

# The Effect of the Characteristics of the Dataset on the Selection Stability

Salem Alelyani, Huan Liu  
 Department of Computer Science and Engineering  
 Arizona State University  
 Tempe, AZ, USA  
 Email: salem.alelyani@asu.edu, huan.liu@asu.edu

Lei Wang  
 School of Computer Science and Software Engineering  
 University of Wollongong  
 NSW, Australia  
 Email: leiw@uow.edu.au

**Abstract**—Feature selection is an effective technique to reduce the dimensionality of a data set and to select relevant features for the domain problem. Recently, stability of feature selection methods has gained increasing attention. In fact, it has become a crucial factor in determining the goodness of a feature selection algorithm besides the learning performance. In this work, we conduct an extensive experimental study using verity of data sets and different well-known feature selection algorithms in order to study the behavior of these algorithms in terms of the stability.

**Index Terms**—Feature selection algorithms, stability, dimensionality reduction, data distribution, Jaccard Index, sample size.

## I. INTRODUCTION

In recent decades, the dimensionality of the data involved in data mining tasks has increased exponentially. Data with extremely high dimensionality presents serious challenges to existing machine learning methods. These kinds of problems are known as "the curse of dimensionality" [9]. To address the problem of curse of dimensionality, dimensionality reduction techniques have been studied, which forms an important branch in machine learning research area. Feature selection is one of major techniques for dimension reduction. It aims to choose a small subset of relevant features from the original set according to certain relevance evaluation criteria [14], [7]. This process usually leads to better learning performance; such as: higher learning accuracy, lower computational cost, and better model interpretability. Hence, the vast majority of feature selection algorithms have been developed and applied to many domains: bioinformatics, pattern recognition, text mining and so on. Feature selection algorithms can be categorized generally with three broad models. These three categories are filter, wrapper, and hybrid model. The filter model is independent to any classifier, i.e. it does not involve any classification in the selection process. Unlike the filter model, wrapper utilizes a classification algorithm in order to select a subset of features with the best classification accuracy. Thus, the wrapper model is more time consuming compared with filter. Also the wrapper model usually produces higher classification accuracy. To overcome the drawbacks in filter model and wrapper model, a hybrid model is introduced. The hybrid model is a combination of filter and wrapper to bridge the gap between these two models by utilizing filter approach to select several candidate subsets of features then select the

best subset using wrapper approach, i.e. by using a classifier to select the best subset.

Stability is a desired character for feature selection algorithms. Our motivation is that since the target concept of a data is fixed, the relevant features should not change across different samples of the data. In real applications, such as genetic analysis, domain experts want stable feature selection results to choose reliable research targets, as unstable feature selection results will confuse them and lower their confidence with the results [3], [6]. The topic of stability of feature selection algorithms has recently gained intensive attention in the research community. The stability of feature selection algorithms is defined as the sensitivity of a feature selection algorithm to perturbation in the training data [12], [15], [19], [6], [10]. Most of the existing work studies the stability from the perspective of the algorithm. We fundamentally disagree with this approach, as we believe that the underlying characteristics of the dataset have a significant impact on the stability. In this work, we will discuss several factors that may effect selection stability. We are going to empirically demonstrate the effect of the dimensionality, the absolute sample size, and the variation of the underlying distribution of the dataset on stability. In addition, we study the effect of the number of selected features on the stability. Finally, we discuss the stability behavior of several well-known feature selection algorithms with a variety of datasets.

## II. LITERATURE REVIEW

The stability of feature selection algorithms is the sensitivity of the selection to variation in the data set [12], [19]. The data variance is usually caused by noise. The existence of noise is ubiquitous; therefore, a good feature selection algorithm should be sufficiently robust to handle noise and can return stable results that contain only relevant features. Stability has gained increasing attention, becoming a hot topic in the feature selection. Furthermore, stability is an important criterion to evaluate the goodness of a feature selection method. One motivation behind this increasing attention to stability is the fact that in domains like bioinformatics, the domain experts would like to see the same or at least similar set of genes, i.e. features to be selected, each time they obtain new samples in the presence of a small amount of perturbation. Otherwise,

they will not trust the algorithm when they get different sets of features while the datasets are drawn for the same problem.

Several stability measures have been proposed to evaluate the similarity among the selected feature subsets [12], [13], [19], [5]. These measures can be broadly categorized into three categories based on their inputs. The first category contains those methods that take the indices of the selected features as an input. A *Jaccard Index* is a representative stability measure in this category [12]. It assesses stability by evaluating the amount of overlap between the results that contains the selected features' indices. Besides *Jaccard Index*, *KI* [13], *Dice Index* [19], and *ANHD* [5] are other proposed measures. For the measures in the second category, the input is the rank of the selected features. These measurers assess stability by evaluating the similarity between two sets of features that are ranked based on their relevance. *Spearman's Rank Correlation Coefficient* [12] is a representative measure for stability by rank that evaluates the correlations between the ranked lists. The third category contains measures that take features' weight as an input. *Pearson's Correlation Coefficient* [12], which assesses the correlation between the weighted results, is a good example.

