Measuring Stability of Feature Selection in Biomedical Datasets

Jonathan L. Lustgarten, MS, PhD,1,2 Vanathi Gopalakrishnan, PhD,1 Shyam Visweswaran, MD, PhD1

1University of Pittsburgh Department of Biomedical Informatics, Pittsburgh, PA
2University of Pennsylvania School of Veterinary Medicine, Philadelphia, PA

Abstract

An important step in the analysis of high-dimensional biomedical data is feature selection. Typically, a feature subset selected by a feature selection method is evaluated for relevance towards a task such as prediction or classification. Another important property of a feature selection method is stability that refers to robustness of the selected features to perturbations in the data. In biomarker discovery, for example, domain experts prefer a parsimonious subset of features that are relatively robust to slight changes in the data. We present a stability measure called the adjusted stability measure that computes robustness of a feature selection method with respect to random feature selection. This measure is useful for comparing the robustness of feature selection methods and is superior to similar measures that do not account for random feature selection. We demonstrate the application of this measure on a biomedical dataset.

Introduction

High-throughput platforms such as DNA microarrays, mass spectrometry and single nucleotide polymorphism microarrays generate high-dimensional transcriptomic, proteomic and genomic data. The analysis of such data typically includes development of classification models to discriminate, for example, between disease and health, and discovery of biomarkers of disease which involves the selection of features that are highly associated with disease. In both situations, an important step in the analysis is the reduction in the dimensionality of the data that is typically accomplished by feature selection. Feature selection, also known as feature subset selection, is the process of selecting a subset of features that optimizes a certain criterion such as classification accuracy.

In addition to classifier performance, another important criterion for evaluating a feature selection method is its stability. Stability is the sensitivity exhibited by a feature selection method to perturbations in the data. Stability is typically more important for knowledge discovery as in biomarker discovery, than in constructing accurate classifiers. In biomarker discovery, domain experts prefer a parsimonious subset of features that are relatively robust to slight changes in the data.

In this paper we present a stability measure that can be used to evaluate and compare feature selection methods. We demonstrate the application of this measure to several classifier based feature selection methods on a biomedical dataset.

Background

A stability measure typically requires defining a similarity measure that assesses the commonality of a pair of feature subsets. Given an appropriate similarity measure, stability is computed as the average similarity over all pairs of feature subsets. Several stability measures that utilize a similarity measure have been described in the literature. Given two feature subsets $s_i$ and $s_j$, Kalousis et al.1 measure similarity between the two subsets as:

$$S_S(s_i,s_j) = 1 - \frac{|s_i| + |s_j| - 2|s_i \cap s_j|}{|s_i| + |s_j| - |s_i \cap s_j|} = \frac{s_i \cap s_j}{s_i \cup s_j}$$

(1)

where, $|s_i \cup s_j|$ is the cardinality of the union of $s_i$ and $s_j$, and $|s_i \cap s_j|$ is the cardinality of the intersection of $s_i$ and $s_j$. The unadjusted stability measure USM for a total of $c$ feature subsets (that may be obtained, for example, from $c$ runs of a feature selection method) is defined as the average of the pairwise similarity for all pairs of feature subsets:

$$USM = \frac{2}{C(C-1)} \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} S_S(s_i,s_j)$$

(2)

The reason for the term unadjusted in USM will become clear later on. Jordan et al.2 developed a analogous similarity and stability measure to compare methods of feature selection for gene expression data. Dunne et al.3 measure similarity as the relative Hamming distance between the two subsets $s_i$ and $s_j$:

$$S_H(s_i,s_j) = 1 - \frac{|s_i \setminus s_j| + |s_j \setminus s_i|}{n},$$

(3)
where ‘\(\setminus\)’ is the set-minus operation and \(n\) is the total number of features.

One deficiency in the above similarity measures (and the corresponding stability measures) is that they are not adjusted for the commonality of a pair of feature subsets that may be obtained purely by chance. These measures do not correct for the artificial increase in the score that occurs with increasingly larger feature subsets. For example, a pair of feature subsets that each contains every feature in the domain attains the maximum score which is not informative. An improved similarity measure would indicate how much better or worse a feature selection method is over one that selects features at random. An improved similarity measure that incorporates the effect of chance was developed by Kuncheva and is given by:

\[
I_C(S_i, S_j) = \frac{r - k^2}{n} = \frac{mn - k^2}{k(n-k)},
\]

(4)

where, \(k = |S_i| = |S_j|\), \(r = |S_i \cap S_j|\), \(n\) is the total number of features, and \(k^2 / n\) is the expected value of \(|S_i \cap S_j|\) that is obtained by modeling the cardinality of the intersection of two equal subsets as a hypergeometric distribution. However, one limitation of this measure is that it is applicable only to feature selection methods that generate feature subsets of fixed cardinality. Feature ranking methods can be adapted to generate feature subsets of a fixed size; however, feature selection methods typically generate feature subsets of different cardinalities.

