2021 年度博士後期課程(ソフトウェア情報学)論文

特徴選択アルゴリズムの安定性解析と堅牢な特徴選択法の提案に関する研究

Study on Stability Analysis of Feature Selection Algorithm and Proposal towards Stable Feature Selection Method

岩手県立大学大学院

ソフトウェア情報学研究科

学籍番号: 2362019001

氏名: Rikta Sen

主任研究指導教員: Basabi Chakraborty 教授
副研究指導教員: Akio Doi 教授
副研究指導教員: Prima O.D. Ardiansyah 准教授

Abstract

Feature selection is one of the essential preprocessing tasks in machine learning and pattern recognition problems for reducing the dimensionality of the data. It removes irrelevant and redundant features leading to simplified classification process and improved accuracy. Several feature selection algorithms have been proposed so far but for any particular problem, the quality of the selected feature subset varies from algorithm to algorithm. Usually, the quality of the feature selection algorithm is evaluated by reduction of cardinality of the selected feature subset, improvement of classification accuracy or the reduction of algorithm complexity (computational cost). But stability of feature selection algorithm is another important characteristic which needs to be considered for evaluation of any feature selection algorithm. Stability refers to the robustness of the selected feature subset to small changes in the training set or set of various parameters of the algorithm. A stable feature selection algorithm is supposed to select the same subset of features for a particular problem irrespective of any changes in the training set of samples or parameters of the algorithm. Selection of stable feature subset is especially required when the physical meaning of the features are important.

Various metrics have been developed so far for measuring the stability of a feature selection algorithm. In this work, an extensive analysis of stability of various types of feature selection algorithms (filter ranked based, filter subset based, and wrapper based algorithms) has been done with various stability measures. It has been found that filter rank based feature selection algorithms possess better stability than others, Jeffries-Matusita (JM) distance based feature selection being the best. JM distance is then verified as an efficient feature selection tool by using the simulation experiment for binary classification problems. A multiclass extension of JM distance has also been proposed as a feature selection algorithm which is found to perform better compared to the previous multiclass extensions of JM distance and other rank based filter approaches. Finally the critical analysis of different stability metrics has been done in which the desired properties of stability metrics are analyzed to determine which stability metrics follow which properties. The limitations of various similarity-based stability metrics are analyzed based on their desired properties. A correction, as well as a novel extension of similarity-based stability metric, Lustgarten measure, an extension of the most popular Kuncheva index, is proposed. The proposed new stability metric fulfills all the desired properties of stability metrics and removes the limitations of other metrics. The proposed stability metric has also been verified and found to be the best among the existing stability metrics by simulation experiments with different bench mark data sets.

Acknowledgements

First and foremost, I want to express my gratitude to my thesis adviser Professor Dr. Basabi Chakraborty, for her guidance, direction, patience, and continuous encouragement during my PhD study. I am indebted to her for her great patience and effort in polishing my English writings and correcting many mistakes in my research papers. I will be forever grateful to her for the support she did during my hard time. Without it, I could not have completed my PhD in due time.

I would also like to thank my committee members, Professor Dr. Akio Doi, Dr. Prima O.D. Ardiansyah, whose valuable comments and suggestions help me a lot to improve my presentation and writings. I am also highly thankful to Dr. Saptarshi Gowasswami for providing me with valuable advice on feature selection and the machine learning area.

In addition, I wish to extend my appreciation to all the members of PRML lab, Graduate School of Software and Information Science, Iwate Prefectural University, who helped me in many ways. I would also like to thank Iwate Prefectural University for wavering my tuition fee and granting scholarships during my PhD study.

Finally, I would want to convey my heartfelt gratitude to my family for always being there for me in both good and bad times. I am grateful for their love, support, and patience, which enabled me to complete this journey smoothly.

Rikta Sen

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Background

Feature selection is the process of selecting relevant features or reducing the number of attributes for model building, which plays a very important role in machine learning, data mining or data analysis. For the analysis of high dimensional datasets coming out from different fields, it is preferable to predict appropriate features with short run time and interpretability to disregard the irrelevant and redundant features. If the selected features are very relevant to predicting the target, then the predictive accuracy of the model will not decrease. Besides this, if redundant or irrelevant features are removed, then the predictive accuracy will also increase.

On the other hand, there are many feature selection algorithms for selecting reliable features, and the selected feature subset varies from algorithm to algorithm. Feature selection algorithms can be broadly classified into two different ways: rank-based approach and feature subset approach. The Rank-based approach ranks each feature using a measure and then selects the top k features from the original feature set. On the other hand, the feature subset aims to find the optimum feature subset using different search approaches. The feature selection approach is also grouped into three different approaches based on feature evaluation, namely filter, wrapper and embedded approach. A feature selection algorithm is often considered best

when its generated feature subset produces better classification accuracy. However, another important parameter for evaluating a feature selection algorithm is its stability in a different run. A feature selection algorithm with good stability means that the algorithm shows consistency in producing key features.

In some practical fields, features are analysed with expensive studies, where it is desirable to keep the number of features as small as possible. In this case, the main objective of feature selection is to select a small set of feature subsets, where all relevant features are present but irrelevant, or redundant features are absent. Another important issue of feature selection is the sensitivity of feature selection algorithms due to the small changes of training data . In general, if the selected feature subset changes radically due to the small change of training data, the feature selection process is unstable. On the other hand, if the selected feature subset is static though the training data changes, the process is stable. In different application areas such as microarray classification [1], molecular profiling [2], biomedical fields [3] and linguistics [4], stability analysis is very critical. In this case, the main objective of feature selection is to select a stable feature subset.

1.2 Problem Statement

The stability measurement of feature selection algorithms plays a critical role in data analysis as it ensures that the feature selection algorithm is trust-able or not. This is because if small changes in training data cause a large change in output of feature selection, then how can we trust that the selected features are appropriate. Therefore, selecting the stable features/genes for further reproducible research in biomedical applications is very important. In the work of Jurman et al., it is stated that a stable gene set is as crucial as to have predictive power [2]. Goh and Wong [5] recommend statistical feature selection with stability analysis as equally important as to improve the quality of the selected features. This is the key point and purpose of studying stability.

One crucial question is whether existing stability algorithms can sufficiently define the stability of feature selections. Another concern is which of the feature selection algorithms is the most stable. It is not straightforward because several issues have to be addressed explicitly to find the appropriate answer. For this reason, different feature selection algorithms are required to study on the basis of stability measurement, and it is necessary to identify a better stability index that can effectively define the stability of a feature selection algorithm.

1.3 Objectives

The aim of the thesis is to analyze the stability of different feature selection algorithms, the details are as follows:

- To explore the stability of different feature selection algorithms using different stability metrics
- To find out the shortcomings of the different stability metrics and improve them
- To develop stable feature selection algorithm

1.4 Contributions

The contributions drawn from the present research are summarized below:

- Assessment of the feature selection algorithms regarding their stability.
- Proposal of an extension of a stable binary class feature selection algorithm to multi-class problems.
- A critical analysis of different stability metrics according to their desirable properties to find out their limitations.

- Proposal of a novel stability measure based on a well-known stability metric to overcome the limitations of the existing measures.

1.5 Overview of the Thesis

The thesis organization is as follows. In Chapter 2, background and related work of feature selection and stability analysis are highlighted. Chapter 3 discusses the stability measure of feature selection algorithms with simulation experiments. At first we calculate the stability of filter ranked based feature selection algorithm. This work reveals that Jeffries-Matusita (JM) distance shows better stability than other measures. After that a comparative study on the stability of both filter based and wrapper based feature selection algorithms have been done. From this work, it is pointed out that filter ranked based feature selection algorithms give better stability than filter subset based and wrapper based approaches. Chapter 4 presents the critical analysis of different stability metrics used for stability measurement of the feature selection algorithms. In this chapter, the limitation of various stability metrics are discussed in detail with the desirable properties of stability measures, and a correction is proposed. While assessing stability measures, it is found that JM distance produces better stability than other filter ranked based stability measures. For this reason, JM distance has been examined as a feature selection tool for binary class problems, and it is discussed in Chapter 5. In this chapter, a new JM distance based feature selection algorithm for multiclass problems has been proposed and compared with previous JM distance based multiclass extensions and other filter based feature selection algorithms. Chapter 6 provides a conclusion and future directions for further research.

Chapter 2

Feature Selection and its Stability

2.1 Introduction

This chapter presents the field of research related to this thesis. First, the preliminaries of feature selection algorithms and the concept of stability are presented. An overview of the literature related to feature selection has been described in brief. This chapter also provides the work related to the stability of feature selection algorithms.

2.2 Feature Selection

With the rapid development of science and technology, the amount of generated data in every sphere of life has been increased tremendously, which causes the classification or mining of data increasingly difficult. These data are often high dimensional, making their analysis more complicated and computationally costly. The data need to be preprocessed to get rid of redundant and irrelevant information. Feature selection is the most important processing step prior to classification or clustering for any pattern recognition or data mining problem [6] [7]. This is a process of dimensionality reduction in which discriminatory and relevant information is retained by discarding redundant and irrelevant information leading to better performance of

the classification model in terms of classification accuracy as well as computational cost [8] [9]. Figure 2.1 shows feature selection process.

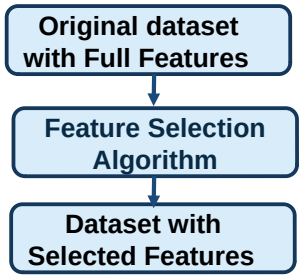
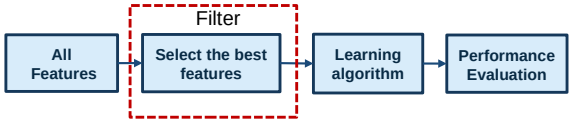
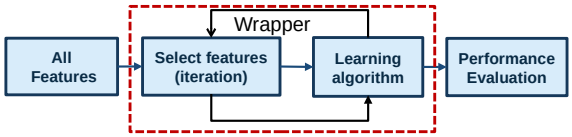


Figure 2.1: Feature Selection process

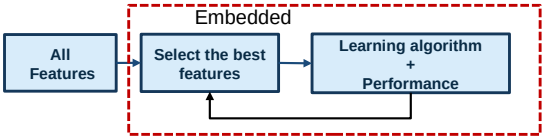
There are mainly three types of feature selection approaches. These are filter-based feature selection, wrapper-based feature selection, and embedded techniques [10]. The filter approach uses an independent evaluation measure for evaluating features subsets without involving the classifier. The wrapper approach uses the classifier accuracy as the evaluation function for selecting the feature subset. Although efficient, the wrapper is computationally expensive compared to the filter approach. Embedded approach, however, selects feature during the training of the classifier. Figure 2.2 shows different type of the feature selection algorithm.



(a) Filter Approach



(b) Wrapper Approach



(c) Embedded Approach

Figure 2.2: Types of Feature Selection approach

Based on the selection process of the optimal feature subset, two main approaches exist, rank based and search based. Rank based approaches evaluate each

feature independently, rank them according to their merit/goodness and then select an appropriate portion of the top ranked features to form the final feature subset. Though simple and computationally light, rank based approaches ignore the interaction between features and cannot guarantee the optimality of the selected subset. Moreover, some strategies need to be adopted to fix the optimum percentage of top ranking features to be selected. According to [11], the best two individual features do not produce the best feature subset of two features. An exhaustive evaluation of all possible feature subsets can only guarantee the optimality of the selected feature subset. But for high dimensional data, it leads to an explosion of computational time with increasing dimension of data. To solve this combinatorial optimization problem, a lot of search algorithms have been developed so far for the selection of optimum feature subset, which include mainly statistical or mathematical and soft computing based techniques.

The feature evaluation measures for filter approaches are generally classified into four categories such as distance based, dependency or relevance based, information theoretic and consistency measure [12]. Distance based evaluation measures include class separability measures such as divergence or Kullback-Liebler distance, Bhattacharyya distance, Jeffries-Matushita distance, Mahalanobis distance. Dependency based measures consider correlation or similarity of a feature to a class. Information theoretic measures determine the information gain of a feature by its inclusion. Consistency based measure penalizes inconsistent features where inconsistency is defined as two instances having the same feature values but different class labels. Some efficient popular filter evaluation measures used for feature ranking are Mutual Information (MI), Information Gain (IG), Gain Ratio (GR), Symmetrical uncertainty (SU), Chi-squared (CS), One-R, Relief, Jeffries-Matusita (JM) distance and Correlation [13].

2.2.1 Feature Selection Techniques

This sub section describes different types of feature selection techniques, many of which are employed in this thesis.

A. Filter based feature selection

- Mutual Information:

  Mutual Information is an information theoretic measure which expresses the dependency of one variable on another variable. If mutual information is used between a feature and the class, then this gives one basis to measure relevance of a feature. Mutual information can be calculated as follows:

$$MI(Class, A) = H(Class) + H(A) - H(Class, A) \tag{2.1}$$

  H indicates entropy, entropy of a random variable is calculated as:

$$H(A) = -\sum_{a} P_a(A) log p_a(A) \tag{2.2}$$

  Mutual Information is one of the most used measures in feature selection. Over the years, there have been several improvements over mutual information. Normalized mutual information maps the value of mutual information between 0 and 1. The work produced by Estévez et al. [14] is an important reference of feature selection using Mutual Information.

  There are other information theoretic measures like information gain and symmetrical uncertainty which are minor variation of the same concepts.

- Information Gain (IG): Information gain is a theoretical measure which shows how much information a feature provides us about the class. It is measured in terms of reduction of entropy achieved by learning a feature A. It can also be defined with mutual information. IG is symmetrical in nature. The information gain of feature A for the class labels Class is as follows [15]:

$$IG(A, Class) = H(A) - H(A|Class) \tag{2.3}$$

  where, H indicates the entropy.

- Gain Ratio (GR):

Gain Ratio (GR) is a non-symmetric measure, which defined as the ratio between the information gain and the entropy of A as [16]:

$$GR = \frac{IG}{H(A)} \tag{2.4}$$

- Symmetrical Uncertainty (SU):

  Symmetrical Uncertainty (SU) compensates for information gain bias toward attributes with more values and normalizes its value to the range [0, 1]. The equation is [17]:

$$SU = 2\frac{IG}{H(A) + H(Class)} \tag{2.5}$$

- One-R:

  One-R applies a simple measuring method from One-R classifier. It ranks attributes by error rate and it treats all numerically-valued attributes as continuous and uses a straightforward method to divide the range of values into several disjoint intervals [18].

- Chi-Squared (CS):

  In Chi-Squared analysis, it is assumed that there is an independency among the feature and the class. Chi-Squared is an analytical technique that compares values from the expected and actual outcomes. CS for two adjacent intervals is given by the following equation [13] :

$$CS = \sum_{i=1}^{2}\sum_{j=1}^{C}\frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{2.6}$$

  Where, C is the number of classes, A_{ij} is the number of instances of the j-th class in the i-th interval and E_{ij} is the expected frequency of A_{ij} given by the formula:

$$E_{ij} = R_i C_j / NT \tag{2.7}$$

Where R_i is the number of instances in the i-th interval and C_j and NT are the number of instances of the j-th class and total number of instances, respectively, in both intervals.

- Bhattacharyya Distance

  Bhattacharya Distance measures the similarity between two probability distributions. It is used to measure the separability or overlap of two classes for each feature in a binary classification problem and rank them to select the best performing features. If the two distributions for two classes c_i and c_j are considered to be Gaussian then Bhattacharya distance B_{ij} can be defined as [19]:

$$B_{ij} = \frac{1}{8}(\mu_i - \mu_j)^T(\frac{\Sigma_i + \Sigma_j}{2})^{-1}(\mu_i - \mu_j) + \frac{1}{2}\ln\frac{|\frac{\Sigma_i+\Sigma_j}{2}|}{\sqrt{|\Sigma_i||\Sigma_j|}} \tag{2.8}$$

  where μ_i, μ_j and Σ_i , Σ_j represent the mean vectors and the covariance matrices, respectively, for the classes c_i and c_j.

- Jeffries-Matushita (JM) Distance

  Jeffries-Matushita (JM) distance is a measure of statistical separability for two classes. This is a distance measure, which improves Bhattacharya Distance by scaling it between 0 and 2. For feature x, it is defined for two classes c_i and c_j as in [20]

$$JM_{ij} = \{\int_x [\sqrt{p(x/c_i)} - \sqrt{p(x/c_j)}]^2 dx\}^{1/2} \tag{2.9}$$

  JM distance is bounded by a range of values from 0 to 2. It is related to Bhattacharyya distance as

$$JM_{ij} = \sqrt{2(1 - e^{-B_{ij}})} \tag{2.10}$$

- Relief and Relief-F:

  The original Relief algorithm is formulated iteratively from an instance based learning approach that evaluates a feature by assigning a weight to the feature.

A weight corresponding to the feature is calculated based on $nearHit$ (closest instance from the same class) and $nearMiss$ (closest instance from a different class). w_i is initialized to 0 and then in each step, it is updated as the following [21]:

$$w_i = w_i - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2 \tag{2.11}$$

Finally, an average of w_i is taken over the iterations.

Relief-F is an extension of the original Relief algorithm that can be used for multiclass problems [22].

- **Fisher score:** The Fisher score algorithm is a feature ranking algorithm, in which features are selected individually in accordance with their scores. In Fisher score, the subset of features are identified in such a way that in the data space spanned by the selected features, the distances between data points in different classes are as large as possible, while the distances between data points in the same class are as small as possible. For a given data set $\{(x_i, y_i)\}_{i=1}^{n}$ where, $x_i \in \mathrm{R}_d$ and $y_i \in \{1,2,...\mathrm{c}\}$, $with\ \mathrm{X} = [\ \mathrm{x}_1, x_2, ...x_n] \in \mathrm{R}^{dn}$ to represent the data matrix, let the selected features be m. So, the input data matrix $X \in \mathrm{R}^{dn}$ reduces to $Z \in \mathrm{R}^{mn}$, where m represents the number of selected features, and n represents the number of samples. Now the Fisher score is represented as the following:

$$F(Z) = tr\{(S_b)(S_t + \gamma I)^{-1}\} \tag{2.12}$$

Where, $tr()$ represents the trace of a matrix, γ is a positive regularization parameter, S_b is called between-class scatter matrix, and S_t is called total scatter matrix, which are defined as:

$$S_b = \sum_{k=1}^{c} n_k(\mu_k - \mu)(\mu_k - \mu)^T \tag{2.13}$$

$$S_t = \sum_{i=1}^{n} (z_i - \mu)(z_i - \mu)^T \tag{2.14}$$

where μ_k and n_k are the mean vector and size of the k-th class respectively in the reduced data space, Z, $\mu = \sum_{k=1}^{c} n_k \mu_k$ is the overall mean vector of the reduced data. Since, the feature selection problem is a combinatorial optimization problem, so to reduce the difficulty the heuristic strategy is used to compute the rank of each feature. Let μ_k and σ_k be the mean and standard deviation of k-th class, corresponding to the j-th feature. Let μ_j and σ_j denote the mean and standard deviation of the whole data set corresponding to the j-th feature. Then the Fisher score of the j-th feature is computed below [23],

$$F(x^j) = \frac{\sum_{k=1}^{c} n_k(\mu_k^j - \mu^j)^2}{(\sigma^j)^2} \tag{2.15}$$

Where, $(\sigma^j)^2 = \sum_{k=1}^{c} n_k(\sigma_k^j)^2$. By using this equation, Fisher score for individual feature is calculated and among them, top ranked features are selected.

For a more broad based understanding of different types of measures like distance measures, information measures, dependency measures and consistency measures the work by Huan Liu [12] can be referenced.

B. Subset evaluation based Filter method

- **CFS**

  Correlation based Feature Selection (CFS) is generally a filter algorithm in which feature subsets are ranked with using a correlation based heuristic evaluation function. CFS was first developed by Hall in 1999 [17]. In general, feature subsets should be consisted of features which are strongly correlated with the target and uncorrelated with each other. As a result, it is very easy to find out irrelevant features by finding the features that have low correlation with the class. Redundant features can also be separated by identifying the features which have correlation with other remaining features. CFS works in two consecutive phases, at first calculating the feature-feature and feature-target correlations in matrix form. After that, a searching procedure is applied for calculating the features space and finally the optimal subset is obtained.

For searching the features space, different types of heuristic search strategies like best first search, forward selection, backward elimination, bi-directional search, and genetic search are appointed to search the features space. CFS uses the Pearson's correlation as an evaluation function of feature subsets as following [24]:

$$M_s = \frac{kr_{cf}}{\sqrt{k(k-1)r_{ff}}} \tag{2.16}$$

Where, M_s is the heuristic merit of a feature subset S containing k features, r_{cf} is the average linear correlation coefficient between feature and class, r_{ff} is the mean linear correlation coefficient between one features to another feature. The numerator of this equation can be thought of as providing an indication of how predictive of the class a set of features is; and the denominator of how much redundancy there is among the features.

- **FCBF**

  Fast Correlation Based Filter (FCBF) is a subset based multivariate feature selection method. FCBF uses symmetrical uncertainty (SU) to measure the relevance of features and uses the backward selection technique with sequential search strategy to find the best feature subset. It has its own stopping criteria to stop the search strategy when there is no feature left to knock out. FCBF discards irrelevant features by ranking the correlation measured by SU between feature and class and between feature and feature. Symmetrical uncertainty (SU) is a modified version of information gain (IG) as defined in the following [25]:

$$SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right] \tag{2.17}$$

  Where, X and Y are the discrete features and H denotes the entropy. $H(X)$ is the entropy of X before observing the Y and $H(X|Y)$ is the entropy of X after observing Y.

  SU has values within a range [0, 1], in which the value of 1 indicates that the knowledge of X can completely predicts the Y or vice versa and the value 0

indicates that both X and Y are independent to each other. Among other correlation based measures, FCBF is very efficient and in general, ran very faster than other subset selection measures [26].

- **Consistency**

  The Consistency-based Filter assesses the value of a subset of features based on the level of consistency in the class values when the training instances are projected onto the subset of attributes [27]. Consistency is being used as an indicator in the consistency-based feature subset selection technique to determine the relevance of a feature subset. This technique yields a minimal feature subset with the same consistency as all of the features [28]. In general, consistency measure is monotonic, fast, able to remove redundant and/or irrelevant features, and capable of handling some noise [29]. In this feature subset selection, the entire feature set of the training data set is partitioned into the maximum feasible combinations of feature subsets and the consistency measure is calculated for each subset to determine the significant feature subset. The consistency criterion ensures that the same combination does not appear in several classes[30].

2.3 Stability of Feature selection

Stability, an important characteristic of any feature selection algorithm, is a measure of the sensitivity of the selected feature subset to the small changes of training set or the parameters of the algorithm. In any application area of data mining involving high dimensional data where the individual feature has a distinct physical meaning, feature selection results cannot be used reliably if it changes with small perturbation in the training set. Also if selected subset from the feature selection algorithm is not always same, robust performance for classification will be not stable and optimum [31]. Figure 2.3 shows how the stability of a feature selection algorithm is calculated.

Based on the evaluation criteria of feature selection algorithms, stability measures are categorized into three groups: Stability by index/subset, Stability by

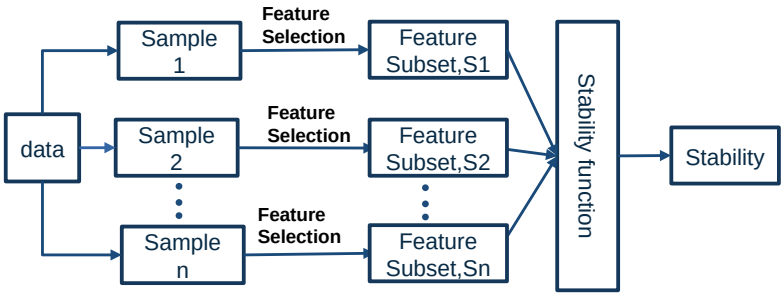


Figure 2.3: Stability Calculation

rank, and Stability by weight [32] (Figure 2.4). Following are the details description of different stability measures used in the literature.

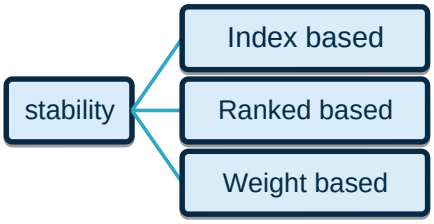


Figure 2.4: Types of Stability

A. Index based stability measures:

In this measure, the amount of overlap between the overall subset of selected features is calculated as stability. There are various index based stability measures which are given below.

- Hamming distance: If S_i and S_j are the two subsets of selected features, then Hamming distance is calculated as the overlap between two subsets. Let the size of the selected subset of features is 'm' and it is a binary vector, in which 1 indicates that feature is present and 0 represents that feature is absent. Now the hamming distance between the two subsets of selected features is calculated as the following:

$$H(S_i, S_j) = \sum_{k=1}^{m} |S_{ik} - S_{jk}| \tag{2.18}$$

If the total number of feature subset is W, then the total hamming distance is calculated as:

$$H_t = \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} H(S_i, S_j) \tag{2.19}$$

Dunne et al. first introduced the Average Normalized Hamming distance as a stability measure, which is given below [33]:

$$H_{AN}(S_i, S_j) = \frac{2 * H_t}{m * |W| * (|W| - 1)} \tag{2.20}$$

This measure determines how much variation there is in the distribution of features present in the subsets selected in different runs of the feature selection algorithm, with 0 indicating no variation and 1 indicating maximum variation. The H_{AN} is in the range [0, 1].