Given different results  $\mathcal{R} = \{R_1, R_2, \dots, R_l\}$  corresponding to  $l$  runs of algorithm  $\mathcal{A}$  on  $l$  different folds of the data set  $\mathcal{D}$ , its stability can be assessed simply by assessing the amount of overlap between the sets in  $\mathcal{R}$ . The evaluation of the stability of an algorithm can be summarized as the following four key steps: (1) generating  $l$  different folds of  $\mathcal{D}$  either by random sampling or cross-validation. Then, (2) a feature selection algorithm is applied to each fold which (3) produces the  $l$  different results shown in  $\mathcal{R}$ . Finally, (4) an average pairwise similarity is calculated from the selection result to obtain the stability using a suitable stability measure. Most of the existing work has used *Jaccard Index* to evaluate stability [12]. In this work, we focus on this representative measure to conduct our experimental study.

The *Jaccard Index* was introduced in [12] to evaluate the stability for subsets of results that contain selected features' indices by evaluating the amount of overlap between the subsets. Equation(2) shows a *Jaccard Index* for two selected subsets.

$$S_J(R_i, R_j) = \frac{|R_i \cap R_j|}{|R_i \cup R_j|}. \quad (1)$$

Now, we can evaluate its stability by evaluating  $S_J$  in a pairwise manner:

$$S_J(\mathcal{R}) = \frac{2}{l(l-1)} \sum_{i=1}^{l-1} \sum_{j=i+1}^l S_J(R_i, R_j) \quad (2)$$

The *Jaccard Index*  $S_J$  returns a value in the interval of [0,1] where 0 means the feature selection results are not stable and 1 means the results are identical, hence very stable.

### III. OUR CONTRIBUTION

As we mentioned above, the recent work in stability focuses on step (4) in the process: how to estimate the stability of

an algorithm. Although this is an important step in order to give a reasonable estimation of how robust an algorithm is, the recent work does not answer some important questions. In this work, we are going to investigate the following questions: (1) what are the factors that may effect the stability of a feature selection algorithm? (2) Are the factors algorithm-related or dataset-related? In addition, we will investigate, given a data set  $\mathcal{D}$ , (3) what is the most suitable feature selection algorithm to select robust, highly predictive and relevant features? To the best of our knowledge, these important questions have not been sufficiently addressed in the literature. In this paper, we are going to present an empirical study of these questions for better understanding of the stability.

#### A. The Effect of The Number of Selected Features $k$

In most real world datasets, e.g. microarrays, the number of truly relevant features  $k_{opt}$  is usually small compared with the dimensionality  $m$ . In general, it is often difficult to know what  $k_{opt}$  exactly is in advance, and the problem of identifying the value of  $k_{opt}$  is not trivial. In practice, there are three possibilities when choosing  $k$ . (1) The first scenario is choosing  $k$  as  $k_{opt}$ . This case is rare in real world problems. Even if  $k_{opt}$  can be known beforehand, it cannot be guaranteed that the optimal feature subset can always be selected due to the noise and outliers in the data sets and due to the fact that finding the optimal subset is *NP-hard* problem. Thus, the selection is not guaranteed to be stable. The second scenario (2) is choosing  $k < k_{opt}$ . Even with an effective feature selection method, choosing a small  $k$  will make the selected features vary in the presence of small variations in the data set. As a result, the selection is not guaranteed to be robust. The third scenario (3) selects more features than the number of relevant features, that is,  $k > k_{opt}$ . Similar to the previous two, this scenario does not guarantee selection stability even in case of selecting all relevant features.

To illustrate the three scenarios, we assume that, for a given data set, features  $f_1$  to  $f_{10}$  are relevant features while  $f_{11}$  to  $f_{100}$  are the irrelevant ones. First, we assume  $k = 10$ . When we run algorithm  $\mathcal{A}$  on  $l$  different folds, we may get different weights for the features each time due to variation across the folds, which may lead to slightly different subset of selected features. In the second scenario, we assume  $k = 5$ . Similarly, a small variation in the data set may result in different relevance weights for the relevant features, which in turn leads to selecting slightly different features at each fold. The last scenario, we assume  $k = 15$ . In this case, even if we select all the 10 relevant features, the other 5 features will be irrelevant and the order of these irrelevant features may vary significantly at each fold, which decreases stability. Evidently, the analysis is tightly related to the characteristics of the data set. We are going to recall this when we talk about the characteristics of the datasets later in this work.

