for weak and unsteady algorithms. Boosting is helpful only for weakly, simple learners which were built on a big size of training instances [48]. The random subspace method is advantageous for weakened and unsteady algorithms that were applied on few number of examples. Other researchers achieved other result [49]. That was the averaging versus the voting measure with multiple models. Where averaging often outperforms voting for Gaussian error of appreciation whereas heavy tail function vote could be winner. This way is used in economic issues. besides, new methods were invented by the other experts. Those approaches are (stacking by extending this technique with probability distribution), and (multi response linear regression). Thus, another researchers suggested a framework to construct hundreds or thousands of algorithms on small data sets. Their results showed that the new approach is scalable, fast and accurate. According all surveys mentioned above.

2.6 ACCURACY OF OUR DATA SETS ON PREVIOUS STUDIES

Some experiments were performed on SVM classifier for **Parkinson's Disease** data set. Where sigmoid kernel function was chosen and all 22 features were used. So, The obtained accuracy was about 86% [51]. While other outcomes were gotten from utilization K-nearest neighbor algorithm with various k values based on euclidean distance. Also, PD single data set was used. But, accuracy was between 80% and 85% [52]. All that work was without applying EL technique. In contrast , the EL accuracy after running our algorithm was 86% [table 3.1]. Moreover, the decreasing on the cost(time, memory) was achieved. Because of features clustering method and selection to the diverse subsets.

Other researches were referred to some experts applied SVM classifier without Ensemble learning approach on **Breast Cancer** single data set. Also, all 34 attributes were employed So, they got next result by optimization the kernel trick. Thus, accuracy was 94 percent when kernel function was linear whereas it was 97 percent with RBF kernel function [53]. Compared to EL accuracy was 90 percent with our algorithm [table 3.2]. But, we decreased the cost(time, memory). Because of clustering the features and choosing the different views. There were another studies show that many classifiers were implemented on **Madelon** data set. So, they got next result [54]:

Table 2.1: Accuracy Of Madelon data set on previous studies

Classifier	BART	CART	LR	NB	RF	SVM	TAN
Accuracy	76%	78.2%	60%	59.8%	67.1%	62%	54.2%

Where, the EL accuracy after applying our algorithm was 73 percent [table 3.3]. It was less than accuracy of other classifiers which mentioned above. But, using EL approach led to minimize the time and memory. Because of combination between accurate and most diverse subsets. Whereas another survey explained that **Madelon** data set gave better accuracy with KNN than SVM algorithm. When the single set with 500 features was fitted to the classifiers. Where accuracy of KNN was between 80 percent, 85 percent. Accuracy of SVM was less than 62 percent [55].

2.7 CLUSTER ENSEMBLE SELECTION

Selection of ensembles gave many different techniques to cluster any data into appropriate groups and perfect performance with small number of points. The approach of ensembles was designed based on two major conditions which were diversity and quality. Where, it could not reach the goal with applying one of those factors. We can obtain divers clustering results by using different ways or the same ways on the same data set with improving for parameters. So, how will we choose the suitable one. Generally, the main idea behind cluster ensembles was that combining their clustering outputs [56].

The first technique supposed a joint objective function that Integrated both conditions. While the second approach assigned different likelihoods into sets. Then, the one with best quality was chosen from each set. The last method was drawing a scatter plot of samples, where each observations corresponded to a pair of clustering suggestions and represented by the quality and diversity average. and then, selected the solution by using convex hull. In supervised learning, the quality and diversity are clearly defined issues where quality measured the accuracy of the ensemble points and diversity specified the difference in the solutions. But, for unsupervised learning, these two concepts were not so clearly clarified. There were no any external objective function such accuracy to measure the quality of the clustering predictions. The simple selection strategies by employing an internal measurement of quality according to another function such that created by Strehl and Ghosh. It computed the diversity by normalizing mutual information between clustering results. Therefore, the ensemble with lower value had a big consistency. As a new strategy, they take diversity into account when ensemble contained one prediction of highest diversity score (quality i.e.) [57].

2.8 FEATURE SELECTION

Feature selection is significant point in machine learning and data mining applications. It applied as preprocessing step. Many studies indicated that relevant attributes has a big impact on the outputs effectively in the practical life [58]. Therefore, many attributes selection ways have been existed to extract the relevant features or subsets to achieve required outcomes of classification and clustering. That by measuring of attributes relevance, general algorithms, evaluation function, and the characters of attributes themselves . As it known, machine learning techniques have difficulty to deal with a big number of data which is a motivating challenge for experts. Feature selection is one of the most sufficient and essential methods in data preprocessing issue where the operation of detection to all relevant features firstly was done. And then, exclusion irrelevant, repeated, or outliers ones were as a next step. This way quickened data mining algorithms and got better performance. There were several approaches to select desired features such as mutual information function and computing the diversity between features with their class. Also, feature choosing methods improves learning, speed of learning, or decreasing the complexity of the hypothesis. It was according to many literature [59].

3. EXPERIMENTS

3.1 DATA 1

Data1 (PD) was Parkinson's Disease Detection data set from Oxford University which worked with National Centre for voice and speech. The experts and professionals recorded the speech signals of patients to get the attributes (features) which represented the columns. The data is consisted of 192 samples which represented rows. Each row indicate to one of 192 voice recording from 31 individuals who are taken their 6 recordings per one. Also, 23 features were provided, 22 of them were recordings. And 23th attributes described the class label where set to 0 for healthy and 1 for PD (i.e. who have disease) .

Some processes were done on the data to be more clear and easy handling, we avoided first row and column.

i. First column is name of person.
ii. First row is name of attribute.

There are more details about our data, they could be explained in the Parkinson UCI repository web site address.

3.2 RESULTS OF PARKINSON'S DISEASE DATA SET

Numerous factors could affect the outcomes such as instances number, how many sets which was divided, number of clusters, distance used in k-means, optimization of svm parameters (box constraint C with linear kernel function, C and Sigma with rbf function), number of features, number of class groups ,and data type itself (integer, real e.g.).

visually from the views plots, we notice that most of the misclassification in training part because of data1 size. That was why, the best training performance with a large amount of instances. Better outcomes were gotten by Optimizing SVM parameters.