P. Somol and J. Novovi cova in 2010 defined Normalize Hamming Index (NHI) as a stability measure based on the hamming distance. Normalize Hamming Index (NHI) is represented by [34]:

$$H_{NHI}(S_i, S_j) = 1 - \frac{H(S_i, S_j)}{m} \tag{2.21}$$

The total stability of all pairwise feature subset in W is defined by Average Normalize Hamming Index.

$$H_{ANHI}(S_i, S_j) = \frac{2 \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} H_{NHI}(S_i, S_j)}{|W| * (|W| - 1)} \tag{2.22}$$

- Jaccard index: Jaccard index, a similarity measure is used as a metric for comparing the diversity of feature subsets [35] [36] [37]. Jaccard index is defined as the cardinality of the intersection divided by the cardinality of the union of the two sets. If S_i and S_j are the two subsets of selected features, then Jaccard index is calculated as the following:

$$SI_J(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{2.23}$$

Where S_i and S_j are the two different feature subsets.

When Jaccard index is used as a stability measure of feature selection, it is defined as the average of similarities across all W runs of the feature selection algorithm, which is shown in the following equation.

The range of this stability index is [0, 1], where the values near to 0 indicate that feature selection result is unstable and the values near to 1 indicate that result is stable.

- Dice-Sorensen index: Dice-Sorensen index, a similarity measure is actually a harmonic mean index can be used as the stability metric of feature selection algorithms. Sorensen similarity or Dice similarity measure introduced by Dice in 1945 and Sorensen in 1948. This index calculates the overlap between two feature subsets by the following equations:

$$SI_{DS}(S_i, S_j) = \frac{2|S_i \cap S_j|}{|S_i| + |S_j|} \tag{2.24}$$

The range of this stability index is also [0, 1], in which 1 indicates that two subsets are identical and 0 indicates that subsets are totally different [38].

- Ochiai index: Ochiai index or geometric mean index is also a similarity index, first introduced in 1957. Ochiai index describing the dissimilarity between two subsets as [39]:

$$SI_{Ochi}(S_i, S_j) = \frac{|S_i \cap S_j|}{\sqrt{(|S_i| * |S_j|)}} \tag{2.25}$$

- Kuncheva index: Kuncheva first introduced the property based stability index in 2007 by mentioning three properties specially the property named correction by chance [40]. This consistency index or Kuncheva index for two subsets, $S_i \subset X$ and $S_j \subset X$, such that the cardinality of subsets, $|S_i| = |S_j| = k$, where $0 < k < |X| = n$, is defined as:

$$SI_{KI}(S_i, S_j) = \frac{r - E[r]}{max(r - E[r])} = \frac{r - \frac{k^2}{n}}{k - \frac{k^2}{n}} = \frac{rn - k^2}{k(n-k)} \tag{2.26}$$

Where, r is the cardinality of intersection of two subsets, n is the total number

of features. The range of Kuncheva index is [-1, 1]. One major drawback of this index is that, it cannot deal with different cardinality of feature subsets.

- Lustgarten's measure: This is a modification of Kuncheva index, introduced by Lustgarten et al. in 2009 [41]. This measure can handle different cardinality of feature subsets. Now if the cardinality of two subsets is $|S_i| = k_i$ and $|S_j| = k_j$, then Lustgarten's measure can be defined as:

$$SI_L(S_i, S_j) = \frac{r - E[r]}{max(r - E[r]) - min(r - E[r])} = \frac{r - \frac{k_i k_j}{n}}{max(r) - min(r)} = \frac{r - \frac{k_i k_j}{n}}{min(k_i, k_j) - max(0, k_i + k_j - n)} \tag{2.27}$$

  The range of this measure is also from -1 to 1, but cannot reach in the exact value of -1 or +1. For random feature selection, the expected value of this measure is 0. Positive values are obtained if the feature selection is more stable than random feature selection and negative values are obtained if the method is less stable than random feature selection.

- Wald's measure: Wald's measure is another modification of Kuncheva index which also deals with different size of feature subsets [42]. This measure was introduced by Wald et al. in 2013 as a modified Kuncheva's consistency index. The range of this measure is [(1- n), 1] [43].

$$SI_W(S_i, S_j) = \frac{r - E[r]}{max(r - E[r])} = \frac{r - \frac{k_i k_j}{n}}{min(k_i, k_j) - \frac{k_i k_j}{n}} \tag{2.28}$$

  This measure provides the maximum value of +1 when the overlap between the two feature subsets is maximum i.e., $k_i = k_j = r$ and it attains the minimum value of -1 for the condition $k_i = k_j = \frac{n}{2}$ and $r = 0$. This measure also provides the value of 0 when overlap between the two subsets is equal to what would be expected by random chance.

- Nogueira's measure: In 2015, Nogueira and Brown proposed an extension of Kuncheva index with variable length of feature subsets [31]. They proposed this measure based on some desirable properties of stability measure identifying from the literature and proved that their proposed measure satisfies all

the desired properties. The range of this is also [-1, 1]. The following equation shows this Nogueira's measure.

$$SI_N(S_i, S_j) = \frac{r - E[r]}{max(|r - E[r]|)} = \frac{r - E[r]}{max[-min(r - E(r)); max(r - E(r)]} = \frac{r - \frac{k_i k_j}{n}}{max[-max(0, k_i + k_j - n) + \frac{k_i k_j}{n}; min(k_i, k_j) - \frac{k_i k_j}{n}]} \tag{2.29}$$

B. Rank based stability measures

Feature ranking is used to quantify the stability of feature selection method by evaluating the correlation between features. In this chapter, two types of rank based stability metrics are used. These are described below:

- Spearman rank correlation coefficient (SRCC):

  Stability of two ranked sets of features R_i and R_j is given by [10]:

$$SRCC(R_i, R_j) = 1 - 6 \sum_m \frac{(R_{im} - R_{jm})}{n(n^2 - 1)} \tag{2.30}$$

  Where n is the total number of feature. The range of $SRCC$ is [-1, 1]. If two ranked subsets are identical, then the value of $SRCC$ is 1 and if exactly opposite, then it is -1. When there is no correlation between subsets, the value will be 0.

- Canberra Distance (CD): This metric represents the absolute dissimilarity between two ranked sets [32]. The value of CD is dependent on the number of features, n. As the number of features becomes larger, the value of CD will be larger. The value of CD for two ranked sets of features R_i and R_j is written as:

$$CD(R_i, R_j) = \sum_{m=1}^{n} \frac{|R_{im} - R_{jm}|}{R_{im} + R_{jm}} \tag{2.31}$$

  By normalizing CD by n, the range of value will be between 0 and 1.

C. Weight based stability measures:

Weight based stability measures only consider the weights of feature sets S and find the correlation between the weights of two feature sets as the stability.

- Pearson's correlation coefficient (PCC): PCC returns the correlation between the weights of the selected subsets of features [44]. The range of PCC is [-1, 1]. 1 indicates weight vectors are perfectly correlated, on the other hand -1 means, they are oppositely correlated. 0 indicates there is no correlation.

$$PCC(W_i, W_j) = \frac{\sum_m (W_{im} - \mu_{W_i})(W_{jm} - \mu_{W_j})}{\sqrt{\sum_m (W_{im} - \mu_{W_i})^2 \sum_m (W_{jm} - \mu_{W_j})^2}} \tag{2.32}$$

  Where, W_i and W_j are weights of two feature subsets S_i and S_j respectively and μ be the mean of feature set S.

2.4 Related Works on Feature Selection Algorithms

Some of the popular filter based feature ranking algorithms are described here in brief. In [45], authors have used the feature ranking methods such as Information Gain (IG), Gain Ratio (GR), correlation, Symmetrical Uncertainty (SU) and Chi-squared (CS) for recognizing the handwritten digits. The effect of feature selection on classification accuracy is analyzed in [46]. This paper uses six feature ranking based filter methods of IG, GR, SU, One-R, CS and Relief-F for the feature selection process and uses three classification models for comparative study. In [47], authors investigated Cancer Classification using feature selection with filter methods of Signal-to-noise statistic, CFS, CS and Relief-F for Probabilistic Neural Networks. Generally, for high dimensional data such as gene expression data, the filter method is extensively used. In [30], the authors proposed an unsupervised feature selection algorithm with feature ranking method of CS for maximizing the classifier performance. This algorithm achieved better prediction accuracy and also reduced the number of features compared to other methods. For handwriting recognition, Cilia et al. in [13], used five univariate feature ranking based methods for ranking the

features while feature subset was chosen by a greedy search approach. They used CS, Relief, GR, IG and SU as the feature ranking method and the Best First (BF) search strategy combined with consistency criterion and correlation based feature selection criterion as for searching feature subsets. Chen et al. in [48], used filter based ranking feature selection (FRFS) methods for Security vulnerability prediction (SVP) and showed that FRFS can improve the performance of SVP compared to others. They also performed the diversity analysis on identified vulnerable modules by using different FRFS methods. In [49], Ghazy et al. used the different ranking and subset-based feature selection techniques for finding the optimum number of features to find an appropriate classifier. They mainly used these feature selection techniques to verify the performance of the intrusion detection system (IDS). In [50], authors proposed a task of feature ranking for multi-target regression (MTR). They studied two types of feature ranking scores for MTR, one was ensemble based, and the other was an extension of the Relief-F method. Lee et al. in [51], proposed an efficient multivariate feature ranking method for gene selection and for improving the accuracy of microarray data classification. In their work, they created a new feature ranking method using the Markov blanket (MB), which embedded with relevance. They showed that the proposed feature ranking method possesses high classification accuracy as well as good efficiency.

In [52], author proposes a novel spectral matching technique by combining the JM distance and the Spectral Angle Mapper (SAM) algorithm in hyperspectral image data. Their proposed JM-SAM approach performs very well than the individual JM distance measure and SAM algorithm with the least average entropy in spectral matching. Dalponte et al. in [53], used the Jeffries-Matusita (JM) distance combined with sequential forward floating selection (SFFS) search strategy for fast and reliable feature selection. In this case, JM distance also has been used for hyperspectral data. In [54], the authors presented an analysis of the linear attenuation coefficients, which were used as a useful feature of mono-spectral and multispectral images using statistical pattern classification tools. In this paper, feature extraction was performed by JM distance and Karhunen-Loeve transformation. Daamouche et al. [55] proposed a particle swarm optimization (PSO) based approach for very high resolution (VHR) image classification, in which JM distance, support vector

machine (SVM), cross-validation (CV) accuracy and normal Bhattacharyya distance were used as the fitness function.

In [56], authors developed a new technique for crop identification by combining the wavelet variance and the JM distance (CIWJ). The proposed CIWJ approach outperforms other approaches for efficient crop mapping, such as agricultural crop identification with high spatial resolution images and classifications for more general or specific land use. In this paper [57], JM distance is applied as an evaluator of image segmentation in the area of remote sensing images. Here authors proposed an unsupervised evaluation method for evaluating the performance of segmentation using the JM distance and the area-weighted variance (WV). Authors in the paper of [58] proposed an extension of the JM distance measure for multiclass problems of feature selection. They formulated an equation for JM distance measure and used optical remote-sensing data for the experiment. They also compared their results with the most familiar weighted average JM distance. Sen et al. [59] studied JM distance as an efficient tool for feature selection in binary classification problems compared to their other feature ranking methods.

2.5 Related Works on Stability Measures for Feature Selection Algorithms

Recently, the stability of feature selection is a parameter which has been shown importance in feature selection literature. Wang et al. [60] studied the stability of three forms of feature selection methods using software engineering data set. They proposed a newly Average Pairwise Tanimoto Index (APTI) to measure the stability of feature selection methods. Somol and Novovi˘cova [61] proposed various new consistency measures for appraising the stability of feature selection algorithms which select a subset of varying sizes. They compared their results with generalized Kalousis measure that estimates pairwise similarities between subsets. Nogueira and Brown [31] provided some comparative studies and identified the desirable/undesirable properties for stability measures of feature selection algorithms

that return feature sets. They also proposed a generalization of Kuncheva's index for feature selection methods that do not return feature sets of the same cardinality. Another work of Nogueira and Brown [62] argued that some desirable properties that were missing should be presented in existing stability measures and found that the simple Pearson's correlation coefficient has all necessary properties than other alternatives. They also guided how this measure in the application can offer better interpretability and more assurance in the model selection process. Lustgarten et al. [41] presented a new stability metric that can be used to evaluate feature subset with robustness and comparable to random feature selection. They evaluated this metric on the biomedical data set with three different classifier based feature selection methods that included Support Vector Machines (SVM), Logistic Regression (LR) and Naïve Bayes (NB). Their proposed metric can be applied directly on methods which have different cardinality feature subsets. Wald et al. [42] proposed another modified Kuncheva's consistency index, which can handle the feature subsets of different size. In their work, they used both the filter-based subset selection and wrapper-based subset selection on the same data sets. They found that consistency-based filter generates the smallest feature subsets with the highest stability but CFS gives more consistent sizes feature subsets with moderate stability. Khaire and Dhanalakshmi [10] worked on a paper where they outlined the feature selection algorithms with instability problem. They considered several stability measures based on index, rank and weight, and also discussed the solutions of the instability problem of feature selection algorithms.

Chapter 3

Stability of Feature Selection Algorithms

3.1 Introduction

The main aim of this chapter is to conduct a comparative study on the stability of feature selection approaches. At first a work has been done on the filter ranked feature selection algorithms which are IG, CS, One-R, GR, SU, JMD and Relief. These feature selection methods are employed on fifteen UCI data sets to measure the stability of the generated feature set with three different types of stability measures. These stability measures include Kuncheva index and Jaccard index as index based stability measures, Spearman ranked correlation coefficient and Canberra distance as rank based stability measures, and the Pearson correlation coefficient as weight based stability measures. After This work has been extended to include the comparative study on the stability of both filter and wrapper based feature selection algorithms. In this work, both filter ranked based and subset based approaches have been used.

3.2 Stability of Filter based Feature Selection Algorithms

3.2.1 Methodology

To measure the stability of a feature selection algorithm, several steps have been considered. Initially, different sub-samples are drawn from the original data set. Each of the feature selection algorithms then generates feature subsets from the ranked features. Top 50% features from the ranked feature list have been selected as the feature subset. Finally, aggregating all the feature subsets, the stability of the feature selection algorithm on the specific data set is measured by several stability metrics. The procedures are described in the following.

Partitioning data sets

As we know from the definition, the stability measure calculates the robustness of feature selection algorithms under the training data variation. In this experiment, the variation of training data is accomplished with perturbation technique. Here 70% of instances of a data set are randomly selected without replacement which results in a single partition. This partition also maintains the original class ratio of a data set. Above process is repeated 10 times to generate 10 partitions, which are the input data of a feature selection algorithm for stability measurement.

3.2.2 Results and Discussions

Table 3.1 shows the summary of the data sets used in the experiments. All of these 15 data sets have two classes with a different number of features and instances. They are collected from the UCI repository [63].

Table 3.2 illustrates the stability of seven feature selection algorithms on fifteen data sets using index based stability approaches. From now to subsequent

tables, the best stability value is pointed in bold. When Kuncheva Index is used for stability measures, CS produces the highest stability for 7 data sets. Both One-R and JMD show the second highest stability for 6 data sets. When Jaccard Index is used for stability measures, JM distance (JMD) produces the highest stability for most of the data sets (8 cases). IG however, shows the second highest stability for 7 data sets. Stability of CS with Jaccard index is not as strong as that of Kuncheva index. It is also observed that Relief does not exhibit stability well in either of the stability evaluation approaches.

Table 3.1: Summary of the Data sets

Data set	No. of features	No. of instances	No. of class
Sonar	61	208	2
Ionosphere	34	345	2
Heart Disease	14	270	2
Apndcts	8	106	2
Breast cancer	32	569	2
Diabetes	9	768	2
Prostate Cancer	10	100	2
Cryotherapy	7	90	2
Japanese Vowels	12	640	2
Indian Liver Patient (ILPD)	10	543	2
Banknote authentication	5	1372	2
Climate Model	21	540	2
SPECTF	45	349	2
Parkinson	23	197	2
Musk (version1)	168	476	2

Table 3.3 presents the stability evaluations outcome of the feature selection algorithms with two different stability measures: Spearman ranked correlation coefficient and Canberra Distance. It is observed that, with the Spearman approach, JMD has the best stability because 9 out of 15 data sets produce the best stability. Other feature selection approaches do not show promising stability. When Canberra distance is applied, JMD also shows better stability compared to other approaches. It produces the highest stability for 7 cases. Again the performance of Relief is the worst among them.

Table 3.4 provides the stability performances of seven feature selection algorithms with the Pearson correlation coefficient, a weight based stability measure.

What is interesting in this table is that JMD is the most stable feature selection algorithm that shows best stabilities for above 73% data sets. Other feature selection algorithms are not as good as JMD in terms of stability. From the analysis of experimental results, it is seen that each of the feature selection algorithms shows different results when different types of stability measures are applied. It is also observed that some data sets do not have better stability for all the feature selection approaches. This is because data set characteristics may have a link with feature stability. The interesting finding is that for the majority data sets JM distance produces the best stability when three different stability measures are applied. On the average, JM distance produces better results than other feature selection methods for the case of stability measures. In the case of stability performance, Relief is not stable, producing different feature subset for the different run.

Table 3.2: Index based stability measures for different feature selection methods

Data set	Kuncheva Index							Jaccard Index						
	IG	**CS**	**GR**	**SU**	**OneR**	**JMD**	**Relief**	**IG**	**CS**	**GR**	**SU**	**OneR**	**JMD**	**Relief**
Sonar	**0.872**	0.864	0.706	0.792	0.842	0.859	0.693	**0.775**	0.617	0.522	0.596	0.672	0.726	0.669
Ionosphere	0.762	0.743	0.799	0.756	0.716	**0.885**	0.562	0.561	0.583	0.663	0.603	0.551	**0.791**	0.385
Heart disease	0.887	0.907	0.862	0.862	**0.969**	0.945	0.591	0.792	0.829	0.752	0.751	0.943	**0.898**	0.728
Apndcts	0.806	**0.812**	0.806	0.806	0.666	0.790	0.760	**0.817**	0.815	0.815	0.815	0.701	0.800	0.680
Breast cancer	**0.946**	0.943	0.883	0.938	0.928	0.945	0.837	**0.945**	0.942	0.884	0.936	0.927	0.944	0.724
Diabetes	0.874	0.822	**1.000**	0.933	0.830	0.880	0.548	0.748	0.644	**1.000**	0.867	0.659	0.659	0.643
Prostate Cancer	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.876	0.806	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.760	0.689
Cryotherapy	0.925	0.925	0.925	**1.000**	**1.000**	**1.000**	0.750	0.867	0.867	0.867	**1.000**	**1.000**	**1.000**	0.736
Japanese Vowels	**1.000**	**1.000**	0.881	0.956	0.793	**1.000**	0.664	**1.000**	**1.000**	0.893	0.960	0.823	**1.000**	0.719
Banknote	**1.000**	**1.000**	0.822	**1.000**	**1.000**	**1.000**	0.900	**1.000**	**1.000**	0.763	**1.000**	**1.000**	**1.000**	0.867
ILPD	0.826	0.826	0.826	0.826	0.736	**0.990**	0.558	0.877	0.877	0.877	0.877	0.816	**0.999**	0.533
Climate model	1.000	**1.000**	**1.000**	**1.000**	**1.000**	0.685	0.534	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.709	0.661
SPECTF	0.845	**0.908**	0.845	0.845	0.814	0.885	0.638	0.859	0.835	0.859	0.859	0.833	0.875	0.475
Parkinson	0.945	**1.000**	0.701	0.817	**1.000**	0.916	0.691	0.867	**1.000**	0.641	0.663	**1.000**	0.909	0.544
Musk (V 1)	0.788	0.782	0.798	0.800	0.762	**0.812**	0.576	0.653	0.649	0.665	0.668	0.628	**0.687**	0.408
Average	0.898	**0.902**	0.857	0.889	0.870	0.898	0.674	0.851	0.844	0.813	0.840	0.837	**0.850**	0.631

Table 3.3: Rank based stability measures for different feature selection methods

Data set	**Spearman ranked correlation coefficient**							**Canberra Distance**						
	IG	**CS**	**GR**	**SU**	**OneR**	**JMD**	**Relief**	**IG**	**CS**	**GR**	**SU**	**OneR**	**JMD**	**Relief**
Sonar	0.832	0.836	0.800	0.822	0.834	**0.880**	0.421	0.829	0.831	0.821	0.825	0.825	**0.868**	0.380
Ionosphere	0.698	0.681	0.787	0.757	0.571	**0.930**	0.185	0.601	0.591	0.637	0.598	0.525	**0.794**	0.510
Heart disease	**1.000**	**1.000**	0.889	0.929	0.956	**1.000**	0.527	0.853	0.855	0.802	0.822	**0.882**	0.833	0.456
Apndcts	**0.814**	0.778	0.613	0.679	0.578	0.604	0.356	**0.696**	0.676	0.622	0.641	0.578	0.688	0.460
Breast cancer	**0.984**	0.985	0.935	0.978	0.972	0.983	0.808	**0.894**	0.892	0.817	0.881	0.851	0.872	0.598
Diabetes	0.887	0.897	**0.943**	0.931	0.765	0.883	0.395	0.786	0.784	**0.827**	0.816	0.718	0.757	0.414
Prostate Cancer	0.952	0.951	0.953	0.951	0.961	**0.964**	0.597	0.944	0.942	0.944	0.942	**0.949**	0.728	0.613
Cryotherapy	0.958	0.946	0.958	0.958	**0.962**	0.865	0.731	**0.908**	**0.908**	**0.908**	**0.908**	0.906	0.782	0.610
Japanese Vowels	0.992	0.994	0.971	0.991	0.855	**0.999**	0.700	0.935	0.945	0.854	0.945	0.738	**0.976**	0.533
Banknote	**1.000**	**1.000**	0.929	**1.000**	**1.000**	**1.000**	0.836	**1.000**	**1.000**	0.911	**1.000**	**1.000**	**1.000**	0.804
ILPD	0.941	0.884	**0.959**	0.954	0.709	0.928	0.300	0.864	0.831	**0.880**	0.872	0.666	0.855	0.392
Climate model	0.938	0.938	0.938	0.938	**1.000**	0.741	0.361	0.925	0.925	0.925	0.925	**1.000**	0.578	0.360
SPECTF	0.906	0.897	0.800	0.886	0.715	**0.927**	0.350	0.782	0.780	0.720	0.763	0.727	**0.783**	0.368
Parkinson	0.742	0.787	0.723	0.783	0.702	**0.922**	0.587	0.644	0.651	0.653	0.667	0.599	**0.808**	0.494
Musk (V 1)	0.701	0.696	0.679	0.726	0.723	**1.000**	0.501	0.661	0.662	0.676	0.669	0.708	**1.000**	0.461
Average	0.890	0.885	0.858	0.886	0.820	**0.908**	0.510	**0.821**	0.818	0.800	0.818	0.778	**0.821**	0.497

Table 3.4: Weight based stability measures for different feature selection methods

Dataset	Pearson Correlation Coefficient						
	IG	**CS**	**GR**	**SU**	**OneR**	**JMD**	**Relief**
Sonar	0.781	0.737	0.687	0.721	0.737	**0.799**	0.489
Ionosphere	0.692	0.614	0.787	0.757	0.614	**0.995**	0.404
Heart disease	0.909	0.897	0.883	**0.925**	0.897	0.915	0.357
Apndcts	**0.811**	0.763	0.581	0.655	0.763	0.613	0.534
Breast cancer	0.993	0.988	0.935	0.978	0.986	**0.994**	0.839
Diabetes	0.939	0.875	0.938	0.925	0.875	**0.968**	0.413
Prostate Cancer	0.983	**0.997**	0.939	0.936	**0.997**	0.964	0.601
Cryotherapy	0.943	0.931	**0.951**	**0.951**	0.931	0.930	0.790
Japanese Vowels	0.999	0.999	0.971	0.991	0.999	**1.000**	0.786
Banknote	0.996	0.997	0.929	**1.000**	0.997	**1.000**	0.880
ILPD	0.959	0.945	0.955	0.950	0.945	**0.983**	0.447
Climate model	0.921	0.867	0.851	0.851	0.867	**0.978**	0.563
SPECTF	0.887	0.856	0.790	0.880	0.856	**0.928**	0.388
Parkinson	0.808	0.773	0.722	0.783	0.773	**0.924**	0.578
Musk (V 1)	0.673	0.637	0.650	0.701	0.637	**0.831**	0.487
Average	0.886	0.858	0.838	0.867	0.858	**0.921**	0.570

3.3 Comparative Study on Stability of Filter and Wrapper Algorithms

In this section, stability of filter based and wrapper based feature selection techniques are explored with using both the subset based and feature ranking approaches. In previous section stability has been calculated for filter ranked based feature selection methods with using only binary datasets. In this part we have extended the work. For filter based feature selection, both feature ranking and feature subset selection approach are explored and for wrapper method only subset based approach are considered with three different learners. Here, eight filter ranked based feature selection (FRFS) methods; three filter subset based feature selection (FSFS) methods and wrapper method with three learners of Decision Tree, K-NN and Linear SVM are applied. Stability are calculated with using seven different stability metrics.

