A Hybrid Approach for Feature Subset Selection to Classify High-Dimensional Data



Sanzida Akter (011 142 032)
Hasib Rashid Pritom (011 141 100)
Antara Anika Reza (011 142 065)
Rakib Hasan Suvo (011 142 016)
Department of Computer Science and Engineering
United International University

A thesis submitted for the degree of
BSc in Computer Science & Engineering

March 2019

Abstract

In this modern era, we need to deal with large dimension of data set every day. Be it on social media, business, medicine or gene expression there are thousands of millions data to be processed everyday. So dealing with big data or high dimensional data is very important. Feature selection is one of the best way to reduce the dimension of the large data down to a point where we can handle them. In this paper we have proposed three approaches for handling high dimensional data. A hybrid approach for high dimensional data sets and two new manner to apply ReliefF and Symmetrical Uncertainty Attribute Evaluation. This new hybrid approach is a combination of filter and wrapper method and it was tested through intrinsic experiments and proved to be quite efficient and good in terms of feature selection. Wrapper method is great for producing the best result but for high dimensional data sets the Wrapper method could take days. But with our approach it took few minutes as most of the features that are irrelevant or redundant were trimmed off. So this approach is less expensive than wrapper but more efficient than filter. Also our manner of using ReliefF and Symmetrical Uncertainty shows us quite good results if we choose the right threshold and work through with proposed manner.

This work is dedicated to our family and friends.

Acknowledgements

We would really like to express our gratitude towards our thesis supervisor Dr. Dewan Md. Farid, Associate Professor, Department of Computer Science and Engineering, United International University. Without his guidance, support and enthusiasm we won't be able to go further. He has motivated us with his life lessons and give us the opportunity to work with him. We would like to thank Prof. Dr. Chowdhury Mofizur Rahman, Vice Chancellor, United International University. And we would also like express our gratitude towards Dr. Swakkhar Swatabda, Associate professor, Department of Computer Science and Engineering, United International University. He has always supported us and helped us to learn a lot throughout our journey in this university.

Contents

List of Figures vi

List of Tables vii

1 Introduction 1

1.1 Motivation . . . 1

1.2 Objectives of the Thesis . . . 2

1.3 Organization of the Thesis . . . 3

2 Feature Selection Methods 4

2.1 Filter Method . . . 4

2.2 Wrapper Method . . . 5

2.3 Embedded Method . . . 6

2.4 Feature selection methods we have used in our research . . . 6

3 Proposed Method 8

3.1 ReliefF Attribute Evaluation . . . 8

3.2 Symmetrical Uncertainty Attribute Evaluation . . . 10

3.3 Our Hybrid Approach . . . 12

4 Experimental Analysis 15

4.1 Datasets . . . 15

4.2 Results . . . 16

5 Conclusions and Future Work 25

Bibliography 26

List of Figures

4.1 Classification accuracy of leukemia using feature selection methods . . . 22
4.2 Classification accuracy of ovarian tumor using feature selection methods 22
4.3 Classification accuracy of dexter using feature selection methods 23
4.4 Classification accuracy of mouse type using feature selection methods . 23

List of Tables

3.1 Ranking values of Urban attributes 10

4.1 Summary of the benchmark data sets 16
4.2 Classification accuracy of Breast Cancer dataset in different feature selection methods and algorithms 17
4.3 Classification accuracy of Phishing data set in different feature selection methods and algorithms 18
4.4 Classification accuracy of Spam base dataset in different feature selection methods and algorithms 18
4.5 Classification accuracy of Urban dataset in different feature selection methods and algorithms 19
4.6 Classification accuracy of Leukemia dataset in different feature selection methods and algorithms 20
4.7 Classification accuracy of Ovarian Tumor dataset in different feature selection methods and algorithms 20
4.8 Classification accuracy of Dexter dataset in different feature selection methods and algorithms 21
4.9 Classification accuracy of Mouse Type dataset in different feature selection methods and algorithms 21

List of Algorithms

1 Proposed method to use ReliefF for feature selection 9
2 Proposed method to use Symmetrical Uncertainty for feature selection . 11
3 Proposed Hybrid approach for feature selection 13

Chapter 1

Introduction

Data Mining has become a phenomenon in recent years. It is a procedure of determining unknown patterns and hidden relations in large data sets. But before we can apply data mining algorithms, we need to preprocess the data set. And feature selection in an essential step of preprocessing. Feature selection is a technique where we choose the most relevant features to enhance the predictive performance of a model. It includes increasing the accuracy, reducing the computational cost, saving the storage and lessening over-fitting. The best application of this technique is on data set that is high-dimensional. There are three types of techniques in feature selection: Filter, Wrapper and Embedded method. using these techniques we proposed three approaches to identify the right threshold to work with ReliefF, Symmetrical Uncertainty. An our hybrid approach is a combination of Symmetrical Uncertainty and Wrapper C4.5.

1.1 Motivation

Dimensionality indicates to the number of attributes or features in a data set. High Dimensional refers to the numbers of dimensions that remains exceedingly high. This makes the calculations extremely difficult. With high dimensional data, the number of attributes can exceed the number of observations. If we consider microarrays, which can contain thousands of samples in which each sample can contain tens of millions of genes. And an observation can have millions of features. For example a human can have millions of possible gene combinations. Other areas where high-dimensional datasets are seen include finance, high resolution imaging, and website analysis . There are some challenges in mining high dimensional data sets. The complexity of many existing data mining algorithms is exponential with respect to the number of dimensions [1]. With

increasing dimensions, these algorithms soon become computationally intractable. So they cannot be applied in many real applications. Secondly, the challenge for modeling in high dimensional data sets is to avoid over-fitting the training data. It is expected that such a model, in addition to performing well on the training set, would also perform equally well on an independent testing set [2]. Often the small number of samples in high dimensional data sets cause the classication model to over fit to the training data, for having poor generalization ability for the model. The more common approaches to addressing these challenges of high dimensional data sets are reducing the dimension of the data set or to apply methods that are independent of data dimensionality. That's why we worked on some methods to find the best feature selection techniques to reduce the dimension of high dimensional data sets.

