ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

ENSEMBLE BASED FEATURE SELECTION WITH HYBRID MODEL

M.Sc. THESIS

Ceylan DEMİR

Department of Mathematical Engineering

Mathematical Engineering Programme

JUNE 2019

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

ENSEMBLE BASED FEATURE SELECTION WITH HYBRID MODEL

M.Sc. THESIS

Ceylan DEMİR
(509161204)

Department of Mathematical Engineering

Mathematical Engineering Programme

Thesis Advisor: Assist. Prof. Dr. İzzet GÖKSEL
Co-advisor: Assoc. Prof. Dr. Süreyya AKYÜZ

JUNE 2019

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

HİBRİT MODELİ İLE TOPLULUK TEMELLİ ÖZNİTELİK SEÇİMİ

YÜKSEK LİSANS TEZİ

Ceylan DEMİR
(509161204)

Matematik Mühendisliği Ana Bilim Dalı

Matematik Mühendisliği Programı

Tez Danışmanı: Dr. Öğr. Üyesi İzzet GÖKSEL
Eş Danışman: Doç. Dr. Süreyya AKYÜZ

HAZİRAN 2019

Ceylan DEMİR, a M.Sc. student of ITU Graduate School of Science Engineering and Technology 509161204 successfully defended the thesis entitled "ENSEMBLE BASED FEATURE SELECTION WITH HYBRID MODEL", which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Assist. Prof. Dr. İzzet GÖKSEL** Istanbul Technical University	.....
Co-advisor :	**Assoc. Prof. Dr. Süreyya AKYÜZ** Bahçeşehir University	.....
Jury Members :	**Prof. Dr. Nalan ANTAR** Istanbul Technical University	.....
	Assoc. Prof. Dr. İlkay BAKIRTAŞ AKAR Istanbul Technical University	.....
	Assist. Prof. Dr. Tarkan AYDIN Bahçeşehir University	.....

Date of Submission : **3 May 2019**
Date of Defense : **17 June 2019**

To my family,

FOREWORD

First, I would like to emphasize my gratitude to my advisor Assist. Prof. Dr. İzzet Göksel for his helpfulness and encouragement. Moreover, I am truly indebted and thankful to my co-advisor Assoc. Prof. Dr. Süreyya Akyüz for her guidance during preparation, encouragement with scientific point of view and high level of enthusiasm. I am also grateful for her valuable feedback and organization as well as the amount of time she has reserved for my queries and development. I would never imagine that this research work would be this exciting and fruitful. Furthermore, I would like to thank Prof. Dr. Nalan Antar, Assoc. Prof. Dr. İlkay Bakırtaş Akar and Assist. Prof. Dr. Tarkan Aydın for participating in my thesis committee and giving me feedback.

Besides, I would like to state my appreciativeness for my friends Ammar, Merve, Esma and Hayriye. You have been great in keeping me up and helping me to overcome this process.

Last but not least, my sincere thanks will be to my family for all their patience and support. I am the luckiest person on earth to have this awesome family.

June 2019

Ceylan DEMİR
(Mathematical Engineer)

TABLE OF CONTENTS

Page

ABBREVIATIONS

CMIM	**:** Conditional Mutual Information Maximization
DISR	**:** Double Input Symmetrical Relevance
DVM	**:** Destek Vektör Makinesi
Eq.	**:** Equation
ICAP	**:** Interaction Capping
JMI	**:** Joint Mutual Information
KKT	**:** Karush-Kuhn-Tucker
MIM	**:** Mutual Information Maximization
MRMR	**:** Minimum Redundancy Maximum Relevance
NMI	**:** Normalized Mutual Information
QP	**:** Quadratic Problem
SNMI	**:** Sum of Normalized Mutual Information
SVM	**:** Support Vector Machine

SYMBOLS

$Acc(.)$: Accuracy function
c : Number of classes
c^i : A signature vector of classifier i
$d(.)$: Distance metric
$Div(.)$: Non-pairwise diversity function
E_T : Ensemble
E_i : Training subsets
$H(X)$: Entropy function of a random variable X
$H(X|Y)$: Conditional entropy function of X given another discrete random variable Y
$I(X;Y)$: Information gain between X and Y
$\mathbb{I}(.)$: Indicator function
J : Feature score
K : Pairwise diversity measure
$K(.)$: Kernel function
L : Library of classification solutions
L_{rest} : The rest of feature selection methods in the library after pruning
m : Margin
n_j : The number of instances from class j
$NM(j)$: Data samples to x_j with the same class label
$NH(j)$: Data samples to x_j with a different class label
$O(TN)$: The time complexity of the orientation ordering method
$P(x_i, y_j)$: Joint probability of x_i and y_j
S : Selected feature set
S_u : Subensemble
$\vec{w}$: Normal vector to hyperplane
$\vec{x}$: Input vector
X, Y : Random variables
Z_{sel} : Selection set
Z_{tr} : Training data set in the study of Kuncheva and Whitaker
α_i : Lagrange multipliers
μ : The mean feature value
μ_j : The mean feature value on class j
ξ_i : Slack variables
σ_j : The standard deviation of feature value on class j
$\Psi(.)$: Objective function

LIST OF TABLES

Page

LIST OF FIGURES

Page

ENSEMBLE BASED FEATURE SELECTION WITH HYBRID MODEL

SUMMARY

Today with the development of technology, especially in the field of information technology, "Big Data" concept emerges. The amount of accumulated data is increasing day by day, for this reason the big data concept has reached an important place. However, the collected big data is not a meaningful collection of information in its raw form, it has to go through a variety of procedures. Therefore, "Machine Learning" techniques are frequently used to obtain meaningful data from big data.

Machine Learning research area has highly significant techniques, one of them are Feature Selection Methods. Feature selection is one of the core concepts in machine learning that extremely impacts the performance of the model, because it serves as a fundamental technique to direct the use of variables to what is most effective and efficient for a given machine learning model. However, utilizing feature selection methods alone is not sufficient to improve the performance of the model. Therefore, ensemble based techniques were proposed in the literature. Combination of several feature selection methods and variation in data set aspects were developed under favour of ensemble based techniques. There are three kind of generation approaches analyzed in the literature to generate a diverse ensemble library: Data variation, function variation and hybrid variation. In this study, the proposed model is consubstantiated with hybrid diversity ensemble learning technique and pruning.

In brief, the task for "Hybrid Variation" method, which includes both "Data Variation" and "Function Variation" methods with multi-class classification especially "Support Vector Machine" (SVM), is proposed. In addition, the study contains "Joint Criterion" ensemble pruning method.

In Chapter 1, general introduction of machine learning and methods in the literature are mentioned. Besides, purpose of the thesis and hypothesis are given.

In Chapter 2, literature reviews about all methods that are utilized in this study are given. In this part, there are three main sections: Ensemble learning, classification and ensemble pruning techniques. Regarding ensemble learning, there are following subsections: Data variation, its definition and methods, function variation, its definition, techniques which are in the literarture and hybrid variation. Regarding classification, several classification methods are mentioned; however, the base classifier of the model, i.e. SVM is extensively described. Finally, ensemble pruning and its several methods are given.

In Chapter 3, the proposed model "Ensemble Based Feature Selection with Hybrid Model" is explained in detail. The combination of data diversity and function diversity which constructs the hybrid model is given firstly, and then integration of joint criterion pruning approach is clarified step by step.

In Chapter 4, the materials which are used in the model and experimental setup are mentioned.

In Chapter 5, experimental results and their explanation are given.

At last, conclusion and recommendations are mentioned. The results of the study and possible future projects are discussed in this chapter.

HİBRİT MODELİ İLE TOPLULUK TEMELLİ ÖZNİTELİK SEÇİMİ

ÖZET

Günümüzde teknolojinin gelişmesiyle, özellikle bilgi teknolojileri alanında, “Büyük Veri” kavramı ortaya çıkmıştır. Biriken veri miktarı gün geçtikçe artmakta, bu nedenle büyük veri kavramı önemli bir yere sahip olmuştur. Bununla birlikte, toplanan büyük verilerin ham formu, anlamlı bir bilgi toplamı değildir; anlamlı hale gelebilmesi için çeşitli işlemlerden geçmesi gerekir. Bu nedenle büyük verilerden anlamlı bilgiler elde etmek için “Makine Öğrenimi” teknikleri sıklıkla kullanılır.

Ham veri, makine öğrenmesi algoritmasına girdi olarak verildiğinde bu makine için kullanılabilir bir veri olmamaktadır. Algoritmanın yorumlayabileceği forma dönüştürmek için literatürde çeşitli yöntemler kullanılmaktadır. "Öznitelik Çıkarımı" bu yöntemlerden biridir. Bir kan verisi ele alınırsa, kan ham haliyle herhangi bir anlam ifade etmemekte, ancak çeşitli testler uygulandıktan sonra ortaya çıkan kandaki kolestrol miktarı, alyuvar sayısı, antikor sayısı gibi daha anlamlı veriler ile kan hakkında yorum yapılabilmektedir. İşte bu örnekte belirtilen kolestrol miktarı, alyuvar sayısı, antikor sayısı gibi daha anlamlı veriler, öznitelik olarak adlandırılmakta, bu öznitelikleri elde etmeye yarayan tekniklere de öznitelik çıkarımı yöntemleri denmektedir. Eğer kullanılan veri kümesinin tahmin edilmesi istenen bilgileri önceden biliniyorsa, yani veri kümesi etiketli ise, öznitelikler çıkarıldıktan sonra çeşitli sınıflandırma yöntemleriyle modelin tahmin sonucu ve performansı hesaplanabilir. Ancak veride etiket bilgisi bulunmuyorsa, bu öznitelikler çeşitli kümeleme yöntemleri için girdi olarak kullanılır ve sonuç elde edilir. Bununla beraber, veri kümesindeki bazı veriler etiketli, bazıları ise etiketsiz olabilir. Bu durumda, etiketli veriler için çeşitli sınıflandırma algoritmaları, etiketsiz veriler için ise çeşitli kümeleme algoritmaları kullanılır ve elde edilen modelin performansı bu şekilde hesaplanır.

Ham veriden çıkartılan her öznitelik, ulaşılmak istenen hedef bilgiyi elde etmede bir anlam ifade etmeyebilir. İşte bu noktada, makine öğrenimi alanındaki bir diğer yöntemin, “Öznitelik Seçme” yöntemlerinin önemi ortaya çıkmaktadır. Öznitelik seçimi, makine öğreniminde modelin performansını önemli ölçüde etkileyen temel kavramlardan biridir. Değişkenlerin kullanımını belirli bir makine öğrenme modeli için en etkili ve en verimli olan yönteme doğru yönlendirmek için öznitelik seçimi yöntemleri sıklıkla kullanılır. Elde edilmek istenen sonuca ulaştıracak özniteliklerin seçimi bu yöntemlerle yapılır, böylelikle kurulan modelin hızı ve performansı önemli ölçüde artar.

Bununla birlikte, sadece öznitelik seçme yöntemlerinin kullanılması, modelin performansını artırmak için her zaman yeterli olmayabilir. Bu nedenle literatürde “Topluluk Temelli Teknikler” önerilmiştir. Topluluk temelli teknikler ile öne sürülen hipoteze göre, model üzerinde bir öznitelik seçimi yöntemi kullanmak yerine birden fazla yöntemin aynı anda kullanılması, modelin sonucunun daha kesin olmasını

sağlamaktadır. Ayrıca kullanılan veri kümesini rastgele bölerek elde edilen alt veri kümelerinin aynı anda kullanımı da model sonucunu etkileyen bir diğer önemli hipotezdir.

Çeşitli öznitelik seçim yöntemlerinin kombinasyonu ve veri kümesi varyasyonu yöntemleri, topluluk temelli teknikler lehinde geliştirilmiştir. Literatürde veri kümesi varyasyonu, fonksiyon varyasyonu ve hibrit varyasyon olarak gruplanabilen üç tür topluluk temelli yaklaşım vardır. Hibrit varyasyonu, aynı anda hem birden fazla öznitelik seçme yöntemi hem de alt veri kümelerinin kullanılmasıyla oluşturulmuş, topluluk temelli bir yöntemdir.

Tüm bunlara ek olarak topluluk temelli model içerisindeki her elemanın, modelin sonucunu iyileştirdiği söylenememektedir. İşte bu noktada modelin performansını kötü etkileyen elemanlar, çeşitli yöntemlerle topluluktan çıkartılır. Bu yöntemler bütününe "Topluluk Budama Yöntemleri" denmektedir. Topluluk budama yöntemleri modelin performansını ve kesinliğini önemli ölçüde etkileyen yaklaşımlardır.

Bu çalışmada, önerilen model hibrit çeşitlilik topluluğu öğrenme tekniği ile geliştirilmiş ve topluluk budama yöntemi ile desteklenmiştir.

Bu tez çalışmasında önerilen model, veri kümesi varyasyonu yöntemi ve fonksiyon varyasyonu yönteminin kombinasyonuyla oluşturulan hibrit modeldir. Hibrit model üzerinde sınıflandırma sonuçlarını elde etmek için “Destek Vektör Makinesi (DVM)” kullanılmış, elde edilen sonuç matrisine “Ortak Kriter” topluluk tabanlı budama yöntemi uygulanıp daha iyi çözümler elde edilmiştir.