In this work, selected features are presented as a subset and stability are calculated from the feature subset with using index based measures only, not considering the rank based or weight based stability measures.

3.3.1 Working Procedure

In this work, 30 different sizes of datasets were used to evaluate the stability of filter and wrapper based feature selection algorithms. Seven different subset based stability measures were used to calculate the stability and then compare those stability measures by finding their merits and demerits.

3.3.2 Simulation Experiment

For simulation experiments, both binary and multiclass datasets were used. Among 30 datasets, 20 were collected from UCI [64] and rest of the datasets were taken from OpenML [65]. Among these datasets, some needs to be preprocessed like datasets have missing values or are in categorical in nature. Here, categorical type missing values in the datasets are replaced with the most frequently used value, and after that, the whole dataset is converted into numeric type. Numeric type missing values are replaced with the average value. Categorical type datasets without missing values are directly converted to numeric type. In this work, three groups of feature selection algorithms are used to measure the stability of these feature selection algorithms such as the filter ranked based, filter subset based and wrapper based. In filter ranked based measures, IG, GR, SU, CS, One-R, JM distance, Fisher Score (FS) and Relief-F are used. For filter subset based approaches, Correlation based Feature Selection (CFS), consistency and First Correlation Based Filter (FCBF) are used. In wrapper based approach, sequential forward search (SFS) with three different classifiers such as decision tree (DT), K-NN and SVM Linear is used.

In filter ranked based approaches, first of all, classification accuracy was calculated at different percentages of selected features, such as 10%, 25%, 50% and 75%. Then maximum classification accuracy was identified from those percentages

of selected features. For stability calculation, we took the percentage of feature in which accuracy was maximum.

3.3.3 Simulation Results

Table 3.5 shows the stability of filter ranked based feature selection algorithms with different types of stability measures. Here, average value (Avg) and standard deviation (SD) of stability measure are taken from the overall 30 datasets. Seven different subset based stability metrics are used for calculating the stability which are Hamming distance, Jaccard index, Dice-Sorensen index, Ochiai index, Lustgarten's measure, Nogueira's measure and Wald's measure. In the stability calculation of this ranked based feature selection, Relief-F shows lowest stability, but other algorithm's stability measure is very much comparable.

Table 3.5: Stability of Filter ranked based feature selection algorithms

		Lustgarten measure	Nogueira measure	Wald measure	Hamming distance	Jaccard index	Dice-Sorensen index	Ochiai index
IG	Avg	0.7921	0.9254	0.9254	0.9460	0.8839	0.9253	0.9253
	SD	0.0722	0.0796	0.0796	0.0656	0.1398	0.0962	0.0962
GR	Avg	0.7852	0.9068	0.9068	0.9312	0.8811	0.9253	0.9249
	SD	0.0672	0.0798	0.0798	0.0567	0.1257	0.0962	0.0947
SU	Avg	0.7958	0.9236	0.9236	0.9411	0.8963	0.9354	0.9354
	SD	0.0699	0.0719	0.0719	0.0605	0.1053	0.0713	0.0713
CS	Avg	0.7866	0.9132	0.9132	0.9322	0.8712	0.9193	0.9193
	SD	0.0723	0.0703	0.0703	0.0601	0.1277	0.0869	0.0869
One-R	Avg	0.8011	0.9360	0.9360	0.9501	0.9007	0.9378	0.9378
	SD	0.0795	0.0753	0.0753	0.0647	0.1396	0.0950	0.0950
JM dist	Avg	0.7825	0.9074	0.9074	0.9296	0.8764	0.9192	0.9192
	SD	0.0780	0.0983	0.0983	0.0750	0.1577	0.1305	0.1305
Fisher score	Avg	0.7573	0.8640	0.8640	0.8990	0.8281	0.8876	0.8876
	SD	0.0936	0.1177	0.1177	0.0920	0.1577	0.1226	0.1226
Relief-F	Avg	0.6724	0.7422	0.7422	0.8199	0.7095	0.8019	0.8019
	SD	0.1025	0.1513	0.1513	0.1146	0.1789	0.1567	0.1567

The stability calculation of filter subset based feature selection algorithm is shown in Table 3.6. Among the three algorithms, CFS shows better stability than FCBF and consistency.

Table 3.7 shows the stability calculation of wrapper based feature selection algorithm with three different classifiers such as DT, KNN and SVM Linear. Among

Table 3.6: Stability of Filter subset based feature selection algorithms

		Lustgarten measure	Nogueira measure	Wald measure	Hamming distance	Jaccard index	Dice-Sorensen index	Ochiai index
CFS	Avg	0.7711	0.9125	0.9198	0.8853	0.7340	0.8257	0.8327
	SD	0.0836	0.0848	0.0828	0.0817	0.1790	0.1266	0.1223
FCBF	Avg	0.7571	0.8550	0.8568	0.8992	0.6790	0.7533	0.7572
	SD	0.0934	0.1310	0.1310	0.0848	0.2758	0.2432	0.2401
Consistency	Avg	0.6910	0.7630	0.7657	0.8228	0.5263	0.6320	0.6403
	SD	0.1033	0.1400	0.1455	0.1010	0.2332	0.2100	0.2090

Table 3.7: Stability of Wrapper based feature selection algorithm with different classifiers

		Lustgarten measure	Nogueira measure	Wald measure	Hamming distance	Jaccard index	Dice-Sorensen index	Ochiai index
DT	Avg	0.6696	0.7219	0.7225	0.8439	0.4247	0.4999	0.5063
	SD	0.1245	0.1589	0.1601	0.0822	0.2883	0.2919	0.2916
KNN	Avg	0.7454	0.7724	0.7705	0.9457	0.5767	0.5786	0.5770
	SD	0.1959	0.2107	0.2131	0.0672	0.4006	0.3967	0.3976
SVM	Avg	0.6823	0.7369	0.7470	0.8027	0.4882	0.5931	0.6041
Linear	SD	0.1159	0.1259	0.1274	0.1207	0.2538	0.2314	0.2282

varieties of DT algorithms, C4.5 is used in this simulation experiment. For KNN classifier, K is set to 5. Table 3.7 shows that stability of wrapper method with KNN classifier is better than using with DT or SVM Linear.

Table 3.8 shows the overall comparison of stability measure among different types of feature selection algorithms. In this table, *Avg Stb* and *SD Stb* are taken from the Table 3.5, 3.6, 3.7 by averaging the *Avg* value and *SD* value of different stability metrics. This table shows that ranked based filter method gives better stability than filter subset based and wrapper based approaches.

3.4 Conclusion

In this chapter, at first five stability measures have been used for assessing the strength of seven feature selection algorithms on binary datasets. In this case rank based feature selection approaches including IG, GR, SU, One-R, CS, JM Distance and Relief have been used. Stability indices which are used here are Kuncheva index,

Table 3.8: Overall comparison of Stability among different types of feature selection algorithms

Feature Selection algorithm		Stability Calculation	
		Avg Stb	SD Stb
Filter ranked based	IG	0.8981	0.0979
	GR	0.8879	0.0920
	SU	0.9018	**0.0813**
	CS	0.8868	0.0892
	One-R	**0.9099**	0.0982
	JM dist	0.8854	0.1159
	Fisher-score	0.8522	0.1224
	Relief-F	0.7457	0.1490
Filter subset based	CFS	0.8285	0.1169
	FCBF	0.7838	0.1786
	Consistency	0.6760	0.1667
Wrapper based	DT	0.6121	0.2033
	KNN	0.7015	0.2773
	SVM Linear	0.6490	0.1753

Jaccard index, Spearman ranked correlation, Canberra distance, and Pearson's correlation coefficient. Comparative results are presented based on several experiments done on 15 UCI datasets and are found that all of the feature selection approaches are not equally stable. It is observed that JM Distance produces the best stability score for most of the datasets, whereas the stability score for Relief is the lowest. After that, a comparative study of the stability of both filter based and wrapper based feature selection algorithms have been performed with simulation experiments. In this case, both rank based and subset based feature selection algorithms are used and simulation experiments are performed with using both binary and multiclass publicly available UCI datasets. Simulation results of stability measures reveal that wrapper method shows the least stability while feature ranking based filter method exhibits the highest stability.

Chapter 4

A Critical Study on Stability Measures of Feature Selection

4.1 Introduction

In many real life domains, especially for medical or business data, identifying the subset of meaningful and interpretable features is of prime importance for further experimental research. Thus, in addition to the effectiveness of the selected feature subset's ability for accurate classification, the other important criterion for the evaluation of feature selection algorithm is its stability. Stability of an algorithm characterizes the repeatability of its outcome given different sets of input from the same data generating process i.e., with the same underlying probability distribution. A stable feature selection algorithm should not produce radically different feature preferences in the form of ranked lists or subsets of features with different groups of the same training data.

The concept of measuring the stability of classification algorithm is examined by Turney [66] in which he introduced a method for quantifying stability, based on a measure of the agreement between classification concepts induced by the algorithm on different sets of training data. The stability of a feature selection algorithm is related to the change in the selected feature subset due to perturbation of training

data or different settings of algorithmic parameters or initialization of the algorithm with different random seeds. A stable feature selection algorithm is more important for knowledge discovery as it exhibits a good confidence level to the domain expert for example, to separate the disease associated genes from microarray studies [67], proteins from mass spectrometry (MS)-based proteomics studies [68], or single nucleotide polymorphism (SNP) from genome wide association (GWA) studies [69]. It is possible that different training sample sets produce different feature subsets which may lead to the same classification concept due to a high level of redundancy in the initial feature set. In this case, contrary to the classification algorithm which can be considered stable, feature selection algorithm produces different outputs. So technically, the concept of stability measurement of a classification algorithm can not be used for stability measurement of any feature selection algorithm. The first published work on the extensive analysis regarding the stability of feature selection algorithm is presented in [35].

Generally, feature selection algorithms provide feature preferences in either a ranked or weighted feature list or an optimum subset of selected features. Depending on the differences of representing feature preferences in the outcome of feature selection algorithms, the assessment of their stabilities is different. Accordingly, various stability measures suitable for evaluating the stability of different categories of feature selection algorithms are developed. Here stability measures related to feature subset-based feature selection algorithms are studied. While there are various stability measures for feature subset selection algorithm, similarity based measures, especially Kuncheva's consistency index [40] is quite popular and widely used. To overcome the main limitation of the Kuncheva index i.e., its inability to cope with feature subsets of different cardinalities, a few modified similarity measures related to the Kuncheva index are also available in the literature. In this work, the Kuncheva index and its existing modifications (Lustgarten, nPOG, Wald, and Nogueira) are studied, their merits, demerits, and limitations are analyzed. One more limitation of the most recent modified similarity measure, Nogueira's measure, has been pointed out. Finally, corrections to Lustgarten's measure have been proposed to define a new modified stability measure that satisfies the desired properties and overcomes the limitations of existing popular similarity based stability measures. The effec-

tiveness of the newly modified Lustgarten's measure has been evaluated with simple toy experiments.

In summary, the contributions of the chapter are highlighted below:

- Critical analysis of existing similarity based stability measures and their desired properties
- Newly pointing out a limitation of Nogueira's measure and a part of Wald's measure
- Proposed correction to Lustgarten's measure to overcome its limitation
- Proposal of a novel extension of Lustgarten's measure which overcomes the limitations of the existing measures

4.2 Stability by Similarity

In similarity based approach, first introduced by Dunne et al. [33], stability is measured by the similarity between two selected feature subsets. For M, the number of feature subsets, stability measure $\Phi(Z)$ is calculated as the average pairwise similarity Φ, between the $M(M-1)$ possible pairs of feature subsets in Z as follows [43, 70]:

$$\Phi(Z) = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1, j \neq i}^{M} SI(S_i, S_j) \tag{4.1}$$

where, SI is a function taking two feature subsets S_i and S_j as inputs and return a similarity value as the output. Similarity can be measured in a variety of ways like the ratio of the intersection to the union of two selected feature subsets, or the amount of overlap between the overall subset of selected features [71, 10]. Dunne et al. [33] proposed relative Hamming Distance between two feature subsets as the similarity measure. Kalousis et al. published the work of stability of feature selection algorithms in 2005 [70] with an extensive discussion on stability measures. Jaccard index was proposed as a similarity based stability measure of feature selection between selected feature subsets in [35]. Other similarity based stability measures used

in the literature are the Dice-Sørenson index, first introduced by Yu et al. in 2008 [72], the Ochiai index [38], the POG (Percentage of Overlapping Genes) index [73]. In 2007, Kuncheva analyzed the performance of different existing stability measures [40] and proposed a new property based similarity measure. A set of 3 properties, which is fundamental for any stability measure, has been introduced in her work. In this chapter, similarity based stability measures, especially Kuncheva index, and some modified measures related to Kuncheva index have been highlighted.

In many research works, ensemble techniques are employed to enhance the stability of feature selection algorithms like, Bayesian model averaging [74, 75], aggregating the results of a collection of feature ranking methods [76, 77], and aggregating the results of the same feature selection method from bootstrapped subsets of samples [78, 1, 79].

4.3 Analysis of Kuncheva Index and Its Extensions

Several similarity based stability measures according to Equation (4.1) are found to be biased by the number of features in the selected feature subset. The stabilities of two feature selection algorithms selecting two identical feature subsets of eight features from a feature set of cardinality 10 and eight features from a feature set of cardinality 100 do not possess the same significance. The later one is more stable having lesser possibility of selecting exactly same 8 features by chance. Kuncheva [40] analyzed this anomaly and, to correct the bias, proposed a similarity measure having the property of correction for chance. Kuncheva's measure has become the most popular and pioneer work on assessing stability of feature subset selection. In the following subsections, Kuncheva index and its popular extensions with their limitations are discussed.

4.3.1 Kuncheva Index

Kuncheva proposed a similarity measure based on the consistency between a pair of feature subsets according to three desirable properties which are monotonicity, limits and correction for chance. Kuncheva index is defined as follows [40]:

$$SI_{KI}(S_i, S_j) = \frac{r - E[r]}{max(r - E[r])} = \frac{r - \frac{k^2}{n}}{k - \frac{k^2}{n}} = \frac{rn - k^2}{k(n - k)} \tag{4.2}$$

where, n represents the total number of features, $r = |S_i \cap S_j|$, is the cardinality of intersection of two selected subsets of features S_i and S_j and $k = |S_i| = |S_j|$, is the cardinality of the selected feature subsets. The maximum limit of Kuncheva index is 1, which is achieved when $r = k$, i.e., when the two selected feature subsets are identical. The minimum value is -1 only when $r= 0$ provided $k = n/2$. For other values of k, with $r = 0$, Kuncheva index does not produce the minimum value -1. Beside this, Kuncheva index is not defined for $k = 0$ and $k = n$, in both the cases Kuncheva index is set to 0. The term, $\frac{k^2}{n}$ is very important part of this measure that corrects the bias due to the chance of selecting the features which are common between the two randomly chosen subsets. In this case, if the stability index is zero, it expresses that the overlap between two subsets is almost due to chance [71].

While Kuncheva index is very efficient for measuring the stability of feature selection algorithms, a major drawback is, it cannot be used for selected feature subsets with different sizes. Several modifications are proposed to overcome the limitation, which are analyzed below.

4.3.2 Extensions of Kuncheva Index

There are three popular extensions of Kuncheva Index for selected feature subsets of different cardinalities. All the measures are of the same general form as Kuncheva, differing in the denominator of the respective measures.

1. Lustgarten's Measure

In 2009, Lustgarten et al. proposed a modification of Kuncheva index by dividing the value of numerator by its range. Lustgarten's measure satisfies the property of correction by chance and is applicable to different cardinality of selected feature subsets [41]. It is popularly used as the modified version of Kuncheva index in different works [80, 10]. In [41], Lustgarten's measure is defined as:

$$SI_L(S_i, S_j) = \frac{r - E[r]}{max(r - E[r]) - min(r - E[r])} \tag{4.3}$$

If two selected feature subsets S_i and S_j are of cardinalities k_i and k_j, respectively, then $E[r] = \frac{k_i k_j}{n}$ and hence the above equation becomes

$$SI_L(S_i, S_j) = \frac{r - \frac{k_i k_j}{n}}{max(r - E[r]) - min(r - E[r])} = \frac{r - \frac{k_i k_j}{n}}{max(r) - min(r)} \tag{4.4}$$

Now $max(r) = min(k_i, k_j)$ and $min(r) = max(0, k_i + k_j - n)$, the above equation reduces to:

$$SI_L(S_i, S_j) = \frac{r - \frac{k_i k_j}{n}}{min(k_i, k_j) - max(0, k_i + k_j - n)} \tag{4.5}$$

This measure has a value in the interval $(-1,1)$. For random feature subset selection, Lustgarten's measure provides a value of 0. Like Kuncheva index, Lustgarten's measure produces a positive value when feature selection method is more stable than random feature selection and produces a negative value when feature selection method is less stable than random feature selection. If S_i or S_j or both have no features or S_i or S_j or both contain all the feature in the domain, then Equation (4.5) is undefined, and in this case it is set to 0, same as in the case of Kuncheva index.

The main drawback of this measure is that Lustgarten's measure does not provide the fixed maximum value of +1 (even when the condition of maximum stability i.e., $k_i = k_j = r$ occurs) rather it depends on the variation of k_i and k_j;

the maximum value close to +1 is achieved when both k_i and k_j are either very small or very close to n. Similarly, it cannot reach the minimum value of -1 for the condition when the cardinality of intersection between feature subsets is zero, i.e., $r = 0$. In above two cases, Kuncheva index provides the maximum and minimum stability value of +1 and -1, respectively.

2. Wald's Measure

Wald et al. in 2013, proposed another modification of Kuncheva's index by dividing the numerator by its maximal value [42] (same as Kuncheva) and is defined as:

$$SI_W(S_i, S_j) = \frac{r - E[r]}{max(r - E[r])} = \frac{r - \frac{k_i k_j}{n}}{min(k_i, k_j) - \frac{k_i k_j}{n}} \tag{4.6}$$

This measure provides the maximum value of +1 when the overlap between the two feature subsets is maximum i.e., $k_i = k_j = r$ and it attains the minimum value of -1 for the condition $k_i = k_j = \frac{n}{2}$ and $r = 0$. This measure also provides the value of 0 when overlap between the two subsets is equal to what would be expected by random chance.

The limitations of Wald's measure are as follows:

1. When one of the feature subset is a proper subset of the other i.e., $S_i \subset S_j$, $k_i < k_j$ and $k_i = r$ or $S_j \subset S_i$, $k_j < k_i$ and $k_j = r$, this measure returns the value of +1. In this case, two feature subsets are not identical and all the elements of two feature subset are not the same. This condition is illustrated by the following example. Suppose, in a feature selection problem, one selected feature set is, $S_i = \{a, c\}$ and other is $S_j = \{a, b, c, d, f, g\}$. Therefore, S_i is a proper subset of S_j and $k_i < k_j$. Let the total number of feature, n equal to 10. The cardinality of intersection of two feature sets is, $r = 2$ and $k_i = r$. Therefore, $min(k_i, k_j) = k_i = 2$, $k_i k_j / n = 12/10 = 6/5$ and the Wald's measure is

   $SI_W(S_i, S_j) = (r - \frac{k_i k_j}{n})/(min(k_i, k_j) - \frac{k_i k_j}{n}) = (2 - 6/5)/(2 - 6/5) = 1.$

2. This measure does not guarantee the lower bound of -1 and depends on k_i, k_j and n. It is -1 only when $k_i = k_j = n/2$. For a given n, the minimum of Wald's measure is $1 - n$, provided, $k_i = n - 1$ and $k_j = 1$ or vice versa with $r = 0$.

   We have defined a generalized lower bound as follows:

   - For the case when $(k_i + k_j) = n$,

     Let us consider, $k_i = q$, $k_j = n - q$, $k_i \leq k_j$ and $r = 0$, and the value of q has the range as $q = 1, 2, 3, \ldots n/2$, then Wald's measure provides, $SI_W(S_i, S_j) = 1 - n/q$.

     This can be proved as following:

     If, $k_i = q = 1$, $k_j = n - q$ and $r = 0$, then $SI_W(S_i, S_j) = 1 - n$

     $k_i = q = 2$, $k_j = n - q$ and $r = 0$, then $SI_W(S_i, S_j) = 1 - n/2$

     $k_i = q = 3$, $k_j = n - q$ and $r = 0$, then $SI_W(S_i, S_j) = 1 - n/3$

     .....

     $k_i = q = n/2$, $k_j = n - q$ and $r = 0$, then $SI_W(S_i, S_j) = 1 - n/(n/2) = -1$

   - For the case when $(k_i + k_j) < n$,

     The value of $SI_W(S_i, S_j)$ is defined within a range as follows,

     $-1 < SI_W(S_i, S_j) < 0$

3. Average nPOG and Average nPOGR

Percentage of overlapping Gene/Features (POG) is defined as the stability measure in [73]. POG is not symmetric, $POG(S_i, S_j) \neq POG(S_j, S_i)$. The measure is defined as:

$$POG(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i|} = \frac{r}{k_i} or, POG(S_j, S_i) = \frac{|S_i \cap S_j|}{|S_j|} = \frac{r}{k_j} \tag{4.7}$$

POG does not consider the correlation between features in the selected feature subsets. $POGR$ is introduced by Zhang et al.which considers the correlated

features, defined as in [81],

$$POGR(S_i, S_j) = \frac{r + Z_{i,j}}{k_i} or, POGR(S_j, S_i) = \frac{r + Z_{j,i}}{k_j} \tag{4.8}$$

where, $Z_{i,j}$ (or $Z_{j,i}$) represents the number of features in feature subset $S_i (or S_j)$, which is significantly positively correlated with at least one feature in feature subset S_j (or S_i). Normalized POG ($nPOG$) and normalized $POGR$ ($nPOGR$) are defined as:

$$\begin{aligned} nPOG(S_i, S_j) &= \frac{POG(S_i, S_j) - E[POG(S_i, S_j)]}{1 - E[POG(S_i, S_j)]} \\ &= \frac{\frac{r}{k_i} - E[r]}{1 - E[r]} = \frac{r - E[r]}{k_i - E[r]} = \frac{r - \frac{k_i k_j}{n}}{k_i - \frac{k_i k_j}{n}} \end{aligned} \tag{4.9}$$

$$\begin{aligned} nPOGR(S_i, S_j) &= \frac{POGR(S_i, S_j) - E[POGR(S_i, S_j)]}{1 - E[POG(S_i, S_j)]} \\ &= \frac{r + Z_{i,j} - E[r] - E[Z_{i,j}]}{k_i - E[r] - E[Z_{i,j}]} \end{aligned} \tag{4.10}$$

It is seen from Equation (4.9) that the measure $nPOG$ is same as Wald's measure, suffering from the same drawbacks as of Wald's measure in addition to being non-symmetric.

4. Nogueira and Brown's Measure

Nogueira and Brown (in 2015) proposed another modification of Kuncheva index by dividing the numerator by its maximal absolute value [31] so that its value belongs to the range $[-1, 1]$ to overcome the limitation of Wald's measure. This measure is defined as follows:

$$\begin{aligned} SI_N(S_i, S_j) &= \frac{r - E[r]}{max(|r - E[r]|)} = \frac{r - E[r]}{max[-min(r - E(r)); max(r - E(r)]} \\ &= \frac{r - \frac{k_i k_j}{n}}{max[-max(0, k_i + k_j - n) + \frac{k_i k_j}{n}; min(k_i, k_j) - \frac{k_i k_j}{n}]} \end{aligned} \tag{4.11}$$

Nogueira's measure can be considered as a generalization of Kuncheva

index for different cardinalities of the selected feature subset and its value for $k_i = k_j = k$ matches with the value of Kuncheva index. The authors in [31] claimed that this measure is bounded by -1 and $+1$ and reaches its maximum value when the two feature subsets are identical. The authors also showed that this measure satisfies the desired properties (1 to 6 of the list in the next subsection) of a stability measure.

However in our experiments with several data sets, we have found the following limitation of this measure:

1. If one feature subset is a proper subset of the other, i.e., $S_i \subset S_j$, $k_i < k_j$ and $k_i = r$ or $S_j \subset S_i$, $k_j < k_i$ and $k_j = r$, this measure returns the maximum value of $+1$, which should not be the case as the two feature subsets are identical. Moreover, we noted that unlike Wald's measure, Nogueira's measure does not produce the maximum value $+1$ for all the cases whenever the condition of proper subset (one of the feature subset is the proper subset of the other) occurs. We have elaborated this findings by toy example and experiment in the next section.

2. Nogueira's measure gives the minimal value of -1, for the conditions $k_i = q, k_j = n - q, k_i \leq k_j$, or vice versa, and $r = 0$ with q in the range $q = 1, 2, 3...n/2$. For other cases, when $k_i + k_j < n$ and $r = 0$, Nogueira's measure, like Wald's measure, lies between -1 and 0 i.e., $-1 < SI_N(S_i, S_j) < 0$.

In the next subsection, the desired properties of any stability measure are listed and Kuncheva index and its modifications are examined.