1.2 Objectives of the Thesis

For mining high dimensional data sets we need some reliable and enhanced methods for feature selection. Wrapper method usually gives us the better result but it is computationally expensive as it selects and tests all possible subsets of features and identify the best subset for the classifier. On the other hand filter method is really fast as it works by ranking features based on some criteria but most of the time the performance is not better than wrapper. So we have proposed two approaches to apply ReliefF and Symmetrical Uncertainty to attain better performance. Finally we have proposed a hybrid approach to use the best of both filter and wrapper method. In case of our proposed approach for ReliefF and Symmetrical Uncertainty we will eliminate features and choose the cut-off point by following some specific steps. And in our hybrid approach we reduced the dimension of the high-dimensional data sets by using filter method and then applied wrapper in very less computational cost to score a better performance. In our experiments we have worked with four classifiers on eight data sets. Four of those data sets have high dimension. The classifiers are - Bagging, Boosting, Random Forest and Decision tree. We have used ReliefF Attribute Evaluation and Symmetrical Uncertainty Attribute Evaluation as Filter methods and Wrapper Subset Evaluation as wrapper method to compare the performances of our proposed approaches and other feature selection techniques.

1.3 Organization of the Thesis

The thesis is organized as follows:

Chapter 2 provides details about feature selection methods.

Chapter 3 presents the proposed method.

Chapter 4 discusses the results and experimental analysis.

Chapter 5 presents the conclusions and discusses the future works.

Chapter 2

Feature Selection Methods

There are three types of techniques in feature selection: Filter, Wrapper and Embedded method. Filter evaluates each feature independently from the classifier, ranks the features and takes the superior ones [3]. This method is classifier independent as it relies on the common aspects of the features. Wrappers takes a subset of the features, evaluates the classifier's performance on this subset, and then another subset is evaluated on the classifier [4]. The subset acquiring maximum performance is selected. The result is always classifier dependent. Embedded techniques on the other hand perform feature selection during learning process like artificial neural networks do [4]. Considering these techniques we can divide feature selection into feature ranking and subset selection. Ranking organizes features according to their predictive scores. And subset selection is a set of features that score a good accuracy altogether.

2.1 Filter Method

Filter method measures the relevance of a feature with its target class and ranks according to it. It assigns a feature selection value for each feature. This method is not biased as it is independent on the learning algorithm. And it lies on the general characteristics of the training dataset and the features are evaluated based on some criteria. One of the advantages of filter is that it is really fast so it is a great choice for high dimensional data sets. And a common disadvantage of this method is that it ignores the interaction with the classifier (the search in the feature subset space is separated from the search in the hypothesis space), and that most proposed techniques are uni-variate [5]. So it ignore any kind of dependencies between features. Filter is further classified into feature ranking and subset selection. And the result usually depend on the search

strategy. There are four ways of feature subset selection. They are - 1) Forward selection, 2) Backward elimination, 3) Bidirectional selection, and 4) Heuristic feature subset selection. Forward selection typically starts with an empty feature set and then considers adding one or more features to the set. Backward elimination typically starts with the whole feature set and considers removing one or more features from the set. Bidirectional search starts from both sides - from an empty set and from the whole set, simultaneously considering larger and smaller feature subsets. Heuristic selection generates a starting subset based on a heuristic (e.g. a genetic algorithm), and then explores it further [6]. We have used two types of filter methods in our experiments.

2.2 Wrapper Method

Wrapper method relies on greedy search algorithms. They find the right set of features for a machine learning algorithm by evaluating all possible combinations of features that performs the best estimate of accuracy for that specific algorithm. There are few downsides to this approach. First of all it requires to test all possible combinations of the features. Which is computationally very expensive, for very large data sets. Secondly the feature sets may not be optimal for every machine learning algorithm. Wrapper methods can be classified into three categories, 1) Step forward, 2) Step Backwards and 3) Exhaustive feature selection. In Step forward feature selection the performance of the classifier is evaluated for each feature. Then the feature that performs the best is selected out of all the features. Then the first feature is tried in combination with all the other features. The combination of two features that gives the best algorithm performance is selected. The process continues until the specified number of features are selected. We need to convert categorical feature values into numeric feature values to apply this method. In Step Backwards Feature Selection first of all one feature of the feature set is removed in round robin approach and classifier's performance is evaluated. And the feature set that gives the best performance is retained. Similarly after evaluating the performance another feature is removed using the same round robin approach. And the performance is evaluated for the features except the two features that were removed. This process continues until the specified number of the features remain in the feature set. In exhaustive feature selection, the performance of a machine learning algorithm is evaluated against all possible combinations of the features in the data set. The feature subset that shows best performance is selected. Since it tries all

the combination of features and selects the best, the exhaustive search algorithm is the greediest algorithm of all the wrapper methods.

2.3 Embedded Method

Embedded methods have combined qualities of filter and wrapper methods and have implemented by algorithms that have their own feature selection methods. Feature selection algorithm is integrated as part of the learning algorithm in embedded techniques. The most typical embedded technique is decision tree algorithm. A feature in each recursive step of the tree growth process is selected. Then the sample set is divided into smaller subsets. The importance of the features are measured by the number of child nodes in a subset which are in the same class. Decision tree generation process is same to the process of feature selection. Common embedded methods include various types of decision tree algorithms: CART, C4.5, random forest [7]. Regularization methods are the most common type of embedded feature selection methods. Regularization methods are also called penalization methods that brings in additional constraints into the optimization of a predictive algorithm by which the model becomes biased toward lower complexity. Examples of regularization algorithms are the LASSO, Elastic Net and Ridge Regression [8].