Bu çalışmanın birinci bölümünde makine öğreniminin genel tanıtımı ve literatürdeki yöntemlerden genel hatlarıyla bahsedilmiştir. Ayrıca, tezin amacı ve hipotezi de bu bölümde verilmiştir.

İkinci bölümünde, bu çalışmada kullanılan tüm yöntemler hakkında literatür araştırmalarına ve geçmişte yapılmış olan çalışmalara yer verilmiştir. Bu bölüm topluluk öğrenmesi, sınıflandırma ve topluluk budaması olmak üzere üç alt başlığa bölünmüştür. Topluluk öğrenmesi bölümünde, veri kümesi varyasyonu, tanımı ve yöntemleri, fonksiyon varyasyonu, tanımı, literatürde olan öznitelik seçimi teknikleri ile hibrit varyasyonu ve genel tanımına yer verilmiştir. Sınıflandırma alt başlığında, birkaç sınıflandırma yönteminden genel hatlarıyla bahsedilmiş; bununla birlikte, önerilen modelin temel sınıflandırma yöntemi olan DVM hakkında geniş tanım ve matematiksel alt yapısı anlatılmıştır. Ve son alt başlık, topluluk budama yöntemleri bölümünde, tanımlar ve literatürde önerilmiş topluluk budama alt-yöntemlerine yer verilmiştir.

Bu tez çalışmasının üçüncü bölümünde, önerilen “Hibrit Model ile Topluluk Tabanlı Özellik Seçimi” ayrıntılı olarak açıklanmıştır. Öncelikle topluluğu oluşturmak için kullanılan “Torbalama” ve sekiz öznitelik seçimi yönteminin kombinasyonuyla oluşturulmuş hibrit modelin yapısı ve sözde kodlarına yer verilmiştir. Buna göre, kullanılan veri kümesi yüzde 80’i eğitim, yüzde 20’si test veri kümesi olarak ayrılmış, eğitim veri kümesinden torbalama yöntemiyle, her torbada yüz veri olmak üzere otuz torba üretilmiştir. Üretilen her torbaya ayrı ayrı sekiz öznitelik seçim yöntemi uygulanmıştır. Bu öznitelik seçim yöntemleriyle torbalardaki her örnek için yüz öznitelik seçilmiştir. Daha sonra DVM çoklu-sınıf sınıflandırma yönteminin entegre edilmesi anlatılmış, ardından sözde kodu belirtilmiştir. Elde edilen çözüm matrisi

üzerine, her torbaya ve her öznitelik seçimine olmak üzere, ortak kriter budama tekniği entegre edilmiş ve alt bir çözüm matrisi elde edilmiştir.

Ortak kriter budama tekniği için literatürde önerilen yöntem, kümeleme problemleri üzerinde uygulanmış, ancak bu çalışmada öznitelik seçimi yöntemlerinin sınıflandırma çözümleri üzerinde uygulanmıştır. Buna ek olarak literatürde kümeleme problemlerine uygulanan budama tekniği, ikili çeşitlilikleri içermektedir. Bu çalışmada hem ikili çeşitlilik içeren ortak kriter budama yöntemi hem de ikili olmayan çeşitlilik ile ortak kriter budama tekniği kullanılmıştır. Topluluk budama yöntemi uygulanırken topluluk alt küme kardinalitesi dört, beş, altı, yedi ve sekiz olacak şekilde seçilerek hangi öznitelik yöntemlerinin kesinlik ve çeşitlilik oluşturma açısından daha iyi olduğu saptanmıştır.

Dördüncü bölümünde, modelde girdi olarak kullanılan veri kümesi özellikleri ve çalışmada kullanılan platformlardan bahsedilmiştir.

Çalışmanın beşinci bölümünde deneysel sonuçlar ve bunların açıklamaları verilmiştir. Bu açıklamalar tablo ve şekillerle desteklenmiştir.

Son bölümünde, sonuç ve önerilerden bahsedilmiştir. Elde edilen sonuçlara göre, topluluk boyutu çok büyük olmadığı için ikili çeşitlilik içeren budama yöntemi ile bir ve sekiz arasındaki kardinalitilerde en kesin sonuç elde edilmemiş, topluluğun kardinalitesi artırıldıkça kesinliğin genel olarak arttığı gözlenmiştir. Ancak literatürde kullanılmamış ikili olmayan çeşitlilik içeren budama yöntemi ile, topluluk boyutu küçük olmasına rağmen istenen sonuç elde edilmiş, böylece literatüre bir katkıda bulunulmuştur. Gelecekte yapılacak çalışmalarda farklı tip veri içeren ve farklı büyüklüklere sahip birden fazla veri kümesine bu yöntemler uygulanabilir. Ayrıca topluluğun boyutu artırılarak, yani daha fazla öznitelik seçme yöntemi kullanılıp daha fazla torba üretilerek, çıkan sonuçlar incelenebilir. Tezde kullanılan yöntem dışında başka torbalama yöntemleri de kullanılarak ortaya çıkan sonuçlar kıyaslanabilir. Buna ek olarak, bu çalışmada kullanılan DVM sınıflandırma yöntemi dışında farklı sınıflandırma yöntemleri kullanılarak modellerin başarımları karşılaştırılabilir.

1. INTRODUCTION

Developments in the assemblage and depot of digital data have caused a tremendous increase of stored data. On the other hand, the rapid digitalization of life, the Internet becoming an integral part of daily life, and the widespread use of technology as an acceptable commodity have increased collected data. This kind of data, the so-called Big Data, can be obtained from various industrial organizations and production sites, banks, educational institutions and organizations, health institutions and government sources, especially in the social media fields. However, it is worthless unless the big data is processed. The process, in which stacked data is processed and converted into meaningful information, is called "Data Mining". Basically, data is divided into predefined classes according to their attributes by using data mining techniques.

By obtaining meaningful data from big data, banks can increase customer satisfaction, take measures against possible fraud, production areas can increase productivity and diversity in production, governments can take measures against possible threats, educational institutions and organizations can determine the best educational model for students, sales oriented firms can increase their profits, improvements can be made in these and many other areas. Therefore, processing of the collected data and getting meaningful results from them has become important in many areas.

Data mining includes many areas of work, such as statistics, database technologies, machine learning, deep learning, artificial intelligence and visualization and it works as interdisciplinary group of methods. Generally, all these disciplinary fields feed each other; especially machine learning and data mining often use the same methods. At this point, the importance of Machine Learning Algorithms, the most well-known techniques for data mining, become apparent [1].

The most fundamental definition of machine learning is a data analytics technique that is used for analysis of diversified kind of data and teaches computers to learn from experience. Basically, it is based on the principle of automatic learning and development. Machine learning, with various algorithms and methods, tries to find

out some patterns in the data and learns by looking at the corresponding labels firstly, after that, develops systems that can make deductions by taking advantage of their experience. This possibility is provided by many algorithms that use various statistical and mathematical approaches. One or more of these methods and algorithms are used together to construct a model. This construction aims to run the model in a more efficient and fast manner.

The major goal in machine learning is to forecast future actions by using prior observations. In the availability of adequately large data and parameters, machine learning can make more correct predictions about future compared to people. Machine learning algorithms are usually grouped into four types: Supervised Learning, Unsupervised Learning, Semi-supervised Learning and Reinforcement Learning. Each group generally uses different approaches; however, their aims are the same.

The aim of supervised learning is to determine the mapping function from the input variable x to the output variable y for training steps [2]. In the testing step, labels of the new points are forecasted by using the relation which is found in training step. Classification and regression are members of supervised learning. In unsupervised learning, there is input data x without corresponding output variables. The purpose for unsupervised learning is to generate the underlying structure or distribution in the data. Clustering and association problems are examples of unsupervised learning. Semi- supervised learning, which sits between supervised and unsupervised learning, has a large amount of input data x and only some of the data is labeled y. The last one, reinforcement learning is a machine learning approach that learns what needs to be done for the aim. Heuristic approach is the most distinguishing property of reinforcement learning. According to the property of the problem, the most suitable machine learning approach is chosen and the best solution is sought.

Besides, in order for any problem to be solved by machine learning methods, the problem must be appropriately represented. The problem to be solved may not always have the qualities that can be given directly to machine learning methods; therefore, the problem must be converted into a form that can be used by the machine. There are various techniques in the machine learning field. One of them is *Feature Extraction*, that is transforming the input data into set of features. Feature extraction is a process of dimension reduction, in other words, it is a mapping of data vector X into a

lower-dimensional feature vector Y [3, 4]. These features, that are extracted from data set, can be different according to the type of data set. A correct feature extraction and a model design for these features influence the success and performance of the result.

Feature extraction techniques improve the accuracy of a machine learning algorithm, model performance for high-dimensional data sets and the interpretability of the model. Nevertheless, many machine learning methods suffer from intractability problems owing to proliferation of large-scale data sets [5]. Data dimensionality and number of attributes are main effects for these recalcitrant problems. To avoid these impacts and potential adverse consequences, other techniques, namely *Feature Selection* are used. The feature selection is defined as the choosing of the best subset that can exemplify the original data set [6]. By using these techniques, less related features can be eliminated and consequently, more beneficial features can be kept; hence, other steps that follow the feature selection can accelerate the model and give more accurate results. Feature selection methods reduce the size of the feature set and increase the algorithm speed, decrease the amount of memory required to store the data, remove non-relevant and noisy data, enhance the data quality and boost the success of the model obtained [6].

Feature Selection is a significant topic in the machine learning field; however, other applications are needed for some problems. Therefore, various applications for feature selection are offered that generate the data pre-processing step of the machine learning problems. At this point, ensemble based feature selection methods are proposed to create an optimal subset of features by consolidating multiple feature selectors based on the discernment behind the ensemble learning. In other words, ensemble is a group of learning models that collectively resolve the problem [7]. Recent researches demonstrate that the decision of an ensemble of feature selection approaches reveals a more accurate estimation than any single feature selection method that is used alone [8]. There are three kinds of ensemble based approaches for feature selection which are Data Variation, Function Variation and Hybrid Variation methods [8, 9].

Another machine learning application area is classification that is one of the most frequently used and oldest data mining techniques. The concept of classification is simply to distribute the data between the various classes defined on a data set. The classification algorithms learn this distribution from the given training set and then try

to classify it correctly, when the unlabelled test data is given as new input. By using classification algorithms, accuracy and error of the constituted model can be attained.

The last mentioned approach is *Ensemble Pruning* that chooses the best subset of the ensemble taking into consideration accuracy and diversity of models synchronically. The main goal of the ensemble pruning method is to investigate for a good subset of ensemble members that fulfills as well as, or more preferable better than, the original ensemble set [7].

1.1 Purpose of Thesis

Classification is one of the supervised learning methods in machine learning field. Researches show that decision of an ensemble gives better consequences than a single classification solution. The accuracy and the diversity of the ensemble are considered vital factors affecting the success of ensemble learning. The significance of these factors derives from the fact that there is a trade-off between the accuracy and the diversity of attributes of different classification resolutions in which the development of one of these attributes induces the corruption of the other one. Exploring the best subset of the ensemble is one of the challenging problems in the literature.

The main purpose of this study is to choose the best classification solutions from an ensemble that optimizes the accuracy and the diversity synchronically with a hybrid model that composes function variation (feature selection) and data variation (bootstrap aggregation) algorithms. In other words, by implementing a hybrid approach on a data set, and afterwards using ensemble pruning algorithm on classification results of the model gives a more precise result. By using the hybrid model with ensemble pruning approach, the best classification solutions are expected to be found. In addition, by implementing different cardinalities of ensemble pruning techniques, best size of the classification solutions is expected to be found.

1.2 Hypothesis

In this thesis, it is expected that the accuracy will be greater by using ensemble learning methods. Especially, it is expected that the accuracy of the model will be greater via joint criterion method with non-pairwise diversity than pairwise type of the method.

2. LITERATURE REVIEW

2.1 Ensemble Learning

Ensemble learning is an approach that can solve a machine learning problem with trained multiple learners. Final conclusion of this approach is made after compounding each output of single learners in accordance with some criteria. "No Free Lunch" theorem indicates that there is no single model that operates best for every problem [10]. For this reason, the purpose of the ensemble learning is to enhance the accuracy of the single classifiers. On the other hand, owing to the potential noise in the data, overlapping data dispersions and outliers; single classifiers cannot generally acquire a determined classification accuracy. All these have boosted the necessities to generate ensemble techniques.

Generating an ensemble model is prepared in two phases. Firstly, a couple of base classifiers are produced in a sequential or parallel manner. In general, in the ordered manner, the structure of a base classifier may influence the structure of the ensuing classifiers. In the second and last part, the emergent classifier outcomes are compounded to get a decision regarding the final classification of a new test sample. At this point, several sorts of combination methods, such as majority voting, are implemented for the classification problem. Then, among the majority of the class labels of the individual classifiers, the class label is chosen by majority voting.

The basic ensemble learning framework is shown in Figure 2.1.

Ensemble methods include two essential phases: Production of diversity and assemblage of the decisions. There exist three sorts of generation approaches investigated in the literature to create a diverse ensemble library: Data variation, function variation and hybrid variation.