Another observation is that the larger the value for  $k$ , the higher the stability will be. This is due to several reasons. First, with large  $k$ , the chance of selecting all relevant features becomes high. In addition, the probability that two selected

feature subsets intersect with each other by chance will become higher too. Assume that  $R_i$  and  $R_j$  are the  $i^{th}$  and  $j^{th}$  selected feature sets, respectively. The prior probability of selecting any feature  $f$  is given by:

$$p(f) = \frac{1}{m}$$

Thus, the probability of selecting  $k$  features in any  $R$  is:

$$p(R_i) = p(R_j) = \frac{1}{\binom{m}{k}}$$

The probability of selecting at least one common feature in  $R_i$  and  $R_j$  is:

$$p(|R_i \cap R_j| \geq 1) = \frac{\binom{m}{m-k}}{\binom{m}{k}} \quad (3)$$

It is obvious that the larger  $k$  is, the larger  $p(|R_i \cap R_j| \geq 1)$  will be.

### B. The Effect of The Dataset Characteristics

Most existing works study the stability from an algorithm perspective while ignoring the effects the dataset exerts on it. Intuitively, following this approach only will not effectively solve the challenging questions on the feature selection stability such as: what are the factors that impact stability and how may we consider these factors in the selection process in order to improve selection stability. Also, observing the behavior of different algorithms on different datasets may help to answer the questions regarding how to choose the most appropriate feature selection algorithm for a given dataset. We believe, as we will empirically demonstrate later, that the underlying characteristics of the dataset  $\mathcal{D}$  can have a significant impact on the stability. Thus, studying the stability from the dataset perspective is necessary.

Let  $\mathcal{D}_1$  and  $\mathcal{D}_2$  denote two different datasets with the number of instances,  $n_1$  and  $n_2$ , and the number of features,  $m_1$  and  $m_2$ , respectively. Also, let  $k_1$  and  $k_2$  denote the number of selected features for the two datasets. We apply the algorithm,  $\mathcal{A}$  to each of the  $l$ -folds of  $\mathcal{D}_1$  and  $\mathcal{D}_2$ , respectively. Then, we assess the stability  $\mathcal{S}_J(\mathcal{R}_1)$  and  $\mathcal{S}_J(\mathcal{R}_2)$ . Do we expect to have the same stability, although we use the same algorithm  $\mathcal{A}$ ? Intuitively, the answer is that the stability is not always the same. The important questions now is why  $\mathcal{A}$  does not behave similarly, in terms of stability, on these two datasets? There must be certain characteristics in  $\mathcal{D}_1$  and  $\mathcal{D}_2$  that may be affecting the stability. In the following, we are going to analyze some of these factors and discuss their potential influences on stability.

1) *The Effect of Dimensionality  $m$* : The larger the dimensionality  $m$ , the lower the probability  $p(R_i = R_j)$ , where  $p(R_i = R_j)$  is the prior probability of selecting the same set of features in  $R_i$  and  $R_j$  by chance. Furthermore, the number of combinations of  $k$  features chosen from  $m$  is known as  $m$

choose  $k$ ,  $\binom{m}{k} = \frac{m!}{k!(m-k)!}$ . Hence, the prior probability of choosing the same set of features twice is:

$$p(R_i = R_j) = \frac{1}{\binom{m}{k}} = \frac{k!(m-k)!}{m!}, \quad (4)$$

According to Equation(4), when  $k$  is close to  $m$ ,  $p(R_i, R_j)$  approaches 1, and equals 1 when  $k = m$ . This, also, gives the same conclusion drawn from Equation(3).

2) *The Effect of Sample Size  $n$* : The effect of sample size on the learning process has been well studied in the literature. It is proven that when a limited number of training samples is available, the potential for overtraining is high, and the learning performance may not generalize to a larger populations reducing learning quality [17], [11]. Unsurprisingly, this fact holds for selection stability too, as we will empirically prove in the experiment.

3) *The Effect of the Underlying Data Distribution*: In addition to the dimensionality and sample size, the underlying distribution of the data has a significant impact on stability. For example, when the  $l^{th}$  fold is significantly different from other folds, this may lead to selecting different subsets of features, which means lowering stability. This is related to the theory of important sampling, which suggests an increasing in the number of samples from regions that contribute more to better performance and decrease the number of samples from less attractive regions [8]. Although this approach will lead to reducing data variance, it may cause the loss of important information in the ignored samples. Therefore, [8] suggests, alternatively, to assign less weight to these undesired samples and higher weights to samples from attractive regions.