When linear kernel function was used (try many values of C as 2e-1,4e-1,...until C=9e-1) with k-means parameters (K=5, correlation and euclidean distance). The classification on training set was more reasonable and accurate with correlation distance, as will be shown in the next two figures (3.1) (3.2).

Figure 3.1: View3- linear kernel, C=5e-1, correlation distance

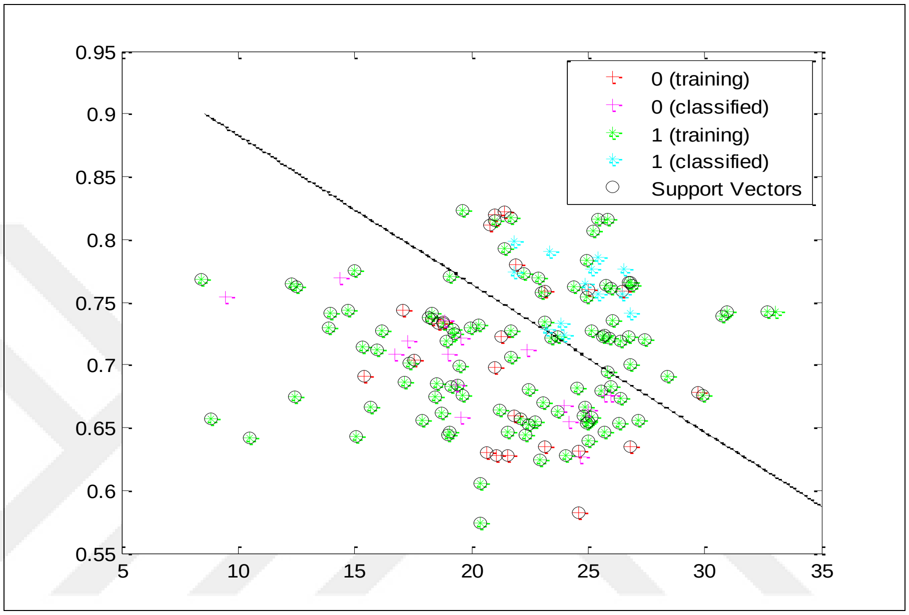
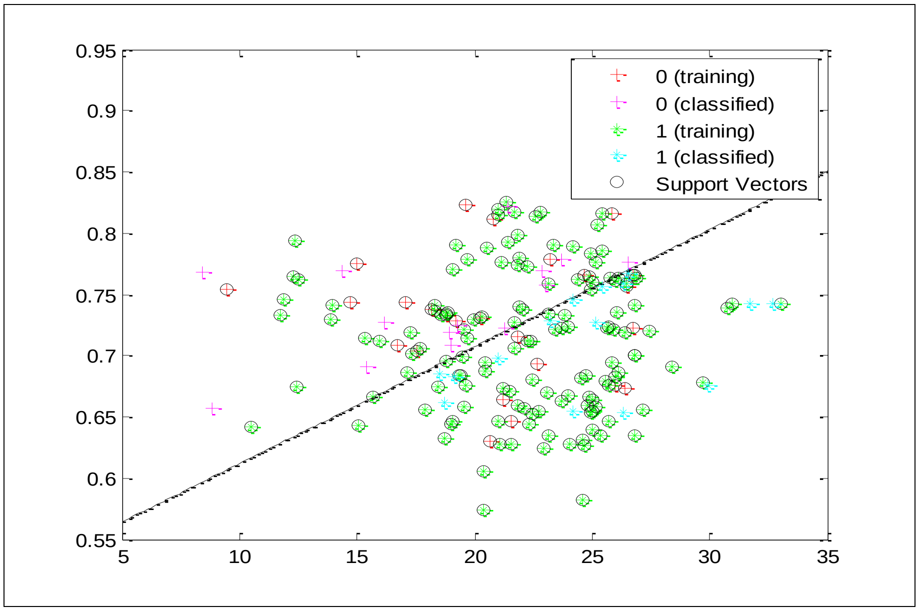


Figure 3.2: View3- linear kernel, C=5e-1, euclidean distance



In contrast to, the results were more perfect with rbf kernel function (where C=8e-1, sigma=0.7 e.g.) than linear kernel like next figures explained (3.3) (3.4). Also, using rbf kernel with correlation distance better than choosing euclidean distance with rbf.