2.4 Feature selection methods we have used in our research

ReleifF Attribute Evaluation It is a statistical feature selection technique of Filter method. It assigns a relevance weight to each feature base on the relevancy of that feature and the target. Given training data S, sample size m, and a threshold of relevancy τ , Relief detects those features which are statistically relevant to the target concept and τ encodes a relevance threshold between 0 to 1 [9]. It measures the distance between instances using Manhattan norm and The basic idea of ReliefF algorithm is to draw instances at random, compute their nearest neighbors, and adjust a feature weighing vector to give more weight to features that discriminate the instance from neighbors of different classes [10]. It rewards the predictor if it assign same value to the neighbor of the same class but impose penalty if it does the opposite.

Symmetrical Uncertainty Attribute Evaluation The basic working mechanism of Symmetrical Uncertainty is correlation. Symmetric uncertainty is used to measure the degree of association between discrete features using information theory. It has derived from entropy [11]. We don't need to normalize the values because it can do it by itself unlike Information Gain. It compensates for information gain's bias towards features with more values and normalizes its values to the range [0, 1] with value 1 indicating that knowledge of the value of either one completely predicts the value of the other and value 0 indicating that X and Y are independent [12]. As this method is entropy based, it can be used on nominal features. But we can also use it on continuous features to measure correlation if we discretize them properly. This method is perfect for high dimensional data set to remove irrelevant and redundant features.

Wrapper Subset Evaluation using C4.5 The Decision Tree Induction algorithm uses the C4.5 algorithm to form a tree by splitting each variable and measuring the information gain provided [13]. It works by recursive partitioning and it stops when all instances are same label value. The wrapper—which is tuned to the specific interaction between an induction algorithm and its training data has been shown to give good results, but in practise may be too slow to be of practical use on large real-world domains containing many features [14]. The induction algorithm search the best subset repeatedly. And using C4.5 as that algorithm gives us some benefits. C4.5 is capable of handling missing values using probabilistic approach and it is computationally fast. So when we apply Wrapper C4.5 then the subsets are evaluated and compared constantly in a loop using information gain of the features and search for the right one.

Chapter 3

Proposed Method

In an individual ranking technique, we need to set a threshold to eliminate less relevant features. But that is not an easy estimation. So we have proposed three different approaches to create a better solution.

3.1 ReliefF Attribute Evaluation

ReliefF is a filter algorithm which ranks feature by assigning a weight according to the relevance of that feature and the target.
Suppose N is the number of attributes and if $N <= 200$ then the cut-off point will be (2/3) of the first ranked attributes. Then we will check if there are any attributes that was ranked as zero weighted and we'll eliminate them too. After that we will start eliminating features in a backward manner using the ranking weight. We will eliminate attributes in a range like in Table 3.1. In this figure we eliminate attributes from 0.13 up to 0.17 rank and test it with classifier C4.5 with 10 fold cross validation in training data. Then we eliminate from 0.18 up to 0.2. After each elimination we will decrease our elimination range and after a specific threshold we will eliminate one by one. We will stop if we see any deterioration in the accuracy. And restore the last set of eliminated features.

And If $N > 200$ then we will only consider the first 50 ranked attributes as the cut-off point. Then we will check if there are any attributes that was ranked as zero and we'll eliminate them too. After that we will eliminate one by one in a backward manner. In our experiments we use this technique with first 100 attributes as cut-off point and then the first 50. And we find the first 50 attribute cut-off point is the better option. But you can also use first 100 as cut-off point depending on your work. We

will stop if we see any deterioration in the accuracy. And restore the last eliminated feature.

The whole procedure has been showed in Algorithm 1.

Algorithm 1: Proposed method to use ReliefF for feature selection

Input: Training data D(F;C) = D(F_1; F_2; : : : ; F_N;C)

Output: Optimal set of selected features S

1: Initialize:
   S ←empty set, X ←initial set of all F features, $Th_1 \leftarrow 0$, $Th_2 \leftarrow 0$, N ←number of features, E ←empty set, $A \leftarrow 0$, $A_n \leftarrow 0$
2: Apply C4.5 on X of D and set A ←resultant accuracy;
3: **for** $i \leftarrow 1$ to N **do**
4: compute weight w_i using ReliefF for all features $X_i \in X$;
5: Organize all $X_i \in X$ using highest w_i in descending order;
6: **end for**
7: **if** $N \leq 200$ **then**
8: **for** $i \leftarrow 1$ to N **do**
9: Set $S \leftarrow X_i$ and stop if $i \leftarrow (2/3)N$;
10: Find feature $S_i \in S$ with $w_i \leftarrow 0$ then set $S \leftarrow S \setminus S_i$ and $N \leftarrow$ number of feature in S;
11: **end for**
11: Apply C4.5 on S and set $A_n \leftarrow$ resultant accuracy
12: **while** $A_n \geq A$ **do**
13: Set the range $Th_1 \leftarrow$ lowest w_i and $Th_2 \leftarrow$ up to w_i to eliminate;
14: **for** $j \leftarrow Th_1$ to Th_2 **do**
15: Set $S = S \setminus S_j$ and $E = S_j$;
16: **end for**
17: Apply C4.5 on S and set $A_n \leftarrow$ resultant accuracy;
18: Decrease difference between Th_1 and Th_2;
19: **end while**
20: Restore the last set of eliminated features and set them in S
21: **else**
22: **for** $i \leftarrow 1$ to N **do**
23: Set $S \leftarrow X_i$ and stop if $i \leftarrow 50$;

24: Find feature $S_i \in S$ with $w_i \leftarrow 0$ then set $S \leftarrow S \setminus S_i$ and $N \leftarrow$ number of feature in S;
25: **end for**
25: Apply C4.5 on S and set $A_n \leftarrow$ resultant accuracy
26: **while** $A_n \geq A$ **do**
27: Eliminate one by one and set $S = S \setminus S_i$ and $E = S_i$;
27: Apply C4.5 on S and set $A_n \leftarrow$ textresultant accuracy;
28: **end while**
29: Restore the last eliminated feature and set it in S;
30: **end if**
31: **return** =0