4.3.3 Desired Properties of Stability Measure

Kuncheva first introduced the consistency based stability measure depending on three desired properties [40]. Beside this, Zucknick et al. also highlighted the three properties of similarity based stability measure in their work [38], which are symmetry, homogeneity and bounds/limits. Later Nogueira identified some properties

from literature and listed in [31, 62, 43]. Based on the research works so far, we have summarized the important desired properties of stability measures as follows:

1. Fully Defined: This property demonstrates that a stability measure should be able to handle any collection of feature subsets, irrespective of its size. Stability measures without this property can not be defined for the class of feature selection algorithms which produce variable size feature subsets.

2. Limits/bounds: The stability measure should be bounded by values that do not depend on the size of the feature subset. The significance of any stability value is much understood when it has a finite range compared to the range of $[-\infty, \infty]$.

3. Maximum-minimum value: The stability measure should reach its maximum value when all the selected feature subsets are identical, the minimum value should be reached when the intersection of the feature subsets is zero. Interestingly, it does not happen for all the measures.

4. Monotonicity: This property is highlighted in Nogueira's work [31, 43]. It states that the stability measure should be an increasing function of the similarity of the feature subsets.

5. Correction for chance: Kuncheva first introduced this property to reduce the effect of size of the selected feature subset. It confirms that the expected value of the stability measure should be constant when the subsets are independently selected at random.

6. Symmetry: Stability measure should be symmetrical irrespective of the order of the feature subsets taken for measurement.

7. Homogeneity: This property represents that, the stability measure should not change if the same constant value is multiplied to the different features in the feature subsets [38].

8. Redundancy awareness: This property reveals that, if the features are redundant in a feature selection problem, then the stability measure of feature

selection should be able to calculate the true amount of redundant information between the feature subsets [31]. In the present work, this property is not considered.

Table 4.1 shows the properties of different similarity based stability measures.

Table 4.1: Properties of Stability measure of feature selection algorithms

Stability measure	**Fully Defined**	**Limits**	**Max-Mini value**	**Monotonicity**	**Correction for chance**	**Symmetry**	**Homogeneity**
Jaccard	✓	✓	✓	✓		✓	✓
Dice-Sørenson	✓	✓	✓	✓		✓	✓
Ochiai	✓	✓	✓	✓		✓	✓
Hamming distance	✓	✓	✓	✓		✓	✓
POG	✓	✓	✓	✓			✓
Kuncheva		✓	✓	✓	✓	✓	✓
Lustgarten	✓	✓		✓	✓	✓	✓
Wald	✓		✓	✓	✓	✓	✓
nPOG	✓		✓	✓	✓		✓
Nogueira and Brown	✓	✓	✓	✓	✓	✓	✓

4.4 Toy Experiment for Illustration of the Drawbacks

In the previous section, we analyzed the merits, demerits and the limitations of different extended version of Kuncheva index. To have a better understanding, we design toy experiments of feature subset selection where different stability measures are used to evaluate similarity between the different pairs of the selected feature subsets S_i, S_j. Here we present the experiments, their results and analysis for the cases arising from different cardinalities of the selected subsets.

1. **For the case when the two selected feature subsets are such that $S_i \subset S_j$ or $S_j \subset S_i$.**

   Let, the total number of features in this experimental problem is $n = 20$. Feature subsets of different cardinalities can be selected from the set of 20 features as a result of the several run of a feature selection algorithm. Among the selected feature subsets from multiple runs of the algorithm, 20 different pairs of feature subsets are considered for stability measurement where each pair contains one feature subset that is a proper subset of the other feature subset. Table 4.2 and Figure 4.1 represent the values of similarity of different measures for different pairs of feature subsets.

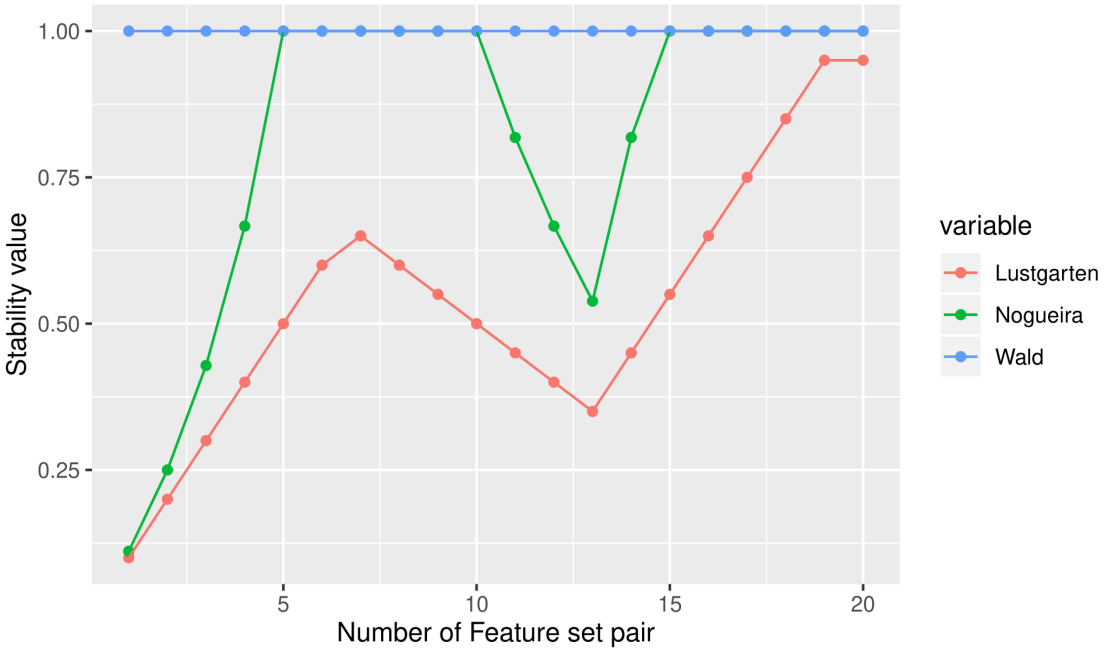


Figure 4.1: Similarity measures for the case when $S_i \subset S_j$ or vice versa.

From Table 4.2 it is found that Wald's measure always produces maximum value +1 while one feature subset is proper subset of the other which means that the two subsets are not identical. Nogueira's measure randomly produces maximum value +1 in some cases but not in all the cases when one subset is proper subset of the other. In this case of stability measurement, Wald's measure and Nogueira's measure produces incorrect result, because this is not the condition for getting maximum stability. Lustgarten's measure shows more consistent result except for two cases (feature subset pair 19 and 20) when the

Table 4.2: Similarity values for the case when $S_i \subset S_j$ or vice versa

Index of feature subset pair	**Cardinality of one feature subset k_i**	**Cardinality of other feature subset k_j**	**Cardinality of intersection of feature subsets, r**	**Lustgarten's measure, $SI_L(S_i, S_j)$**	**Nogueira's measure, $SI_N(S_i, S_j)$**	**Wald's measure, $SI_W(S_i, S_j)$**
1	18	1	1	0.1	0.11	1
2	16	2	2	0.2	0.25	1
3	14	3	3	0.3	0.43	1
4	12	4	4	0.4	0.67	1
5	10	5	5	0.5	1	1
6	8	6	6	0.6	1	1
7	6	7	6	0.65	1	1
8	4	8	4	0.6	1	1
9	2	9	2	0.55	1	1
10	1	10	1	0.5	1	1
11	3	11	3	0.45	0.81	1
12	5	12	5	0.4	0.67	1
13	7	13	7	0.35	0.54	1
14	9	14	9	0.45	0.81	1
15	11	15	11	0.55	1	1
16	13	16	13	0.65	1	1
17	15	17	15	0.75	1	1
18	17	18	17	0.85	1	1
19	19	19	19	0.95	1	1
20	1	1	1	0.95	1	1

two feature subsets are identical and the value should be +1. Figure 4.2 also highlights this condition. In our next experiment, we considered the case when two feature subsets taken for similarity measurement are completely identical with different cardinalities.

2. **For the case when S_i and S_j are identical.**

   Here we design another experiment for feature subset selection in which each selected feature subset pair consists of two identical feature subsets. The total number of features is same as before, $n = 20$ and we considered 19 different pairs of the selected feature subsets with different cardinalities.

   Table 4.3 shows the values of the different similarity measures for the case considered here. It is found that as the two stability measures, Nogueira's measure and Wald's measure provide the accurate result for similarity calculation as expected. The other measure, Lustgarten's measure, cannot provide the maximum stability of +1. While Lustgarten's measure cannot provide the exact value of +1, it provides a value within a known finite range [0.5, +1).

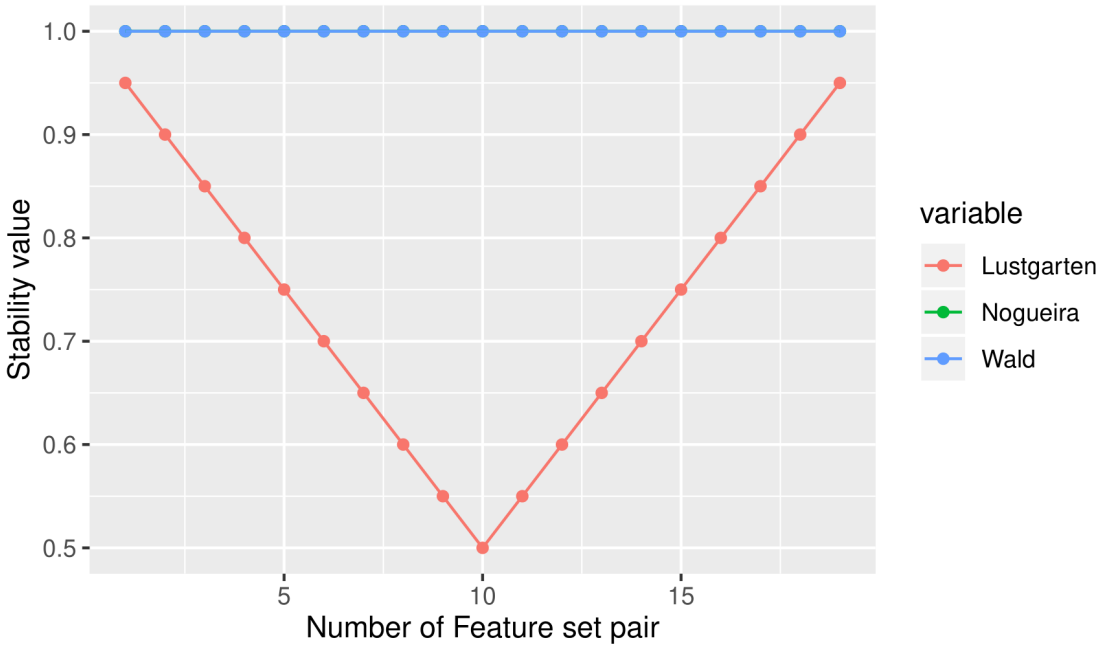


Figure 4.2: Similarity values when two feature subsets are identical.

The graphical representation of Table 4.3 is shown in Figure 4.2. It is noted that Nogueira's measure provides the same values as the Wald's measure, resulting overlap of this two lines in the figure. The next experiment has been conducted for the case when the similarity value between two feature subsets is minimal i.e., there is no common feature between the two subsets.

3. **For the case when $S_i \cap S_j$ is null ($r = 0$)**

As before, the total number of feature in this experiment is, $n = 20$. We considered 19 different feature subset pairs with the condition $r = 0$. Table 4.4 represents the similarity values of different measures.

It is seen that, in line with the analysis in the previous section, Nogueira's measure and Wald's measure reach the minimum value of -1, but does not show the value of -1 for all the cases when $r = 0$. For Wald's measure, the minimum value is achieved only when $k_i = k_j = n/2$ with $r = 0$. The values of Wald's measure in Table 4.4 also supports the fact we mathematically proved in the previous section, for example, if $k_i = q$, $k_j = n - q$, $k_i < k_j$ or vice versa, $r = 0$ and q has the range $q = 1, 2, 3, \ldots n/2$, then Wald's measure provides the value $(1 - n/q)$. Figure 4.3 represents the graphical view of Table 4.4. As expected according to our analysis, Nogueira's stability measure gives the

Table 4.3: Similarity values for the case when S_i and S_j are identical

Index of feature subset pair	**Cardinality of one feature subset** k_i	**Cardinality of other feature subset** k_j	**Cardinality of intersection of feature subsets,** r	**Lustgarten's measure,** $SI_L(S_i, S_j)$	**Nogueira's measure,** $SI_N(S_i, S_j)$	**Wald's measure,** $SI_W(S_i, S_j)$
1	1	1	1	0.95	1	1
2	2	2	2	0.9	1	1
3	3	3	3	0.85	1	1
4	4	4	4	0.8	1	1
5	5	5	5	0.75	1	1
6	6	6	6	0.7	1	1
7	7	7	7	0.65	1	1
8	8	8	8	0.6	1	1
9	9	9	9	0.55	1	1
10	10	10	10	0.5	1	1
11	11	11	11	0.55	1	1
12	12	12	12	0.6	1	1
13	13	13	13	0.65	1	1
14	14	14	14	0.7	1	1
15	15	15	15	0.75	1	1
16	16	16	16	0.8	1	1
17	17	17	17	0.85	1	1
18	18	18	18	0.9	1	1
19	19	19	19	0.95	1	1

minimal value of -1, for $k_i + k_j = n, k_i = q$, $k_j = n - q$, $k_i < k_j$ or vice versa, with $r = 0$ and q has the range $q = 1, 2, 3, \ldots n/2$. For other cases, when $k_i + k_j < n$ and $r = 0$, both Nogueira's measure and Wald's measure have the same value between -1 and 0. For minimal stability condition, Lustgarten's measure provides a value between -1 to 0, but never reaches -1.

From the results of the above toy experiments it can be stated that, Lustgarten's stability measure provides more systematic results than other extended version of Kuncheva index except for two conditions, one is when the two selected feature subsets are identical or stability value should be a fixed maximum value of $+1$ and another is when intersection between the feature subsets is zero or the stability value should be a fixed minimum value of -1. While the Lustgarten's stability values in these two cases are not appropriate, the values are bounded by finite numbers. In the next section we propose corrections to the Lustgarten's measure to make it appropriate for the conditions of maximal and minimal stability. The detail proposal is described in the next section.

Table 4.4: Similarity values for the case when$S_i \cap S_j$ is null ($r = 0$)

Index of feature subset pair	**Cardinality of one feature subset k_i**	**Cardinality of other feature subset in k_j**	**Cardinality of intersection of feature subsets, r**	**Lustgarten's measure,** $SI_L(S_i, S_j)$	**Nogueira's measure,** $SI_N(S_i, S_j)$	**Wald's measure,** $SI_W(S_i, S_j)$
1	19	1	0	-0.95	-1	-19
2	18	2	0	-0.9	-1	-9
3	17	3	0	-0.85	-1	-5.67
4	16	4	0	-0.8	-1	-4
5	15	5	0	-0.75	-1	-3
6	14	6	0	-0.7	-1	-2.33
7	13	7	0	-0.65	-1	-1.86
8	12	8	0	-0.6	-1	-1.5
9	11	9	0	-0.55	-1	-1.22
10	10	10	0	-0.5	-1	-1
11	9	9	0	-0.45	-0.82	-0.82
12	8	7	0	-0.4	-0.67	-0.67
13	7	7	0	-0.35	-0.54	-0.54
14	6	6	0	-0.3	-0.43	-0.43
15	5	4	0	-0.25	-0.33	-0.33
16	4	4	0	-0.2	-0.25	-0.25
17	3	2	0	-0.15	-0.18	-0.18
18	2	2	0	-0.1	-0.11	-0.11
19	1	1	0	-0.05	-0.05	-0.05

4.5 Proposed Correction of Lustgarten's Measure

The main shortcomings of Lustgarten's measure are that it cannot reach its maximum value of +1, when the feature subsets are identical and similarly cannot reach its minimum value of −1 when the cardinality of intersection between feature subsets is zero. Lustgarten's measure possesses all the desired properties except the property of maximum-minimum value. Here we have proposed corrections to remove the drawbacks.

4.5.1 Proposed Correction Value for Different Conditions

Different possible cases are considered for correction and are stated below:

1. The correction for maximum value:

   The maximum similarity value for the stability measure should occur when the

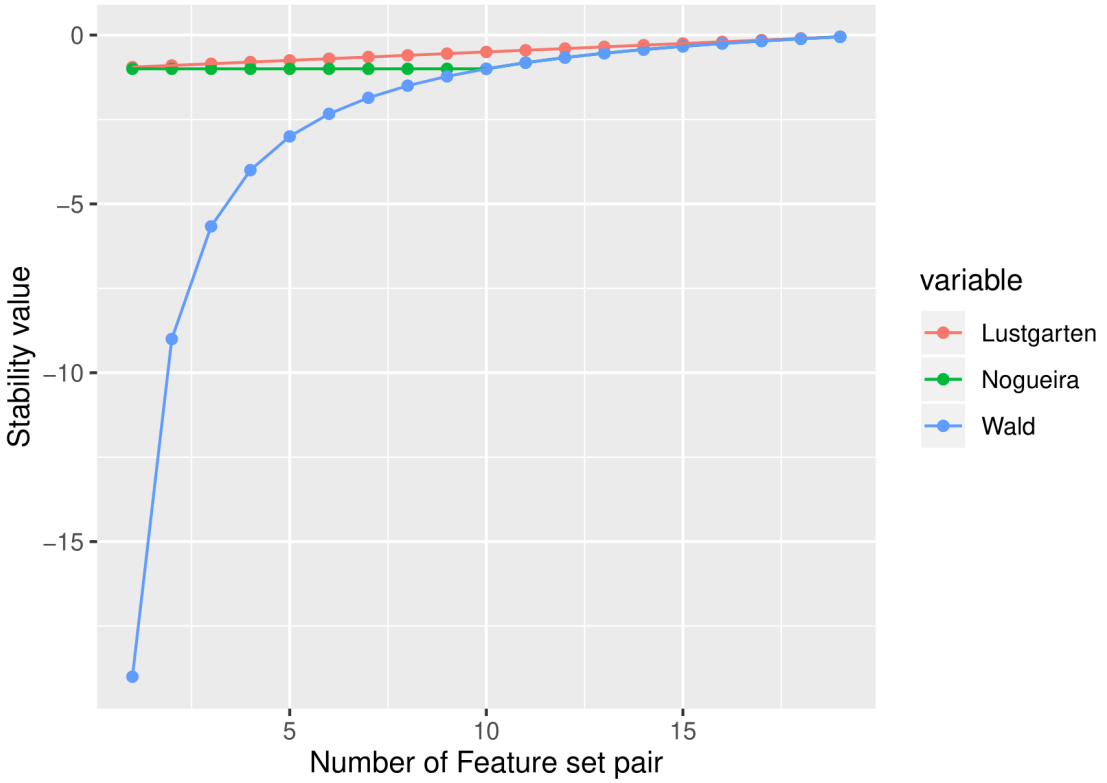


Figure 4.3: Similarity values when the intersection of the feature subsets is null.

two feature sets are identical, i.e., $k_i = k_j = r$. In this case, Kuncheva index and other stability measures provide the maximum value of +1, but Lustgarten's measure provides different values which are less than +1, depending on the cardinality of the selected feature subsets. In this work, we propose the correction of the measure based on three different cases for the cardinality of r.

- Case 1: When $0 < r < n/2$

  The Lustgarten's measure for the feature subsets in this case can be written as:

  $SI_L(S_i, S_j) = \frac{r - \frac{k_i k_j}{n}}{min(k_i,k_j) - max(0, k_i + k_j - n)} = \frac{r - \frac{r*r}{n}}{r - 0} = 1 - \frac{r}{n}.$

  where n is the number of all features.

  $Correction\ value = Ideal\ value – Lustgarten's\ measure = 1 - (1 - \frac{r}{n}) = \frac{r}{n}$

- Case 2: When $r = n/2$

  In this case, Lustgarten's measure can be written as:

  $SI_L(S_i, S_j) = \frac{r - \frac{k_i k_j}{n}}{min(k_i,k_j) - max(0, k_i + k_j - n)} = \frac{r - \frac{n/2 n/2}{n}}{r - 0} = \frac{n/2 - n/4}{n/2} = \frac{1}{2}.$

where n is the number of all features.

$Correction\ value = Ideal\ value – Lustgarten's\ measure = 1 - \frac{1}{2} = \frac{1}{2}$

- Case 3: When $n/2 < r < n$

  For this case, Lustgarten's measure can be written as:
  $SI_L(S_i, S_j) = \frac{r - \frac{k_i k_j}{n}}{min(k_i,k_j) - max(0, k_i + k_j - n)} = \frac{r - \frac{r*r}{n}}{r - (k_i + k_j - n)} = \frac{r(n-r)}{n(n-r)} = \frac{r}{n}$.
  where n is the number of all features.

  $Correction\ value = Ideal\ value – Lustgarten's\ measure = 1 - \frac{r}{n} = \frac{n-r}{n}$

2. The correction for minimum value:

In this case, selected feature subsets have no common feature, i.e., $r = 0$. In this condition, Kuncheva index and some other extension of Kuncheva index should provide the minimum value of -1. However, for the Kuncheva index and Wald's measure, this is satisfied only when $k_i = k_j = k = n/2$. We assessed the correction for the other cases of cardinalities of k_i and k_j as follows:

- Case 1: When $k_i + k_j = n$.

  In this case, let us consider, $k_i = n - p$ and $k_j = p$, or vice versa, where $p = 1, 2, 3...n/2$ Lustgarten's measure can be written as:
  $SI_L(S_i, S_j) = \frac{r - \frac{k_i k_j}{n}}{min(k_i,k_j) - max(0, k_i + k_j - n)} = \frac{0 - \frac{p(n-p)}{n}}{p - 0} = -\frac{(n-p)}{n} = -\frac{max(k_i,k_j)}{n}$
  $Correction\ value = Ideal\ value – Lustgarten's\ measure$
  $= -1 - (-\frac{max(k_i,k_j)}{n}) = \frac{max(k_i,k_j)}{n} - 1$

- Case 2: When $k_i + k_j < n$.

  In this case, let us consider $k_i > k_j$, or vice versa, Lustgarten's measure can be written as: $SI_L(S_i, S_j) = \frac{r - \frac{k_i k_j}{n}}{min(k_i,k_j) - max(0, k_i + k_j - n)} = \frac{0 - \frac{k_i k_j}{n}}{k_j - 0} = -\frac{k_i}{n} = -\frac{max(k_i,k_j)}{n}$
  $Correction\ value = Ideal\ value – Lustgarten's\ measure$
  $= -1 - (-\frac{max(k_i,k_j)}{n}) = \frac{max(k_i,k_j)}{n} - 1$

In all the cases, correction value for the condition $r = 0$ is same.

4.5.2 Proposed Corrected Lustgarten's Measure

Based on the above analysis, here we summarize our newly proposed corrected Lustgarten's measure $SI_{Lnew}(S_i, S_j)$ in Equation (4.12) for defining similarity between two selected feature subsets S_i and S_j having cardinalities k_i and k_j, respectively, while r, n being the cardinality of intersection of the selected feature subsets and total number of features. In Equation (4.12), $r = k_i = k_j$, when r is defined within the range $0 < r < n$.

$$SI_{Lnew}(S_i, S_j) = \begin{cases} \frac{r - \frac{k_i k_j}{n}}{min(k_i,k_j) - max(0, k_i + k_j - n)} + \frac{r}{n}, & \text{if } 0 < r < n/2 \\ \frac{r - \frac{k_i k_j}{n}}{min(k_i,k_j) - max(0, k_i + k_j - n)} + \frac{1}{2}, & \text{if } r = n/2 \\ \frac{r - \frac{k_i k_j}{n}}{min(k_i,k_j) - max(0, k_i + k_j - n)} + \frac{n-r}{n}, & \text{if } n/2 < r < n \\ \frac{r - \frac{k_i k_j}{n}}{min(k_i,k_j) - max(0, k_i + k_j - n)} + \frac{max(k_i,k_j)}{n} - 1, & \text{if } r = 0 \\ 0, & \text{if } r = n \end{cases} \tag{4.12}$$

4.5.3 Toy Experiment for Verification

An experimental illustration have been done, similar to our experiments in the previous section, for verification of the proposed corrected Lustgarten's measure. As before, the total number of features is $n = 20$. Selected feature subset pairs of different cardinalities are considered. Various measures along with our proposed corrected Lustgarten's measures are used for stability measurement. The results are shown in Table 4.5.