Figure 3.3: View7- rbf kernel, C=8e-1, sigma=0.7, correlation distance

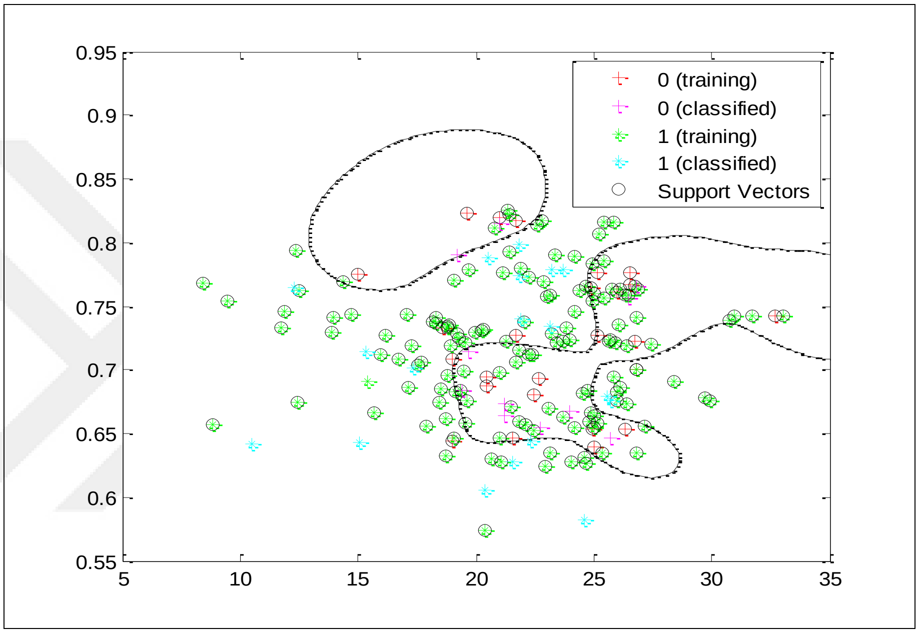
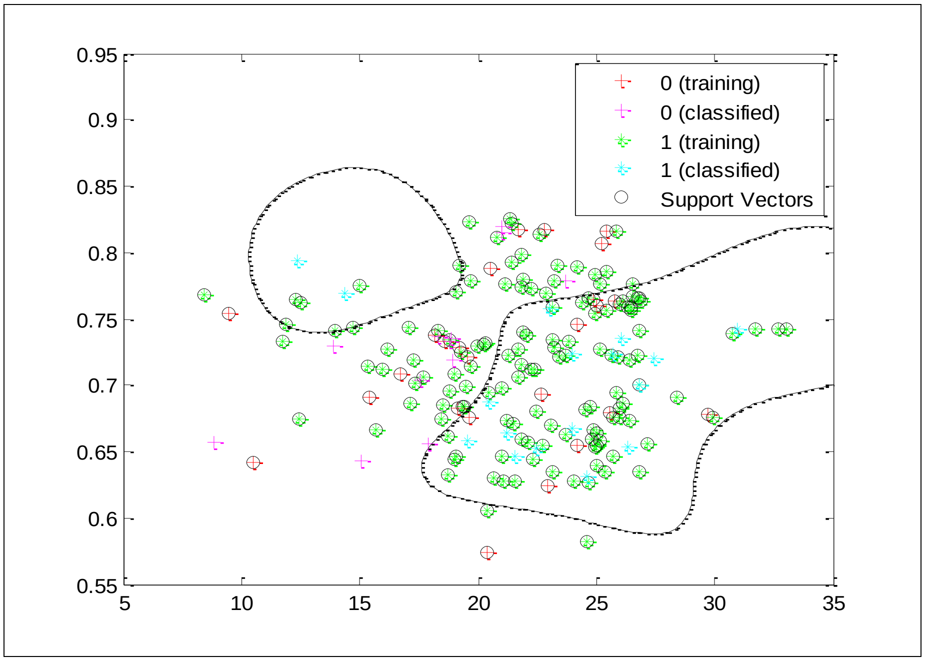
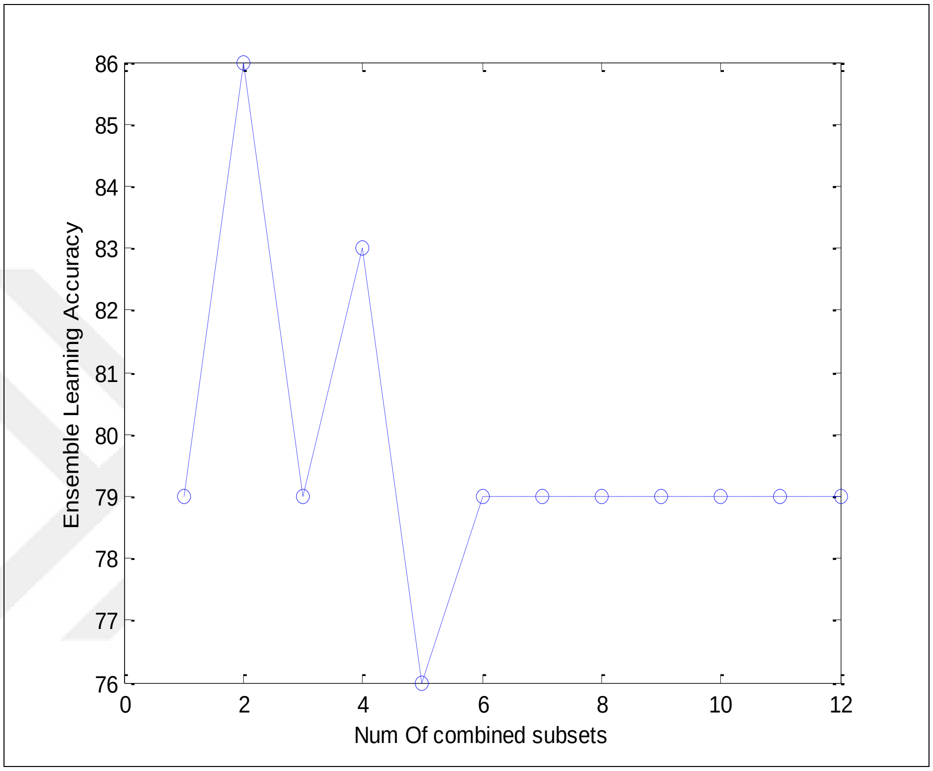