## IV. EXPERIMENT

In this experimental study, we aim to investigate several important questions regarding the impact of a dataset's characteristics and the number of selected features  $k$  to a feature selection algorithm. Also, we aim to highlight the importance of choosing a feature selection method that can work well on a given dataset. In this section, we are going to empirically study these issues to gain better understanding of stability. In order to achieve this goal, we conduct the following experiments from 23 different datasets, five feature selection algorithms and a Jaccard Index to assess the stability.

### A. Datasets

In this experiment, we use the 32 datasets listed in Table I, which are publicly available online. We divide these datasets to three different groups. The *first group* contains 5 microarrays, which can be downloaded from *ASU Feature Selection Repository*<sup>1</sup>. The datasets in this group share similar characteristics in terms of the number of features  $m$  and the number of samples  $n$ . They have a large number of features ranging from 5748 to 49151. In addition, the number of samples is also relatively large for a microarray, ranging from 85 to 180 samples. The

<sup>1</sup><http://featureselection.asu.edu/>

TABLE I  
DATASETS STATISTICS

		Dataset Name	#Samples $n$	Dimensionality $m$	#Classes
<i>first group</i>	1	CLL-SUB	111	11340	3
	2	GLA-BRA	180	49151	4
	3	TOX	171	5748	4
	4	GLI	85	22283	2
	5	PRO-CAN	171	11302	4
<i>second group</i>	1	ovarian-gilks	23	36534	2
	2	headneck-pyeon-2	23	54675	2
	3	oral-odonnell	27	22283	2
	4	leukemia-wei	27	21481	2
	5	renal-williams	29	17776	2
	6	colon-watanabe	30	54675	2
	7	colon-laiho	31	22283	2
	8	lung-bild	33	54675	2
	9	pancreas-ishikawa	36	22645	2
	10	breast-farmer	37	22215	3
	11	lung-barret	39	22283	2
	12	sarcoma-detwiller	40	22283	2
	13	prostate-true-2	40	12783	2
	14	lymphoma-booman	40	14362	2
<i>third group</i>	1	breasttissue	106	9	6
	2	dermatology	358	34	6
	3	ecoli	336	7	8
	4	glass	214	9	6
	5	heart	270	13	2
	6	iris	150	4	3
	7	liver-disorders	345	6	2
	8	post-operative	87	8	2
	9	soybean	47	35	4
	10	swissbank	200	6	2
	11	wdbc	569	30	2
	12	wine	178	13	3
	13	yeast	205	20	4

*second group* contains 14 different datasets. Similar to the *first group*, the *second group* has a large number of features, yet, a small number of samples  $n \leq 40$ . Finally, the *third group* has generally a large number of samples and small dimensionality.

### B. Feature Selection Algorithms

We chose five well-known feature selection algorithms to conduct this experiment. These algorithms are: ReliefF [18], ChiSquared [18], Information Gain [2], Fisher [4], and L1SVM [1]. All algorithms except L1SVM are publicly available at *ASU Feature Selection Repository*. All algorithms, except L1SVM, are filter-based, so they do not involve any classifier during the selection process. ReliefF and L1SVM both attempt to maximize the margin where the former is related to hypothesis margin maximization, and the latter is sample margin maximization [16]. Yet, Fisher score assigns higher scores to the features that are able to better discriminate the samples from different classes. ChiSquare assesses whether a particular feature is independent of the class label. Similar to ChiSquare, Information Gain assesses the independence between a feature and the class label by the difference between the entropy of the feature and the conditional entropy given the class label.

### C. Experiment Methodology

In order to demonstrate the effect of different factors, mentioned above, we run the algorithms on the datasets using 10 cross-validation, that is,  $l = 10$ . At each run, each algorithm assigns weights to all the features, then, we select the features with the higher weights to be the feature selection results at the  $i^{th}$  fold, denoted by  $R_i$ . All  $R$ s from the  $l$  runs will form the set  $\mathcal{R}$ . Finally, we calculate the average pairwise *Jaccard Index* based on Equation(2) for  $\mathcal{R}$ .

### D. Results

We empirically evaluate the effect of the underlying characteristics of the dataset on the stability of the feature selection algorithm. We use 32 datasets that are grouped to three different sets. The datasets in each group share common characteristics in terms of dimensionality  $m$  and sample size  $n$ . For a clearer illustration, we will divide the results based on the observations.

1) *The Effect of  $k$* : Figure 1 shows the stability of six datasets with different values of  $k$ . If the number of features in the datasets is less than 1000, we evaluate the stability for  $k = \{1, \dots, m\}$ , otherwise,  $k = \{1, \dots, 1000\}$ . Here, we chose only 6 representative datasets because the rest of the datasets behave similarly almost all the time. Figure 1 shows that the larger the value of  $k$  is, the higher the stability

will be. Furthermore, the stability on the microarrays, *CLL-SUB*, *GLI*, and *PRO-CAN*, will increase with the increase of  $k$  until reaching the maximum, where  $k = m$ . Nevertheless, this increase of stability does not necessarily reflect the real robustness of the selection due to the small number of the relevant feature  $k_{rel}$  compared with  $k$ .