In Table 4.5, the 1st to 6th feature subset pairs are formed in such way that each pair have identical feature subsets, i.e., $k_i = k_j = r$. Corrected Lustgarten's measure, Nogueira's and Wald's measure produce the correct maximum value. For the 7th to 12th feature subset pairs, the intersection between feature subsets for each pair is zero, so the stability should be minimum, with value of -1. Corrected Lustgarten's measure only provides this minimum stability value for all the feature subset pairs (7th to 12th). For the last eight feature subset pairs (13th to 20th), one

Table 4.5: Comparison of stability measures with proposed corrected Lustgarten's measure

Index of feature subset pair	**Cardinality of one feature subset k_i**	**Cardinality of another feature subset k_j**	**Cardinality of intersection of feature subsets r**	**Lustgarten's measure, $SI_L(S_i, S_j)$**	**Correction value in Lustgarten's measure**	**Corrected Lustgarten's measure $SI_{Lnew}(S_i, S_j)$**	**Nogueira's measure, $SI_N(S_i, S_j)$**	**Wald's measure, $SI_W(S_i, S_j)$**
1	1	1	1	0.95	0.05	1	1	1
2	2	2	2	0.90	0.10	1	1	1
3	10	10	10	0.50	0.50	1	1	1
4	19	19	19	0.95	0.05	1	1	1
5	12	12	12	0.60	0.40	1	1	1
6	7	7	7	0.65	0.35	1	1	1
7	19	1	0	-0.95	-0.05	-1	-1	-19
8	15	5	0	-0.75	-0.25	-1	-1	-3
9	10	10	0	-0.50	-0.50	-1	-1	-1
10	5	4	0	-0.25	-0.75	-1	-0.33	-0.33
11	3	2	0	-0.15	-0.85	-1	-0.18	-0.18
12	1	1	0	-0.05	-0.95	-1	-0.05	-0.05
13	18	1	1	0.10	0	0.10	0.11	1
14	10	5	5	0.50	0	0.50	1	1
15	4	12	4	0.40	0	0.40	0.67	1
16	14	3	3	0.30	0	0.30	0.42	1
17	1	10	1	0.50	0	0.50	1	1
18	3	11	3	0.45	0	0.45	0.81	1
19	15	17	15	0.75	0	0.75	1	1
20	9	14	9	0.45	0	0.45	0.81	1

feature subset is proper subset of the other feature subset i.e., the two subsets are not identical. Wald's measure gives a stability value of +1. Nogueira's measure also gives the stability value of +1 for 3 cases, and less than 1 for rest of the five cases. Lustgarten's measure produces value less than 1 for all the cases which is more appropriate than Nogueira's measure or Wald's measure. It can be verified that corrected Lustgarten's measure can produce appropriate values in all the different possible cases.

4.6 Experiments with Benchmark Data Sets

To conduct the experiment, we have collected fifteen benchmark data sets which are taken from UCI [64]. Table 4.6 summarises the data sets, including the data set name, the total number of features, the total number of instances and the total number of classes. Among the 15 data sets, nine are binary-class, and the rest are multi-class. Some data sets need to be pre-processed due to their categorical nature or having missing values. We converted categorical type features into numeric types. We also replaced all missing values of numeric features with the average from the data.

Table 4.6: Dataset Description

Datasets	Total no. of features	Total no. of instances	Total no. of classes
dna	180	3186	3
iris	4	150	3
Page-blocks	10	5473	5
Pen-digits	16	10992	10
splice	60	3190	3
Waveform-5000	40	5000	3
Banknote	4	1372	2
Climate-model	18	540	2
Cryotherapy	6	90	2
diabetes	8	768	2
heart	13	270	2
Japanese-vowels	12	9961	2
Prostate-cancer	8	100	2
apndcts	7	106	2
sonar	60	208	2

4.6.1 Experimental Process

In the simulation experiment, we first generate M perturbed sample sets of a given data set using a re-sampling technique. Afterwards, we apply a feature selection algorithm on each of the M sample sets, obtaining a feature subset S_i for i_{th} sample set. Therefore, we get M feature subsets $\phi = \{S_1, S_2, ..., S_M\}$. The feature selection algorithm that we have used is the filter subset based feature selection named correlation-based feature selection (CFS) [17]. A stability measure algorithm then takes feature subsets $\phi = \{S_1, S_2, ..., S_M\}$ as input and calculate the stability among different feature subset pairs. If the total number of feature subsets is M, then there are $\frac{M(M-1)}{2}$ possible pairs of feature subsets that are used to calculate the stability. In this experiment, we choose the value of M to 10 and therefore, we have 45 feature subset pairs. After that we have calculated the stability with Lustgarten measure, Wald measure, Nogueira measure and proposed corrected Lustgarten measure and compared the results for all datasets. We also examined whether our proposed corrected Lustgarten measure satisfies the properties that other modified Kuncheva index-based approaches can not.

4.6.2 Results and Discussion

Table 4.7 represents the feature selection result of *apndcts* dataset for 10 feature subsets using CFS based feature selection algorithm. This table shows that 'Feature Subset 1' and 'Feature Subset 10' are identical. Similarly, 'Feature Subset 6' and 'Feature Subset 9' are also identical. Other feature subsets are not identical to one another. In this table, the cardinality of the each feature subset is also presented.

Table 4.7: Ten Selected feature subsets of *apndcts* dataset

Feature Subset	**Selected feature subset**	**Cardinality of selected feature subset**
Feature Subset 1	At3, At2, At5, At7	4
Feature Subset 2	At1, At2, At6, At5, At3	5
Feature Subset 3	At3, At1, At2, At7, At5	5
Feature Subset 4	At3, At2, At1, At7	4
Feature Subset 5	At3, At7, At2	3
Feature Subset 6	At1, At7, At6, At3	4
Feature Subset 7	At3, At1, At6	3
Feature Subset 8	At1, At7, At5, At6, At3	5
Feature Subset 9	At3, At7, At1, At6	4
Feature Subset 10	At3, At5, At2, At7	4

It is required to identify the type of feature subset pair to explain the stability value clearly. Table 4.8 shows the pair of feature subsets matrix for *apndcts* data set that compares all the feature subset pairs. It is observed that with ten feature subsets, ${}_{10}C_2$ or 45 feature subset pairs can be constructed. In the table, each identical feature subset pair is denoted as 'I', each feature subset pair with proper subset is denoted as 'P', and each feature subset pair of neither proper nor identical is denoted as 'N'. After counting each type of subset pair for this data set, we get P for 13 cases, I for two cases, and N for 30 cases.

After repeating the same procedure for the rest of the data sets, we have summarized the results in Table 4.9. The table represents the types of feature subset pairs for 15 data sets using a CFS-based feature selection algorithm. Among 15 data sets, five data sets such as *dna*, *Page-blocks*, *Banknote*, *Climate-model*, *Japanese-vowels* give 45 identical feature subset pairs out of 45 feature subset pairs, indicating all feature subsets in the different samples are the same. Therefore, regarding these five data sets, the stability results must be maximum value of 1. As not all feature

Table 4.8: Pair of feature subsets matrix for *apndcts* dataset

Feature Subset	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1		N	P	N	P	N	N	N	N	I
2			N	N	N	N	P	N	N	N
3				P	P	N	N	N	N	P
4					P	N	N	N	N	N
5						N	N	N	N	P
6							P	P	I	N
7								P	P	N
8									P	N
9										N
10										

Table 4.9: Types of feature subset pair obtained for 15 datasets

Datasets	**No. of Feature subsets**	**No. of total feature subset pair**	**No. of feature subset pair with proper subset, P**	**No. of identical feature subset pair, I**	**No. of neither proper nor identical feature subset pair, N**
dna	10	45	0	45	0
iris	10	45	9	36	0
Page-blocks	10	45	0	45	0
Pen-digits	10	45	16	29	0
splice	10	45	16	29	0
Waveform-5000	10	45	26	19	0
Banknote	10	45	0	45	0
Climate-model	10	45	0	45	0
Cryotherapy	10	45	23	22	0
diabetes	10	45	9	36	0
heart	10	45	36	9	0
Japanese-vowels	10	45	0	45	0
Prostate-cancer	10	45	21	24	0
apndcts	10	45	13	2	30
sonar	10	45	1	0	44

Table 4.10: Comparison of four stability measures

Datasets	**Lustgarten measure**	**Nogueira measure**	**Wald measure**	**Corrected Lustgarten (proposed)**
dna	0.9889	1	1	1
iris	0.7500	1	1	0.9694
Page-blocks	0.8500	1	1	1
Pen-digits	0.8944	1	1	0.9778
splice	0.9468	1	1	0.9826
Waveform-5000	0.8158	1	1	0.9325
Banknote	0.7500	1	1	1
Climate-model	0.9722	1	1	1
Cryotherapy	0.7574	1	1	0.9315
diabetes	0.8125	1	1	0.9681
heart	0.7556	0.9416	1	0.8479
Japanese-vowels	0.8750	1	1	1
Prostate-cancer	0.8417	1	1	0.9486
apndcts	0.5984	0.6697	0.6893	0.6460
sonar	0.7378	0.8280	0.8280	0.7378

subset pairs for the rest of the data sets are identical, stability results must be less than 1. For *sonar* data set, there is no identical feature subset pair. Only one feature set pair has a proper subset, and the rest of the 44 feature subset pairs are neither proper nor identical. Only the data set *apndcts* has three different types of feature subset pairs: two identical feature subset pairs, 13 feature subset pairs with proper subset, and 30 neither proper nor identical feature subset pairs. Due to the variation of types of different feature subset pairs, the stability results of CFS based feature selection algorithm is different.

Table 4.10 highlights the overall comparison among the four different stability measures in calculating the stability of the feature selection algorithm. Results show that Wald's measure provides the maximum stability value of 1 for 13 data sets, and Nogueira's measure provides the maximum stability value of 1 for 12 data sets. However, not a single case, Lustgarten measure provides the maximum stability value. It is observed that our corrected Lustgarten measure yields maximum stability value for five cases. It was shown in the previous table (Table 4.9) that only five data sets give identical feature subset pairs. As a result, the stability value should be the maximum of 1 for these five data sets, not for other data sets. On

the other hand, Lustgarten measure does not provide the maximum stability value of 1 for these five data sets, which should be the maximum. Although Wald's and Nogueira's measures yield the maximum stability value of 1 for these five data sets, they both provide a maximum of 1 for other data sets. Instead, the value should be less than 1 (seven cases for Nogueira, and eight cases for Wald). This is because, for these data sets, there are proper subset pairs along with identical feature subset pairs. In other words, Wald's measure and Nogueira's measure give incorrect results for some data sets when feature subset pairs have proper subset pair, i.e. $S_1 \subset S_2$ or vice versa.

One common limitation of Wald's measure and Nogueira's measure is that for a feature subset pair, if one feature subset is a proper subset of another, the stability value is maximum. Unlike Wald's measure, Nogueira's measure does not provide the maximum value for all the cases whenever the condition of proper subset occurs. For this reason, Wald's measure gives the maximum stability value of 1 for *heart* data set, but in that data set, Nogueira's measure gives the stability value 0.9436. Therefore, it can be said that our proposed corrected Lustgarten's measure gives more appropriate stability results.

4.7 Conclusion

In this chapter, at first, we have investigated Kuncheva index and its modifications and extensions meticulously, highlighting their merits and limitations. To overcome the shortcoming of Kuncheva index, several modifications and extensions of Kuncheva index are proposed by different researchers like Lustgarten's measure, Wald's measure, nPOG and the most recent Nogueira's measure. We have summarized the required properties of a stability measure and examined whether these are satisfied by the existing popular measures. After that we have proposed a new modified measure based on the correction of Lustgarten's stability measure. It is found by toy experiments that, with the proposed new correction, corrected Lustgarten's measure can overcome the limitations of the other measures and satisfy all the tabulated properties. After conducting the toy experiments, we have conducted

another experiment with using fifteen benchmark datasets. This experiment also verifies the limitations of Wald's measure, Nogueira's measure and Lustgarten measure. In addition, this study reveals that the proposed corrected Lustgarten measure overcomes the limitations of Lustgarten measure and gives better stability values than other modified Kuncheva indices.The error in Lustgarten's stability measure is found to be very specific and systematic compared to erratic behaviour of other extensions of Kuncheva index like Wald's measure or Nogueira's measure. So we attempted to correct Lustgarten's measure to define the new proposed measure and could be able to achieve a new measure which produces consistent values.

Chapter 5

Jeffries-Matusita (JM) distance based Feature Selection

5.1 Introduction

For binary classification problems the class labels divide each of the features into two distributions and hence measures like Bhattacharya Distance can be used to measure the separability between these feature class distributions. For binary classification problems Bhattacharya Distance can also be used for feature ranking. Jeffries-Matusita (JM) distance is an improvement over Bhattacharya Distance, which standardized the distance between 0 to 2 for an easy comparison across datasets. As per literature study, though Bhattacharya Distance has been used for feature ranking [82], there does not seem to be any research effort where JM distance is used for feature ranking.

In this chapter, at first JM distance has been used as a feature ranking measure for binary classification problem over 24 publicly available datasets. The results have been compared with three popular ranking based measures including Information Gain, Relief and Chi-Squared. An analysis of the JM value with the classification accuracy has been done, to understand if such analysis reveals any more information over and above the selection of the features. After that, an efficient

feature subset selection algorithm for multiclass problems based on JM distance has been proposed. The proposed approach consists of two steps. In the first step, similar to existing JM distance based feature selection approaches, features are ranked according to the JM distances for all class pairs and multiple ranked feature lists are created corresponding to each class pair. The second step is the novel contribution of this work in which a heuristic approach is developed to select the final optimum feature subset from the multiple ranked feature lists (corresponding to each class pair) based on the average JM distance values of the top ranking features of each class pair. Unlike traditional approaches, the proposed algorithm does not use any explicit search mechanism to find out the optimum feature subset. The proposed algorithm has been evaluated by comparing with other multiclass JM distance based feature selection as well as with some other popular filter based ranking methods by simulation experiment with benchmark data sets.

5.2 JM Distance based Feature Selection Algorithm for Binary Classification

In this section, a simulation experiment has been done where JM distance can be used as a filter ranked based feature selection algorithm for binary class problems. Working process and simulation results have been shown in the following.

5.2.1 Material and Methods

In this section, different details of the empirical study that we conducted are furnished which can be used to reproduce the results. Datasets are taken from the public UCI data repository [63]. Dataset description is shown in Table 5.1.

- 'R' has been as the computational environment [83].
- Naive Bayes have been used as the machine learning classifier.

- 'R' Packages SpatialEco and FSelector have been used for the calculation of JM Distance, Information Gain (IG), Chi-Square (CS) and Relief.
- Default parameters have been used for computation of relief.
- 80% of each the datasets has been taken as training and 20% as testing. This has been repeated 10 times with different seeds and the average value has been reported.
- The classification accuracies have been computed for 10%, 25%, 50% and 75% for each of the methods and the maximum of them have been reported.

Table 5.1: Description of Data sets

Datasets	**No. of features**	**No. of classes**	**No. of Instances**
Sonar	61	2	208
Ion (Ionosphere)	34	2	351
Bupa (Liver Disorders)	7	2	345
Heart	14	2	270
Biodeg (QSAR biodegradation)	41	2	1055
Apndcts (Appendicitis)	8	2	106
Mcg (MAGIC Gamma Telescope)	11	2	19020
Twonorm	21	2	7400
Best cancer (Breast Cancer Wisconsin)	32	2	569
Diabetes (Pima Indians Diabetes)	9	2	768
Prostate Cancer	10	2	100
Lung Cancer	7	2	59
Cryotherapy	7	2	90
Fertility diagnosis	10	2	100
Indian Liver Patient dataset (ILPD)	10	2	583
Banknote authentication	5	2	1372
Faults (Steel Plates Faults)	27	2	1941
kc2 (KC2 Software defect prediction)	22	2	522
Phoneme	5	2	5404
pc1 (PC1 Software defect prediction)	23	2	1109
Climate model	21	2	540
SPECTF	45	2	349
Satellite	37	2	5100
Japanese Vowels	12	2	640

5.2.2 Results and Analysis

The classification accuracies of the four methods JM distance, IG, CS and Relief for feature selection algorithms are compared. These methods are also compared in terms of execution time. Finally, an investigation has been performed with the top 10% JM values of the datasets with the classification accuracy, to understand if these values give any general idea of the separability of the classes in the dataset.

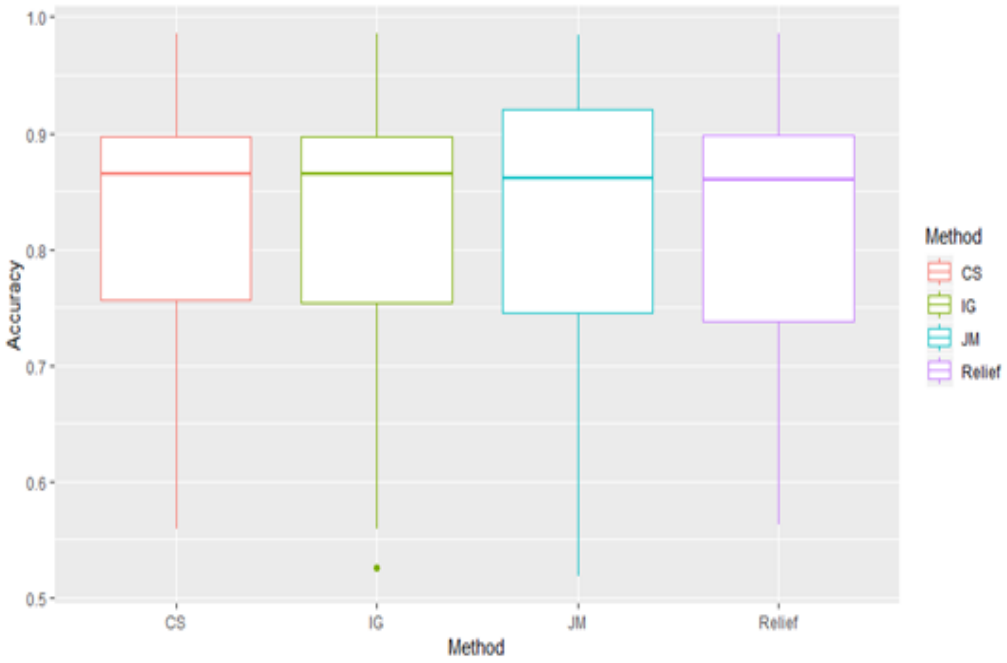


Figure 5.1: Classification Accuracy of all the methods

A. Comparison on classification accuracy

In Table 5.2, the maximum classification accuracy of these methods is reported. It can be observed from the Table 5.2 that

- In terms of maximum wins, JM distance is 3rd followed CS and IG.

  A more detailed comparison is enclosed in Figure 5.1 below. From Figure 5.1, it can be concluded that

- All the methods are quite comparable, though JM distance does not win in terms of number of datasets, the range of classification accuracies produced by JM distance is very much comparable with all other methods.

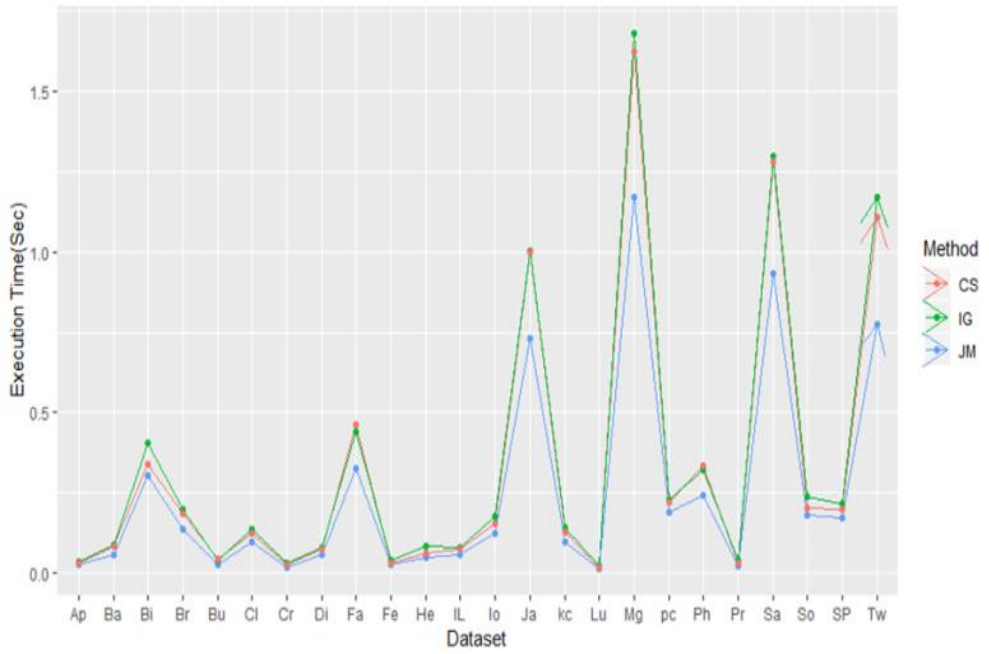


Figure 5.2: Comparison of JM, CS and IG on the basis of execution time

B. Comparison on execution time

In this section, analysis of the comparison on the basis of execution time has been enclosed. This is enclosed in the below Table 5.3.

From Table 5.3, it can be observed that

- Relief is the most expensive of all the four.
- IG, JM distance and CS are comparable.

The comparison of these three methods has been shown in Figure 5.2. It can be observed that JM distance which is the blue line consistently takes the lowest time as compared to other datasets.

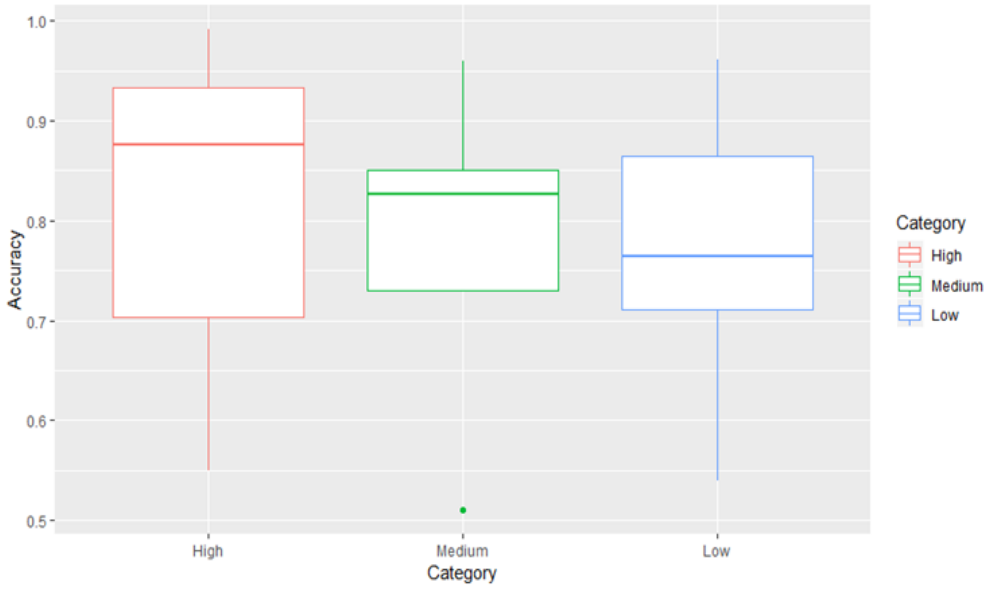


Figure 5.3: Comparison of top JM distance values across datasets

C. Comparison of classification accuracy and top JM values of the datasets

An investigation on the classification accuracies achieved on the datasets has been done with the average of top 10% JM values. The comparison has been demonstrated in the Figure 5.3 below. The datasets having a value greater than one have been marked as 'high', the ones having a value greater than 0.5 and less than 1 have been marked as medium and ones having less than 0.5 has been marked as 'low' as far as class separability is concerned.

As observed from Figure 5.3

- The datasets marked as high have displayed higher classification accuracy on average as compared to Medium.
- Similarly, datasets marked as medium showed higher classification accuracy on average as compared to Low.
- The maximum values for 'high' marked datasets are also much more than the 'Medium' class.

Table 5.2: Comparison of Classification Accuracy

Datasets	**JMD**	**IG**	**Relief**	**CS**
Sonar	0.1813	0.2402	24.2600	0.2052
Ion	0.1239	0.1782	15.3400	0.1548
Bupa	0.0286	0.0412	3.3550	0.0434
Heart	0.0492	0.0840	4.7760	0.0619
Biodeg	0.3029	0.4050	52.1800	0.3405
Apndcts	0.0262	0.0348	1.4180	0.0328
Mgc	1.1680	1.6800	314.6800	1.6210
Twonorm	0.7747	1.1700	212.5200	1.1070
Brest cancer	0.1382	0.1989	23.1900	0.1878
Diabetes	0.0571	0.0822	9.1740	0.0745
Prostate Cancer	0.0246	0.0408	1.4640	0.0318
Lung Cancer	0.0145	0.0241	0.5226	0.0177
Cryotherapy	0.0196	0.0304	1.0064	0.0275
Fertility Diagnosis	0.0270	0.0389	1.5300	0.0323
ILPD	0.0582	0.0804	10.9800	0.0757
Banknote authentication	0.0585	0.0894	9.5260	0.0854
Faults	0.3262	0.4407	66.1800	0.4637
kc2	0.0969	0.1419	13.6200	0.1299
Phoneme	0.2430	0.3236	41.3000	0.3334
pc1	0.1891	0.2307	28.5200	0.2219
Climate Model	0.0991	0.1383	12.4800	0.1257
SPECTF	0.1711	0.2152	21.7900	0.1994
Satellite	0.9317	1.2960	237.9000	1.2800
Japanese Vowels	0.7298	1.0048	158.9100	0.9991

- Similarly, the minimum values of the 'low' marked datasets are much less than the 'Medium' marked datasets.