Figure 3.4: View7- rbf kernel, C=8e-1, sigma=0.7, euclidean distance



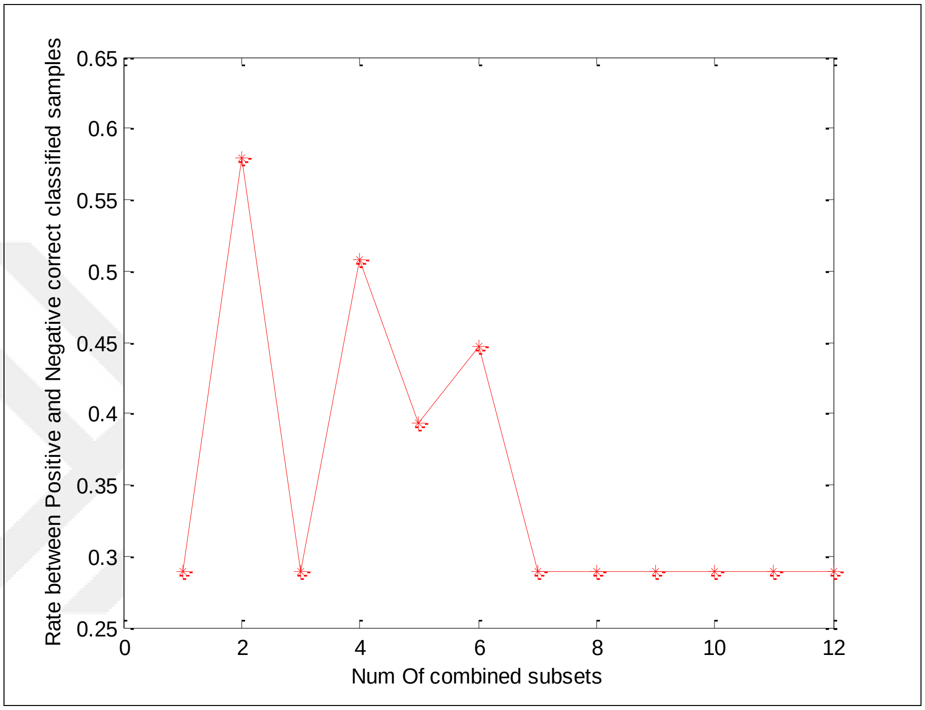
In addition, the significant idea of this research was reached. Where the best ensemble accuracy with combination two views together such figure (3.5) described.

Figure 3.5: Accuracy of ensemble learning



There is another measurement to see good outputs . So, Matthews correlation coefficient (MCC) was applied.

Figure 3.6: MCC with ensemble learning



Improved score was when merging for four views was done whereas the best with merging two subsets together (compared to Acc. of EL). And better balance between positive and negative samples as it was shown in previous plot.

Table 3.1: Results of Parkinson's data set

Subset No	Features Number	Features No	Individual Accuracy	Combined subsets Num	EL Accuracy	TN Rate	TP Rate	MCC
1	6	4, 5, 6, 7,8 15	79%	1	79%	91.3043%	8.6956%	0.2894
2	4	1, 2, 3, 16	48%	2	86%	84.0000%	16.0000%	0.5797
4	6	9, 10, 11, 12, 13, 14	69%	3	79%	91.3043%	8.6956%	0.2894
6	3	18, 19, 22	62%	4	83%	83.3333%	16.6666%	0.5076
16	10	4, 5, 6, 7,8, 15, 17 19, 20, 22	79%	9	79%	91.3043%	8.6956%	0.2894
21	3	1, 2 , 3	48%	10	79%	91.3043%	8.6956%	0.2894
19	2	16, 18	62%	11	79%	91.3043%	8.6956%	0.2894
Overall Accuracy						79%		

The result was obtained where an applying EL on Parkinson's disease data was beneficial. Therefore, an EL accuracy improved and the optimum accuracy by combing (1st and 2nd subsets).

Moreover, the features (4th, 5th, 6th, 7th, 8th, 9th, 10th, 11th, 12th, 13th, 14th, 15th) were the most effective and influential on the class among all attributes according to the individual accuracies of their subsets.

Also, based on the TP,TN rates in the previous table, the correct classification was on the negative samples higher than positive ones according the data itself.

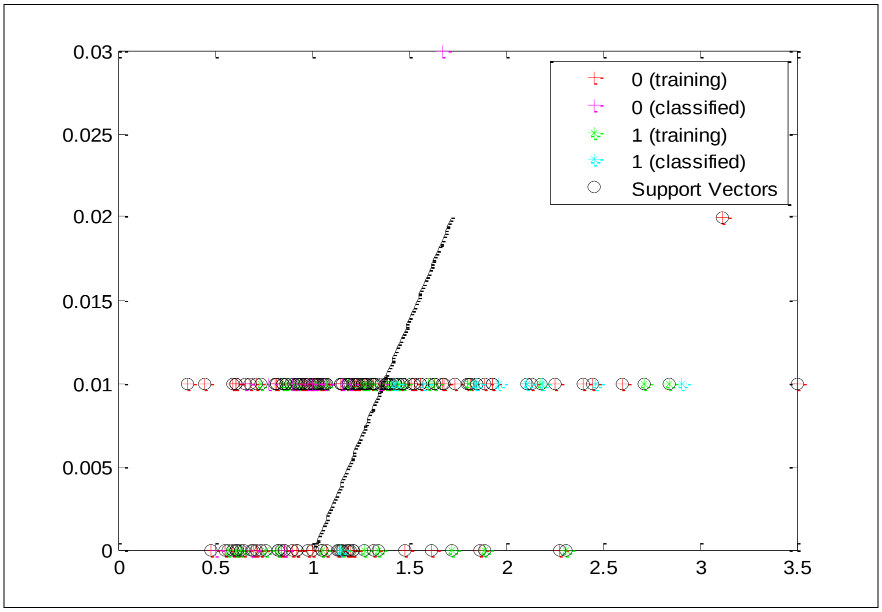
3.3 DATA 2

Data2 (BC) was Breast Cancer Wisconsin (Prognostic) data set which created by following doctors from Wisconsin University: Dr. William H. Wolberg, General Surgery Dept., W. Nick Street, Computer Sciences Dept., Olvi L. Mangasarian, Computer Sciences Dept. Where the first 30 attributes were obtained from a digital image of a fine needle aspirate (FNA). And the last 4 attributes were gotten from medical tests. The data is consisted of 198 instance which represent rows. Each row indicate to one of 198 recordings. Specific recordings and medical tests were done by doctors for patients to get the features (variables) which represent the columns. So, 34 attributes were provided. 33 of them were features and 34th for the class label, where set to 0 if the disease (non recur) and 1 if (recur). For more information look at link of Breast Cancer Wisconsin (Prognostic).

3.4 RESULTS OF BREAST CANCER DATA SET

Different outputs have been obtained with this data. Thus, when linear kernel function was selected (try many values as 2e-1,4e-1,...until C=9e-1) and k-means parameters (K=5, correlation and euclidean distance). The classification on training set was more reasonable and accurate with k- means correlation distance, as will be seen in the next figure (4.7).