Another interesting observation is noticed when a small number of features are selected, as we can obtain high stability during the selection of the first a few numbers of features, especially for the microarrays when  $m$  is very large. This might indicate that these features are significantly relevant to the problem. Accordingly, the stability here could be taken as a criterion to select a suitable  $k$ . Figure 2 shows *CLL-SUB*, *GLI*, and *PRO-CAN* with  $k = \{1, \dots, 100\}$  where the peaks of the stability, indicated by the black arrows mean that the features from 1 to the peak are more frequently selected in all folds than in others. These three plots show different behaviors. For example, *GLI* shows only one peak, which makes it easy to pick an appropriate  $k$  according to this criterion, while *CLL-SUB* has two and *PRO-CAN* has several peaks that make it more complicated to pick the best one. In this case, more constraints are needed. For example,  $k$  should be greater than a certain minimum value. The stability plot of *CLL-SUB* in Figure 2 shows the first peak at  $k = 10$  with stability around 0.73. This means that the algorithms running at each fold agree with each other on a subset of features 73% of the time, which makes these features more frequent to occur at the top of the list. For features beyond the tenth, stability starts to degrade rapidly, which indicates the algorithms running on each fold become less agreeable.

2) *The Effect of Sample Size  $n$* : First of all, it is important to mention that this study focuses on the absolute sample size. Figure 3 shows the stability of all 32 datasets. Figure 3(a) shows the stability of the *first group* where the number of samples is relatively large for microarrays. The red denotes average stability. Figure 3(b), on the other hand, corresponds to the *second group* of the datasets. Similarly, Figure 3(c) shows the stability of the *third group*, where the number of samples  $n$  is very large. From these figures, we can observe the significant difference between the average stability of each group. We can conclude from this observation that the sample size correlates positively with selection stability.

3) *The Effect of the Dimensionality  $m$* : In contrast to the sample size, large dimensionality adversely affects stability. Figure 3(a)(b)(c) clearly shows the impact of the huge dimensionality of the microarrays in (a) and (b) compared with that of (c). However, Figure 3 does not show which factor has more impact on the stability: the sample size or the dimensionality. To answer this question, we run another experiment using a *TOX* dataset because it has large  $m$  and  $n$ . We first run each algorithm on the whole dataset to obtain the weight assigned to each feature. By sorting the weight values, we observe a long tail which corresponds to very low weights or simply zero weights, as shown in Figure 4. We found that the first 100 features have the largest weights; hence, we consider them as the relevant features. Yet, the features

from 1,465 to the end are assigned weights equal to zero, so they are considered as the irrelevant features here. The 1,365 remaining features in between that are moderately relevant are not used in the following experiment. Then, we partition the new version of *TOX* dataset into different partitions, varying both the number of samples and features. The first partition contains 11 samples and 1,100 features. Where the relevant features are always included and 1,000 features are randomly selected from the irrelevant features. Then, we run each of the five feature selection algorithms on the first partition and assess the stability. Next, we increase the number of samples by 10 for the second partition and run the algorithms and evaluate the stability again. When the total number of samples is approached, we add 100 more irrelevant features and start again with 11 samples and so on until we reach the total number of features and samples. Figure II shows stability as a surface where each sub-figure represents the stability of one algorithm. It is observed that the stability becomes weakened, or lower, when the number of samples is smaller and the number of features is larger. This experiment also shows that the the number of samples impact is more significant than the impact of the number of features. Furthermore, the impact of the number of features vanishes when the sample size is significantly large. As the figure shows, the difference in stability between the full set of features and the partition with 1,100 features is at maximum with the smaller sample size, decreasing as larger sample size increases.

4) *The Effect of the Underlying Distribution*: To demonstrate the effect of the underlying distribution of the dataset, we perform a controlled experiment as follows. We sample 11 folds of each dataset. Among the 11 folds sampled from the original training set, ten are allowed to overlap each other, whereas the 11<sup>th</sup> fold (called last fold or  $D_{11}$ ) has no overlapping with any of the ten folds. In this case, we can expect that the difference among the underlying distribution represented by each of the ten folds (denoted as  $D_1, D_2, \dots, D_{10}$ ) is smaller, while the difference between the underlying distribution represented by the 11<sup>th</sup> fold (denoted as  $D_{11}$ ) and any of  $D_1, D_2, \dots, D_{10}$  is larger. By examining the results in Figure 5, it is found that the feature selection results on the first ten folds agree more with each other, but the result on the 11<sup>th</sup> fold is quite different. This shows the affect of variation of underlying distribution to feature selection. We show this only with *CLL-SUB* dataset due to page constraints. All other datasets behave similarly. Thus, we believe that the underlying distribution is an important factor that may effect the stability; consequently, sampling techniques may improve it.