5.3 JM Distance based Feature Subset Selection Approach for Multiclass Problems

In the previous section JM distance was used as a feature selection tool only for binary class problem. In this section, we have worked on the JM distance for multiclass problem and an efficient feature subset selection algorithm for multiclass problems based on JM distance has been proposed. Until recently, JM based approaches that have been proposed for feature selection in multiclass problems are all univariate

Table 5.3: Comparison of Execution Time for all data sets

Datasets	**JMD**	**IG**	**Relief**	**CS**
Sonar	0.1813	0.2402	24.2600	0.2052
Ion	0.1239	0.1782	15.3400	0.1548
Bupa	0.0286	0.0412	3.3550	0.0434
Heart	0.0492	0.0840	4.7760	0.0619
Biodeg	0.3029	0.4050	52.1800	0.3405
Apndcts	0.0262	0.0348	1.4180	0.0328
Mgc	1.1680	1.6800	314.6800	1.6210
Twonorm	0.7747	1.1700	212.5200	1.1070
Brest cancer	0.1382	0.1989	23.1900	0.1878
Diabetes	0.0571	0.0822	9.1740	0.0745
Prostate Cancer	0.0246	0.0408	1.4640	0.0318
Lung Cancer	0.0145	0.0241	0.5226	0.0177
Cryotherapy	0.0196	0.0304	1.0064	0.0275
Fertility Diagnosis	0.0270	0.0389	1.5300	0.0323
ILPD	0.0582	0.0804	10.9800	0.0757
Banknote authentication	0.0585	0.0894	9.5260	0.0854
Faults	0.3262	0.4407	66.1800	0.4637
kc2	0.0969	0.1419	13.6200	0.1299
Phoneme	0.2430	0.3236	41.3000	0.3334
pc1	0.1891	0.2307	28.5200	0.2219
Climate Model	0.0991	0.1383	12.4800	0.1257
SPECTF	0.1711	0.2152	21.7900	0.1994
Satellite	0.9317	1.2960	237.9000	1.2800
Japanese Vowels	0.7298	1.0048	158.9100	0.9991

feature ranking strategy in which average JM distance for all the class pairs are used for final feature ranking whereas we propose here a heuristic approach to select the final optimal feature subset. This proposed algorithm has been evaluated by comparing with other multiclass JM distance as well as with some other popular filter based ranking methods by simulation experiment with benchmark data sets.

5.3.1 JM Distance Extensions for Multiclass Problems

An extended definition of JM distance for the multiclass problem, the weighted average JM_{ave}, has been reported in [84]. For m number of classes, it is defined according to [84] as:

$$JM_{ave} = \frac{1}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} JM_{ij} \tag{5.1}$$

In [58], the authors proposed another multiclass extension of JM distance, JM_{Bh}, which is an equivalent of Bhattacharyya bound to Bayes error, and justified its efficiency over JM_{ave} in feature selection for multiclass problems by simulation experiments. For m class problem, JM_{Bh} is defined as:

$$JM_{Bh} = \sum_{i=1}^{m} \sum_{j \geq i}^{m} JM_{ij}^2 \tag{5.2}$$

5.3.2 Proposed Feature Selection Approach with JM Distance for Multiclass Problems (JM_{mc})

In this part, a novel optimum feature subset selection approach JM_{mc}, for a multiclass problem based on the feature evaluation by JM measure is proposed, which is described in detail in this section. Traditionally optimum feature subset selection approaches require a measure for evaluating feature or feature subset and a search strategy for finding out the best feature subset from possible feature subsets. The proposed approach is composed of two steps. In the first step (algorithm 1), similar to existing JM distance based feature selection approaches, features are ranked according to the JM distances for all class pairs and multiple ranked feature lists are created corresponding to each class pair. Other multiclass JM distance based approaches for feature selection use some kind of average JM measure, averaged over all class pairs, which is described in the previous section by Equation 5.1 and Equation 5.2, for final feature ranking. In our work, a novel heuristic approach for selecting an optimum feature subset from the multiple lists of ranked features corresponding to each class pair, obtained after the first step, is proposed in the second step (algorithm 2). Moreover, our approach does not include traditional search based methods for final feature subset selection. The heuristics proposed in our work is based on the following core concept. If the average JM values of the top ranking features for a class pair is high, the classes are considered to be well separated and

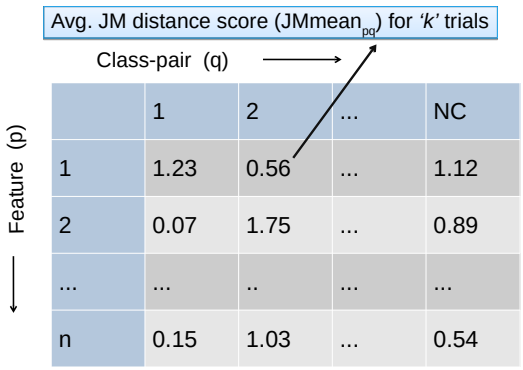


Figure 5.4: Class pair-Feature Table

few number of top ranking features are needed for good classification accuracy. The case is opposite for low average JM values of the top ranking features. The classes are not well separated, and we need to consider features in the final feature subset to provide better classification accuracy. The detailed algorithm is presented below.

- *Algorithm 1*: For any multiclass data set, let the number of features be n and the number of classes be m where $m > 2$. The number of class pairs (NC) will be $\frac{m(m-1)}{2}$. JM distance is calculated for each feature and for each pair of classes according to Eq. (2.9). Each set of calculations are repeated for K times with K different seed values and average value of K trials is taken as the JM distance $JMmean_{pq}$ of qth class pair of pth feature where $p = 1, \cdots, n$ and $q = 1, \cdots, NC = \frac{m(m-1)}{2}$. The procedure is shown in Algorithm 5.1. Figure 5.4 illustrate an example of class pair- feature table after the execution of Algorithm 5.1. In this table, the value of a particular cell indicates a $JMmean_{pq}$ for K trials.

- *Algorithm 2*: In Algorithm 5.2, for qth class pair, features are ranked in descending order according to the values of $JMmean_{pq}$ distance of all the features ($JMmean_{pq}, p = 1, \cdots, n$). This is done for all class pairs. So we have now feature lists (FL_q, where $q = 1, \cdots, NC$) of ranked features corresponding to each class pair. The number of features in all the feature lists are the same as the total number of features n. Now our proposed approach selects the most important feature subset from the (FL_q, where $q = 1, \cdots, NC$) ordered feature lists for the multiclass problem. The underlying concept of the selection of features from the feature lists is presented below.

Algorithm 5.1 Calculation of average JM distance for all class-pairs and features

```
1: procedure JMClassPair(classPair, featureList, K)
2:     JFC_k : JM distances for all class-pairs of all features of the k^th iteration
3:     JMmean : Average JM distance for all class- pairs of all features for K trials
4:     for each q of classPair do
5:         for each p of featureList do
6:             sum← 0
7:             for each k of K do
8:                 sum← sum+JFC_k[p, q]
9:             end for
10:            JMmean[p, q] ← sum/K
11:        end for
12:    end for
13:    return JMmean
14: end procedure
```

- $JMmean_{pq}$ distance is a separability measure, which determines the average class separability of a feature p for a class pair q. The larger $JMmean_{pq}$ distance value of a feature p indicates that the feature can separate the class pair q very well, which implies an increase in classification accuracy. For multiclass problems, $JMmean_{pq}$ distance is calculated for each class pair q and different class pair shows a different value of $JMmean_{pq}$ distance for a single feature p. If for a particular class pair q, the $JMmean_{pq}$ distances of the top ranking features are near to 2 (the upper limit of standard JM distance is theoretically 2), then the features are very strong to easily separate classes. In this case, a lesser number of features can provide good classification accuracy. On the other hand, if the $JMmean_{pq}$ distance values of the top ranking features are low, the features have not good separability, then comparatively more features are required to provide moderate classification accuracy. Based on this concept, different percentages of features are selected from different class pairs for improvement of classification accuracy.

  The final feature subset selection from the ranked feature lists of all class pairs need to be done according to the following rules.

  1. If the top $JMmean_{pq}$ distance value ($FLmax_q$) of the feature list of the qth class pair (FL_q) is greater than 1, then top $\alpha\%$ of features are selected from the feature list of that class pair FL_q . (α has to be predefined).

2. If the top $JMmean_{pq}$ distance value of the feature list ($FLmax_q$) of any class pair is greater than 0.5 but less than or equal to 1, then top $2\alpha\%$ of features are selected from the feature list of that class pair.

3. If the top $JMmean_{pq}$ distance value of the feature list ($FLmax_q$) of the class pair is less than or equal to 0.5 , then top $3\alpha\%$ of features are selected from the feature list of that class pair.

- The selected features from all the class pairs are put in a list (selected feature list SFL). As there is a possibility that a particular feature might be selected more than once, the frequency of occurrence of each feature p in the selected feature list SFL is counted and let it be $|SFL_p|, p = 1, \cdots, n$. Let the median value be SFL_{med}. Now for finding out the final feature subset, depending on the relative number of features and the number of class pairs, two cases are considered.

  1. If the total number of features n in the data set is less than or equal to the total number of class pairs ($n \leq NC$), then the final feature subset will constitute the features (from the selected feature list SFL) whose occurrence frequency is more than or equal to SFL_{med}

  2. If the total number of features is greater than the total number of class pairs ($n > NC$), then all the features in SFL will be selected as the final feature subset.

  The proposed heuristics selection process is presented clearly in Algorithm 5.2.

5.3.3 Simulation Experiment

The implementation of the proposed algorithm has been done with benchmark data set for simulation experiments. The data set description is presented in the next subsection.

Algorithm 5.2 Selection of final feature subset

1: **procedure** JMMULTICLASS($classPair, featureList, JM$)
2: n : len($featureList$)
3: NC : len($classPair$)
4: SFL : candidate features list
5: $SFL \leftarrow$ empty list
6: **for** each q of $classPair$ **do**
7: $FLmax[q] \leftarrow$ find the feature with the highest $JMmean[p, q], p = 1, \cdots, n$
8: $SortedfeatureList \leftarrow$ sort the features descending order of $JMmean[p, q]$, $p = 1, \cdots, n$
9: F_s= empty list
10: **if** $FLmax[q] > 1$ **then**
11: $F_s \leftarrow$ Take α% of top features in $SortedfeatureList$
12: **else if** $FLmax[q] > 0.5$ AND $FLmax[q] <= 1$ **then**
13: $F_s \leftarrow$ Take 2α% of top features in $SortedfeatureList$
14: **else if** FLmax[q] <= 0.5 **then**
15: $F_s \leftarrow$ Take 3α% of top features in $SortedfeatureList$
16: **end if**
17: **for** each Feature of F_s **do**
18: SFL.insert(Feature)
19: **end for**
20: **end for**
21: Freq $\leftarrow$ dictionary for count the frequency of features in SFL
22: **for** each feature of SFL **do**
23: Freq[feature] $\leftarrow$ SFL.count(feature)
24: **end for**
25: **if** n<= NC **then**
26: $SFL_{med} \leftarrow$ find the median value of frequency in Freq
27: selectedFeature $\leftarrow$ all features in Freq >= SFL_{med}
28: **else if** n > NC **then**
29: selectedFeature $\leftarrow$ features in Freq
30: **end if**
31: *return* selectedFeature
32: **end procedure**

A. Data set Description

In this work, 37 data sets are used for performing a simulation experiment to validate the proposed approach. Among them, 25 data sets are collected from UCI repository [64] and rest are collected from OpenML [65]. Some data sets need to be preprocessed, which have missing values or are categorical in nature. Here, categorical type missing values in the data sets are replaced with the most frequently used value, and after that, the whole data set is converted into numeric type. Numeric type missing values are replaced with the average value. Categorical type data sets without missing values are directly converted to numeric type. Table 5.4 represents the summary of data sets which includes the number of features, the number of instances, the number of classes and a short description of each data set.

B. Experimental Method

For the simulation experiment, a 10-fold cross validation method is used. The training set samples are used for feature selection. JM distance for each feature and for each class-pair is calculated according to Eqs. (2.9) and (2.8). The proposed approach in Algorithm 5.2 is used to select the subset of features based on average JM values for each class pair and each feature. The selected feature subset is evaluated by its performance for supervised classification using the Naive Bayes classifier. The same training samples are used for training the classifier, and the test samples are used for measuring classification accuracy, F-measeure and AUC of the classifier of the feature subset. The classification experiment is also repeated 10 times, and average classification accuracy, F-measeure and AUC are taken as the performance measure of the selected feature subset.

In order to set the value of parameter α, a preliminary experiment is conducted with 10 datasets. The value of α is changed from 1 to 20, and in each case, the classification performance of the proposed approach on these datasets is observed. Results show that classification accuracy for most of the datasets is the highest when the α value is near 10. Therefore, the value of α is here fixed as 10.

Table 5.4: Summary of Data sets

Datasets	Feature	Instance	Class
analcat-authorship	70	841	4
analcat-marketing	32	364	5
breast-tissue	9	106	6
Bridges	11	105	6
Cars	7	406	3
Cmc	9	1473	3
Dermatology	34	366	6
Dna	180	3186	3
eye-movements	27	10936	3
gas-drift	128	13910	6
Har	561	10299	6
indian-pines	220	9144	8
Iris	4	150	3
mfeat-factors	216	2000	10
Mfeat-fourier	76	2000	10
mfeat-karhunen	64	2000	10
mfeat-morph	6	2000	10
mfeat-pixel	240	2000	10
mfeat-zernike	47	2000	10
mice-protein	77	1080	8
Page-blocks	10	5473	5
Pasture	21	36	3
Pendigits	16	10992	10
Satimage	36	6430	6
Seeds	7	210	3
Segment	18	2310	7
Splice	60	3190	3
squash-stored	24	52	3
Squash-unstored	23	52	3
Synthetic-control	60	600	6
Teaching-assistant	5	151	3
Vehicle	18	846	4
vertebra-column	6	310	3
Vowel	12	990	11
waveform-5000	40	5000	3
Wine	13	178	3
Wine-quality	11	4898	7

For comparative evaluation of the proposed algorithm, feature subset selection has been done with two other available multiclass extensions of JM distance, weighted average JM distance (JM_{ave}) and (JM_{Bh}) another one equivalent to Bhattacharyya bound according to Eq. (5.1) and Eq. (5.2) respectively and the results

have been compared. The performance of the proposed algorithm also has been compared with other filter based feature ranking algorithms such as IG, GR, SU, CS, One-R and Relief-F by simulation experiment. The classification accuracy, F-measure, and AUC of Naive Bayes Classifier with the selected features by those algorithms, the number of features being same as the number of features selected by the proposed multiclass JM distance based algorithm, are used for performance comparison.

For all simulation experiments, Intel(R) Core(TM) i5-4590 CPU @3.30GHz Processor and 8GB RAM with a 64 bit operating system of Windows 8.1 Pro is used. R (version 3.5.3) is used with several key standard packages such as SpatialEco, FSelector, varSel, MLmetrics, caret, and e1071 for implementation of the algorithms.

5.3.4 Performance Measures for Simulation Experiment

- **Percentage of Selected Features**

  Since our proposed method selects a specific number of features from the full set of feature, the percentage of feature selection is important. The percentage of feature selection is calculated by

$$Selection\,Rate(\%) = \frac{Num.\,of\,Selected\,Features}{Total\,Feature} \times 100 \tag{5.3}$$

- **Classification Accuracy**

  The proposed feature selection approach with multiclass JM distance is evaluated by the classification accuracy as a performance measure which is defined as [85]

$$Accuracy(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{5.4}$$

  Where TP, true positive, TN, true negative, FP, false positive, and FN, false negative, represent the number of positive cases correctly detected, the number of negative cases correctly detected, the number of negative cases detected as positive and the number of positive cases detected as negative respectively.

- **Precision**

  Precision (also called positive predictive value) is a measure of the correctness of a positive prediction. For any classification task, the precision of a class (target value) is defined as the number of true positives divided by the total number of elements labelled as belonging to the positive class. It is calculated by using the following formula:

$$Precision(p) = \frac{TP}{TP + FP} \tag{5.5}$$

- **Recall**

  Recall is the measure of how many true positives get predicted out of all the positive class elements. It is sometimes also called sensitivity. The measure is collected by the following formula:

$$Recall(r) = \frac{TP}{TP + FN} \tag{5.6}$$

- **F-measure**

  F-measure combines precision and recall. It can be defined as the (weighted) harmonic mean of precision and recall by the following equation [86]:

$$F = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.7}$$

- **Area under the ROC Curve (AUC)**

  The area under a receiver operating characteristic (ROC) curve, or AUC, is a single scalar value that calculates the general performance of a binary classifier [87]. The range of AUC is [0.5, 1], where the minimum value indicates that the performance of the classifier is random, and the maximum value indicates that the classifier is perfect with a zero error rate. The AUC is an important measure to evaluate the overall performance of a classifier because its calculation relies on the complete ROC curve, which involves all possible classification thresholds.

Table 5.5: Feature Selection with Proposed Approach (JM_{mc})

Dataset	Total Feature (TF)	Selected Feature (SF)	Selection rate (%)	Selection time (second)
analcat-authorship	70	24	34.29	0.063
analcat-marketing	32	25	78.13	0.063
breast-tissue	9	5	55.56	0.063
Bridges	11	4	36.36	0.078
Cars	7	3	42.86	0.055
Cmc	9	6	66.67	0.062
Dermatology	34	24	70.59	0.063
Dna	180	32	17.78	0.070
eye-movements	27	13	48.15	0.074
gas-drift	128	85	66.41	0.102
Har	561	318	56.68	0.109
indian-pines	220	174	79.10	0.102
Iris	4	2	50.00	0.052
mfeat-factors	216	183	84.72	0.061
mfeat-fourier	76	51	67.11	0.070
mfeat-karhunen	64	45	70.31	0.063
mfeat-morph.	6	4	66.67	0.052
mfeat-pixel	240	185	77.08	0.107
mfeat-zernike	47	43	91.49	0.070
mice-protein	77	56	72.72	0.078
page-blocks	10	7	70.00	0.063
Pasture	21	6	28.57	0.061
Pendigits	16	2	12.50	0.075
Satimage	36	20	55.56	0.070
Seeds	7	3	42.86	0.059
Segment	18	2	11.11	0.055
Splice	60	11	18.33	0.070
squash-stored	24	6	25.00	0.063
squash-unstored	23	7	30.43	0.065
synthetic-control	60	33	55.00	0.077
teaching-assistant	5	4	80.00	0.067
Vehicle	18	12	66.67	0.068
vertebra-column	6	3	50.00	0.066
Vowel	12	7	58.33	0.057
waveform-5000	40	15	37.50	0.073
Wine	13	3	23.08	0.070
wine-quality	11	4	36.37	0.062

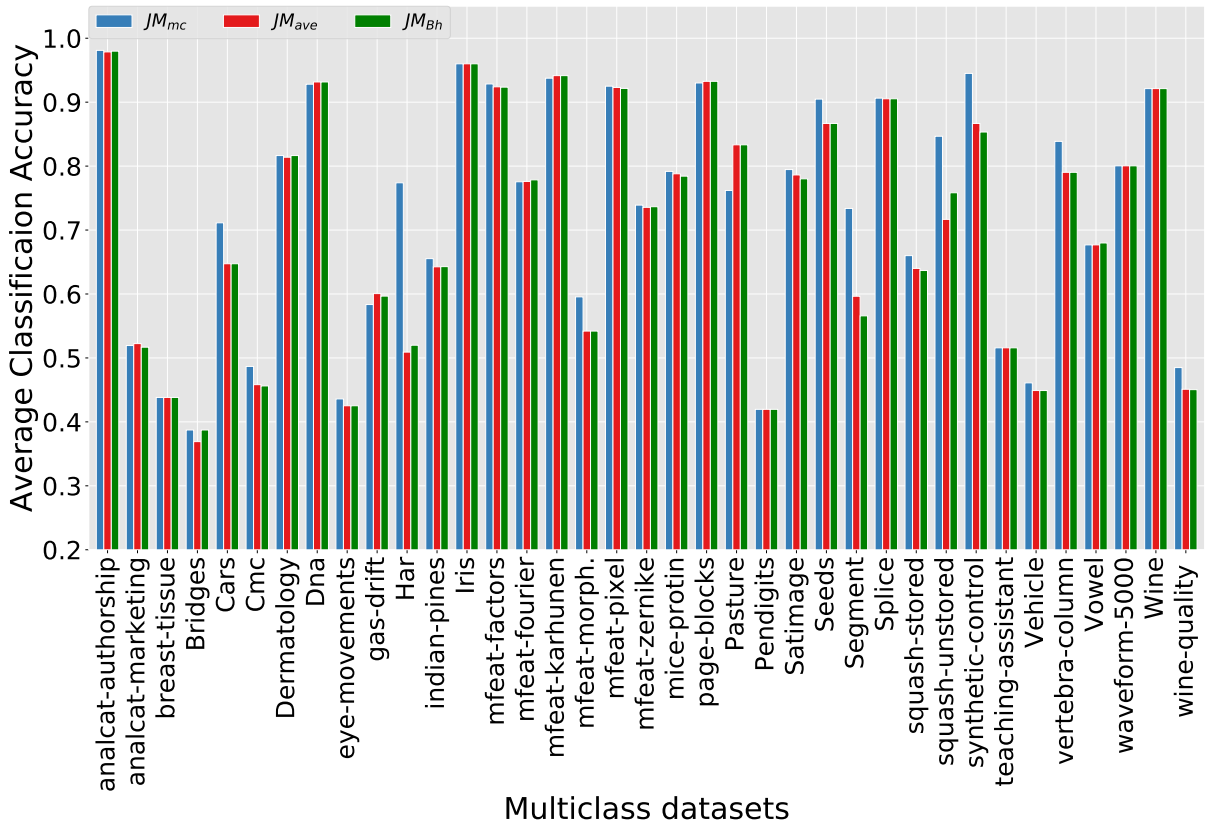


Figure 5.5: Classification Accuracy using various JM measures for all datasets

5.3.5 Simulation Results and Discussion

Table 5.5 represents the selected feature subset with our proposed approach (JM_{mc}) for 37 multiclass datasets. For different datasets, the percentage of selected features is different. For the 'Segment' dataset, the percentage of feature selection is about 11.11%, and it is the lowest among the 37 datasets. For the 'Pendigits' dataset, the percentage is the second lowest and is about 12.50%. Among the 37 datasets, 15 datasets have feature selection rate lower than 50%, and for 13 datasets, the rate is between 50-70% and rest of the 9 datasets, the feature selection rate is more than 70%. Table 5.5 also highlights the feature selection time in seconds for 37 datasets. For all these datasets, time is very short, and the range is about 0.050 to 0.110 seconds.

Figure 5.5 shows the average classification accuracy (Avg) of 37 multiclass datasets using various JM distance measures. From the figure, it is clearly expressed that the classification accuracy of JM distance with our proposed approach (JM_{mc})

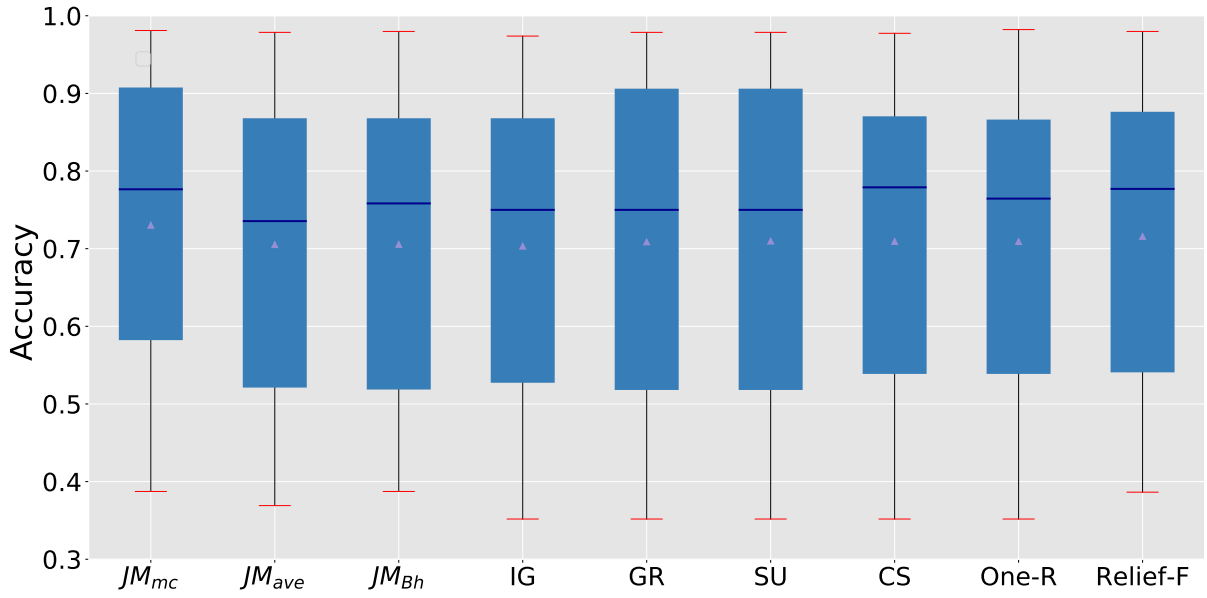


Figure 5.6: Classification performance over all datasets with different methods

is very much comparable to the other two multiclass JM distance measures. For some datasets, our approach performs much better than the other two measures such as JM_{ave} and JM_{Bh}. For almost 22 datasets over 37, classification accuracy of JM_{mc} is higher than JM_{ave} and JM_{Bh}.