Table 3.1: Ranking values of Urban attributes

Ranking value	Attribute
0.179	28 ShpIndx_40
0.177	91 ShpIndx_100
0.173	129 Area_140
0.173	108 Area_120
0.171	27 Compact_40
0.169	43 GLCM3_40
0.166	105 BordLngth_100
0.166	112 ShpIndx_120
0.166	133 ShpIndx_140
0.166	87 Area_100
0.164	128 BrdIndx_140
0.164	107 BrdIndx_120
0.161	80 GLCM2_80
0.154	46 Round_60
0.146	67 Round_80
0.144	55 SD_NIR_60
0.143	86 BrdIndx_100
0.142	101 GLCM2_100
0.139	64 GLCM3_60
0.138	57 GLCM1_60
0.137	16 Rect

3.2 Symmetrical Uncertainty Attribute Evaluation

Symmetric Uncertainty can be used to calculate the fitness of features for feature selection by calculating between feature and the target class [15]. Suppose N is the number of attributes and if $N <= 200$ then we will use first (2/3) of first ranked attributes as

the cut-off point. After that we will check if there are any attributes that was ranked as zero and we'll eliminate them too. After that we will start eliminate them in a backward manner using the ranking value. We will eliminate attributes in a range like we did in our ReliefF approach. After each elimination we will decrease our elimination range and compare its performance using classifier C4.5. After a specific threshold we will eliminate one by one. We will keep going until we find the right amount of attributes that will give us a good result.We will stop if we see any deterioration in the accuracy. And restore the last set of eliminated features.

And if $N > 200$ then we will only consider the first 150 ranked attributes. Then we will eliminate those attribute that has ranked 0 as feature selection value. After that we will start eliminating attributes in a ranking range. We will gradually decrease the eliminating range and compare the performances using C4.5. After continuing for a while we will start eliminating one by one. We will stop if we see any deterioration in the accuracy. And restore the last set of eliminated features.

The whole procedure has been showed in Algorithm 2.

Algorithm 2: Proposed method to use Symmetrical Uncertainty for feature selection

Input: Training data D(F;C) = D(F_1; F_2; : : : ; F_N;C)

Output: Optimal set of features S

1: Initialize;
   S ←empty set, X ←initial set of all F features, $Th_1 \leftarrow 0$, $Th_2 \leftarrow 0$, N ←number of features, E ←empty set, $A \leftarrow 0$, $A_n \leftarrow 0$
2: Apply C4.5 on D and set A ←resultant accuracy;
3: **for** $i \leftarrow 1$ to N **do**
4: compute relevancy value su_i using Symmetrical Uncertainty for all features $X_i \in X$;
5: Organize all $X_i \in X$ using highest su_i in descending order;
6: **end for**
7: **if** $N \leq 200$ **then**
8: **for** $i \leftarrow 1$ to N **do**
9: Set $S \leftarrow X_i$ and stop if $i \leftarrow (2/3)N$;
10: Find feature $S_i \in S$ with $su_i \leftarrow 0$
then set $S \leftarrow S \setminus S_i$ and $N \leftarrow$ number of feature in S;
11: **end for**

Apply C4.5 on S and set $A_n \leftarrow$ resultant accuracy;
while $A_n \geq A$ **do**
Set the range $Th_1 \leftarrow$ lowest su_i and $Th_2 \leftarrow$ up to the su_i to eliminate;
for $j - Th_1$ to Th_2 **do**
Set $S - S \setminus S_j$ and $E - S_j$;
end for
Apply C4.5 on S and set $A_n \leftarrow$ resultant accuracy;
Decrease difference between Th_1 and Th_2;
end while
Restore the last set of eliminated features and set them in S;
else
for $i \leftarrow 1$ to N **do**
Set $S \leftarrow X_i$ and stop if $i \leq 150$;
Find feature $S_i \in S$ with $su_i \leftarrow 0$ then set $S \leftarrow S \setminus S_i$ and $N \leftarrow$ number of feature in S;
end for
Apply C4.5 on S and set $A_n \leftarrow$resultant accuracy;
while $A_n \geq$A **do**
Set the range $Th_1 \leftarrow$lowest su_i and $Th_2 \leftarrow$ up to the su_i to eliminate
for $j = Th_1$ to Th_2 **do**
Set $S = S \setminus S_j$ and $E = S_j$;
end for
Apply C4.5 on S and set $A_n \leftarrow$ resultant accuracy;
Decrease difference between Th_1 and Th_2;
end while
Restore the last set of eliminated features and set them in S;
end if
return =0

3.3 Our Hybrid Approach

Hybrid approach is one of the best methods of feature selection as it follows filter method to select a small number of features and then we can apply wrapper to find

the best performed subset. So the hybrid approach is less expensive than wrapper but more efficient than filter.

In our experiments, we have used Symmetrical Uncertainty as our filter method. Suppose N is the number of attributes and if $N <= 200$, we will apply Symmetrical Uncertainty and use first (3/4) of the attributes as cut-off points. After that we will check if there any attributes left which have ranked zero and we will eliminate them too. Now on the rest of the data set we will apply Wrapper C4.5 in which for the learning scheme we used 10 fold cross validation with greedy step wise search strategy to produce a better result.

If $N > 200$, then we will keep the first 700 ranked attributes at most. Then we will eliminate those that ranked 0. After that we will eliminate attribute by a heuristic range like 50 or 100 in a backward manner and use classifier C4.5 to compare the predictive performance. We will stop if we see any deterioration in the accuracy. And restore the last set of eliminated features. Because we need to have the right amount of attributes to apply wrapper. You can keep more features if you see the amount of instances is less. After that we will apply Wrapper C4.5 using 10 fold cross validation as a learning scheme with greedy step wise search strategy.