#### E. Algorithms Behaviors

It is not easy to predict the performance of an algorithm on a given dataset. In this section we attempt to observe the behavior of the five feature selection methods on the datasets. Figure 3 shows how the performance of the algorithms varies in terms of stability. Each algorithm behaves differently on different datasets, and with respect to the datasets' groups. In addition, the algorithms behave differently on a single dataset.

For example, we find that with microarray datasets, i.e. *the first and the second groups*, where the dimensionality is huge, the stability difference between algorithms is much larger than *the third group* of datasets. We find that the algorithms give almost equally good stability results with datasets in *the third group*. However, the difference is higher with *the first and second groups*. In addition, we find that *LISVM* selects stable results compared with other methods when dimensionality is high, *LISVM* beats all other methods in having a high dimensionality and a small sample size, in *the second group* of datasets. However, *LISVM* is a clear loser with *the third group*. It is obvious that *ChiSquare* and *Information Gain* are not the preferred methods when the dimensionality is high but they become more competitive in *the third group* which has no clear winner. Finally, *Fisher* and *ReliefF* are almost equally good with the three groups of the dataset. Based on these results, we prefer *LISVM* in case of higher dimensionalities and smaller numbers of samples, although, it becomes less preferred in other cases. Still yet, we prefer either *Fisher* or *ReliefF* in other cases, or in case of a lack of information about the datasets.

## V. CONCLUSION AND FUTURE WORK

The researcher in the realm of feature selection should pay more attention to the underlying characteristics of the dataset for better understanding of selection stability. As the distribution of the dataset is not always known, we believe more attention to sampling techniques should be strengthened. In addition, choosing the appropriate method to perform selection on a given dataset is an interesting problem where algorithms do not perform equally well on different datasets. An algorithm that perform the best on a dataset may not perform good on another one.

In the future, we are going to investigate in more depth the above issues to propose a framework for a feature selection process that takes into account the factors mentioned above in effort to improve selection stability methods. We are going to study different forms of sample variation on the dataset, i.e. different distribution of the samples.

To conclude this paper, we studied several factors that affect the stability of selecting a subset of features. We empirically prove that the stability is dataset dependent, yet not completely algorithm independent. We found that the dimensionality and the sample size of the dataset have significant impact on the selection stability. The larger sample size has a positive impact while a larger dimensionality negatively impacts stability. We found that with a large enough sample size, the impact of dimensionality vanishes. Sample size and dimensionality, aside, the underlying distribution of the dataset plays an important role in stability. We found that the fold, with no overlap with other folds, tends to have different relevant features even when this fold is sampled from the same dataset. In addition, we studied the impact of the number of selected features  $k$  on the stability. We found that determining  $k$  that is close to the optimal number is an important key to reflect the ultimate stability of the algorithm. Finally, we discuss the different

behaviors of the feature selection algorithms on datasets with different characteristics. We showed that Fisher is less sensitive to the characteristics of the dataset when it provides good stability with the three groups of the datasets.

## VI. ACKNOWLEDGEMENTS:

This work is, in part, supported by an NSF grant. We would like to thank The Ministry of Higher Education of Saudi Arabia for the full scholarship given to the first author. We are grateful to the members of DMML at ASU, and anonymous reviewers for their constructive comments.