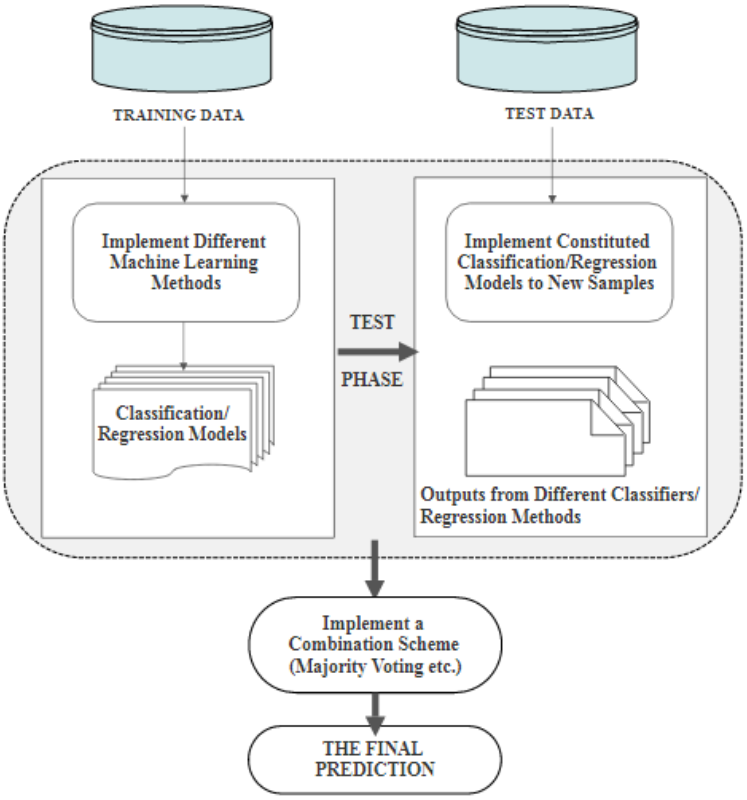

Figure 2.1 : Framework of Ensemble Learning.

2.1.1 Data variation

Creating various and certain individual classifiers for an ensemble is the aim of ensemble learning methods. By voting the decisions of the individual classifiers in the ensemble, the results are aggregated to reach accurate classification decisions [10].

On the other hand, the diversity in the data set is highly important. Therefore, data variation approach is used for generating different sub-data sets from the original data set.

In the ensemble learning literature, there are several popular approaches, two of which are Bootstrap Aggregation (Bagging) and Boosting [5]. These two techniques are the most known methods in the literature for data variation.

2.1.1.1 Bootstrap aggregation (Bagging)

Bootstrap Aggregation, in other words Bagging algorithm is the first ensemble learning technique and was proposed by L. Breiman in 1996 [11]. Bagging method is very useful for high dimensional data set problems and powerful for ensemble method to

Table 2.1 : The pseudocode of bagging ensemble learning.

Algorithm 1 Bootstrap aggregation (bagging) ensemble learning method

```
Set E_1 = E_2 = ... = E_n = 0
for i = 1,2,...,n do
    for j = 1,2,...,m do
        index = m * rand()
        E_i = E_i ∪ E^index
    end for
end for
```

improve the performance of the model. It is a method of retraining the basic learner by deriving new training sets from an original training set. The training set is produced by random selection by putting a sample set consisting of n samples in bagging. Each selected sample is put back into the training set. In this case, some examples are not included in the new training set while others may take place more than once. Outputs of these randomly selected sub-data sets are aggregated with voting or averaging. Classification is used in voting and regression is used in averaging [12].

The pseudocode of bagging algorithm that is used after train-test splitting is given in Table 2.1 where n is the number of bags and m is the number of training instances in each bag. This algorithm takes the training set Z_{tr} as input and outputs the generated training subsets $E_1, E_2, \ldots, E_n$. [5]

2.1.1.2 Boosting

Boosting algorithm was proposed by Freund and Schapire in 1996 [13]. Boosting expresses group of algorithms that use averages of weights to turn weak learners into stronger learners. Each working model defines what features the next model will focus on, which indicates that boosting is all about teamwork unlike bagging algorithm. This procedure is improved for classification; however, it can be used for regression to enhance the performance of a learning algorithm. The most famous boosting algorithm is recognized as "Adaptive Boosting" approach that is also shortly known as "AdaBoost" [14]. AdaBoost algorithm is a meta-heuristic algorithm to attain more preferable performance of decision trees on binary classification samples [15].

Table 2.2 : The pseudocode of AdaBoost ensemble learning.

Algorithm 2 AdaBoost ensemble learning method

1: **for** $i = 1, 2, \ldots, M$ **do**
2: Fit a classifier $T^m(x)$ to the training data using weights w_i
3: $err^m = \sum_{i=1}^{n} w_i \cdot \mathbb{I}(c_i \neq T^m(x_i)) / \sum_{i=1}^{n} w_i$
4: $a^m = \frac{\log(1-err^m)}{err^m}$
5: **for** $i = 1, 2, \ldots, n$ **do**
6: $w_i = w_i \cdot \exp(\alpha^m \cdot \mathbb{I}(c_i \neq T^m(x_i)))$
7: **end for**
8: Re-normalize w_i
9: **end for**

Its algorithm is given in Table 2.2 [15, 16]. This algorithm takes the observation weights w_i=$\frac{1}{n}$ where $i = 1, 2, \ldots, n$ as input and outputs a classification rule.

2.1.2 Function variation

Data variation and function variation approaches use entirely different methodological analysis during the process. Although data variation approach uses different sub-data sets that are composed of original data set, function variation uses the same data set. The function variation method provides the diversity of data by using more than one feature selection method [8]. The consequences are aggregated into a single feature ranking after the performance of all feature selection methods.

Function variation process is demonstrated in Figure 2.2 [5].

There are several ways to divide feature selection algorithms into some groups by using different division approaches. Feature Selection methods can be divided as filter, wrapper and embedded feature selection methods in terms of dissimilar selection strategies. Filter methods typically collect individual variables and manipulate some before creating a model [17]. In addition, Filter methods have the advantage of being fast and independent of the classification model due to operating on the data set directly, and ensuring a feature weighting, ranking or subset as output [9]. Guided by the result of model, Wrapper methods fulfill a search in the area of feature subsets. In contrast with Filter methods, Wrapper methods frequently report better consequences; however, at the price of an increased computational cost [9]. Lastly, Embedded

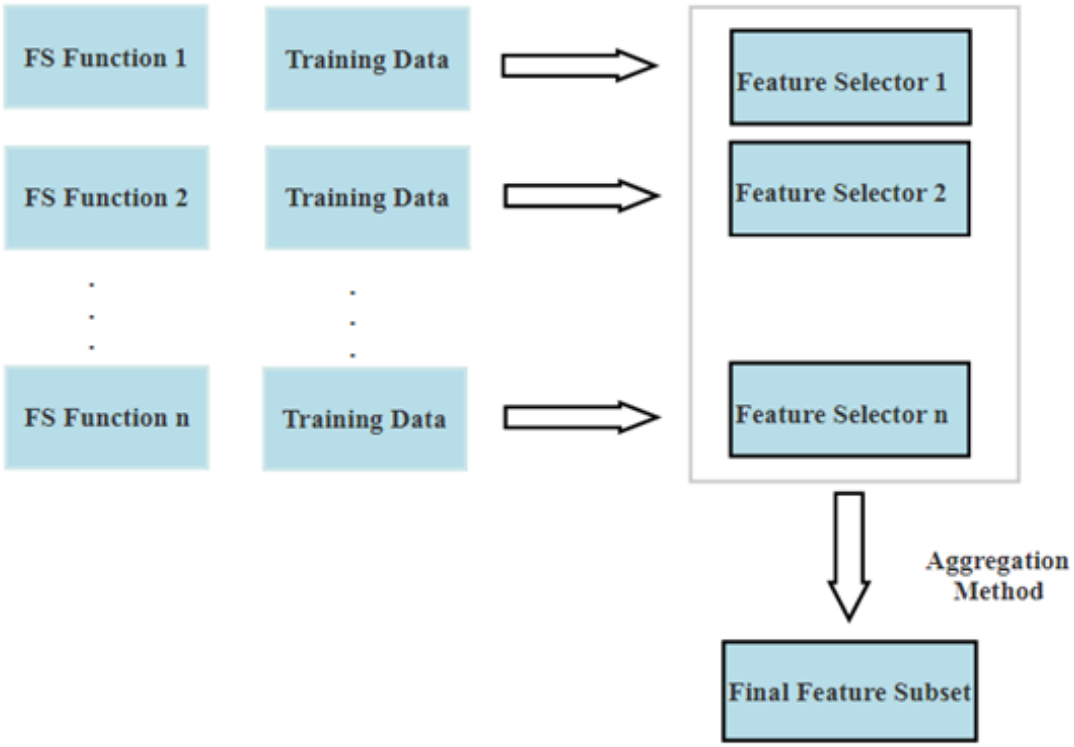

Figure 2.2 : Function variation and aggregation methods.

methods utilize all the variables to create a model and after that analyze the model to deduce the significance of the variables. As a result, the significance of variable is connected directly to the learner used to model the relationship [17]. Additively, between performance and computational cost, a good trade-off is supplied by the Embedded methods [9].

In accordance with the presence of label information, feature selection algorithms can be extensively grouped as supervised, unsupervised and semi-supervised techniques [18]. Supervised feature selection is usually modelled for classification or regression issues. A general framework of supervised feature selection methods is exemplified in Figure 2.3 [18].

In contrast to supervised learning algorithms, unsupervised feature selection methods are ordinarily created for clustering problems [18]. The framework of unsupervised feature selection approaches is shown in Figure 2.4 [18].

When adequate label information is present, supervised feature selection methods are used, while any label information is not needed by unsupervised feature selection techniques. However, there are small number of labeled instances and a large number of unlabeled instances in several real-world implementations. Not only supervised

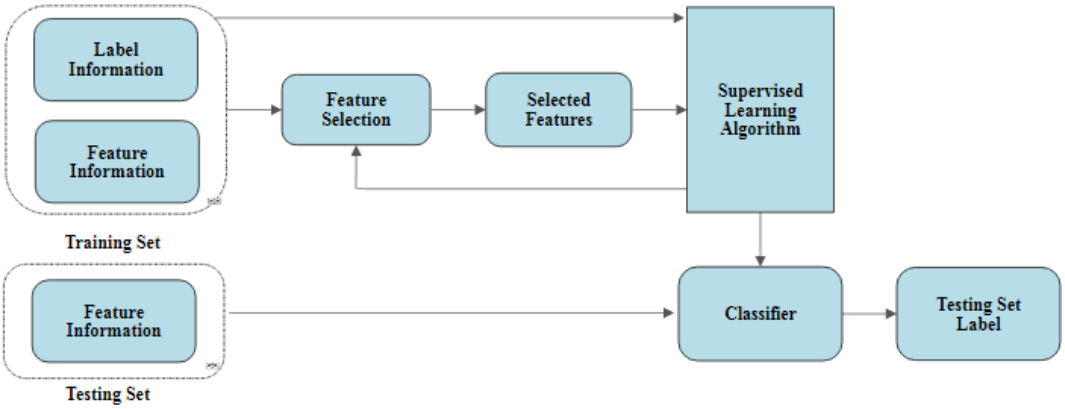

Figure 2.3 : A general framework of supervised feature selection.

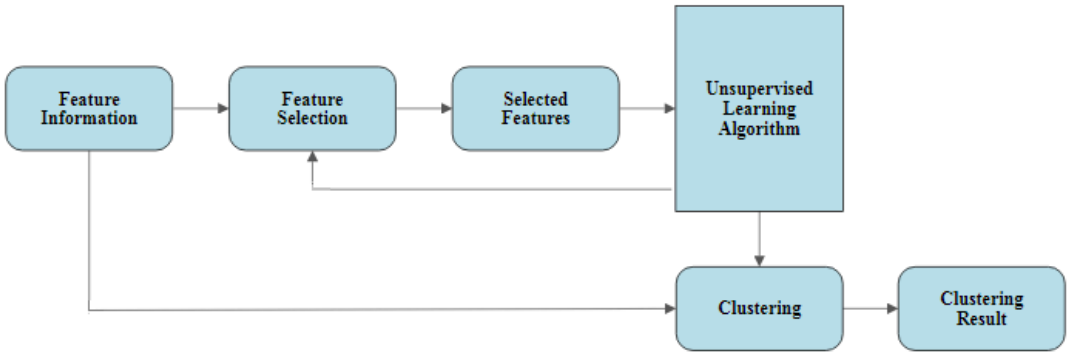

Figure 2.4 : A general framework of unsupervised feature selection.

feature selection but also unsupervised feature selection methods are not appropriate in this scenario [18]. Thus, improving semi-supervised techniques by utilizing both labeled and unlabeled instances are desired. The semi-supervised learning algorithm is demonstrated in Figure 2.5 [18].

With respect to used data set perspective, feature selection can be grouped into two main classes as static data perspective and streaming data perspective [18]. In addition, there is some other sub-data set for static data perspective and streaming data perspective which leads to other feature selection algorithms. The distribution of data perspective feature selection methods is shown in Figure 2.6 [18].