The whole procedure has been showed in Algorithm 3.

Algorithm 3: Proposed Hybrid approach for feature selection

Input: Training data D(F;C) = D(F_1; F_2; : : : ; F_N;C)

Output: Optimal set of features S

1: Initialize:
   $S \leftarrow$empty set, $X \leftarrow$initial set of all F features, $Th \leftarrow 0$, $N \leftarrow$number of features, $E \leftarrow$empty set, $A \leftarrow 0$, $A_n \leftarrow 0$
2: Apply C4.5 on D and set $A \leftarrow$resultant accuracy;
3: **for** $i \leftarrow 1$ to N **do**
4: compute value su_i using Symmetrical Uncertainty for all features $X_i \in X$;
5: Organize all $X_i \in X$ using highest su_i in descending order;
6: **end for**
7: **if** $N \leq 200$ **then**
8: **for** $i \leftarrow 1$ to N **do**
9: Set $S \leftarrow X_i$ and stop if $i \leftarrow (3/4)N$;

10: Find feature $S_i \in S$ with $su_i \leftarrow 0$ then set $S \leftarrow S \setminus S_i$ and $N \leftarrow$ number of feature in S;
11: **end for**
11: Apply Wrapper C4.5 on S and set $S \leftarrow$ resultant subset;
12: **else**
13: **for** $i \leftarrow 1$ to N **do**
14: set $S \leftarrow X_i$ and stop if $i \leq 700$;
15: Find feature $S_i \in S$ with $su_i \leftarrow 0$ then set $S \leftarrow S \setminus S_i$ and $N \leftarrow$ number of features in S;
16: **end for**
16: Apply C4.5 on S and set $A_n \leftarrow$ resultant accuracy;
17: **while** $A_n \geq A$ **do**
18: Set the range $Th \leftarrow$ number of attribute we want to eliminate ;
19: **for** $i = 1$ to Th **do**
20: Set $S - S \setminus S_j$ and $E - S_j$;
21: **end for**
21: Apply C4.5 on S and set $A_n \leftarrow$ resultant accuracy;
22: **end while**
23: Restore the last set of eliminated features and set them in S;
24: **end if**
24: Apply Wrapper C4.5 on S and set $S \leftarrow$ resultant subset;
25: **return** =0

Chapter 4

Experimental Analysis

In our experiments, we have worked with four classification algorithms to compare the predictive performance of the feature selection methods. We have used WEKA data mining tool in its default parameters for all the experiments. The classification algorithms were - Bagging, Boosting, Random Forest and C4.5. Every performance was evaluated using 10 fold cross validation on the training dataset. The feature selection methods we have used on the data sets which have attributes less than 200 were - Wrapper Subset Evaluation, ReliefF Attribute Evaluation, Symmetrical Uncertainty Attribute Evaluation and Our Hybrid Approach. ReliefF and Symmetrical Uncertainty are Filter methods. And Wrapper Subset Evaluation is a wrapper method. For the all subset evaluation methods we use greedy step wise search strategy. And in the Wrapper Subset Evaluation we use C4.5 algorithm with 10 fold cross validation. Also on the high dimensional data set we have applied ReliefF and Symmetrical Uncertainty Attribute Evaluation. We didn't use Wrapper Subset Evaluation on high dimensional data sets as it were taking days to produce results. The Hybrid method is a combination of Filter and Wrapper method. In most of the data sets we have applied supervised discretization to preprocess it. And for measuring the predictive performance of the feature selection methods we have used 10 fold cross validation on the training data set.

4.1 Datasets

The data sets were collected from two websites. The data sets with attributes less than 200 were collected from UCI Machine Learning Repository and the data sets which have attributes greater than 7000 were collected from OpenML. All of the data sets

Table 4.1: Summary of the benchmark data sets

Name	Description	Missing Values	Att #	Ins #	Feature Types
Breast Cancer	Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.	No	32	569	Multivariate
Phishing Website	This dataset are included various important web features to detect phishing	No	31	11055	Nominal
Spam Base	The spam e-mails came from our postmaster and individuals who had filed spam. The non-spam e-mails came from filed work and personal e-mails	No	58	4601	Multivariate
Urban	Contains data for classifying a high resolution aerial image into 9 types of urban land cover. Multi-scale spectral, size, shape, and texture information are used for classification.	No	148	168	Multivariate
leukemia	Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.	No	7130	72	Multivariate
Ovarian tumor	All attributes are numeric and represent measurements from DNA microarrays.	No	54682	283	Multivariate
Dexter	DEXTER is a text classification problem in a bag-of-word representation.	No	20001	600	Multivariate
Mouse Type	gene transcription data from DNA microarrays	No	45102	214	Multivariate

are multivariate except one that is nominal. There are no missing attributes. The high dimensional data sets have instances less than 700. All of the information about the datasets were presented in Table II

4.2 Results

In our work we have applied four classifiers to compare the performance of Without FS as Without Feature Selection, Wrapper C4.5, ReliefF, Symmetrical Uncertainty and Our Hybrid Approach. We have considered eight data sets as benchmark. Classification accuracy is given under each algorithm's column for each feature selection method. And also for each data set the selected attribute number that have been used in each

feature selection method have been showed as Selected Att. The total time depicts the time taken for scoring the highest accuracy between four algorithms. That means the total time includes the time feature selection methods took and the time taken by the classifiers.