Figure 3.7: View9- linear kernel, C=7e-1, correlation distance



In the next plot, the performance was optimized with rbf kernel function also (C=7e-1, sigma=0.8). The result was visually clear to explain that data2 with rbf function is better than linear kernel.

Figure 3.8: View9- rbf kernel, C=7e-1, sigma=0.8, correlation distance

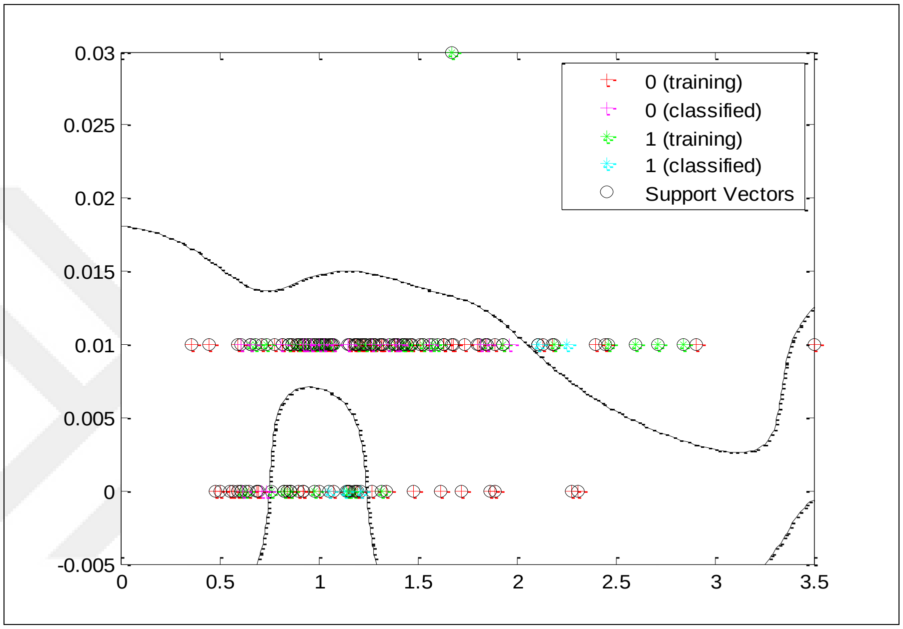
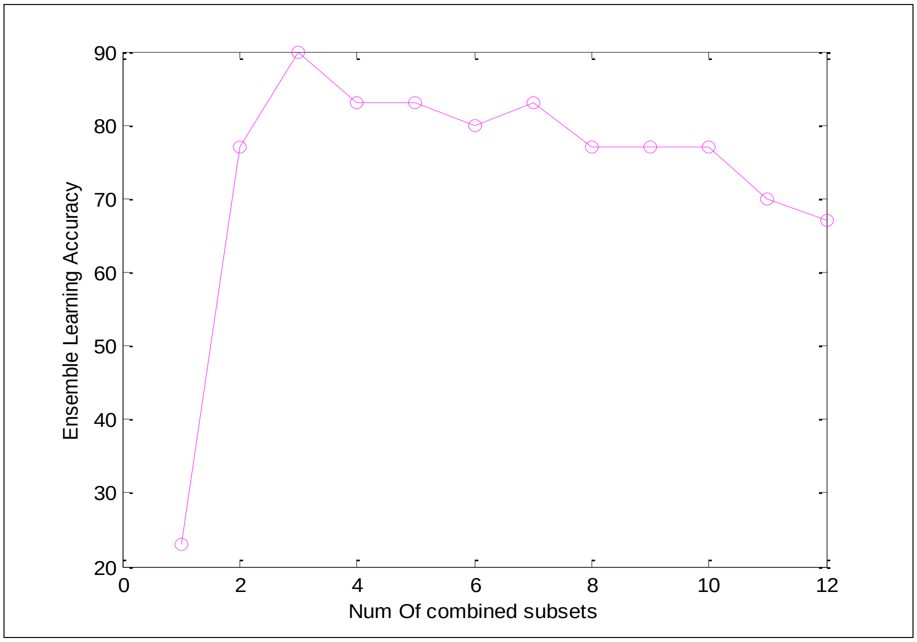


Figure 3.9: Accuracy of ensemble learning



The perfect ensemble outcome was with combination of first three views such the previous figure explained .

Also, the best balance between positive and negative samples by computing MCC like the next figure will show. The highest score was when first three subsets were merged.