Hundreds of feature selection algorithms have been offered in the last two decades [18]. However, only four main groups and their subgroups are mentioned in this section. These are information theoretical based, sparse learning based, statistical based and similarity based feature selection methods.

These four groups are shown in Figure 2.7 [18].

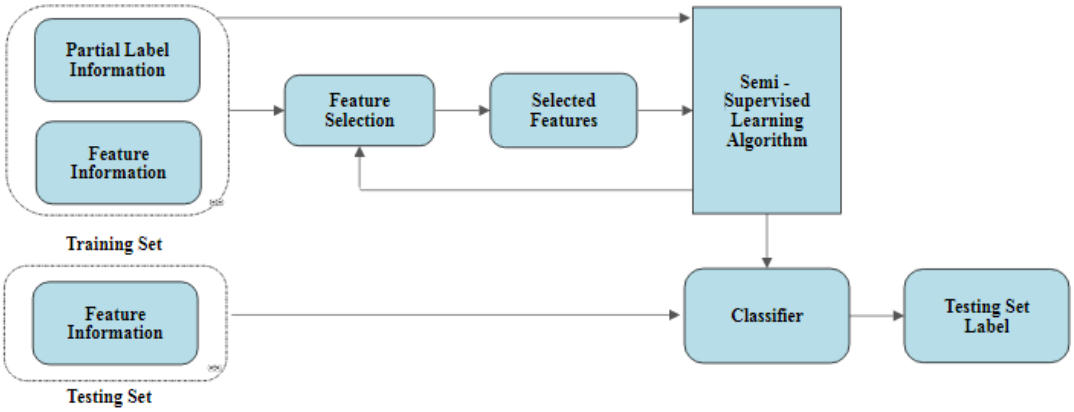

Figure 2.5 : A general framework of semi-supervised feature selection.

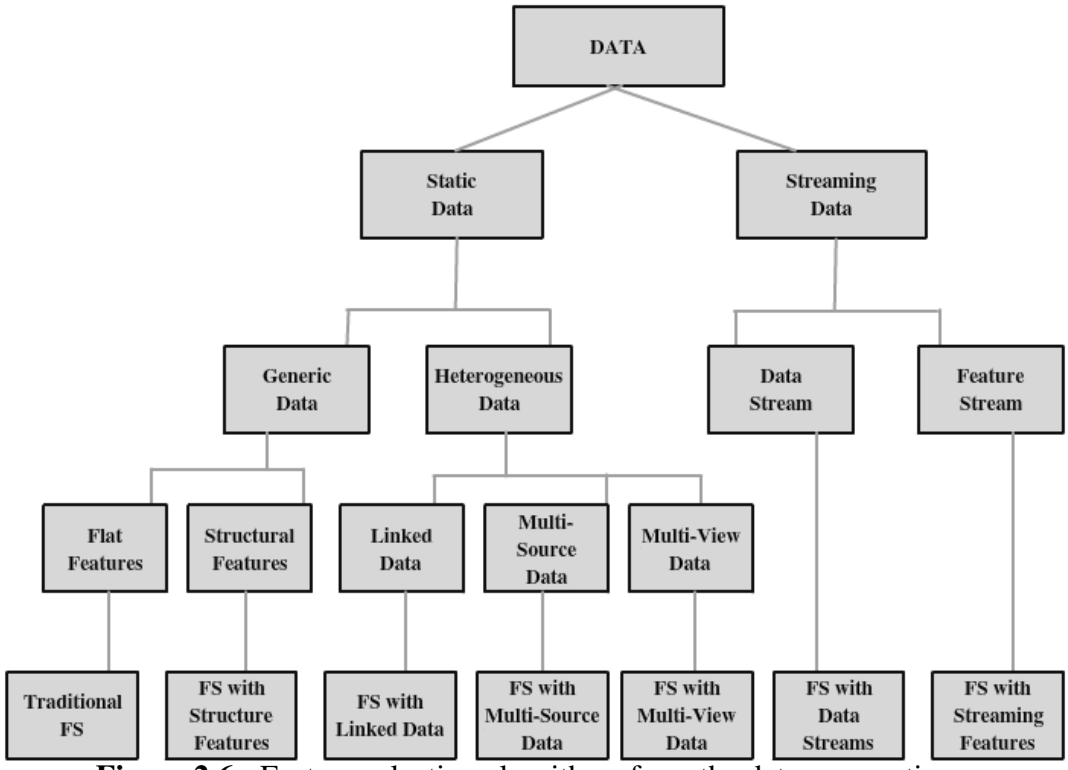

Figure 2.6 : Feature selection algorithms from the data perspective.

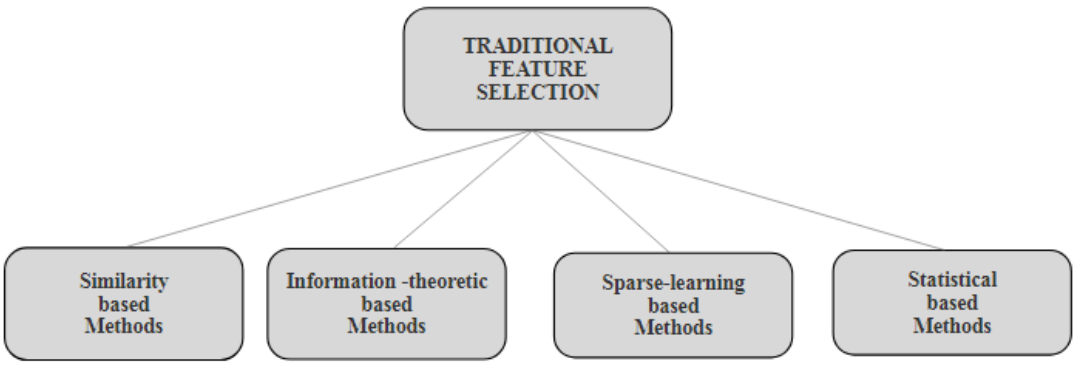

Figure 2.7 : Feature selection algorithms in four groups.

2.1.2.1 Information theoretical based feature selection methods

Information theoretical based methods are the major family of existing feature selection algorithms [18]. Members of this family use dissimilar heuristic filter criteria to gauge the significance of the attributes that maximize the relationship of the attributes and minimize the redundancy of the attributes, and also, these feature selection algorithms can only work with discrete data [18,19]. Some data discretization methods are necessary for numerical attribute values.

Between discrete random variables X and Y, there is a concept called information gain [18, 20] to evaluate their dependance with entropy and conditional entropy. The information gain between X and Y is computed as

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x_i \in X} \sum_{y_j \in Y} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \tag{2.1}$$

where $P(x_i, y_j)$ is the joint probability of x_i and y_j, *H(X)* is the entropy of a random variable X and *H(X|Y)* is the conditional entropy of X given another discrete random variable Y. If random variables X and Y are independent, information gain will be zero; otherwise, information gain is symmetric, i.e. I(X, Y) = I(Y, X) [18].

Following are some feature selection methods based on information theory:

- Mutual Information Maximization (or Information Gain) – MIM
- Mutual Information Feature Selection – MIFS
- Minimum Redundancy Maximum Relevance – MRMR
- Conditional Infomax Feature Extraction – CIFE
- Joint Mutual Information – JMI
- Conditional Mutual Information Maximization – CMIM
- Informative Fragments
- Interaction Capping – ICAP
- Double Input Symmetrical Relevance – DISR
- Fast Correlation Based Filter – FCBF

2.1.2.2 Sparse learning based feature selection methods

The second method is called sparse learning based feature selection methods which minimize the experimental error by inducing regularization term to the objective function so that some of feature coefficients are small or exactly zero [18].

Some sparse learning based feature selection methods are listed as follows:

- Feature Selection with l_1-norm Regularizer
- Feature Selection with $l_{2,1}$-norm Regularizer
- Efficient and Robust Feature Selection – REFS
- Multi-Cluster Feature Selection – MCFS
- $l_{2,1}$-norm Regularized Discriminative Feature Selection
- Feature Selection Using Nonnegative Spectral Analysis – NDFS
- Feature selection via joint embedding learning and sparse regression – JELSR

2.1.2.3 Statistical based feature selection methods

This kind of algorithms is based on several statistical mensurations. Most of these feature selection methods are filter based, because they depend on statistical criteria [18].

Some statistical based feature selection methods are listed as follows:

- Low Variance
- T-score
- F-score
- Chi-Square Score
- Gini Index
- Correlation Based Feature Selection – CFS

2.1.2.4 Similarity based feature selection methods

Principally, feature selection methods utilize a variety of criteria; to illustrate, correlation, dependency, information, distance, separability, and reconfiguration error to identify attribute suitability [18]. In all feature selection methods, similarity based algorithms appraise the significance of features [18].

Some similarity based feature selection methods are listed as follows:

- Laplacian Score
- SPEC
- Fisher Score
- Trace Ratio Criterion
- ReliefF

In this thesis, 8 feature selection methods are used which are given among above methods:

CMIM (Conditional Mutual Information Maximization)

CMIM is the member of information theoretic feature selection group that can only be reduced to a nonlinear combination of Shannon information terms unlike other methods of this group [18]. This method iteratively chooses the features which can maximize the mutual information with the class labels given the selected features. Even if the foreseeable power for class labels is powerful, CMIM does not select a feature similar to the preselected ones [21, 22].

The formula of feature score of each new unchosen feature is given below:

$$J_{CMIM}(X_k) = \min_{X_j \in S} [I(X_k; Y|X_j)] \tag{2.2}$$

where X_k is new unselected feature in all k selected features. In addition, I represents the information gain and S represents selected feature set that includes k selected features.

It can obviously be said that, if X_k is unnecessary when S is known or unless X_k is strongly correlated with the class label Y, I(X_k; $Y|X_j$) value is small.

After some derivations, the criterion of CMIM is equivalent to the following structure [18]:

$$J_{CMIM}(X_k) = I(X_k;Y) - \max_{X_j \in S}[I(X_j;X_k) - I\left(X_j;X_k|Y\right)] \ . \tag{2.3}$$

For this reason, it can be said that this method is a special example of the conditional likelihood maximization framework [18]:

$$J_{cmi}(X_k) = I(X_k;Y) + \sum_{X_j \in S} g[I(X_j;X_k), I\left(X_j;X_k|Y\right)] \tag{2.4}$$

where g is a function of two variables $I(X_j;X_k)$ and $I\left(X_j;X_k|Y\right)$.

MIM (Mutual Information Maximization)

MIM is the member of information theoretic feature selection group that evaluates the significance of a feature by its correlation with the class label [18,23]. The assumption of this method is that if a feature has a powerful correlation with the class label, the classification performance will be better.

The score of mutual information for a new unselected feature X_k is given below [18]:

$$J_{MIM}(X_k) = I(X_k;Y) \ . \tag{2.5}$$

This feature score is individually evaluated independent of other features. Thus, while the redundancy quality of feature is entirely disregarded, just the feature correlation is taken into consideration in MIM [18].

JMI (Joint Mutual Information)

A JMI criterion is recommended by authors to increase shared information between the new unselected attribute and the selected attributes given the class labels [24,25]. The fundamental idea of JMI composes of adding new features which are complemental to available features for given class labels [18].

The formula of the criterion of JMI is given below:

$$J_{JMI}(X_k) = \sum_{X_j \in S} I(X_k, X_j;Y) \ . \tag{2.6}$$

In contradiction to other feature selection methods, which can be shown via the linear combination of Shannon information terms, JMI approach cannot be decreased to the condition likelihood maximization framework unswervingly [18].

The rewritten version of JMI criterion was proposed by some other authors as follows [18, 26]:

$$J_{JMI}(X_k) = I(X_k;Y) - \frac{1}{|S|}\sum_{X_j \in S} I(X_j;X_k) + \frac{1}{|S|}\, sum_{X_j \in S} I\left(X_j;X_k|Y\right) \ . \tag{2.7}$$

ICAP (Interaction Capping)

ICAP is the member of information theoretic feature selection category that is similar to feature selection criterion CMIM [18, 27].

Its formula is given below [18]:

$$J_{CMIM}(X_k) = I(X_k;Y) - \sum_{X_j \in S} \max[0, I(X_j;X_k) - I\left(X_j;X_k|Y\right)] \ . \tag{2.8}$$

DISR (Double Input Symmetrical Relevance)

Another member of information theoretic feature selection group is DISR that performs normalization techniques to normalize mutual information [18, 28].

The formula of feature score of each new unselected feature is given by [18, 29]:

$$J_{DISR}(X_k) = \sum_{X_j \in S} \frac{I(X_j X_k;Y)}{H\left(X_j X_k Y\right)} \ . \tag{2.9}$$

F-score

F-score is the member of statistical based feature selection methods that can accomplish the multi-class condition by testing; however, to realize this situation, samples from different classes should be well distinguished [18, 30].

The f-score of a feature f_i can be calculated as follows [18]:

$$\text{f-score}(f_i) = \frac{\sum_j \frac{n_j}{c-1}(\mu_j - \mu)^2}{\frac{1}{n-c}\sum_j (n_j - 1)\sigma_j^2} \ . \tag{2.10}$$

Here, f_i are given features, n_j, μ, μ_j, σ_j emblematize the number of instances from class j, the mean feature value, the mean feature value on class j, the standard deviation of feature value on class j, sequentially [18].