Methods	Bagging	Boosting	Random forest	Decision Tree	Att. Select	Total Time of the best accuracy (min)
Without FS	95.7821 %	95.2548 %	96.8366 %	95.9578 %	31	less than 0.01
Wrapper	95.6063 %	94.9033 %	95.7821 %	97.188 %	5	0.02 & 0.01
Relief	95.6063 %	96.3093 %	97.0123 %	96.3093 %	12	less than 0.01 & less than 0.01
Symmetrical Uncert	95.7821%	95.7821%	96.8366%	96.3093%%	20	less than 0.01 & less than 0.01
Hybrid Approach	95.7821%	94.9033 %	95.7821 %	97.188 %	24 –> 5	0.01 & 0.02 & 0.01

Table 4.2: Classification accuracy of Breast Cancer dataset in different feature selection methods and algorithms

In TABLE 4.1, we have worked with breast cancer data set. In the Without FS method, the performance of the classifiers were measured without removing any attribute. Using that method, Random Forest took less than 0.01 min. The Wrapper method took 0.02 min and C4.5 scored the best with using only 5 attributes taking less than 0.01 min. ReliefF and Symmetrical Uncertainty both took less than 0.01 min to reduce attributes number down to 12 and 20 accordingly with a classification time less than 0.01 min. And our hybrid method score the same as Wrapper using C4.5 classifier. It took less than 0.01 min for Symmetrical Uncertainty to reduce the attributes down to 24 then the Wrapper took 0.02 min and reduce the attribute number down to 5.

In TABLE 4.2, using Without FS method Random Forest took less than 0.14 min. The Wrapper method took 4.41 min to reduce attributes down to 19 in which Random Forest scored the best taking 0.11 min. ReliefF took 0.52 min to reduce attribute down to 20 with a classification time of 0.11 min. Symmetrical Uncertainty took less than

Methods	Bagging	Boosting	Random forest	Decision Tree	Att. Select	Total Time of the best accuracy (min)
Without FS	96.2008%	92.5825%	97.2592%	95.8752%	30	0.14
Wrapper	96.0199%	92.5825%	96.9516%	96.2189%	19	4.41 & 0.11
Relief	95.7214 %	92.5825 %	96.9607 %	95.7123 %	20	0.52 & 0.11
Symmetrical Uncert	95.9114 %	92.5825 %	96.9245 %	95.8933 %	18	less than 0.01 & 0.09
Hybrid Approach	95.8933 %	92.5825%	96.7888 %	96.1737 %	20 –> 18	0.01 & 2.08 & 0.11

Table 4.3: Classification accuracy of Phishing data set in different feature selection methods and algorithms

0.01 to reduce attribute down to 18 with a classification time of 0.09 min. And our hybrid method took less than .01 min for Uncertainty and reduce the attribute down to 20 then the Wrapper took 2.08 min to reduce the attribute number down to 18 and done quite well.

Methods	Bagging	Boosting	Random forest	Decision Tree	Att. Select	Total Time of the best accuracy (min)
Without FS	93.9796%	90.0674%	95.4792 %	92.9798 %	57	0.18
Wrapper	93.2623%	89.5023%	93.8926%	93.1102%	16	20.6 & 0.14
Relief	93.4145 %	88.7416%	95.0228 %	92.1321 %	38	0.16& 0.15
Symmetrical Uncert	94.2621 %	90.0456 %	95.1098 %	92.6755 %	38	0.01 & 0.15
Hybrid Approach	93.9796 %	89.85%	94.8924 %	93.7405 %	38 –> 19	0.01 & 6.39 & 14.15

Table 4.4: Classification accuracy of Spam base dataset in different feature selection methods and algorithms

In TABLE 4.3 using Without FS method Random Forest took less than 0.18 min. The Wrapper method took 20.6 min to reduce attributes down to 16 in which Random Forest scored the best taking 0.14 min. ReliefF took 0.16 min to reduce attribute down to 38 with a classification time of 0.15 min. Symmetrical Uncertainty took less than

0.01 to reduce attribute down to 38 with a classification time of 0.15 min. And our hybrid method took less than .01 min for Uncertainty and reduce the attribute down to 38 then the Wrapper took 6.39 min and reduce the attribute number down to 19.

Methods	Bagging	Boosting	Random forest	Decision Tree	Att. Select	Total Time of the best accuracy (min)
Without FS	85.119 %	26.7857 %	89.881 %	85.7143 %	147	0.01
Wrapper	89.2857 %	30.9524 %	88.0952 %	89.881 %	7	0.09 & 0.01
Relief	87.5 %	26.7857 %	88.6905 %	82.1429 %	67	less than 0.01 & 0.01
Symmetrical Uncert	84.5238 %	26.7857 %	89.2857 %	85.119 %	70	less than 0.01 0.01 & 0.01
Hybrid Approach	89.881 %	26.7857 %	91.6667 %	91.6667 %	110 -> 7	less than 0.01 & 0.08 & 0.01

Table 4.5: Classification accuracy of Urban dataset in different feature selection methods and algorithms

In TABLE 4.4, using Without FS method Random Forest took less than 0.1 min. The Wrapper method took 0.09 min to reduce attributes down to 7 in which C4.5 scored the best taking less than 0.01 min. ReliefF and Symmetrical Uncertainty both took less than .01 min to reduce attributes number down to 67 and 70 accordingly with a classification time less than 0.01 min. And our hybrid method took less than .01 min for Uncertainty and reduce the attribute down to 110 then the Wrapper took 0.08 min and reduce the attribute number down to 7.

In TABLE 4.5, using Without FS method Random Forest took 0.06 min with 7129 attributes. In ReliefF and Symmetrical Uncertainty both took .01 min to reduce attributes number down to 16 and 20 accordingly with a classification time less than 0.01 min. And our hybrid method took .01 min for Uncertainty and reduce the attribute down to 50 then the Wrapper took 0.01 min and reduce the attribute number down to 1.