Figure 3.10: MCC with ensemble learning

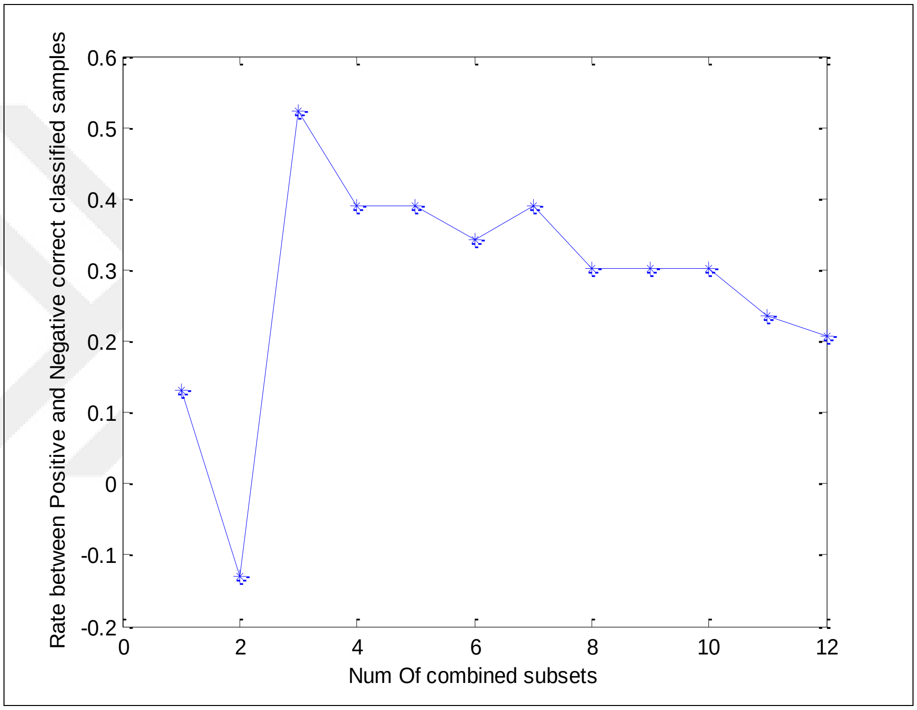


Table 3.2: Results of Breast Cancer data set

Subset No	Feature Num	Features No	Individual Accuracy	Combined subsets Num	EL Accuracy	TP Rate	TN Rate	MCC
1	2	8, 10	13%	1	23%	42.8571%	57.1428%	0.1307
4	1	7, 9, 22	67%	2	77%	0%	100%	-0.1307
6	7	1, 5, 6 , 16 17, 20, 21	80%	3	90%	7.4074%	92.5925%	0.5229
11	5	1, 6, 16, 17, 21	80%	4	83%	8%	92%	0.3888
21	8	2, 3, 4, 11 12, 15, 18 19	37%	6	80%	8.3333%	91.6666%	0.3415
26	5	5, 6, 17, 20, 21	80%	7	83%	8%	92%	0.3888
31	6	5, 6, 17 , 18, 20, 21	77%	8	77%	8.6956%	91.3043%	0.3015
41	4	2, 3, 4, 19	43%	11	70%	9.5238%	90.4761%	0.2357
Overall Accuracy						78%		

At the end, outcomes were obtained after using EL on Breast Cancer data and that was useful. Because we have reached the improved accuracy compared to overall accuracies, when combing (1st, 4th, 6th subsets) was completed. Also, these subsets were the most strongest views based on their features and MRMR score.

Moreover, the features (1st, 5th, 6th, 16th, 17th, 20th, 21th) were the most effective and relevant with their class among all attributes according to the individual accuracies of their subsets. Also, according to the TP, TN rates in the previous table, the highest classification percentage was on the negative instances versus the positive samples .

3.5 DATA 3

Data3 was Madelon data set. It was an artificial dataset and it was related to Cancer distinguish. The data was grouped in specific number of clusters and randomly labeled 1 or -1. It was taken from design of experiments for the NIPS 2003 variable selection benchmark report.

Also, the data was consisted of four thousand four hundred instances which represent rows. Each row indicate to one of them, and all columns represent the five hundred and one features. All of them were features and the last one for the class label. Some pre processes were done on the data to be a little bit simple and fast operations. So, we avoided first row, first column, and changed each -1 of class value to 0. And then, the algorithm was applied on some instances with all features to see the performance compared to data1 and data2 which had few number of attributes. More details could be available in UCI\ Madelon web site.

3.6 RESULTS OF MADELON DATA SET

we have used our algorithm on Madelon data. Then, MCC and EL accuracy were computed to see the effectiveness of the algorithm even with very large number of attributes. Therefore, all generated subsets were very diverse where there were no any redundant features in most views. That was clear in next figures and table.