MRMR (Minimum Redundancy Maximum Relevance)

MRMR is the member of information theoretic feature selection category that considers both attribute relevance and attribute redundancy at the same time [18, 31].

This feature selection approach is disposed to choose features with a high correlation with the output (i.e. class) and a low correlation between themselves [32]. In other words, in accordance with the minimal-redundancy-maximal-relevance criterion that is based on mutual information, this approach orders features.

MRMR criterion formula is given below:

$$J_{MRMR}(X_k) = I(X_k;Y) - \frac{1}{|S|}\sum_{X_j \in S} I(X_k;X_j) \tag{2.11}$$

where β is a nonnegative parameter between zero and one.

It can be easily said that the impact of feature redundancy is progressively decreased when more features are chosen. It is happening harder for new attributes to be unnecessary to the features that have already been in set S when more non-redundant features are chosen.

ReliefF

ReliefF feature selection method is a supervised filter algorithm that is an improved version of the Relief statistical model [18, 33]. This method handles a sample from the data set and performs the feature selection process by creating a model that is related to its closeness to other samples in its class and that is based on its distance to different classes.

The feature score of f_i in relief can be shown below when l data samples are arbitrarily chosen among all n samples:

If l data samples are arbitrarily chosen among all n samples, the feature score of f_i in Relief can be given as follows:

$$\text{Relief-score}(f_i) = \frac{1}{2}\sum_{j=1}^{l} d(X(j,i)) - X(NM(j),i)) - d(X(j,i) - X(NH(j),i)) \tag{2.12}$$

where $NM(j)$ and $NH(j)$ demonstrate the most proximate data samples to x_j with the same class label and different class, respectively [18]. In addition, d(.) is a distance metric that is usually set to be the Euclidean distance [18].

However, Relief can only be used for binary classification. Therefore, the feature score equation above is extended in ReliefF to handle the multi-class classification problem:

$$\begin{aligned}\text{ReliefF-score}(f_i) = &\frac{1}{c}\sum_{j=1}^{l}\frac{-1}{m_j}\cdot\sum_{x_r\in NH(j)}[d(X(j,i))-X(r,i)] \\ &+\sum_{y\neq y_j}\frac{1}{h_{jy}}\cdot\frac{p(y)}{1-p(y)}\cdot\sum_{x_r\in NM(j,y)}[d(X(j,i))-X(r,i))]\end{aligned} \tag{2.13}$$

where c is number of class, $NH(j)$ and $NM(j,y)$ point out the most proximate data samples to x_j in the identical class and a dissimilar class y, respectively, and their sizes are h_{jy} and m_j. $p(y)$ is the proportion of samples with class label y.

ReliefF is equipollent to choosing attributes that maintain a special form of data similarity matrix that can be obtained from class labels [18]. Suppose that the dataset has the identical number of samples in each of the c classes, there are q samples in not only $NM(j)$ but also $NH(j,y)$, the Euclidean distance is applied and all attribute vectors have been normalized. Then, ReliefF criterion is the same to the following with the above supposition [18, 34]:

$$\begin{aligned}\text{ReliefF-score}(f_i) = \sum_{j=1}^{n}\Bigg(&\sum_{s=1}^{q}\frac{1}{q}X(j,i)-X(NM(j)_s))^2 \\ &-\sum_{y\neq y_j}\frac{(\sum_{s=1}^{q}X(j,i)-X(NH(j,y)_s)^2}{(c-1)q}\Bigg)\end{aligned} \tag{2.14}$$

where $NM(j)_s$ represents the s^{th} nearest hit of x_j and $NH(j,y)_s$ indicates the s^{th} nearest miss of x_j in class y [18].

2.1.3 Hybrid variation

Data variation and function variation methods use different methodologies to attain their diversities; moreover, they are not enough to provide the diversity for the ensemble. Thus, Hybrid Variation approach aggregates these two steps of methods [5]. Furthermore, hybrid variation methods generate a higher classification performance than other approaches.

2.2 Classification

Classification is one of the simplest types of supervised learning methods. Basically what it does is to examine the attributes of a new object and incorporate them into

predefined labels. The used data set may simplistically be binary-class or it may be multi-class, too. In training step, the model is trained by using training data sets that have specific labels. After the training phase, the test step is launched and with respect to classification algorithm, the category of test data sets are found. Then, the value of accuracy and error of the model can be obtained. Some types of classification algorithms in machine learning are given as follows.

2.2.1 Logistic regression

Logistic regression is one of the most widely used models in the industry. This classification technique is a linear classifier that is a statistical method for analyzing a data set in which there are one or more than one independent variable that reveal consequences [35].

2.2.2 Naive Bayes classifier

Like logistic regression, Naive Bayes classification technique is a linear classifier that is based on Bayes' Theorem with an assumption of independence among predictors [36]. Even though it is a very difficult method in terms of computation, it is a kind of classification algorithm that works very fast once the data set is trained and works according to the probability of a condition of being the highest.

2.2.3 Decision trees

One of the subjects of machine learning is decision tree learning method. This method constructs classification or regression models in the tree structure form [37]. For this method, there are several algorithms that can be utilized such as Boosted Trees, Rotation Forest and Chi-Square Automatic Interaction Detector.

2.2.4 Random forests

This method is an ensemble learning method for classification using multiple decision trees [38]. It is aimed to boost the classification value by using more than one decision tree during the classification process.

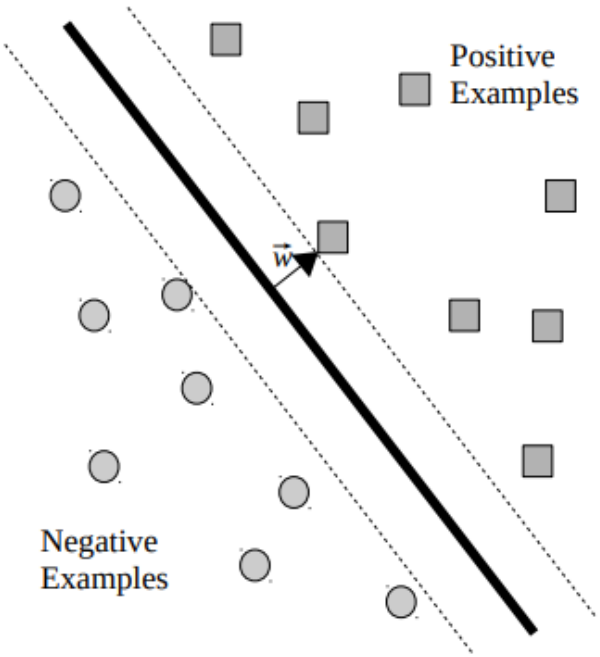

Figure 2.8 : A linear Support Vector Machine.

2.2.5 Support vector machines

Support vector machine (SVM, also support vector network) is used as a base classifier for the model in this study.

In machine learning, SVMs are supervised learning models with interrelated learning algorithms that analyze the data utilized for classification and regression analysis. Given a set of training instances, each signified as related to one or the other of two groups, an SVM training algorithm constructs a model that appoints new samples to one category or the other, making it a non-probabilistic binary linear classifier (even though methods such as Platt scaling exist to use SVM in a stochastic classification setting). An SVM model is a presentment of the instances as points in space, mapped so that the samples of the discrete groups are disunited by a clear gap which is as broad as possible. Novel instances are then mapped into that identical space and forecasted to belong to a group based upon that side of the gap they fall.

Besides fulfilling linear classification, SVMs can efficaciously perform a non-linear classification using what is called the kernel trick, indirectly mapping their inputs into high-dimensional attribute spaces.

SVMs were propounded by Vladimir Vapnik in 1979 [39]. In its common, linear form, an SVM is a hyperplane that divides a set of positive samples from a set of negative instances with maximum margin (see Figure 2.8). In the linear state, the margin is described by the distance of the hyperplane to the nearest of the negative and positive instances. The outcome formula of a linear SVM is

$$u = \vec{w} \cdot \vec{x} - b \tag{2.15}$$

where $\vec{x}$ is the input vector, $\vec{w}$ is the normal vector to the hyperplane, b is a scalar and u is the output of the SVM. The contradistinguishing hyperplane is the plane $u = 0$. The most proximate points lie on the planes $u = \pm 1$. Hence, the margin m is given by

$$m = \frac{1}{\|w\|_2} \, . \tag{2.16}$$

Maximizing this margin can be stated as an optimization problem as follows:

$$\min_{\vec{w},b} \frac{1}{2} \|w\|^2 \text{ subject to } y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, \ \forall \ i \tag{2.17}$$

where x_i is the i^{th} training sample and y_i is the right output of the SVM for the i^{th} training sample. The value y_i is $+1$ for the positive instances and -1 for the negative instances. By use of a Lagrangian, this optimization problem can be transformed into a dual form that is a Quadratic Problem (QP) where the objective function Ψ is merely dependent on a set of Lagrange multipliers α_i:

$$\min_{\vec{\alpha}} \Psi(\vec{\alpha}) = \min_{\vec{\alpha}} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j (\vec{x}_i \cdot \vec{x}_j) \vec{\alpha}_i \vec{\alpha}_j - \sum_{i=1}^{N} \vec{\alpha}_i \tag{2.18}$$

(where N is the number of training samples), subject to the inequality constraints

$$\alpha_i \geq 0, \ \forall \ i \tag{2.19}$$

and one linear equality constraint

$$\sum_{i=1}^{N} y_i \alpha_i = 0 \, . \tag{2.20}$$

Between each Lagrange multiplier and each training instance, there is a one-to-one relation. Once the Lagrange multipliers are identified, the normal vector $\vec{w}$ and the threshold b can be obtained from the Lagrange multipliers:

$$\vec{w} = \sum_{i=1}^{N} y_i \alpha_i \vec{x}_i \ , \ b = \vec{w} \cdot \vec{x}_k - y_k \text{ for some } \alpha_k > 0 \tag{2.21}$$

Because $\vec{w}$ can be calculated by use of Eq. (2.21) from the training data, the amount of calculation required to form an estimate of a linear SVM is constant in the number of non-zero support vectors.

Here, not all data sets are linearly separable. There may be no hyperplane which separates the positive samples from the negative samples. In the formulation hereinabove, the non-separable state would match up to an endless resolution. In addition to this, a modification to the original optimization expression (2.17) that allows, but punishes, the failure of an sample to arrive the correct margin was offered by Cortes and Vapnik in 1995 [40]. That alteration is:

$$\min_{\vec{w},b,\vec{\xi}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \text{ subject to } y_i(\vec{w}\cdot\vec{x}_i - b) \geq 1 - \xi_i,\ \forall\ i \tag{2.22}$$

where ξ_i are slack variables that allow margin failure and C is a parameter that trades off wide margin with a small number of margin failures. When this novel optimization issue is converted into the dual form, it elementarily alters the constraint (2.19) into a box constraint:

$$0 \leq \alpha_i \leq C,\ \forall\ i\ . \tag{2.23}$$

The variables ξ_i do not appear in the dual formulation at all.

SVMs can be even further generalized to nonlinear classifiers [41].

From the Lagrange multipliers, the output of a non-linear SVM is demonstrably calculated:

$$u = \sum_{j=1}^{N} y_j \alpha_j K(\vec{x}_j, \vec{x}) - b \tag{2.24}$$

where K is a kernel function which gauges the resemblance or distance between the input vector $\vec{x}$ and the stored training vector $\vec{x}_j$. Instances of K include Gaussians, polynomials, and neural network nonlinearities [42]. If K is linear, then the equation for the linear SVM (2.15) is recuperated.

The Lagrange multipliers α_i are still calculated by use of a quadratic program. The nonlinearities change the quadratic form; nevertheless, the dual objective function Ψ

is still quadratic in α:

$$\begin{aligned} \min_{\vec{\alpha}} \Psi(\vec{\alpha}) = & \min_{\vec{\alpha}} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j K(\vec{x}_i, \vec{x}_j) \alpha_i \alpha_j - \sum_{i=1}^{N} \alpha_i \\ & 0 \leq \alpha_i \leq C, \; \forall \; i \\ & \sum_{i=1}^{N} y_i \alpha_i = 0 \end{aligned} \tag{2.25}$$

To obtain the QP in Eq. (2.25) as positive definite, the kernel function K must comply Mercer's criteria [42].

The Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient for an optimal point of a positive definite QP. The KKT conditions for the QP in Eq. (2.25) are especially basic. The QP is resolved when, for all i:

$$\begin{aligned} \alpha_i = 0 & \Leftrightarrow y_i u_i \geq 1 \\ 0 \leq \alpha_i \leq C & \Leftrightarrow y_i u_i = 1 \\ \alpha_i = C & \Leftrightarrow y_i u_i \leq 1 \end{aligned} \tag{2.26}$$

where u_i is the output of the SVM for the i^{th} training instance.

In this thesis, multi-class SVM is used; however, the mathematical logic behind the binary class and multi-class classification are the same. In brief, multi-class SVM is based on combining many binary classification decision functions [43].