Figure 5.6 represents the comparison of classification accuracy on the average of all datasets with nine different measures of ranking based filter approaches. This figure depicts that JM distance with our proposed approach (JM_{mc}) produced the average highest classification accuracy of 73.02% for all the data sets compared to other methods. The weighted average JM distance JM_{ave} and the JM distance equivalent to Bhattacharyya bound JM_{Bh} have classification accuracy 70.53% and 70.55% respectively. Among other feature ranking methods excluding our approach, Relief-F produced the highest value of 71.59%.

Table 5.6 shows the detailed comparison of classification accuracy with the selected feature subset using the proposed feature selection technique with other methods for multiclass datasets. Here, average classification accuracy (Avg) and standard deviation (SD) are calculated for ten iterations. The highest average value and the lowest SD value are represented in boldface in the table. From this table, it is highlighted that proposed JM_{mc} can achieve the highest classification accuracy for 15 data sets; on the other hand, JM_{ave} and JM_{Bh} have the highest accuracy for

four datasets and three datasets respectively. Other feature ranking measures such as IG, GR, SU and CS show the highest accuracy for three datasets, five datasets, four datasets and three datasets respectively. One-R has the highest accuracy for four datasets, and Relief-F shows the highest accuracy for five datasets. In the case of four datasets such as Iris, Pendigits, teaching-assistant and waveform-5000, we got the same average classification accuracy as well as standard deviation (SD) values using all the methods. In the case of SD values, JM_{mc} provides the lowest SD for 12 datasets compared to other approaches. CS possesses very poor SD of classification accuracy, and for only two data sets, the SD value is minimum. The rest of the measures, such as JM_{ave}, JM_{Bh}, IG, GR, SU, One-R and Relief-F have the lowest SD values for three datasets, three datasets, seven datasets, five datasets, ten datasets, four datasets and four datasets respectively. From this result, we can infer that our proposed approach JM_{mc} produces a stable output compared to others.

Table 5.7 shows the comparison of F-measure among all the nine methods for multiclass datasets. Results show that, for 15 out of 37 datasets, JM_{mc} provides the highest F-measure. For other two JM measures JM_{ave} and JM_{Bh}, each approach provides the highest F-measure value only for two datasets. The rest of the measures, such as IG, GR, SU, CS, One-R and Relief-F, have the highest F-measure values for five datasets, eight datasets, five datasets, seven datasets, nine datasets and eight datasets respectively.

Table 5.8 represents the comparison among the nine methods in terms of AUC. Again, it is observed that JM_{mc} produces the highest AUC value for 15 in 37 datasets. For seven datasets, both JM_{ave} and JM_{Bh} outperform other approaches regarding AUC value. In addition, each of SU, CS, and One-R exhibits the highest AUC results for eight datasets, whereas IG, GR, and Relief-F show the highest AUC values for five datasets, 12 datasets, and six datasets respectively.

Table 5.9 illustrates a comparison among nine methods based on the average execution time (in seconds) for all the datasets. The execution times of three different JM distance measures are comparable and slightly higher than other methods except for Relief-F. In this table, the lowest execution time is highlighted in

boldface. For 17 datasets, CS method shows the lowest execution time. JM_{mc}, have the lowest execution time for three datasets, and the other two JM distance measures, JM_{ave} and JM_{Bh} have the lowest execution time for nine datasets and six datasets respectively. One-R method takes the lowest time for 3 data sets, and Relief-F needs the highest time, which is much more than others. Relief-F's execution time is approximately 212 times greater or more than other methods.

Table 5.10 represents a summary of results over all the datasets for all the methods regarding classification accuracy, F-measure, AUC and computational time. In this table, the computed average rank of the different approaches is shown. The nine methods are ranked (from the best to the worst as 1 to 9) based on the value of the evaluation metric individually for all the datasets. If multiple methods show the same effectiveness, they are given the same ranking value. This ranking process is performed for all datasets, and finally, the average rank value over all the data sets is calculated for all the methods. It is found that JM_{mc} achieves the highest rank among all the approaches in terms of the classification accuracy, F-measure and AUC, which are shown in boldface in the table. For execution time, the average ranking over the data set is not suitable as computational time depends on the size of the data set. It seems that the average computational time for JM_{mc} is the third lowest, losing to other JM measures.

Table 5.11 represents the results of pair wise t-test with the proposed approach, JM_{mc} and each of the other approaches regarding classification accuracy, F-measure, AUC and execution time over all the data sets. In this t-test, p-value less than 0.05 indicates JM_{mc} is significantly better than other approaches. Otherwise, there is no significant difference between the approaches. The results on classification accuracy and F-measure indicate that the proposed approach performs significantly better than other approaches. Regarding AUC, proposed JM_{mc} shows significantly better performance for four methods, including JM_{ave}, JM_{Bh}, One-R, and Relief-F. On the other hand, there is no significant variation among the approaches regarding execution time because JM_{mc} shows significantly better results only for Relief-F.

It is found that the average classification accuracy, F-measure, AUC (over all the data sets) of the proposed approach JM_{mc}, is the highest compared to all

other methods. The computational cost of all the methods except Relief-F is very comparable. Relief-F's computational cost is almost 234 times higher than our proposed approach JM_{mc}. The analysis of the standard deviation of the results over 10 independent trials reveals that our proposed approach is quite stable. In our work, feature -feature interaction is not considered to restrict the computational time. The features are ranked according to their individual goodness, and all the methods selected for comparison also work in the same way. We are now working on extending this work in specific application areas such as gene expression data, where the number of features is large and computationally efficient algorithm is important. In our proposed approach, after some trial and error, we fixed the value of α at 10 for all the data sets.

Table 5.6: Classification accuracy of proposed approach and other methods

Dataset		JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
Dataset	tt	JM-mc	JM-ave	JM-Bh	IG	GR	SU	CS	One-R	Relief-F
analcat-authorship	Avg	0.981	0.979	0.980	0.974	0.979	0.979	0.977	**0.982**	0.980
	SD	**0.014**	0.019	0.015	0.019	0.019	0.019	0.017	0.015	0.019
analcat-marketing	Avg	0.519	**0.522**	0.517	0.519	0.519	0.519	0.519	0.519	0.517
	SD	**0.068**	0.076	0.078	0.085	0.085	0.085	0.085	0.085	0.078
breast-tissue	Avg	**0.438**	**0.438**	**0.438**	0.434	0.434	0.434	0.434	0.434	0.395
	SD	0.167	0.167	0.167	0.174	0.174	0.174	0.174	0.174	**0.132**
Bridges	Avg	**0.387**	0.369	**0.387**	0.352	0.352	0.352	0.352	0.352	0.386
	SD	0.146	0.148	0.146	**0.129**	**0.129**	**0.129**	**0.129**	**0.129**	0.177
Cars	Avg	**0.711**	0.647	0.647	0.647	0.647	0.647	0.645	0.657	0.647
	SD	0.068	0.062	0.062	0.062	0.062	0.062	0.067	**0.062**	0.062
Cmc	Avg	**0.487**	0.458	0.456	0.481	0.481	0.481	0.481	0.486	0.481
	SD	0.045	0.043	**0.033**	0.045	0.045	0.045	0.045	0.045	0.045
Dermatology	Avg	0.827	0.814	0.817	**0.836**	0.833	**0.836**	0.781	0.822	0.836
	SD	0.058	0.055	0.058	**0.044**	0.050	**0.044**	0.066	0.058	0.051
Dna	Avg	0.928	**0.932**	**0.932**	0.927	0.928	0.927	0.927	0.854	0.915
	SD	0.019	0.018	0.018	**0.017**	0.018	**0.017**	0.017	0.023	0.022

Table 5.6 Continued from previous page

Dataset		JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
eye-movements	Avg	0.436	0.425	0.425	0.434	0.433	0.434	0.436	0.440	**0.449**
	SD	0.012	0.012	0.012	0.013	0.011	0.013	0.013	0.010	**0.008**
gas-drift	Avg	0.584	**0.601**	0.597	0.589	0.600	0.595	0.591	0.586	0.561
	SD	0.034	0.029	0.032	0.029	**0.027**	0.028	0.032	0.035	0.030
Har	Avg	0.774	0.509	0.520	0.529	0.500	0.504	0.690	0.708	**0.846**
	SD	0.011	0.009	**0.007**	0.008	0.010	0.010	0.011	0.007	0.010
indian-pines	Avg	**0.655**	0.643	0.643	0.642	0.642	0.643	0.642	0.642	0.647
	SD	**0.009**	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.012
Iris	Avg	**0.960**	**0.960**	**0.960**	**0.960**	**0.960**	**0.960**	**0.960**	**0.960**	**0.960**
	SD	**0.047**	**0.047**	**0.047**	**0.047**	**0.047**	**0.047**	**0.047**	**0.047**	**0.047**
mfeat-factors	Avg	**0.929**	0.924	0.924	0.923	0.920	0.922	0.922	**0.929**	0.927
	SD	**0.018**	0.022	0.022	0.022	0.021	0.021	0.023	0.020	0.020
mfeat-fourier	Avg	0.777	0.776	0.779	0.779	0.773	0.780	0.779	**0.781**	0.775
	SD	**0.018**	0.021	0.022	0.021	0.019	0.019	0.021	0.022	0.023
mfeat-karhunen	Avg	0.939	0.942	0.942	0.941	0.941	0.941	**0.943**	0.939	0.939
	SD	0.019	0.018	0.018	0.019	0.017	0.019	**0.016**	0.019	0.022
mfeat-morph.	Avg	0.596	0.542	0.542	**0.608**	0.542	0.608	0.542	0.589	0.542
	SD	**0.019**	0.035	0.035	0.024	0.035	0.024	0.035	0.021	0.035

Table 5.6 Continued from previous page

Dataset		JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
mfeat-pixel	Avg	**0.928**	0.923	0.922	0.926	0.926	0.926	0.923	0.928	0.925
	SD	0.022	0.024	0.026	0.024	0.022	0.025	0.024	0.019	**0.018**
mfeat-zernike	Avg	0.740	0.736	0.737	0.740	0.737	0.740	0.737	0.739	**0.741**
	SD	0.030	0.027	0.024	**0.024**	0.024	**0.024**	0.024	0.024	0.027
mice-protein	Avg	0.792	0.788	0.784	0.792	0.785	0.792	0.787	0.781	**0.813**
	SD	0.024	**0.019**	0.023	**0.019**	0.023	**0.019**	0.020	**0.019**	0.019
page-blocks	Avg	0.931	0.933	0.933	**0.943**	0.936	0.936	0.869	0.931	0.900
	SD	0.007	0.008	0.008	0.008	**0.005**	**0.005**	0.029	0.007	0.015
Pasture	Avg	0.792	0.833	0.833	0.750	0.750	0.750	0.792	0.833	**0.875**
	SD	0.197	0.197	0.197	0.284	0.284	0.284	0.249	0.197	**0.129**
Pendigits	Avg	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**
	SD	**0.012**	**0.012**	**0.012**	**0.012**	**0.012**	**0.012**	**0.012**	**0.012**	**0.012**
Satimage	Avg	0.795	0.786	0.780	0.786	0.780	0.780	0.787	**0.800**	0.777
	SD	0.011	0.010	**0.009**	0.010	**0.009**	**0.009**	0.011	0.010	0.010
Seeds	Avg	**0.905**	0.867	0.867	0.867	**0.905**	**0.905**	**0.905**	0.867	0.867
	SD	0.050	**0.040**	0.067	0.067	0.050	0.050	0.050	0.067	0.067
Segment	Avg	**0.734**	0.597	0.566	0.566	0.731	0.731	0.731	0.566	0.549
	SD	**0.029**	0.057	0.051	0.051	0.034	0.034	0.034	0.051	0.048

Table 5.6 Continued from previous page

Dataset		JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
Splice	Avg	**0.906**	0.905	0.905	0.905	0.906	0.905	0.905	0.904	0.904
	SD	**0.021**	0.022	0.022	0.022	**0.021**	0.022	0.022	0.022	0.024
squash-stored	Avg	**0.660**	0.640	0.637	0.540	0.540	0.540	0.540	0.540	0.653
	SD	0.200	**0.178**	0.189	0.200	0.200	0.200	0.200	0.200	0.199
squash-unstored	Avg	0.847	0.717	0.758	0.808	**0.908**	**0.908**	0.808	0.808	0.808
	SD	**0.087**	0.191	0.172	0.143	0.098	0.098	0.143	0.143	0.143
synthetic-control	Avg	**0.945**	0.867	0.853	0.862	0.840	0.853	0.832	0.865	0.867
	SD	**0.031**	0.044	0.044	0.045	0.057	0.052	0.047	0.048	0.044
teaching-assistant	Avg	**0.516**	**0.516**	**0.516**	**0.516**	**0.516**	**0.516**	**0.516**	**0.516**	**0.516**
	SD	**0.124**	**0.124**	**0.124**	**0.124**	**0.124**	**0.124**	**0.124**	**0.124**	**0.124**
vehicle	Avg	**0.461**	0.449	0.449	0.414	0.442	0.407	0.407	0.414	0.397
	SD	**0.031**	0.061	0.061	0.053	0.051	0.040	0.060	0.053	0.055
vertebra-column	Avg	**0.839**	0.790	0.790	0.765	0.790	0.790	0.790	0.765	0.803
	SD	**0.034**	0.061	0.061	0.066	0.061	0.061	0.061	0.066	0.036
Vowel	Avg	0.677	0.677	0.680	0.663	**0.693**	0.663	0.663	0.663	0.673
	SD	0.035	0.035	0.037	**0.033**	0.035	**0.033**	0.039	**0.033**	0.042
waveform-5000	Avg	**0.800**	**0.800**	**0.800**	**0.800**	**0.800**	**0.800**	**0.800**	**0.800**	**0.800**
	SD	**0.021**	**0.021**	**0.021**	**0.021**	**0.021**	**0.021**	**0.021**	**0.021**	**0.021**

Table 5.6 Continued from previous page

Dataset		JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
Wine	Avg	0.921	0.921	0.921	0.944	0.887	0.899	**0.966**	0.944	0.921
	SD	0.054	0.054	0.054	0.059	0.071	**0.035**	0.039	0.059	0.054
wine-quality	Avg	**0.485**	0.451	0.450	0.446	0.450	0.446	0.450	0.481	0.481
	SD	0.026	0.021	0.026	**0.020**	0.026	**0.020**	0.026	0.026	0.021

Table 5.7: F-measure of proposed approach and other methods

Dataset		JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
Dataset	tt	JM-mc	JM-ave	JM-Bh	IG	GR	SU	CS	One-R	Relief-F
analcat-authorship		**0.986**	0.971	0.985	0.970	0.979	0.978	0.983	0.982	0.981
analcat-marketing		**0.388**	0.372	0.343	0.343	0.343	0.343	0.343	0.343	0.348
breast-tissue		0.414	0.414	0.414	**0.425**	**0.425**	**0.425**	**0.425**	**0.425**	0.372
Bridges		0.482	0.469	0.482	0.443	0.443	0.443	0.443	0.443	**0.489**
Cars		**0.622**	0.553	0.553	0.553	0.553	0.553	0.568	0.546	0.553
Cmc		**0.486**	0.459	0.458	0.478	0.478	0.478	0.478	0.481	0.478
Dermatology		0.842	0.840	0.842	0.863	**0.864**	0.863	0.818	0.812	0.836
Dna		0.921	**0.925**	**0.925**	0.919	0.921	0.919	0.920	0.842	0.907
eye-movements		0.414	0.400	0.400	0.410	0.409	0.410	0.410	0.418	**0.441**
gas-drift		0.585	**0.600**	0.597	0.585	0.596	0.590	0.587	0.581	0.554
Har		0.767	0.483	0.497	0.509	0.473	0.478	0.673	0.692	**0.844**
indian-pines		**0.586**	0.550	0.554	0.550	0.551	0.551	0.550	0.550	0.560
Iris		**0.962**	**0.962**	**0.962**	**0.962**	**0.962**	**0.962**	**0.962**	**0.962**	**0.962**
mfeat-factors		0.928	0.925	0.925	0.924	0.921	0.923	0.923	**0.930**	0.928
mfeat-fourier		0.779	0.779	0.781	0.782	0.776	0.783	0.782	**0.784**	0.778
mfeat-karhunen		0.939	0.943	0.943	0.942	0.943	0.942	**0.945**	0.940	0.940

Table 5.7 Continued from previous page

Dataset	JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
mfeat-morph.	0.570	0.541	0.541	0.577	0.541	0.577	0.541	**0.595**	0.541
mfeat-pixel	0.927	0.925	0.924	0.928	0.929	0.928	0.926	**0.930**	0.927
mfeat-zernike	**0.741**	0.737	0.738	0.741	0.738	0.741	0.738	0.740	0.739
mice-protein	0.793	0.786	0.783	0.792	0.784	0.792	0.785	0.778	**0.817**
page-blocks	0.663	0.653	0.653	**0.691**	0.679	0.679	0.583	0.663	0.625
Pasture	0.773	0.744	0.748	0.792	0.792	0.792	0.749	0.782	**0.833**
Pendigits	**0.339**	**0.339**	**0.339**	**0.339**	**0.339**	**0.339**	**0.339**	**0.339**	**0.339**
Satimage	0.777	0.767	0.761	0.767	0.761	0.761	0.769	**0.781**	0.758
Seeds	**0.908**	0.870	0.870	0.870	**0.908**	**0.908**	**0.908**	0.870	0.870
Segment	**0.746**	0.575	0.570	0.570	0.705	0.705	0.705	0.570	0.542
Splice	**0.898**	0.897	0.897	0.896	0.898	0.896	0.896	0.896	0.896
squash-stored	**0.729**	0.692	0.703	0.520	0.520	0.520	0.520	0.520	0.711
squash-unstored	0.802	0.631	0.565	0.772	**0.825**	0.804	0.758	0.772	0.712
synthetic-control	**0.946**	0.869	0.856	0.863	0.841	0.856	0.834	0.868	0.869
teaching-assistant	**0.510**	0.508	**0.510**	0.508	0.508	0.508	0.508	0.508	0.508
Vehicle	**0.444**	0.416	0.416	0.368	0.402	0.355	0.356	0.368	0.347
vertebra-column	**0.797**	0.742	0.742	0.709	0.742	0.742	0.742	0.709	0.738
Vowel	0.678	0.678	0.680	0.662	**0.691**	0.662	0.663	0.662	0.675

Table 5.7 Continued from previous page

Dataset	JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
waveform-5000	**0.789**	**0.789**	**0.789**	**0.789**	**0.789**	**0.789**	**0.789**	**0.789**	**0.789**
Wine	0.927	0.927	0.927	0.949	0.899	0.912	**0.969**	0.949	0.927
wine-quality	**0.268**	0.208	0.226	0.219	0.226	0.219	0.226	0.268	0.235

Table 5.8: AUC of proposed approach and other methods

Dataset	JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
analcat-authorship	**0.989**	0.985	0.989	0.982	0.989	0.987	0.986	0.987	0.987
analcat-marketing	0.730	**0.742**	0.732	0.715	0.715	0.715	0.715	0.715	0.714
breast-tissue	0.826	0.826	0.826	**0.853**	**0.853**	**0.853**	**0.853**	**0.853**	0.843
Bridges	**0.842**	0.816	**0.842**	0.816	0.816	0.816	0.816	0.816	0.802
Cars	**0.762**	0.711	0.711	0.711	0.711	0.711	0.718	0.706	0.711
Cmc	**0.646**	0.635	0.633	0.644	0.644	0.644	0.644	0.643	0.644
Dermatology	0.919	0.937	0.919	0.936	**0.941**	0.936	0.925	0.894	0.915
Dna	0.938	**0.942**	**0.942**	0.934	0.938	0.934	0.935	0.865	0.928
eye-movements	**0.632**	0.630	0.630	0.628	0.630	0.628	0.632	0.632	0.574
gas-drift	0.813	**0.821**	0.818	0.814	0.820	0.817	0.815	0.815	0.801
Har	0.914	0.843	0.845	0.851	0.839	0.841	0.868	0.878	**0.939**
indian-pines	**0.859**	0.857	0.857	0.857	0.857	0.857	0.857	0.857	0.855
Iris	**0.980**	**0.980**	**0.980**	**0.980**	**0.980**	**0.980**	**0.980**	**0.980**	**0.980**
mfeat-factors	0.946	0.947	0.947	0.946	0.942	0.945	0.943	**0.949**	0.946
mfeat-fourier	0.847	0.844	0.847	0.848	0.846	0.851	0.849	**0.851**	0.845
mfeat-karhunen	0.945	0.949	0.949	0.948	0.947	0.948	**0.951**	0.946	0.945

Table 5.8 Continued from previous page

Dataset	JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
mfeat-morph.	0.823	**0.860**	**0.860**	0.839	**0.860**	0.839	**0.860**	0.845	**0.860**
mfeat-pixel	0.936	0.933	0.933	0.936	0.935	0.936	0.933	**0.938**	0.937
mfeat-zernike	0.872	0.872	0.869	0.872	0.869	0.872	0.869	0.871	**0.873**
mice-protein	0.835	0.834	0.834	0.839	0.835	0.839	0.836	0.832	**0.850**
page-blocks	0.781	0.783	0.783	0.817	**0.824**	**0.824**	0.784	0.781	0.802
Pasture	0.889	0.903	0.917	**0.931**	**0.931**	**0.931**	0.903	0.889	0.847
Pendigits	**0.698**	**0.698**	**0.698**	**0.698**	**0.698**	**0.698**	**0.698**	**0.698**	**0.698**
Satimage	0.843	0.829	0.823	0.829	0.823	0.823	0.830	**0.846**	0.822
Seeds	**0.919**	0.870	0.870	0.870	**0.919**	**0.919**	**0.919**	0.870	0.870
Segment	**0.827**	0.722	0.709	0.709	0.823	0.823	0.823	0.709	0.674
Splice	**0.941**	0.940	0.940	0.940	**0.941**	0.940	0.940	0.940	0.939
squash-stored	0.808	0.803	0.778	**0.835**	**0.835**	**0.835**	**0.835**	**0.835**	0.790
squash-unstored	0.885	0.677	0.844	0.844	**0.927**	**0.927**	0.844	0.844	0.802
synthetic-control	**0.979**	0.945	0.941	0.945	0.940	0.942	0.935	0.945	0.945
teaching-assistant	**0.643**	0.661	**0.643**	0.661	0.661	0.661	0.661	0.661	0.661
Vehicle	0.667	**0.699**	**0.699**	0.699	0.699	0.693	0.690	0.699	0.677
vertebra-column	**0.740**	0.673	0.673	0.649	0.673	0.673	0.673	0.649	0.693

Table 5.8 Continued from previous page

Dataset	JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
Vowel	0.826	0.826	0.823	0.832	**0.845**	0.832	0.828	0.832	0.818
waveform-5000	**0.821**	**0.821**	**0.821**	**0.821**	**0.821**	**0.821**	**0.821**	**0.821**	**0.821**
Wine	0.930	0.930	0.930	0.931	0.867	0.881	**0.958**	0.931	0.930
wine-quality	**0.739**	0.699	0.702	0.701	0.702	0.701	0.702	0.691	0.671

Table 5.9: Average Execution Time (Seconds).