## REFERENCES

- [1] P.S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. pages 82–90. Morgan Kaufmann, 1998.
- [2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.
- [3] Chad A Davis, Fabian Gerick, Volker Hintermair, Caroline C Friedel, Katrin Fundel, Robert Kfner, and Ralf Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22(19):2356–2363, Oct 2006.
- [4] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2 edition, 2001.
- [5] Kevin Dunne, Pdraig Cunningham, and Francisco Azuaje. Solutions to instability problems with sequential wrapper-based approaches to feature selection. Technical Report TCD-CD-2002-28, Department of Computer Science, Trinity College, Dublin, Ireland, 2002.
- [6] Gokhan Gulgezen, Zehra Cataltepe, and Lei Yu. Stable and accurate feature selection. In *ECML/PKDD (1)*, pages 455–468, 2009.
- [7] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [8] Y. Han and L. Yu. A Variance Reduction Framework for Stable Feature Selection. In *2010 IEEE International Conference on Data Mining*, pages 206–215. IEEE, 2010.
- [9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.
- [10] Zengyou He and Weichuan Yu. Stable feature selection for biomarker discovery, 2010.
- [11] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158, 1997.
- [12] Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12(1):95–116, May 2007.
- [13] Ludmila I. Kuncheva. A stability index for feature selection. In *Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications*, pages 390–395, Anaheim, CA, USA, 2007. ACTA Press.
- [14] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic Publishers, 1998.
- [15] Pavel Pavel Krek, Josef Kittler, and Vclav Hlav. *Computer Analysis of Images and Patterns*, chapter Improving Stability of Feature Selection Methods, pages 929–936. Springer Berlin / Heidelberg, 2007.
- [16] Ran Gilad-Bachrach Ranb, Amir Navot, and Naftali Tishby. Margin based feature selection - theory and algorithms. In *In International Conference on Machine Learning (ICML)*, pages 43–50. ACM Press, 2004.
- [17] T.W. Way, B. Sahiner, L.M. Hadjiiski, and H.P. Chan. Effect of finite sample size on feature selection and classification: A simulation study. *Medical physics*, 37:907, 2010.
- [18] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Pub, 2005.
- [19] Lei Yu, Chris Ding, and Steven Loscalzo. Stable feature selection via dense feature groups. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 803–811, New York, NY, USA, 2008. ACM.

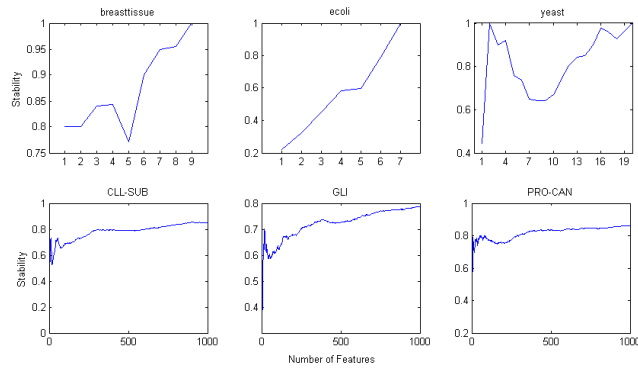


Fig. 1. The effect of large  $k$ .

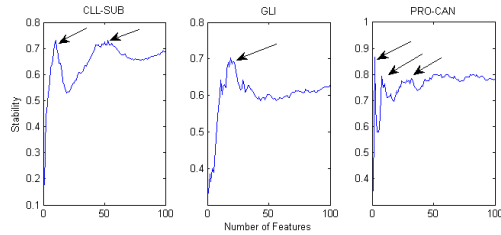


Fig. 2. Demonstration of the stability as a potential criterion to choose the appropriate  $k$ .

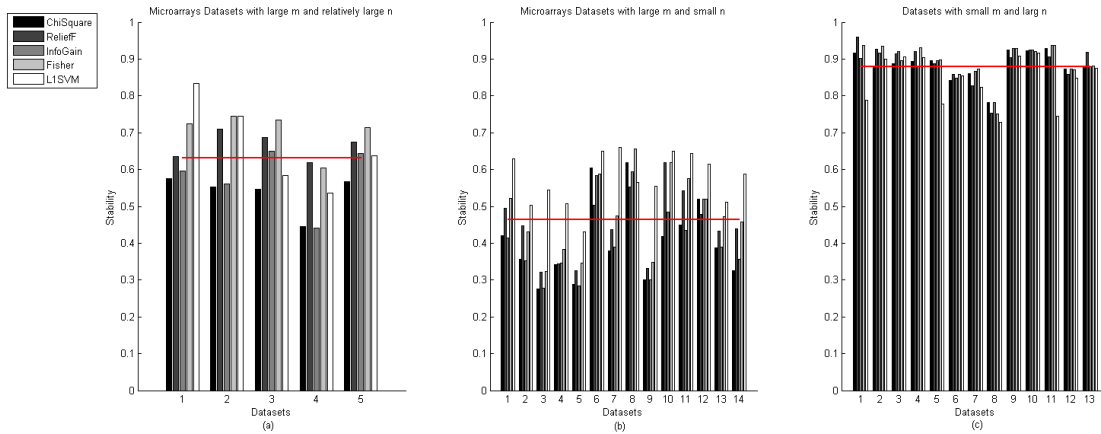


Fig. 3. Jaccard Index stability for all algorithms and all datasets. (a) shows the stability on the *first group* of the datasets. (b) shows the stability on the *second group* of datasets. And (c) shows the stability in the datasets in the *third group*. For the detail of each dataset, see Table I. Note: the x-axis numbering corresponds to the numbering in Table I.