2.3 Ensemble Pruning Methods

Ensemble pruning methods are excessively used in data mining and machine learning. Ensemble pruning methods relate to the reduction of ensemble of models to increase the efficiency and predictive performance of models; therefore, this approach is quite significant [44]. To enhance ensemble performance and obtain more clever ensembles, these methods are quite beneficial [45].

In the previous studies, different approaches for ensemble pruning have been proposed; however, they can be categorized into four sub categories such as ordering-based, clustering-based, optimization based and other ensemble pruning methods [44].

2.3.1 Ordering-based pruning method

Ordering-based ensemble pruning method was proposed by Margineantu and Dietterich [46]. The simplest methods are in this category. The models of the ensemble

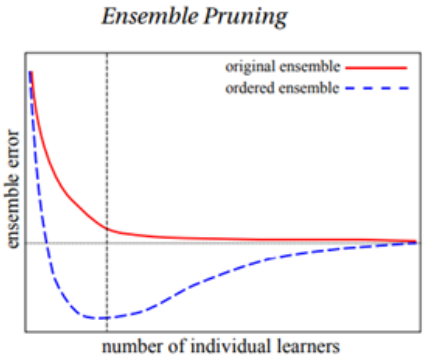

Figure 2.9 : Error curves of the original ensemble (aggregated in random order) and ordered ensemble.

are ranked by ordering-based methods once in accordance with an evaluation function and they are chosen in this stable order [44].

Figure 2.9 compares original ensemble and ordered ensemble [47].

In the literature, there are several ordering-based ensemble pruning methods. Kappa pruning, kappa-error diagram pruning, orientation pruning and complementary measure method are some of the ordering-based ensemble pruning methods.

2.3.1.1 Kappa pruning method

Kappa pruning method uses a diversity measure for appraisal [44, 46]. Selecting the subset of most diversified classifiers from an ensemble is the aim of kappa pruning approach [46]. The diversity is gauged by a statistic value which corresponds to agreement of classifiers on the selection set.

Kappa pruning formula is given below:

$$S_u = \arg\max_k KZ_{tr}(h_k, H_{S_{u-1}}) \quad (k \in E_T \backslash S_{u-1}) \tag{2.27}$$

where K exemplifies the pairwise diversity measure, E_T is ensemble and Z_{tr} represents the training data set in the study of Kuncheva and Whitaker [48].

2.3.1.2 Kappa-error diagram pruning method

Kappa-error diagram pruning method creates a convex hull of the points in the diagram. These points can be considered as a brief of the whole diagram and contain not only the most exact but also the most diverse couples of individual learners [47]. The ensemble, that is pruned, composes of any individual learner which seems in a pair corresponding

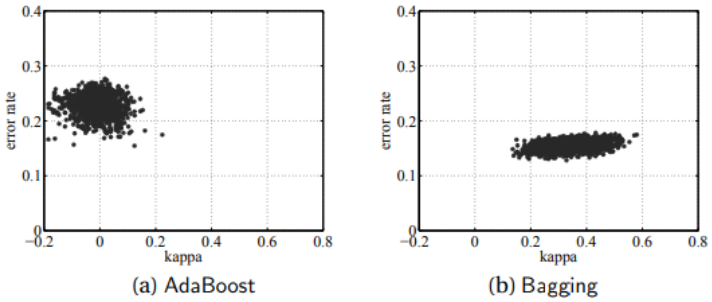

Figure 2.10 : Instances of kappa-error diagrams on credit-g data set.

to a point on the convex hull. This pruning method contemporaneously considers the accuracy along with the diversity of each learner [47].

This method is based on the kappa-error diagram which is shown in Figure 2.10 [47].

As shown in Figure 2.10, visualizing the classifiers ensemble is provided by the Kappa-Error diagram [46].

2.3.1.3 Orientation pruning method

Orientation ordering ensemble pruning method is an active and productive ranking-based pruning approach for classifier ensembles [44, 49]. The angle between a reference vector and a signature vector is increased by orientation ordering method in order that the ensemble classifiers are ranked according to value of this angle [44].

Here, the signature vector is defined below. To illustrate, the signature vector c^i of the i^{th} individual learner h_i is a $|V|$-dimensional vector where the j^{th} element is

$$c_j^{(i)} = 2\mathbb{I}(h_i(x_j) = y_j) - 1 \tag{2.28}$$

where $(x_j, y_j) \in V$.

For ensemble pruning, this technique is a quickest way which has a time complexity of $O(TN)$ that is the time complexity of the orientation ordering method [44].

In this thesis, Joint Criterion Method that is improved in the study of Fern and Lin is used [50]. This technique intends to optimize the joint criterion function by compounding accuracy and diversity in the same objective function. In the research of Fern and Lin, the joint criterion was used for clustering problems. According to that study, to choose an ensemble with a size K, K clustering solutions are selected to

optimize the objective function below:

$$\alpha \sum_{i=1,2,\ldots,K} SNMI(C_i, L) + (1-\alpha) \sum_{i \neq j} [1 - NMI(C_i, C_j)] \tag{2.29}$$

where the first term defines the quality, the second term computes the pairwise diversity and α determines the importance of objectives [50]. In the above-mentioned objective function, L exemplifies a large library of clustering resolutions, $SNMI(C_i, L)$ gauges the quality of clustering solution C_i that can be calculated as

$$SNMI(C_i, L) = \sum_{i=1}^{r} NMI(C, C_i) \tag{2.30}$$

for a dedicated ensemble E of r clustering solutions remarked by E= $\{C_1, C_2, \ldots, C_r\}$, where $NMI(C, C_i)$ is the normalized mutual information between clustering resolutions C and C_i [50]. The technique begins with choosing the highest-quality solution and continually combining one resolution to the ensemble that maximizes the objective function in Eq. (2.29) [50].

In this thesis, accuracy is used as classification result instead of quality that is the first term of objective function and classification solution is utilized as non-pairwise diversity of each resolution. In contrast with the averaging pairwise measures, non-pairwise measurements try to evaluate the ensemble diversity directly [47]. In the literature, several approaches were proposed to obtain non-pairwise diversity in ensemble such as interrater agreement, Kohavi-Wolpert variance, generalized diversity and coincident failure [47].

2.3.1.4 Complementary measure method

Complementary measure pruning method involves the classifier which has the performance of most supplementary to the selected sub-ensemble [51]. This method begins with elements in the selection set Z_{sel} which is the first classifier with the smallest validation error [30]. In addition, the sub-ensemble is composed of S_{u-1} associated with combination of the highest classification accuracy classifier [51].

The formula of sub-ensemble is

$$S_u = \arg\max_{k} \sum_{(x,y) \in Z_{sel}} I(y = h_k(x) \text{ and } H_{S_{u-1}}(x) \neq y) \tag{2.31}$$

where $k \in E_T \backslash S_{u-1}$.

2.3.2 Clustering-based pruning method

This pruning method is called "clustering-based" since the most simplistic method to specify the prototypes is to use clustering approaches [47]. There are several significant factors to select the best clustering algorithm. Detecting the distance measure, the stage of pruning every cluster and excerpting the convenient number of clusters influence the performance in this pruning approach [44].

Clustering-based pruning methods contain two phases in general: Firstly, clustering algorithms are used to find similar groups of models. Afterwards, the distance measure is used to reduce the repetitive error probability in a different validation set. This distance measure is identical to the measure of diversity [48].

In the previous studies, different clustering techniques have been exploited such as hierarchical agglomerative clustering that considers the possibility that each learner does not make validation errors coincidently as the distance, deterministic annealing for clustering and k-means clustering based on Euclidean distance [47].

2.3.3 Optimization-based pruning method

Ensemble pruning can be regarded as an optimization problem. There are three optimization approaches: Genetic algorithms, semi-definite programming and hill climbing [44].

In optimization and machine learning, genetic algorithms have extensive applications. GASEN is an example of application in genetic algorithms which was proposed by Zhou et al. to create selective ensembles [52]. By using different genetic operators or different coding schemes, there are dissimilar GASEN applications. For instance, in 2003, "bit coding scheme" that directly gets 0-1 weights and refrains the problem of setting a suitable threshold to make decision which individual learner should be exempted was used by Zhou and Tang [47, 52].

2.3.4 Other pruning methods

The methods in this category are the methods which are not included in other groups, such as methods based on statistical techniques for directly choosing classifiers' subset, based on boosting or based on reinforcement learning.

3. THE PROPOSED MODEL

In this thesis, the proposed model is constituted via combination of hybrid variation and joint criterion ensemble pruning algorithm.

According to the proposed model, ensemble based feature selection is utilized with a hybrid model. The components of hybrid model are bagging method regarding data variation and eight feature selection methods regarding function variation which are CMIM, MIM, JMI, ICAP, DISR, f-score, MRMR and reliefF algorithms [see Section 2.1.2 for details]. With the hybrid model, one of the supervised classification algorithms, multi-class SVM, is used to classify the consequences of the model. In the training stage, prediction results of all bags are obtained by multi-class SVM. At this point, pruning phase is started for the training data set.

In the study of Fern and Lin [50], joint criterion ensemble pruning method was utilized, as mentioned in the previous chapter. According to joint criterion pruning, quality and pairwise diversity were used for clustering solutions. However in this study, not only pairwise diversity, but also non-pairwise diversity is combined with accuracy; comparison of these two approaches is done and a better result with non-pairwise type is obtained. Moreover, in this thesis, different α parameter values are examined and their comparisons are presented. For all trial results, majority voting is implemented on all bags, then the most preferred feature selection methods are determined with different subset sizes.

After the detection of the sub-ensemble, the testing phase is launched. The feature selection methods selected by joint criterion ensemble pruning method are implemented on the test data set. Finally, the classification solutions are selected by majority voting and accuracy of the model is designated.

Table 3.1 : The pseudocode of Hybrid Variation.

Algorithm 3 Combining bagging and several feature selection algorithms

```
1: for i = 1,2,...,n do
2:     for j = 1,2,...,k do
3:         selectedFeatures[i, j] = fs[j](bags[i])
4:     end for
5: end for
```

3.1 Hybrid Model

Bagging by means of data variation is the most largely utilized ensemble learning technique that retains a significant role in finding the subset of samples and features to acquire diverse classifiers given data instances. In the proposed model, the number of bags is 30 and the number of samples in each bag is 100.

Additionally, function diversity methods that make use of multiple feature selection methods at the same time are commonly used ensemble learning techniques to gain more diverse classifiers. Unlike these two approaches, hybrid variation aggregates both data variation and function variation phases; as it is claimed that involving data variation or function variation alone is not sufficient to generate a good ensemble.

The pseudocode of hybrid variation is given in Table 3.1. This algorithm takes n bags as input and outputs selected features.

After combining of bags and several feature selection algorithms which construct the hybrid model, the classification phase is launched. Accordingly, hybrid diversity is utilized as a base ensemble model with multi-class SVM as a base classifier of the model.

The pseudocode of the hybrid model with multi-class SVM is given in Table 3.2. This algorithm takes the selected features as input and gives predictions as output.

These pseudocodes all belong to hybrid diversity approach. At the end of the classification, the prediction values of all samples are obtained and ensemble pruning phase can start.

Table 3.2 : The pseudocode of multi-class SVM with Hybrid Model.

Algorithm 4 Multi-class SVM with Hybrid Model

```
for i = 1,2,...,n do
    for j = 1,2,...,k do
        predictedValues[i, j] =SVM(selectedFeatures[i, j])
    end for
end for
```

3.2 Hybrid Model with Joint Criterion Ensemble Pruning Method

Joint Criterion ensemble pruning method is the one of the ordering based pruning techniques. In the literature, Joint Criterion is used for clustering problems and is developed with quality and pairwise diversity in the study of Fern and Lin [50]. However, this ensemble pruning method did not use classification solutions of hybrid variation in previous studies. In this thesis, it is proposed to identify the best number of classification solutions that optimizes the trade-off between accuracy and diversity by utilizing the hybrid diversity model to fill the gap in the literature. And also, two kinds of diversity are combined with accuracy as pairwise and non-pairwise diversity.

In the study of Fern and Lin, the objective function is given by

$$\alpha \sum_{i=1,2,\ldots,K} SNMI(C_i,L) + (1-\alpha) \sum_{i\neq j} [1 - NMI(C_i,C_j)] \tag{3.1}$$

where the first part of equation is the sum of the quality of the selected clustering resolutions, the second part gauges their pairwise diversity and α detects the significance of objectives as mentioned previously [50]. In addition, $NMI(C_i,C_j)$ represents the normalized mutual information between two clustering solutions and $SNMI(C_i,L)$ represents the sum of normalized mutual information of clustering solutions in the library L.