Figure 3.11: Accuracy of ensemble learning

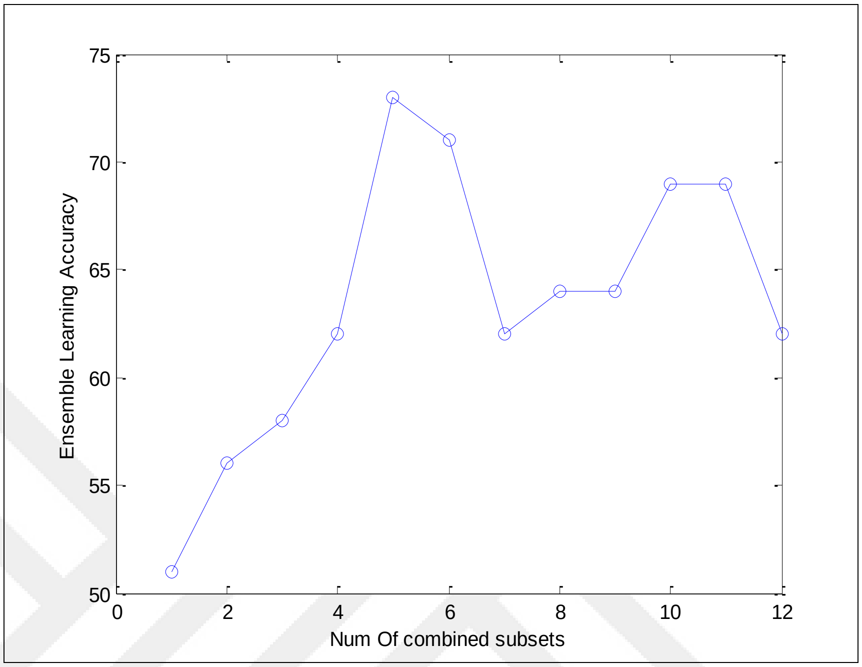


Figure 3.12: MCC with ensemble learning

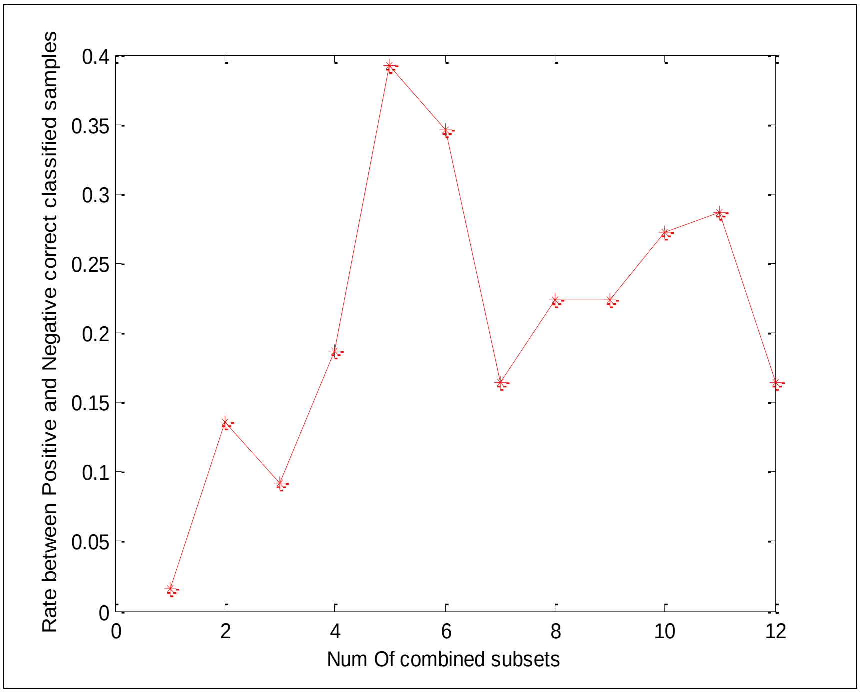


Table 3.3: Results of Madelon data set

Subset No	Feature Num	Individual Accuracy	Combined Subsets Num	EL Accuracy	TP Rate	TN Rate	MCC
2	103	33%	1	51%	65.2173%	34.7826%	0.0165
5	97	44%	2	56%	60%	40%	0.1365
4	99	51%	3	58%	73.0769%	26.9230%	0.0915
7	90	62%	4	62%	71.4285%	28.5714%	0.1872
12	100	58%	5	73%	75.7575%	24.2424%	0.3919
17	103	42%	6	71%	75%	25%	0.3460
10	129	44%	7	62%	75%	25%	0.1641
9	111	56%	8	64%	72.4137%	27.5862%	0.2241
22	94	53%	9	64%	72.4137%	27.5862%	0.2241
Overall Accuracy					62%		

Consequently, the result in table 3 show that using EL on Madelon data was sufficient even it had a large number of features. Also, we have reached the best accuracy with applying EL by combining first diverse five subsets. Because the EL accuracy was higher than individual accuracy after sending the single data set to the classifier. As an overall accuracy was clear in the previous table .

Moreover, the views (4th, 22th, 9th, 12th, 7th,) were the most strongest subsets and relevant with their class among all subsets according to the individual accuracies.

4. CONCLUSION

To summarize in brief sentences, all benefits from using of more than one approach were reached at the end of this research by looking to the previous outcomes such as data1, data2, and data3 tables. Also, we have noticed that ensemble learning most time gave us sufficient and perfect outputs especially, with big number of features **m** where all generated subsets were diverse. Although, EL process with large number of attributes could be complicated calculations. As a result, EL was influential and beneficial when **m** was not very big variables like Parkinson's data, Breast Cancer data whereas it was very effective, and useful to apply with the data which has huge number of features such Madelon data.

Therefore, using clustering, SVM classifier, MRMR algorithm, and EL helped us to obtain accurate and diverse subsets. The details of the algorithm steps were simply demonstrated in methods chapter. Thus, thanks to flexible manner and efficient algorithms were jointly worked, the major standpoint was completely carried out.

Consequently, diverse, accurate, and sufficient subsets were produced as it should. So, they were chosen. In other words, this achievement was the thesis target.

In a future work, as a view of engineering, real data set could be employed with some modifications on the thesis code. The used methods May be changed. Also, this work extensionally could be PhD thesis with more additions of course .

REFERENCES

[1] Settles, B., 2010. Computer sciences technical report. University of Wisconsin-Madison.

[2] https://en.wikipedia.org/wiki/Active_learning_(machine_learning)

[3] Steven, C. H., Michael, R., & Zhu, J. The Chinese University of Hong Kong., & Jin ,R. Michigan State University, USA. *Batch Mode Active Learning and Its Application to Medical Image Classification.*

[4] Alpaydin, E., 2010. *Introduction to machine learning*- 2nd ed. Cambridge, London.

[5] https://en. Wikipedia. org/ wiki/ Statistical_ classification

[6] Phyu, T., N., 2009. *Survey of Classification Techniques in Data Mining*. Hong Kong.

[7] Ramdass, D., & Seshasai, Sh., 2009. *Document Classification for Newspaper Articles*.

[8] Wei, Di., & Melba, M. Crawford, *Active Learning via Multi-View and Local Proximity Co-Regularization for Hyper spectral Image Classification.*

[9] Schapire, R. *Machine Learning Algorithms for Classification*. Princeton University.

[10] Abello, J., & Cormode, G., 2006. *Report on DIMACS Tutorial on Data Mining and Epidemiology*. Rutgers University. America.

[11] https://azure.microsoft.com/en-us/documentation/articles/machine-

learning-algorithm-choice/

[12] D. Michie, D.J. Spiegelhalter, & C.C. Taylor, 1994. *Machine Learning, Neural and Statistical Classification*.

[13] Prof. Carla P. Gomes. *Foundations of Artificial Intelligence*.

[14] Srinet, A., & Snyder, D. *Bagging and Boosting Slides*. A. Krogh and J. Vedelsby, 1995. *Ensembles, Cross Validation and Active Learning*.

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., 2011. *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.

[16] Hastie, T. *Trees, Bagging, Random Forests and Boosting Slides*, Stanford University. & Hastie, T., Tibshirani, R. & Friedman, J., 2009. The *Elements Of Statistical learning*, 2nd ed.

[17] Holmes, P. *IC1 of Simple Decisions*. Princeton University., Bevers, M. *IC2 of Optimization Models*. USDA. & Ping Lo, Y. *MS3 Two- Dimensional*, Loughborough University.

[18] David, M., *2003. Chapter 20. An Example Inference Task Clustering* PDF. *Information Theory, Inference and Learning Algorithms*. Cambridge University.

[19] https://en.wikipedia.org/wiki/K-means_clustering.