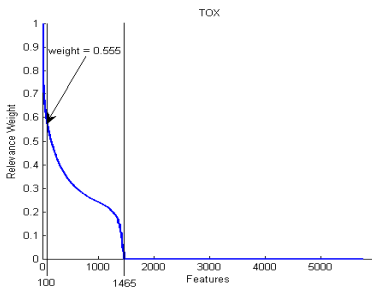


Fig. 4. Features' relevance weight of TOX dataset using ChiSquare (sorted in an descending order).

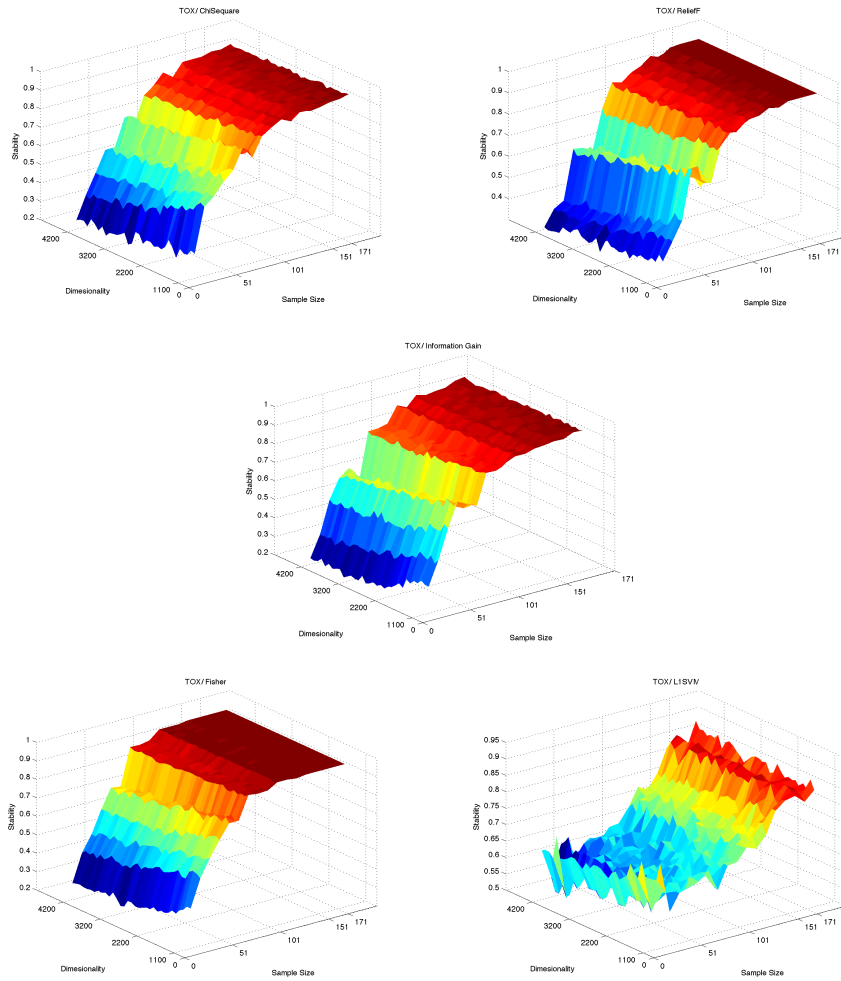


TABLE II  
THE STABILITY OF THE FIVE FEATURE SELECTION ALGORITHMS VS. THE SAMPLE SIZE AND THE DIMENSIONALITY

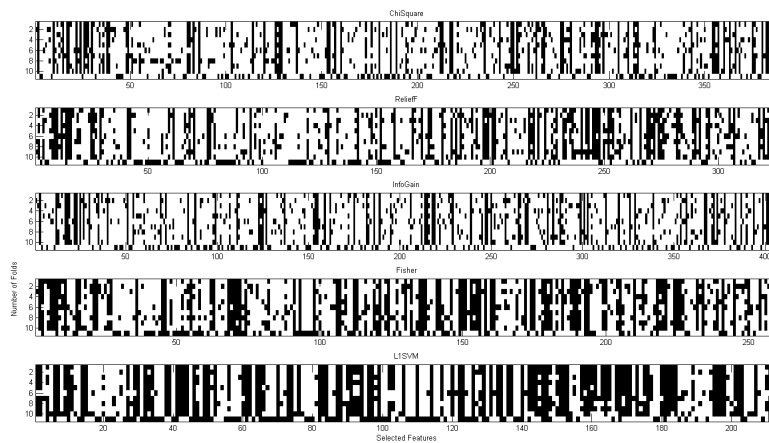


Fig. 5. The frequency of selected features shows the impact of the sample variance on the selection, where the last fold has huge variation which leads to different sets of selected features. All these subplots show the same dataset, CLL-SUB. *Note: We show all features that were selected at least once in all folds.*