However in this study, joint criterion is combined with hybrid model as the first component measures the sum of the accuracy of the classification results and the second part measures the non-pairwise diversity of the prediction results of the classification. According to this,

$$\alpha \sum_{i=1,2,\ldots,m} \frac{Acc(K_i,L)}{i} + (1-\alpha) \sum_{i=1,2,\ldots,m} (1 - Div(K_i,L)) \tag{3.2}$$

where $Acc(.)$ is called accuracy function of classification solutions of K_i in the library L and $Div(.)$ is called non-pairwise diversity function of classification solutions of K_i in the library L. According to this, the first part of the equation is the arithmetic mean of the accuracy of the selected classification results and the second one is the non-pairwise diversity of the prediction of the classification results. Here, the reason for taking the arithmetic mean of the accuracy is to reduce the value of summation. If the arithmetic mean was not utilized to reduce of the summation of the accuracy, this summation would be in the higher interval and the value of the diversity would still remain in the $(0,1)$ interval for each iteration, so the importance of diversity could not be enough to select the next iteration of the process. Therefore, arithmetic mean should be used. L is the library of the classification solutions: $L = (L_1, L_2, \ldots, L_m)$. In this study, α is chosen as 0.5 for pairwise and non-pairwise cases like in previous research in clustering. In addition, different choices for α are examined to investigate its sensitivity and for comparison of the results of the pairwise case.

For the first iteration, when $i = 1$; it cannot be calculated any diversity by using only one classification result. Thus, the result of the first step only depends on accuracy. Moreover, there are 30 bags, so majority voting is needed among all bags. After majority voting among all bags, the first chosen feature selection technique is always MRMR for each case of diversity.

For the second iteration, the accuracy and diversity should be calculated with MRMR with the other methods one by one for each case of the pruning. Therefore, there is a summation for non-pairwise case as follows:

$$\alpha \sum_i \frac{Acc(K_{MRMR}, L_{rest1})}{2} + (1-\alpha) \sum_i [1 - Div(K_{MRMR}, L_{rest1})] \ . \tag{3.3}$$

As mentioned above, if the arithmetic mean was not utilized to reduce the summation of the accuracy, this summation would be in the $(1,2)$ interval and the value of the diversity would still remain in the $(0,1)$ interval, so the importance of diversity could not be enough to select for the second iteration. For the second iteration, selected feature selection method is MIM after majority voting is implemented on each bag.

For the third iteration, the accuracy and diversity should be measured with MRMR and MIM with the other methods one by one. Hence, the next summation becomes:

$$\begin{aligned}&\alpha \sum_{i} \frac{Acc(K(MRMR), K(MIM), L_{rest2})}{3} \\ &+ (1-\alpha) \sum_{i} [1 - Div(K(MRMR), K(MIM), L_{rest2})]\end{aligned} \tag{3.4}$$

For the third iteration, selected feature selection method is reliefF after implementation of majority voting on each bag.

This procedure can be continued until $i = 8$ because there are eight feature selection methods used to generate the ensemble library. In this process, the next selected methods will be DISR, JMI, CMIM, f-score and ICAP.

4. MATERIALS AND EXPERIMENTAL SETUP

4.1 Data Set

In order to calculate the classification performance of the proposed method, a labelled data set that was obtained from Twitter users of different age groups by Antonio A. Morgan-Lopez, Annice E. Kim, Robert F. Chew and Paul Ruddle via accumulating publicly available birthday announcement tweets by using the Twitter Search application programming interface (API) is used in this study [53]. According to this data set, birthday tweets between the ages of 13 - 50 were gathered on August 22, 2014, September 29, 2014, April 2, 2015, and June 21, 2015 by authors who are mentioned above.

In this data set, there are 3184 samples with 3 classes: 1036 samples for young people who are in 13 - 17 range, 1634 samples for young adults who are in 18 - 24 range and 514 adults who are 25 or older. The category "1" belongs to 13 - 17 range, the category "2" belongs to 18 - 24 range and the last category "3" belongs to 25 or older people. These are all classes of this data set; therefore, multi-class SVM is used to measure the classification performance. The distribution of data density is demonstrated in Table 4.1.

These 3184 samples are split as eighty percent for training data set and twenty percent for test data set. According to this splitting, 2548 samples are in the training set and 636 samples are in the test set.

From 3184 birthday tweets, several features were extracted that are language features only, meta-data features only, language and meta-data features, and World Well-Being Project (WWBP) words and phrases [53]. In total, 38.536 features were collected; but in this study, 38.529 are utilized by excluding the non-numeric ones.

Table 4.1 : Number of Twitter users described from birthday tweets by age category.

Age Group	Number
Youth: 13–17	1036
Young adults: 18–24	1634
Adults: 25 or older	514

4.2 Software

In order to run the experiments, hybrid variation approach with multi-class SVM and Joint Criterion ensemble pruning method, data analysis is done in MatLab 2018a by using FEAST Library which is a feature selection toolbox for MatLab.

5. EXPERIMENTAL RESULTS

In this study, the proposed model with joint criterion ensemble pruning method by non-pairwise diversity is implemented on the test data set.

On the test phase, the accuracies of the pruned-hybrid model with non-pairwise diversity are calculated with $\alpha = 0.5$ and the resuls of all subset are given in Table 5.1. In addition, for subset-size $= 1$, this gives MRMR data variation results. And also, for subset-size $= 8$, this gives full ensemble; hence, it can be called as hybrid variation without pruning.

The graph of the proposed model is demonstrated in Figure 5.1. It can be interpreted that when the subset size is equal to six or seven, the performance of the model is stable and the most accurate among all subset sizes. Additively, the model is saturated with these subset sizes.

Then, accuracies are calculated one by one for eight feature selection methods. These can be called as data variation because for all results, there are bags and only one feature selection. The whole accuracies are calculated via implemented majority voting on each bag whose results are shown in Table 5.2.

The accuracy comparison graph of these two results is given in Figure 5.2. It can be said that the proposed model is more stable than data variation. Additionally, the results of the proposed model are more accurate for almost each subset size.

Now, there are several tables with different α value for pruned ensemble with pairwise diversity. The results of pruned ensemble with pairwise diversity and $\alpha = 0.5$ are demonstrated in Table 5.3. The results of pruned ensemble with pairwise diversity and $\alpha = 0.6$ are demonstrated in Table 5.4. The results of pruned ensemble with pairwise diversity and $\alpha = 0.7$ are demonstrated in Table 5.5. The results of pruned ensemble with pairwise diversity and $\alpha = 0.8$ are demonstrated in Table 5.6.

Table 5.1 : The accuracy of hybrid variation with non-pairwise diversity of Joint Criterion for different size of subset.

Pruned Ensemble Subset Size	Accuracy
1 (MRMR)	0.897
2 (MRMR, MIM)	0.903
3 (MRMR, MIM, reliefF)	0.893
4 (MRMR, MIM, reliefF, DISR)	0.915
5 (MRMR, MIM, reliefF, DISR, JMI)	0.9386
6 (MRMR, MIM, reliefF, DISR, JMI, CMIM)	0.9386
7 (MRMR, MIM, reliefF, DISR, JMI, CMIM, f-score)	0.9323
8 (MRMR, MIM, reliefF, DISR, JMI, CMIM, f-score, ICAP)	0.9261

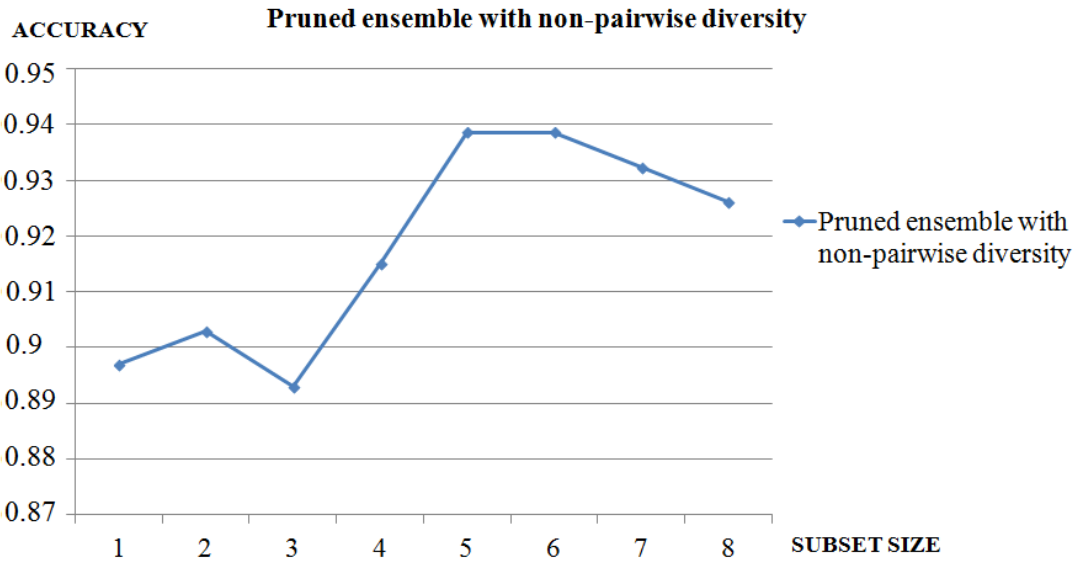

Figure 5.1 : The graph of pruned ensemble model with non-pairwise diversity.

Table 5.2 : The accuracy of data variation for utilized feature selection methods.

Feature Selection Methods	Accuracy of test set (one by one)
CMIM	0.88
DISR	0.91
f-score	0.56
ICAP	0.86
JMI	0.92
MIM	0.90
MRMR	0.897
reliefF	0.59

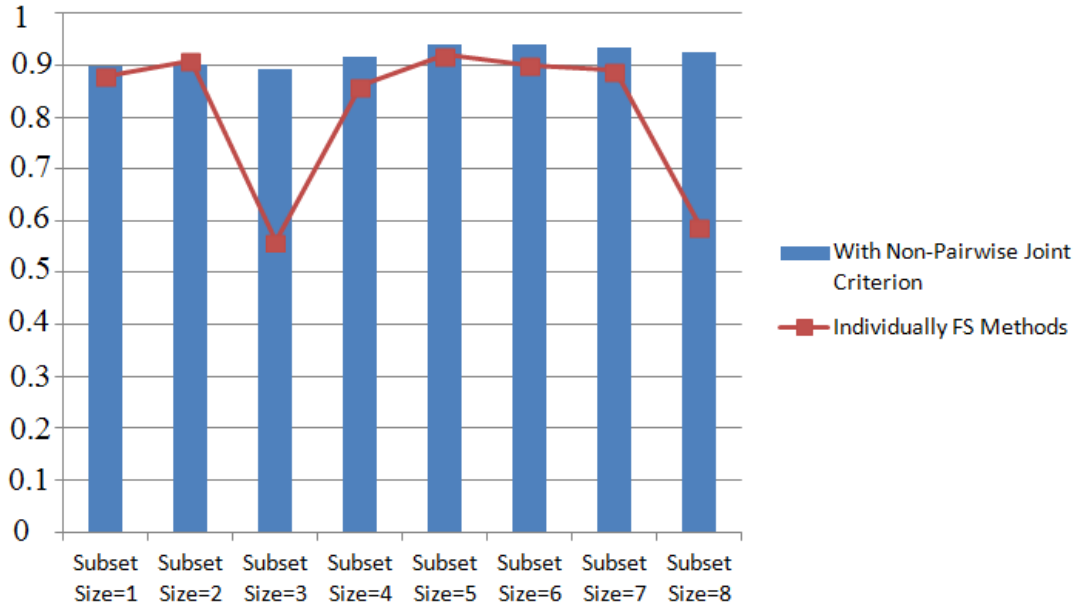

Figure 5.2 : Graph of the comparison of Joint Criterion with non-pairwise diversity and Data Variation.

Table 5.3 : The accuracy of hybrid variation with pairwise diversity of Joint Criterion when $\alpha = 0.5$ for different size of subset.

Pruned Ensemble Subset Size	Accuracy
1 (MRMR)	0.897
2 (MRMR, DISR)	0.909
3 (MRMR, DISR, reliefF)	0.907
4 (MRMR, DISR, reliefF, ICAP)	0.909
5 (MRMR, DISR, reliefF, ICAP, f-score)	0.893
6 (MRMR, DISR, reliefF, ICAP, f-score, CMIM)	0.9104
7 (MRMR, DISR, reliefF, ICAP, f-score, CMIM, JMI)	0.923
8 (MRMR, DISR, reliefF, ICAP, f-score, CMIM, JMI, MIM)	0.9261

Table 5.4 : The accuracy of hybrid variation with pairwise diversity of Joint Criterion when $\alpha = 0.6$ for different size of subset.

Pruned Ensemble Subset Size	Accuracy
1 (MRMR)	0.897
2 (MRMR, DISR)	0.909
3 (MRMR, DISR, reliefF)	0.907
4 (MRMR, DISR, reliefF, f-score)	0.8302
5 (MRMR, DISR, reliefF, f-score, ICAP)	0.893
6 (MRMR, DISR, reliefF, f-score, ICAP, CMIM)	0.9104
7 (MRMR, DISR, reliefF, f-score, ICAP, CMIM, MIM)	0.923
8 (MRMR, DISR, reliefF, f-score, ICAP, CMIM, MIM, JMI)	0.9261

Table 5.5 : The accuracy of hybrid variation with pairwise diversity of Joint Criterion when $\alpha = 0.7$ for different size of subset.