In TABLE 4.6, using Without FS method C4.5 took 1.7 min with 54681 attributes. ReliefF took 2.38 min to reduce attributes number down to 50 with a classification time of 0.01 min. Symmetrical Uncertainty took 0.01 min to reduce attributes number

Methods	Bagging	Boosting	Random forest	Decision Tree	Att. Select	Total Time of the best accuracy (min)
Without FS	95.8333 %	97.2222 %	98.6111 %	90.2778 %	7129	0.06
Relief	91.6667 %	100%	100 %	90.2778 %	16	0.01 & less than .01
Symmetrical Uncert	93.0556 %	100%	100 %	90.2778 %	20	0.01 & less than .01
Hybrid Approach	95.8333 %	95.8333 %	95.8333 %	95.8333 %	50 –> 1	.01 & 0.01 & 0.01

Table 4.6: Classification accuracy of Leukemia dataset in different feature selection methods and algorithms

Methods	Bagging	Boosting	Random forest	Decision Tree	Att. Select	Total Time of the best accuracy (min)
Without FS	92.9329 %	89.7527 %	92.2261 %	93.2862 %	54681	1.7
Relief	93.2862 %	89.3993 %	95.053 %	92.2261 %	50	2.38 & 0.01
Symmetrical Uncert	93.9929 %	89.3993 %	93.6396 %	91.5194 %	44	0.01 & 0.01
Hybrid Approach	94.3463 %	89.7527 %	94.3463 %	95.053 %	500 –>4	0.01 & 0.17 & 0.01

Table 4.7: Classification accuracy of Ovarian Tumor dataset in different feature selection methods and algorithms

down to 44 with a classification time of 0.01 min. And our hybrid method took .01 min for Uncertainty and reduce the attribute down to 500 then the Wrapper took 0.17 min and reduce the attribute number down to 4.

Methods	Bagging	Boosting	Random forest	Decision Tree	Att. Select	Total Time of the best accuracy (min)
Without FS	90 %	86.5 %	95.1667 %	88.1667 %	20000	3.46
Relief	89.6667 %	87.1667 %	94.3333 %	91.5 %	43	0.01 & 0.01
Symmetrical Uncert	90.3333 %	86.5 %	94.3333 %	92.3333 %	45	0.01 & 0.01
Hybrid Approach	90%	81.8333 %	92.8333 %	92.6667 %	119 – >13	0.01 & 1.57 & 0.01

Table 4.8: Classification accuracy of Dexter dataset in different feature selection methods and algorithms

In TABLE 4.7, using Without FS method Random Forest took 3.46 min with 20001 attributes. In ReliefF and Symmetrical Uncertainty both took 0.01 min to reduce attributes number down to 43 and 45 accordingly with a classification time less than 0.01 min. And our hybrid method took 0.01 min for Uncertainty and reduce the attribute down to 119 then the Wrapper took 1.57 min and reduce the attribute number down to 13.

Methods	Bagging	Boosting	Random forest	Decision Tree	Att. Select	Total Time of the best accuracy (min)
Without FS	75.7009 %	42.5234 %	82.243 %	68.2243 %	45101	2.26
Relief	80.8411 %	43.4579 %	85.514 %	75.7009 %	47	1.36 & 0.01
Symmetrical Uncert	78.972 %	42.5234 %	85.514 %	74.2991 %	90	0.01 & 0.01
Hybrid Approach	77.1028 %	42.0561 %	87.3832 %	89.2523 %	600 – > 17	0.01 & 3.56 & 0.01

Table 4.9: Classification accuracy of Mouse Type dataset in different feature selection methods and algorithms

In TABLE 4.8, using Without FS method Random Forest took 2.26 min with 45101 attributes. The ReliefF took 1.36 min to reduce attributes number down to 47 with a

classification time of 0.01 min. The Symmetrical Uncertainty took 0.01 min to reduce attributes number down to 90 with a classification time of 0.01 min. And our hybrid method took 0.01 min for Uncertainty and reduce the attribute down to 600 then the Wrapper took 3.56 min and reduce the attribute number down to 17.
We have also shown line graphs of high dimensional data sets in Figure 1, 2, 3, 4 to compare our approaches.

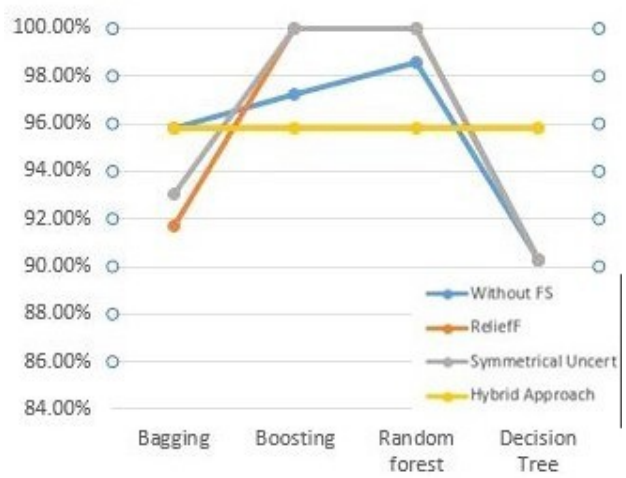


Figure 4.1: Classification accuracy of leukemia using feature selection methods

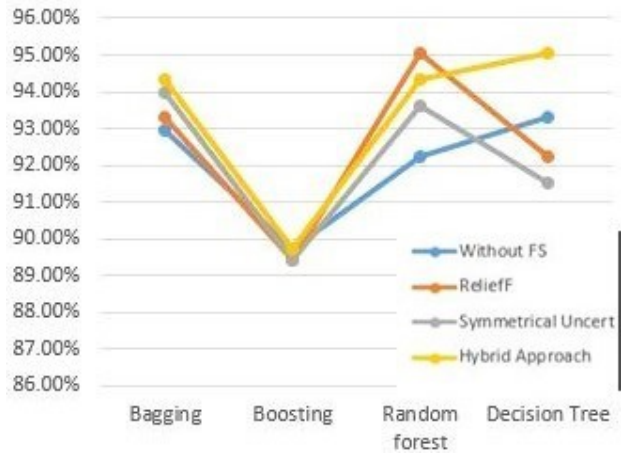


Figure 4.2: Classification accuracy of ovarian tumor using feature selection methods

After these experiments we have seen that using our hybrid approach the C4.5 classifier