[20] Osmar, R. Z., 1999. Chapter 8. Data Clustering book. Alberta University. from source jiawei han data mining book.

[21] Jain, A.K. Michigan State University, Murty, M.N. Indian Institute of Science, & Flynn, P.J. The Ohio State University, 1999. *Data Clustering: A Review PDF*.

[22] Kiranpreet, & Verma, P. *Clustering Amelioration and Optimization with Swarm Intelligence for Color Image Segmentation*. International Journal of Database Theory and Application. India.

[23] http://www.mathworks.com/help/stats/support-vector-machines-for-binary classification .html

[24] Cortes, C., & Vapnik, V., 1995. *Support vector Machine Learning*.

[25] Statnikov, A., Hardin, D. & Aliferis, A. Using SVM Weight- Based Methods. Vanderbilt University, Nashville, USA.

[26] Elsevier, 2011. *Philosophy of Statistics Science*. London.

[27] Guestrin, C., 2007. *VC- Dimension Machine Learning*. Carnegie Mellon University.

[28] Hinton, G., 2008. *Lecture 10 Slides- Support Vector Machines*. King's College, Toronto-Canada.

[29] Chih, W., Chih, C., & Chih, J., 2003. *A Practical Guide to Support Vector Classification*. National Taiwan University. Taiwan.

[30] Muslea, I., Minton, S., & Knoblock, C., 2002. *Adaptive view validation: A first step towards automatic view detection*. International Conference on Machine Learning. USA.

[31] Muslea, I., Minton, S., & Knoblock, C., 2006. *Active Learning with Multiple Views.* Journal of Artificial Intelligence Research. USA.

[32] Di, W. & Crawford, M., 2010. *Multi- view adaptive disagreement based active learning for hyper spectral image classification*. Purdue University, USA

[33] Zhu, X., 2008. *Semi- Supervised Learning Literature Survey*. University of Wisconsin – Madison.

[34] Crawford, M. M., Fellow- IEEE, Tuia, D., Member- IEEE, Yang, H. L., Student Member- IEEE., 2013. *Active Learning for Classification of Remotely Sensed Data.*

[35] Quanz, B., & Huan, J. *Feature Generation for Multi-View Semi-Supervised Learning with Partially Observed Views*. University of Kansas- Lawrence.

[36] Akusok, A., Lendasse, A. Iowa University, USA., Bjork, K., Arcada University, Finland., & Miche, Y. Nokia Group, Finland, 2015. *High-Performance Extreme Learning Machines.*

[37] Devis, T., Volpi, M., Copa, L., Kanevski, M., & Munoz-Marı, J., 2013. *Active learning algorithms for supervised remote sensing image classification*.

[38] Auffarth, B. Institute for Bioengineering of Catalonia, Spain., Lopez, M., & Cerquides, J. Barcelona Universitat, Spain. *Comparison of Redundancy and Relevance Measures for Feature Selection in Tissue Classification of CT images.*

[39] Acid, S., Luis M., & Fernandez, M. *Minimum redundancy maximum relevancy versus score based methods for learning Markov boundaries*. Granada, Spain

[40] Ding, C., & Peng, H. *Minimum Redundancy Feature Selection from Microarray Gene Expression Data*. University of California, USA.

[41] Peng, H., 2007. *mRMR Feature Selection using mutual information computation.*

[42] Ray, N. *Clustering*. CMPUT 466/551.

[43] Teknomo, K., 2007. *K-Means Clustering Tutorial*.

[44] Cai, X., Nie, F., & Huang, H. *Multi-View K-Means Clustering on Big Data. University of Texas*. Proceedings of the Twenty- Third International Joint Conference on Artificial Intelligence.

[45] Bottou, L. Paris, France ., & Bengio, Y. Montreal university, Canada. *Convergence Properties of the K-Means Algorithms.*

[46] Ghani, R., 2000. *Ensemble Classification Methods*.

[47] Thomas G. D. *Ensemble Methods in Machine Learning*. Oregon University, USA.

[48] Sewell, M., 2007. *Ensemble Learning*. University College London.

[49] Polikar, R., 2009**.** *Ensemble learning***.** Scholarpedia**.**

[50] Hernandez, D., Martınez, G., & Suarez, A. Out of *Bootstrap Estimation of Generalization Error Curves in Bagging Ensembles*. Autonoma university, Madrid, Spain.

[51] Dr. R. Ramani, G., & Sivagami, G., 2011. *Parkinson Disease Classification using Data Mining Algorithms.* Rajalakshmi Engineering College, Chennai, INDIA. International Journal of Computer Applications.

[52] Ma, C. M., Yang, W. S., & Cheng, B. W., 2014. *How the Parameters of K-nearest Neighbor Algorithm Impact on the Best Classification Accuracy of Parkinson Data set*. Journal of Applied Sciences. Taiwan, China.

[53] Menaka, K. PSGR Krishnammal College for Women., & Karpagavalli, S. GR Govindarajulu School of Applied Computer Technology. 2013. *Breast Cancer Classification using Support Vector Machine and Genetic Programming*. International Journal of Innovative Research. Coimbatore, India2.

[54] Du, J., Securitie, H., Liu, J. S., & Krakovna V. *Interpretable Selection and Visualization of Features and Interactions Using Bayesian Forests*. Harvard University, Cambridge.

[55] Nomin, B., Tay, B. & Oh, S. Boosting *Classification Accuracy Using Feature Fusion*. Dankook University, Cheonan, South Korea. International Conference on Information and Network Technology 2012. Singapore.

[56] Fern, X. Z. & Lin, W. *Cluster Ensemble Selection pdf*.

[57] Caruana, R., Munson, A. & Niculescu- Mizil, A., 2006. *Getting the most out of ensemble selection*. In Proceedings of the Sixth international Conference on Data Mining.
[58] Kumar, V., & Minz, S., 2014. *Feature Selection, Smart Computing Review*. Jawaharlal Nehru University, New Delhi.

[59] Segen, J. *Feature selection and constructive inference. T*he Seventh International Conference on Pattern Recognition 1984.