Pruned Ensemble Subset Size	Accuracy
1 (MRMR)	0.897
2 (MRMR, DISR)	0.909
3 (MRMR, DISR, CMIM)	0.9324
4 (MRMR, DISR, CMIM, reliefF)	0.912
5 (MRMR, DISR, CMIM, reliefF, f-score)	0.9025
6 (MRMR, DISR, CMIM, reliefF, f-score, ICAP)	0.9104
7 (MRMR, DISR, CMIM, reliefF, f-score, ICAP, JMI)	0.923
8 (MRMR, DISR, CMIM, reliefF, f-score, ICAP, JMI, MIM)	0.9261

Table 5.6 : The accuracy of hybrid variation with pairwise diversity of Joint Criterion when $\alpha = 0.8$ for different size of subset.

Pruned Ensemble Subset Size	Accuracy
1 (MRMR)	0.897
2 (MRMR, DISR)	0.909
3 (MRMR, DISR, MIM)	0.9308
4 (MRMR, DISR, MIM, CMIM)	0.9277
5 (MRMR, DISR, MIM, CMIM, f-score)	0.9261
6 (MRMR, DISR, MIM, CMIM, f-score, reliefF)	0.9104
7 (MRMR, DISR, MIM, CMIM, f-score, reliefF, ICAP)	0.923
8 (MRMR, DISR, MIM, CMIM, f-score, reliefF, ICAP, JMI)	0.9261

6. CONCLUSIONS AND RECOMMENDATIONS

In this thesis, a novel ensemble based feature selection with hybrid diversity approach is developed by joint criterion ensemble pruning technique with pairwise and non-pairwise diversity. The proposed model is validated on the age verified Twitter data set that was created by Antonio A. Morgan-Lopez, Annice E. Kim, Robert F. Chew and Paul Ruddle as mentioned before. In addition, the result of the proposed model is compared with joint criterion with pairwise diversity with different α parameters. And also, the proposed model is compared with eight data variation results.

First of all, the non-pairwise situation is investigated, and then it is observed that there are two subset sizes which are saturating sizes as expected. Secondly, the pairwise condition is examined, but there is no saturating size between 1 and 8, because of the ensemble size. Therefore, it can be said that these two conditions are novel; however, the non-pairwise situation is better than the other.

Afterwards, the results of feature selection which can be called data variation, are investigated, and compared with other results which are mentioned above. Again, it can be said that the proposed model gives better results.

For the future studies, one can try other classification algorithms to compare the results with several data sets and larger ensemble sizes.

REFERENCES

[1] **Han, J.**, **Kamber, M. and Pei, J.** (2011). Data mining concepts and techniques third edition, *Morgan Kaufmann*.

[2] **Suthaharan, S.** (2016). Machine learning models and algorithms for big data classification, *Integr. Ser. Inf. Syst*, *36*, 1–12.

[3] **Guyon, I. and Elisseeff, A.**, (2006). An introduction to feature extraction, Feature extraction, Springer, pp.1–25.

[4] **Liu, H. and Motoda, H.** (1998). *Feature extraction, construction and selection: A data mining perspective*, volume453, Springer Science & Business Media.

[5] **Guan, D.**, **Yuan, W.**, **Lee, Y.K.**, **Najeebullah, K. and Rasel, M.K.** (2014). A review of ensemble learning based feature selection, *IETE Technical Review*, *31*(3), 190–198.

[6] **Budak, H.** Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım, *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *22*, 21–31.

[7] **Zhang, Y.**, **Burer, S. and Street, W.N.** (2006). Ensemble pruning via semi-definite programming, *Journal of Machine Learning Research*, *7*(Jul), 1315–1338.

[8] **Dittman, D.J.**, **Khoshgoftaar, T.M.**, **Wald, R. and Napolitano, A.** (2012). Comparing two new gene selection ensemble approaches with the commonly-used approach, *2012 11th International Conference on Machine Learning and Applications*, volume 2, IEEE, pp.184–191.

[9] **Saeys, Y.**, **Abeel, T. and Van de Peer, Y.** (2008). Robust feature selection using ensemble feature selection techniques, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, pp.313–325.

[10] **Xu, H.**, **Caramanis, C. and Mannor, S.** (2011). Sparse algorithms are not stable: A no-free-lunch theorem, *IEEE transactions on pattern analysis and machine intelligence*, *34*(1), 187–193.

[11] **Breiman, L.** (1996). Bagging predictors, *Machine learning*, *24*(2), 123–140.

[12] **Sewell, M.** (2008). Ensemble learning, *RN*, *11*(02).

[13] **Quinlan, J.R.** *et al.* (1996). Bagging, boosting, and C4. 5, *AAAI/IAAI, Vol. 1*, pp.725–730.

[14] **Schapire, R.**, (1999), EA brief introduction to boosting, proceedings of the 16th International Joint Conference on Artificial Intelligence.

[15] **Freund, Y. and Schapire, R.E.** (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of computer and system sciences*, *55*(1), 119–139.

[16] **Hastie, T.**, **Rosset, S.**, **Zhu, J. and Zou, H.** (2009). Multi-class adaboost, *Statistics and its Interface*, *2*(3), 349–360.

[17] **Tuv, E.**, **Borisov, A.**, **Runger, G. and Torkkola, K.** (2009). Feature selection with ensembles, artificial variables, and redundancy elimination, *Journal of Machine Learning Research*, *10*(Jul), 1341–1366.

[18] **Li, J.**, **Cheng, K.**, **Wang, S.**, **Morstatter, F.**, **Trevino, R.P.**, **Tang, J. and Liu, H.** (2018). Feature selection: A data perspective, *ACM Computing Surveys (CSUR)*, *50*(6), 94.

[19] **Duda, R.O.**, **Hart, P.E. and Stork, D.G.** (2012). *Pattern classification*, John Wiley & Sons.

[20] **Shannon, C.E.** (2001). A mathematical theory of communication. ACM SIGMOBILE Mob, *Comput. Commun. Rev*, *5*(1), 3–55.

[21] **Vidal-Naquet, M. and Ullman, S.** (2003). Object Recognition with Informative Features and Linear Classification., *ICCV*, volume 3, p.281.

[22] **Fleuret, F.** (2004). Fast binary feature selection with conditional mutual information, *Journal of Machine learning research*, *5*(Nov), 1531–1555.

[23] **Lewis, D.D.** (1992). Feature selection and feature extraction for text categorization, *Proceedings of the workshop on Speech and Natural Language*, Association for Computational Linguistics, pp.212–217.

[24] **Yang, H.H. and Moody, J.** (2000). Data visualization and feature selection: New algorithms for nongaussian data, *Advances in neural information processing systems*, pp.687–693.

[25] **Meyer, P.E.**, **Schretter, C. and Bontempi, G.** (2008). Information-theoretic feature selection in microarray data using variable complementarity, *IEEE Journal of Selected Topics in Signal Processing*, *2*(3), 261–274.

[26] **Brown, G.**, **Pocock, A.**, **Zhao, M.J. and Luján, M.** (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, *Journal of machine learning research*, *13*(Jan), 27–66.

[27] **Jakulin, A.** (2005). Machine learning based on attribute interactions: phd dissertation, *Ph.D. thesis*, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko.

[28] **Meyer, P.E. and Bontempi, G.** (2006). On the use of variable complementarity for feature selection in cancer classification, *Workshops on applications of evolutionary computation*, Springer, pp.91–102.

[29] **Guyon, I.**, **Gunn, S.**, **Nikravesh, M. and Zadeh, L.A.** (2008). *Feature extraction: foundations and applications*, volume207, Springer.

[30] **Wright, S.** (1965). The interpretation of population structure by F-statistics with special regard to systems of mating, *Evolution*, *19*(3), 395–420.

[31] **Peng, H.**, **Long, F. and Ding, C.** (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (8), 1226–1238.

[32] **Radovic, M.**, **Ghalwash, M.**, **Filipovic, N. and Obradovic, Z.** (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data, *BMC bioinformatics*, *18*(1), 9.

[33] **Robnik-Šikonja, M. and Kononenko, I.** (2003). Theoretical and empirical analysis of ReliefF and RReliefF, *Machine learning*, *53*(1-2), 23–69.

[34] **Zhao, Z. and Liu, H.** (2007). Spectral feature selection for supervised and unsupervised learning, *Proceedings of the 24th international conference on Machine learning*, ACM, pp.1151–1157.

[35] **Faul, F.**, **Erdfelder, E.**, **Buchner, A. and Lang, A.G.** (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses, *Behavior research methods*, *41*(4), 1149–1160.

[36] **Mukherjee, S. and Sharma, N.** (2012). Intrusion detection using naive Bayes classifier with feature reduction, *Procedia Technology*, *4*, 119–128.

[37] **Vens, C.**, **Struyf, J.**, **Schietgat, L.**, **Džeroski, S. and Blockeel, H.** (2008). Decision trees for hierarchical multi-label classification, *Machine learning*, *73*(2), 185.

[38] **McDonald, A.D.**, **Lee, J.D.**, **Schwarz, C. and Brown, T.L.** (2014). Steering in a random forest: Ensemble learning for detecting drowsiness-related lane departures, *Human factors*, *56*(5), 986–998.

[39] **Vapnik, V.**, (1982), Estimation of Dependences Based on Empirical Data Berlin.

[40] **Cortes, C. and Vapnik, V.** (1995). Support-vector networks, *Machine learning*, *20*(3), 273–297.

[41] **Boser, B.E.**, **Guyon, I.M. and Vapnik, V.N.** (1992). A training algorithm for optimal margin classifiers, *Proceedings of the fifth annual workshop on Computational learning theory*, ACM, pp.144–152.

[42] **Burges, C.J.** (1998). A tutorial on support vector machines for pattern recognition, *Data mining and knowledge discovery*, *2*(2), 121–167.

[43] **Weston, J. and Watkins, C.** (1998). Multi-class support vector machines, **Technical Report**, Citeseer.

[44] **Tsoumakas, G.**, **Partalas, I. and Vlahavas, I.**, (2009). An ensemble pruning primer, Applications of supervised and unsupervised ensemble methods, Springer, pp.1–13.

[45] **Guo, H.**, **Liu, H.**, **Li, R.**, **Wu, C.**, **Guo, Y. and Xu, M.** (2018). Margin & diversity based ordering ensemble pruning, *Neurocomputing*, *275*, 237–246.

[46] **Dietterich, T. and Margineantu, D.** (1997). Pruning Adaptive Boosting, *14th Int'l Conf. Mach. Learn.*, pp.211–218.

[47] **Zhou, Z.H.** (2012). *Ensemble methods: foundations and algorithms*, Chapman and Hall/CRC.

[48] **Kuncheva, L.I. and Whitaker, C.J.** (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, *Machine learning*, *51*(2), 181–207.

[49] **Martínez-Muñoz, G. and Suárez, A.** (2006). Pruning in ordered bagging ensembles, *Proceedings of the 23rd international conference on Machine learning*, ACM, pp.609–616.

[50] **Fern, X.Z. and Lin, W.** (2008). Cluster ensemble selection, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *1*(3), 128–141.

[51] **Martínez-Muñoz, G.**, **Hernández-Lobato, D. and Suárez, A.** (2008). An analysis of ensemble pruning techniques based on ordered aggregation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 245–259.

[52] **Zhou, Z.H. and Tang, W.** (2003). Selective ensemble of decision trees, *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, Springer, pp.476–483.

[53] **Morgan-Lopez, A.A.**, **Kim, A.E.**, **Chew, R.F. and Ruddle, P.** (2017). Predicting age groups of Twitter users based on language and metadata features, *PloS one*, *12*(8), e0183537.

CURRICULUM VITAE

Name Surname: Ceylan Demir

Place and Date of Birth: Bursa, 15.05.1993

E-Mail: demircey@itu.edu.tr

EDUCATION:

- **B.Sc.:** 2016, Istanbul Technical University, Faculty of Science and Letters, Mathematical Engineering
- **M.Sc.:** 2019, Istanbul Technical University, Graduate School of Science Engineering and Technology, Mathematical Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- Jul. 2018 - Present: TÜBİTAK 1001 Project Assistant, Bahçeşehir University, Istanbul.
- Dec. 2017 - Jul. 2018: Business Analyst, Garanti Technology - Digital Channels, Call Center and Web Applications, Istanbul.
- Mar. 2017 - Oct. 2017: Long Term Intern, Şişecam - Network Technology Group of IT Development Center, Istanbul.
- Feb. 2016 - Apr. 2016: Intern, Pinnera - Software Department, Istanbul.
- Aug. 2015 - Sep. 2015: Intern, AXA Insurance - Corporate and Commercial Technical Department, Istanbul.
- May 2015 - Jun. 2015: Intern, Istanbul Technical University - Faculty of Science and Letters, Istanbul.
- Nov. 2014 - Feb. 2015: Student Assistant, Istanbul Technical University - Faculty of Science and Letters, Istanbul.
- Apr. 2013 - May. 2014: Student Assistant, Istanbul Technical University - Department of Information Technologies, Istanbul.
- May 2012 - Oct. 2012: Counsellor, Istanbul Technical University - Science and Society Research Center, Istanbul.

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- Demir, C., Akyüz, S., and Göksel, İ., 2019: Ensemble Feature Selection for Sentiment and Semantic Analysis. 30. European Conference on Operational Research Conference, June 23-26, 2019 Dublin, Ireland.