Dataset	JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
analcat-authorship	0.3848	0.3855	0.3886	0.3768	0.3671	0.3737	**0.3308**	0.3512	78.5809
analcat-marketing	0.3032	0.2948	0.2910	0.1695	0.1711	0.1661	**0.1585**	0.1851	19.4746
breast-tissue	0.1340	0.1340	0.1558	0.0684	0.0781	0.0669	**0.0629**	0.0706	2.5687
Bridges	0.1593	0.1608	0.1593	0.1048	0.0678	0.0710	**0.0551**	0.0616	3.0848
Cars	0.0421	**0.0405**	0.0436	0.0557	0.0500	0.0630	0.0501	0.0493	4.2216
Cmc	**0.0763**	0.0798	0.0783	0.1083	0.1081	0.1030	0.0964	0.0990	16.4370
Dermatology	0.4671	0.4643	0.4648	0.4966	0.2019	0.1952	**0.1674**	0.2260	22.2093
Dna	**0.7068**	0.7200	0.7296	1.6520	1.6399	1.6526	1.4597	1.4746	657.5220
eye-movements	0.7304	**0.7237**	0.7450	1.3826	1.3465	1.2486	1.2457	1.2543	317.4830
gas-drift	6.2794	5.6701	**5.6232**	17.1375	17.0589	16.8604	15.8028	18.0575	1,820.1120
Har	20.2938	20.1048	**19.9776**	33.9376	33.9989	32.0043	30.3931	29.0529	8,216.7800
indian-pines	10.6209	10.3521	10.3141	10.8417	10.1151	10.1829	**9.8049**	10.4443	2,476.8263
Iris	0.0276	**0.0213**	0.0260	0.0374	0.0337	0.0328	0.0291	0.0285	1.3149
mfeat-factors	8.8999	9.1327	9.0809	2.9920	2.9463	2.9463	**2.8187**	2.9289	960.5638
mfeat-fourier	3.0015	3.0041	3.0515	1.1353	1.0719	1.0836	**1.0022**	1.0648	212.2076
mfeat-karhunen	2.5280	2.5284	2.5638	0.9327	0.9773	0.9361	0.8939	**0.8893**	173.6740
mfeat-morph.	0.3698	0.3784	0.3651	0.2244	**0.2174**	0.2206	0.2303	0.2287	18.8695

Table 5.9 Continued from previous page

Dataset	JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
mfeat-pixel	10.2853	10.3834	10.2991	3.6476	3.5441	3.5571	3.3424	**3.2469**	1,151.7145
mfeat-zernike	2.0013	1.9880	1.9813	0.7956	0.7804	0.8120	0.7811	**0.7346**	124.1485
mice-protein	1.8360	1.8371	1.8224	0.6889	0.6754	0.6904	**0.6469**	0.6493	139.1284
page-blocks	0.2932	**0.2799**	0.2897	0.3569	0.3494	0.3691	0.3502	0.3491	61.2673
Pasture	0.0604	0.0604	0.0604	0.0834	0.0728	0.0815	**0.0556**	0.0765	3.1010
Pendigits	1.1599	1.1685	1.2143	0.9925	0.9629	0.9476	**0.8977**	0.9076	197.3686
Satimage	1.2023	1.1983	1.1312	1.2697	1.1986	1.1876	**1.1122**	1.1756	278.5883
Seeds	0.0359	**0.0312**	0.0343	0.0484	0.0474	0.0390	0.0469	0.0462	2.4187
Segment	0.3856	0.3966	0.3838	0.2463	0.2724	0.2481	**0.2316**	0.2549	47.5422
Splice	**0.3068**	0.3123	0.3282	0.6517	0.6343	0.6256	0.5730	0.5771	202.4043
squash-stored	0.0689	**0.0627**	0.0674	0.0824	0.0868	0.0849	0.0684	0.0746	4.1449
squash-unstored	0.0651	**0.0612**	0.0651	0.0804	0.0800	0.0808	0.0732	0.0629	3.5465
synthetic-control	0.7500	0.7471	0.7436	0.3824	0.3669	0.3424	**0.3297**	0.3453	61.7121
teaching-assistant	0.0514	0.0436	0.0436	0.0967	0.0446	0.0393	**0.0342**	0.0349	1.5811
Vehicle	0.1383	0.1378	**0.1360**	0.1769	0.1504	0.1591	0.1390	0.1361	18.5617
vertebra-column	0.0428	**0.0366**	**0.0366**	0.0476	0.0533	0.0482	0.0417	0.0456	3.0889
Vowel	0.6452	0.6192	0.6202	0.2151	0.2157	0.2106	**0.2077**	0.2104	17.4169
waveform-5000	0.3672	0.3697	**0.3737**	0.7156	0.7099	0.7201	0.6417	0.6700	201.8161

Table 5.9 Continued from previous page

Dataset	JM_{mc}	JM_{ave}	JM_{Bh}	IG	GR	SU	CS	One-R	Relief-F
Wine	0.0486	**0.0455**	**0.0455**	0.0800	0.0672	0.0547	0.0530	0.0474	3.6554
wine-quality	0.3869	0.3886	0.3909	0.3275	0.3311	0.3428	**0.3097**	0.3464	63.9681

Table 5.10: Overall comparison between the methods

Methods	Classification Accuracy		F-measure		AUC		Execution Time	
	Avg (%)	Rank	Avg (%)	Rank	Avg (%)	Rank	Avg (Sec)	Rank
JM_{mc}	**73.04**	**2.324**	**70.60**	**2.568**	**83.76**	**3.405**	2.031	4.719
JM_{ave}	70.53	3.973	67.40	4.703	82.26	4.405	2.010	4.378
JM_{Bh}	70.55	4.027	67.28	4.297	82.57	4.135	**2.003**	4.324
IG	70.33	4.108	67.52	4.054	82.86	3.730	2.234	5.865
GR	70.88	4.135	67.98	3.946	83.50	3.486	2.191	5.027
SU	70.98	3.838	67.91	3.946	83.43	3.514	2.131	4.784
CS	70.94	4.514	67.88	4.324	83.32	3.459	2.015	**2.784**
One-R	70.92	3.946	67.80	4.297	82.47	4.432	2.066	3.622
Relief-F	71.59	4.243	68.56	4.622	82.18	5.595	475.381	9.000

Table 5.11: Result of Pair wise t-tests

Methods	Classification Accuracy		F-measure		AUC		Execution Time	
	t-value	p-value	t-value	p-value	t-value	p-value	t-value	p-value
JM_{ave}	2.806	**0.008**	3.273	**0.002**	2.085	**0.044**	1.077	0.289
JM_{Bh}	2.811	**0.008**	3.161	**0.003**	2.400	**0.022**	1.267	0.213
IG	3.043	**0.004**	2.962	**0.005**	1.637	0.110	−0.373	0.712
GR	2.433	**0.020**	2.577	**0.014**	0.535	0.596	−0.295	0.770
SU	2.357	**0.024**	2.668	**0.011**	0.717	0.478	−0.197	0.845
CS	3.651	**0.001**	3.699	**0.001**	1.171	0.249	0.036	0.972
One-R	3.119	**0.004**	3.345	**0.002**	2.407	**0.021**	−0.073	0.942
Relief-F	2.044	**0.048**	2.632	**0.012**	2.694	**0.011**	−2.046	**0.048**

5.4 Conclusion

In this chapter, first we have examined the efficiency of JM distance as an effective filter ranking based feature selection tool compared to some other well known filter ranking measures for binary problems. JM distance has been compared with standard feature ranking measures like information gain, chi-squared, relief etc. over benchmark data sets in terms of classification accuracy, feature reduction and computational cost with simulation experiments. The results in classification accuracy is quite comparable with other methods, however JM distance takes much lesser time as compared to all other methods.

After that, we have proposed an efficient feature subset selection algorithm for multiclass problems based on JM distance. Here we have also evaluated our

proposed approach for multiclass problems with 37 benchmark data sets with regard to classifier accuracy, F-measure, AUC, execution time and percentage of feature selection compared to two different extensions of JM distance measures for multiclass problems and six different popular feature evaluation measures for rank based feature selection. In fact, the elegance of the proposed approach lies in the fact that it integrates the selection of final feature subset from the ranked feature lists and the extension of the JM distance measure for multiclass problems in a unified process. In our proposed approach, after some trial and error, we fixed the value of α at 10 for all the data sets.

Chapter 6

Conclusion

6.1 Introduction

Feature selection is an important step prior to the classification stage of machine learning, pattern recognition and data mining problems for addressing high dimensional data. It removes irrelevant and redundant features, which lead to simplifying the classification process and improving accuracy. A stable feature selection algorithm is crucial for identifying the relevant feature subset of meaningful and interpretable features, which is extremely important in the task of knowledge discovery. In this thesis, we have dealt with the stability of feature selection algorithms, appropriate feature selection algorithm that produces better stability, the extension of this feature selection algorithm, and the critical analysis stability measures.

6.2 Summary of the Study

In chapter 3 , a comparative study of the stability of several well-known filter based feature selection algorithms, producing ranked feature subset, has been done. Fifteen benchmark data sets from the UCI repository have been used for simulation experiments. Three types of stability measures, index-based, rank-based and weight based are used to evaluate the stability of feature selection algorithms. Simulation

results demonstrate that for most of the data sets, Jeffries-Matusita (JM)-based feature selection algorithm exhibits more stability irrespective of all types of stability measures. In this chapter stability of filter based and wrapper based feature selection techniques are also explored with using both the subset based and feature ranking approaches. For filter based feature selection, both feature ranking and feature subset selection approach are explored and for wrapper method only subset based approach are considered with three different learners. Here, eight filter based feature ranking (FFR) methods; three filter based feature subset selection (FFSS) methods and wrapper method with three learners of Decision Tree, K-NN and Linear SVM are applied. Stability are calculated with using seven different stability metrics such as, Lustgarten's measure, Wald's measure, Nogueira's measure, Jaccard index, Hamming distance, Dice-Sorensen's index and Ochiai index. A comparative study of the stability of different filter based and wrapper based methods are focused in which simulation experiments are performed with using 30 publicly available bench mark datasets. Simulation result of stability measure reveals that wrapper method shows the least stability but feature ranking based filter method exhibits the highest stability.

Chapter 4 presented the critical analysis of the stability measure of feature selection algorithms. As Kuncheva index and its modifications are widely used in practical problems, in this work, the merits and limitations of the Kuncheva index and its existing modifications (Lustgarten, Wald, nPOG/nPOGR, Nogueira) are studied and analysed with respect to the requisite properties of stability measure. One more limitation of the most recent modified similarity measure, Nogueira's measure, has been pointed out. Finally, corrections to Lustgarten's measure have been proposed to define a new modified stability measure that satisfies the desired properties and overcomes the limitations of existing popular similarity based stability measures. The effectiveness of the newly modified Lustgarten's measure has been evaluated with simple toy experiments and benchmark datasets.

In chapter 5 , we have ranked the features based on JM distance. The results are comparable with mutual information, Relief and Chi Squared based measures as per experiments conducted over 24 public datasets but in much lesser time.

JM distance also provide some intuition about the dataset prior to any feature selection or machine learning algorithm. A comparison has been done on classification accuracy and JM scores of these datasets, which can provide a good intuition on how good a dataset is for classification and point out the need of or lack of further feature collection. In this Chapter, we also proposed a novel heuristic approach for finding out the optimum feature subset from JM distance based ranked feature lists for multiclass problems without explicitly using any specific search technique. The proposed approach integrates the extension of JM measure for multiclass problems and the selection of the final optimal feature subset in a unified process. The performance of the proposed algorithm has been evaluated by simulation experiments with benchmark data sets in comparison with two other previously developed multiclass JM distance measures (weighted average JM distance and another multiclass extension equivalent to Bhattacharyya bound) and some other popular filter based feature ranking algorithms. It is found that the proposed algorithm performs better in terms of classification accuracy, F-measure, AUC with a reduced set of features and computational cost.

6.3 Future Works

In the future, our work can be extended in the following way:

1. We can explore the stability of other feature selection algorithms and find the link between the data set and the feature selection algorithm.

2. In this thesis, we developed the JM distance measures for multi-class problems. The different parameters we have used for proposed JM measures are needed to tune. For example, we can investigate the role of α over the individual data set.

3. We proposed corrections to Lustgarten's measure, but we can explore more on this measure by employing it on feature selection algorithms for a particular data set.

Bibliography

[1] C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Küffner, and R. Zimmer, "Reliable gene signatures for microarray classification: assessment of stability and performance," *Bioinformatics*, vol. 22, no. 19, pp. 2356–2363, 2006.

[2] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello, "Algebraic stability indicators for ranked lists in molecular profiling," *Bioinformatics*, vol. 24, no. 2, pp. 258–264, 2008.

[3] H. W. Lee, C. Lawton, Y. J. Na, and S. Yoon, "Robustness of chemometrics-based feature selection methods in early cancer detection and biomarker discovery," *Statistical applications in genetics and molecular biology*, vol. 12, no. 2, pp. 207–223, 2013.

[4] S. Wichmann and D. Kamholz, "A stability metric for typological features," *Language Typology and Universals*, vol. 61, no. 3, pp. 251–262, 2008.

[5] W. W. B. Goh and L. Wong, "Evaluating feature-selection stability in next-generation proteomics," *Journal of bioinformatics and computational biology*, vol. 14, no. 05, p. 1650029, 2016.

[6] R. Duda, P. Hart, and D. Stork, "Pattern classification 2nd edition: A wiley-interscience publication," 2000.

[7] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*, vol. 454. Springer Science & Business Media, 2012.

[8] C. Wang, Y. Huang, M. Shao, Q. Hu, and D. Chen, "Feature selection based on neighborhood self-information," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 4031–4042, 2019.

[9] W. Gao, L. Hu, P. Zhang, and J. He, "Feature selection considering the composition of feature relevancy," *Pattern Recognition Letters*, vol. 112, pp. 70–74, 2018.

[10] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University-Computer and Information Sciences*, 2019.

[11] T. M. Cover, "The best two independent measurements are not the two best," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 1, pp. 116–117, 1974.

[12] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on knowledge and data engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[13] N. D. Cilia, C. De Stefano, F. Fontanella, and A. S. di Freca, "A ranking-based feature selection approach for handwritten character recognition," *Pattern Recognition Letters*, vol. 121, pp. 77–86, 2019.

[14] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on neural networks*, vol. 20, no. 2, pp. 189–201, 2009.

[15] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.

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

[17] M. A. Hall, *Correlation-based feature selection for machine learning*. PhD thesis, University of Waikato Hamilton, 1999.

[18] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine learning*, vol. 11, no. 1, pp. 63–90, 1993.

[19] I. Thomas, N. Ching, V. Benning, and J. D'aguanno, "Review article a review of multi-channel indices of class separability," *International Journal of Remote Sensing*, vol. 8, no. 3, pp. 331–350, 1987.

[20] S. M. Davis, D. A. Landgrebe, T. L. Phillips, P. H. Swain, R. M. Hoffer, J. C. Lindenlaub, and L. F. Silva, "Remote sensing: the quantitative approach," *New York*, 1978.

[21] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of biomedical informatics*, vol. 85, pp. 189–203, 2018.

[22] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine learning proceedings 1992*, pp. 249–256, Elsevier, 1992.

[23] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," *arXiv preprint arXiv:1202.3725*, 2012.

[24] O. S. Soliman and A. Rassem, "Correlation based feature selection using quantum bio inspired estimation of distribution algorithm," in *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, pp. 318–329, Springer, 2012.

[25] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 856–863, 2003.

[26] B. Senliol, G. Gulgezen, L. Yu, and Z. Cataltepe, "Fast correlation based filter (fcbf) with a different search strategy," in *2008 23rd international symposium on computer and information sciences*, pp. 1–4, IEEE, 2008.

[27] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and information systems*, vol. 34, no. 3, pp. 483–519, 2013.

[28] B. Ghotra, S. McIntosh, and A. E. Hassan, "A large-scale study of the impact of feature selection techniques on defect classification models," in *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, pp. 146–157, IEEE, 2017.

[29] M. Dash, H. Liu, and H. Motoda, "Consistency based feature selection," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 98–109, Springer, 2000.

[30] D. A. A. G. Singh, S. A. A. Balamurugan, and E. J. Leavline, "An unsupervised feature selection algorithm with feature ranking for maximizing performance of the classifiers," *International Journal of Automation and Computing*, vol. 12, no. 5, pp. 511–517, 2015.

[31] S. Nogueira and G. Brown, "Measuring the stability of feature selection with applications to ensemble methods," in *International Workshop on Multiple Classifier Systems*, pp. 135–146, Springer, 2015.

[32] P. Mohana Chelvan and K. Perumal, "A survey on feature selection stability measures," *International Journal of Computer and Information Technology*, vol. 05, pp. 98–103, 2016.

[33] K. Dunne, P. Cunningham, and F. Azuaje, "Solutions to instability problems with sequential wrapper-based approaches to feature selection," *Journal of Machine Learning Research*, pp. 1–22, 2002.

[34] P. Somol and J. Novovičová, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1921–1939, 2010.

[35] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and information systems*, vol. 12, no. 1, pp. 95–116, 2007.

[36] S. Alelyani, Z. Zhao, and H. Liu, "A dilemma in assessing stability of feature selection algorithms," in *2011 IEEE International Conference on High Performance Computing and Communications*, pp. 701–707, IEEE, 2011.

[37] D. Peteiro-Barral, V. Bolon-Canedo, A. Alonso-Betanzos, B. Guijarro-Berdinas, and N. Sanchez-Maroono, "Scalability analysis of lter-based methods for feature selection," *Advances in Smart Systems Research*, vol. 2, no. 1, p. 21, 2012.

[38] M. Zucknick, S. Richardson, and E. A. Stronach, "Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods," *Statistical applications in genetics and molecular biology*, vol. 7, no. 1, 2008.

[39] P. Kalgotra, R. Sharda, and A. Luse, "Which similarity measure to use in network analysis: Impact of sample size on phi correlation coefficient and ochiai index," *International Journal of Information Management*, vol. 55, p. 102229, 2020.

[40] L. I. Kuncheva, "A stability index for feature selection.," in *Artificial intelligence and applications*, pp. 421–427, 2007.

[41] J. L. Lustgarten, V. Gopalakrishnan, and S. Visweswaran, "Measuring stability of feature selection in biomedical datasets," in *AMIA annual symposium proceedings*, vol. 2009, p. 406, American Medical Informatics Association, 2009.

[42] R. Wald, T. M. Khoshgoftaar, and A. Napolitano, "Stability of filter-and wrapper-based feature subset selection," in *2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, pp. 374–380, IEEE, 2013.

[43] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6345–6398, 2018.

[44] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural computation*, vol. 4, no. 1, pp. 1–58, 1992.

[45] A. Roy, N. Das, A. Saha, R. Sarkar, S. Basu, M. Kundu, and M. Nasipuri, "A comparative study of feature ranking methods in recognition of handwritten numerals," in *Artificial Intelligence and Evolutionary Algorithms in Engineering Systems*, pp. 473–479, Springer, 2015.

[46] 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.

[47] C.-J. Huang and W.-C. Liao, "A comparative study of feature selection methods for probabilistic neural networks in cancer classification," in *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*, pp. 451–458, IEEE, 2003.

[48] X. Chen, Z. Yuan, Z. Cui, D. Zhang, and X. Ju, "Empirical studies on the impact of filter-based ranking feature selection on security vulnerability prediction," *IET Software*, 2020.

[49] R. A. Ghazy, E.-S. M. El-Rabaie, M. I. Dessouky, N. A. El-Fishawy, and F. E. Abd El-Samie, "Feature selection ranking and subset-based techniques with different classifiers for intrusion detection," *Wireless Personal Communications*, vol. 111, no. 1, pp. 375–393, 2020.

[50] M. Petković, D. Kocev, and S. Džeroski, "Feature ranking for multi-target regression," *Machine Learning*, vol. 109, no. 6, pp. 1179–1204, 2020.

[51] J. Lee, I. Y. Choi, and C.-H. Jun, "An efficient multivariate feature ranking method for gene selection in high-dimensional microarray data," *Expert Systems With Applications*, vol. 166, p. 113971, 2021.

[52] S. Padma and S. Sanjeevi, "Jeffries matusita based mixed-measure for improved spectral matching in hyperspectral image analysis," *International journal of applied earth observation and geoinformation*, vol. 32, pp. 138–151, 2014.

[53] M. Dalponte, H. O. Ørka, T. Gobakken, D. Gianelle, and E. Næsset, "Tree species classification in boreal forests with hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 5, pp. 2632–2645, 2012.

[54] M. Homem, N. Mascarenhas, and P. Cruvinel, "The linear attenuation coefficients as features of multiple energy ct image classification," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 452, no. 1-2, pp. 351–360, 2000.

[55] A. Daamouche, F. Melgani, N. Alajlan, and N. Conci, "Swarm optimization of structuring elements for vhr image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 6, pp. 1334–1338, 2013.

[56] B. Qiu, Z. Fan, M. Zhong, Z. Tang, and C. Chen, "A new approach for crop identification with wavelet variance and jm distance," *Environmental monitoring and assessment*, vol. 186, no. 11, pp. 7929–7940, 2014.

[57] Y. Wang, Q. Qi, and Y. Liu, "Unsupervised segmentation evaluation using area-weighted variance and jeffries-matusita distance for remote sensing images," *Remote Sensing*, vol. 10, no. 8, p. 1193, 2018.

[58] L. Bruzzone, F. Roli, and S. B. Serpico, "An extension of the jeffreys-matusita distance to multiclass cases for feature selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, no. 6, pp. 1318–1321, 1995.

[59] R. Sen, S. Goswami, and B. Chakraborty, "Jeffries-matusita distance as a tool for feature selection," in *2019 International Conference on Data Science and Engineering (ICDSE)*, pp. 15–20, IEEE, 2019.

[60] H. Wang, T. M. Khoshgoftaar, and A. Napolitano, "Stability of three forms of feature selection methods on software engineering data.," in *SEKE*, pp. 385–390, 2015.

[61] P. Somol and J. Novovičová, "Evaluating the stability of feature selectors that optimize feature subset cardinality," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 956–966, Springer, 2008.

[62] S. Nogueira and G. Brown, "Measuring the stability of feature selection," in *Joint European conference on machine learning and knowledge discovery in databases*, pp. 442–457, Springer, 2016.

[63] A. Frank, A. Asuncion, *et al.*, "Uci machine learning repository, 2010," *URL http://archive. ics. uci. edu/ml*, vol. 15, p. 22, 2011.

[64] D. Dua and C. Graff, "Uci machine learning repository, university of california, school of information and computer science, irvine, ca, 2019," 2019.

[65] J. Vanschoren, J. N. Van Rijn, B. Bischl, and L. Torgo, "Openml: networked science in machine learning," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 2, pp. 49–60, 2014.

[66] P. Turney, "Technical note: Bias and the quantification of stability," *Machine Learning*, vol. 20, p. 23–33, 1995.

[67] G. Stiglic and P. Kokol, "Stability of ranked gene lists in large microarray analysis studies," *Journal of biomedicine and biotechnology*, 2010.

[68] I. Levner, "Feature selection and nearest centroid classification for protein mass spectrometry," *BMC bioinformatics*, vol. 6, no. 1, p. 68, 2005.

[69] K. Zhang, Z. S. Qin, J. S. Liu, T. Chen, M. S. Waterman, and F. Sun, "Haplotype block partitioning and tag snp selection using genotype data and their applications to association studies," *Genome Research*, vol. 14, no. 5, pp. 908–916, 2004.

[70] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pp. 8–pp, IEEE, 2005.

[71] L. Yu, Y. Han, and M. E. Berens, "Stable gene selection from microarray data via sample weighting," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 1, pp. 262–272, 2011.

[72] L. Yu, C. Ding, and S. Loscalzo, "Stable feature selection via dense feature groups," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 803–811, 2008.

[73] M. Zhang, C. Yao, Z. Guo, J. Zou, L. Zhang, H. Xiao, D. Wang, D. Yang, X. Gong, J. Zhu, *et al.*, "Apparently low reproducibility of true differential

expression discoveries in microarray studies," *Bioinformatics*, vol. 24, no. 18, pp. 2057–2063, 2008.

[74] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick, "Gene selection: a bayesian variable selection approach," *Bioinformatics*, vol. 19, no. 1, pp. 90–97, 2003.

[75] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery, "Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data," *Bioinformatics*, vol. 21, no. 10, pp. 2394–2402, 2005.

[76] J. Dutkowski and A. Gambin, "On consensus biomarker selection," *BMC bioinformatics*, vol. 8, no. S5, 2007.

[77] Y. H. Yang, Y. Xiao, and M. R. Segal, "Identifying differentially expressed genes from microarray experiments via statistic synthesis," *Bioinformatics*, vol. 21, no. 7, pp. 1084–1093, 2005.

[78] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2010.

[79] K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple svm-rfe for gene selection in cancer classification with expression data," *IEEE transactions on nanobioscience*, vol. 4, no. 3, pp. 228–234, 2005.

[80] T. M. Khoshgoftaar, A. Fazelpour, H. Wang, and R. Wald, "A survey of stability analysis of feature subset selection techniques," in *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*, pp. 424–431, IEEE, 2013.

[81] M. Zhang, L. Zhang, J. Zou, C. Yao, H. Xiao, Q. Liu, J. Wang, D. Wang, C. Wang, and Z. Guo, "Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes," *Bioinformatics*, vol. 25, no. 13, pp. 1662–1668, 2009.

[82] P. Drotár, J. Gazda, and Z. Smékal, "An experimental comparison of feature selection methods on two-class biomedical datasets," *Computers in biology and medicine*, vol. 66, pp. 1–10, 2015.

[83] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2018.

[84] P. Swain and R. King, "Two effective feature selection criteria for multispectral remote sensing," *LARS technical reports*, p. 39, 1973.

[85] M. Al Rawi, M. Loey, and H. M. El-Bakry, "Machine learning in gene expression profile for central nervous system tumor classification," *Journal of Convergence Information Technology*, vol. 14, no. 1, pp. 49–60, 2019.

[86] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Information Fusion*, vol. 52, pp. 1–12, 2019.

[87] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

List of Publications

- Journal papers:

  1. **Sen R.**, Mandal A. K., and Chakraborty B. (2021). A Critical Study on Stability Measures of Feature Selection with a Novel Extension of Lustgarten Index. Machine Learning and Knowledge Extraction, Vol 3, No. 4, pp. 771-787. DOI: 10.3390/make3040038. (ESCI index)
  2. **Sen R.**, Mandal A. K., and Chakraborty B. (2021). An effective feature subset selection approach based on Jeffries-Matusita distance for multi-class problems. Journal of Intelligent and Fuzzy Systems, Vol 42, no. 4, pp. 4173-4190. DOI:10.3233/JIFS-202796. (SCIE index)

- Conference proceedings:

  1. **Sen R.**, Mandal A. K., Goswami S., Chakraborty B. "A Comparative Study of the Stability of Filter based Feature Selection Algorithms", (2020). 10th IEEE International Conference on Awareness Science and Technology (iCAST), Japan pp. 1-6, doi: 10.1109/ICAwST.2019.8923245. (IEEE, Scopus index)
  2. **Sen R.**, Goswami S. and Chakraborty B., "Jeffries-Matusita distance as a tool for feature selection," 2019 International Conference on Data Science and Engineering (ICDSE), 2019, pp. 15-20, doi: 10.1109/ICDSE47409.2019.8971800. (IEEE, Scopus index)
  3. **Sen R.**, Mandal A. K., Chakraborty B. "Performance Analysis of Extended Lustgarten Index for Stability of Feature Selection", The 15th IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI 2021), 11-12 December, 2021.