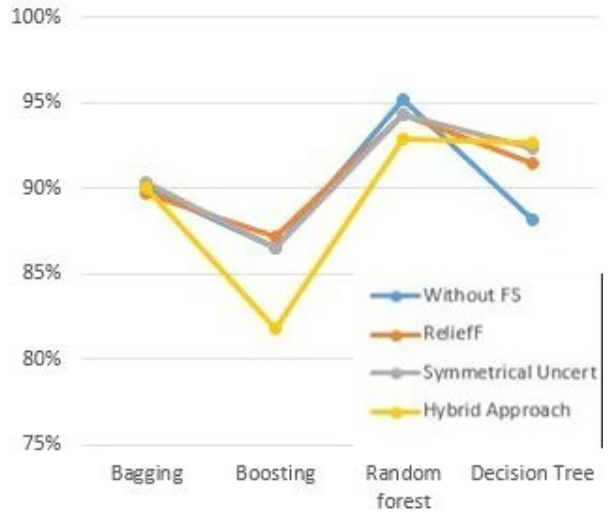


Figure 4.3: Classification accuracy of dexter using feature selection methods

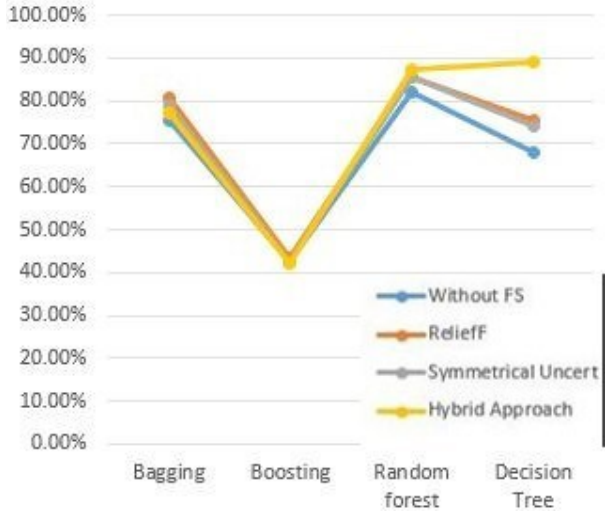


Figure 4.4: Classification accuracy of mouse type using feature selection methods

gives us the best result.

Chapter 5

Conclusions and Future Work

In this paper, we propose a hybrid approach for high dimensional data sets and a new manner to apply ReliefF and Symmetrical Uncertainty Attribute Evaluation. Our main aim was to reduce the computational cost and increase accuracy of the dataset using less attributes. This new hybrid approach was tested through intrinsic experiments and proved to be quite efficient and good in terms of feature selection as it scored great using a small amount of attributes. Wrapper is great for subset selection in most of the cases. But for high dimensional data sets the Wrapper method could take days if it works on thousands of features. But with our approach it took few minutes as most of the features that are not important were trimmed off. The hybrid approach was computation cost efficient and showed effectiveness in C4.5. And also our manner of using ReliefF and Symmetrical Uncertainty shows us quite good results. If we choose the right threshold and work according to proposed approaches then we will score great classification accuracy.

We will start working on more high dimensional data sets to make our approaches perfect and research on other feature selection methods to find in which manner they work the best. Also we will explore big data sets as they have more complexity. We will try to find new techniques to reduce to volume of big data and make it more easy to predict target.

Bibliography

[1] W. Wang and J. Yang, "Mining high-dimensional data," in *Data Mining and Knowledge Discovery Handbook*. Springer, 2009, pp. 803–808. 1

[2] V. Pappu and P. Pardalos, *High Dimensional Data Classification*, 12 2013, p. 34. 2

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003. 4

[4] E. M. Karabulut, S. A. Özel, and T. Ibrikci, "A comparative study on the effect of feature selection on classification accuracy," *Procedia Technology*, vol. 1, pp. 323–327, 2012. 4

[5] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btm344 4

[6] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on*. IEEE, 2015, pp. 1200–1205. 5

[7] M. Sandri and P. Zuccolotto, *Variable Selection Using Random Forests*, 01 2006, pp. 263–270. 6

[8] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics," *Briefings in Bioinformatics*, vol. 9, no. 5, pp. 392–403, 2008. [Online]. Available: http://dx.doi.org/10.1093/bib/bbn027 6

[9] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine Learning Proceedings 1992*. Elsevier, 1992, pp. 249–256. 6

[10] Y. Wang and F. Makedon, "Application of relief-f feature filtering algorithm to selecting informative genes for cancer classification using microarray data," in *Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE*. IEEE, 2004, pp. 497–498. 6

[11] Y. Chen, Y. Li, X.-Q. Cheng, and L. Guo, "Survey and taxonomy of feature selection algorithms in intrusion detection system," in *International Conference on Information Security and Cryptology*. Springer, 2006, pp. 153–167. 7

[12] S. Sasikala, S. A. alias Balamurugan, and S. Geetha, "Multi filtration feature selection (mffs) to improve discriminatory ability in clinical data set," *Applied Computing and Informatics*, vol. 12, no. 2, pp. 117–127, 2016. 7

[13] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986. 7

[14] M. A. Hall, "Correlation-based feature selection for machine learning," 1999. 7

[15] B. Singh, N. Kushwaha, and O. P. Vyas, "A feature subset selection technique for high dimensional data using symmetric uncertainty," 2014. 10



This document was created with the Win2PDF “print to PDF” printer available at
http://www.win2pdf.com

This version of Win2PDF 10 is for evaluation and non-commercial use only.

This page will not be added after purchasing Win2PDF.

http://www.win2pdf.com/purchase/



This document was created with the Win2PDF “print to PDF” printer available at
http://www.win2pdf.com

This version of Win2PDF 10 is for evaluation and non-commercial use only.

This page will not be added after purchasing Win2PDF.

http://www.win2pdf.com/purchase/