We developed the adjusted stability measure (ASM) based on a similarity measure that generalizes the similarity measure developed by Kuncheva. While the Kuncheva similarity measure is applicable only to feature subsets of fixed sizes, our similarity measure is applicable to feature subsets of different cardinalities. Thus, the ASM can be applied to feature selection methods that are likely to select feature subsets of differing cardinalities, and also to feature ranking methods where the cardinality of the selected feature subset is typically fixed by the user.

We first describe our similarity measure in detail and then define ASM at the end of this section. Let \(S_i\) and \(S_j\) be two subsets of features with cardinalities \(k_i\) and \(k_j\) respectively obtained from a dataset containing \(n\) features. Let \(r\) be the cardinality of the intersection \(S_i \cap S_j\) of the two subsets. We derive the expected value of \(r\) that would be obtained by chance alone for fixed \(k_i\), \(k_j\) and \(n\). Using combinatorics, the probability of obtaining an intersection of cardinality \(r\) where the feature subsets are of cardinality \(k_i\) and \(k_j\) is given by:

\[
P(X = r | k_i, k_j, n) = \binom{n}{r} \binom{n-r}{k_i-r} \binom{n}{k_j-r} \binom{n}{k_j}.
\]

(5)

After algebraic manipulation, Equation 5 is rewritten as:

\[
P(X = r | k_i, k_j, n) = \frac{\binom{k_i}{r} \binom{n-k_i}{k_j-r}}{\binom{k_j}{k_i}}.
\]

(6)

This expression defines the hypergeometric distribution of a random variable \(X\) with parameters \(n\), \(k_i\) and \(k_j\). Consider a set of \(n\) objects in which \(k_i\) are defective. The hypergeometric distribution describes the probability that exactly \(r\) objects are defective in a sample of \(k_j\) distinct objects drawn from the set. There are a total of \(\binom{n}{k_j}\) possible ways to obtain a sample.
of \( k_j \) objects from \( n \) objects, there are \( \binom{k_i}{r} \) ways to obtain \( r \) defective objects in a sample of size \( k_p \) and there are \( \binom{n-k_j}{k_j-r} \) ways to obtain the non-defective objects in the sample. The probability is positive when \( r \) is between \( \max(0, k_i + k_j - n) \) and \( \min(k_i, k_j) \), where the former expression gives the smallest possible value for \( r \) and the latter expression gives the largest possible value. The expected value of \( r \) is given by:

\[
E(X = r) = \frac{k_i k_j}{n} .
\]

Note that when \( k = k_i = k_j \) the expected value for \( r \) as given by Expression 7 reduces to \( k^2/n \) which is used in Kuncheva’s similarity measure. We define the similarity for a pair of feature subsets as:

\[
S_A(s_i, s_j) = \frac{r - \frac{k_i k_j}{n}}{\min(k_i, k_j) - \max(0, k_i + k_j - n)} .
\]

In the numerator, we subtract the expected value for \( r \) from the actual value of \( r \) and in the denominator we normalize the score to the range of \( r \). This measure has a range of \((-1, 1)\). A value of 0 is what would be expected from random feature selection; a positive value indicates that the feature selection method is more stable than random feature selection and a negative value indicates that the feature selection method is less stable than random feature selection. When \( s_i \) or \( s_j \) or both are empty sets or \( s_i \) or \( s_j \) or both contain every feature in the domain, equation 8 is undefined, and we assume the value to be 0. Figure 1 shows the maximum values attained by \( S_A \) at different values of \( k_i \) and \( k_j \) and a total of 100 features. The maximum value varies with \( k_i \) and \( k_j \); maximum values close to 1 are obtained when both \( k_i \) and \( k_j \) are either very small or very close to \( n \) and maximum values close to 0 are obtained when either of \( k_i \) or \( k_j \) takes on a small value and the other takes on a value close to \( n \). A similar plot is obtained for the minimum values attained by \( S_A \) at different values of \( k_i \) and \( k_j \), except that it is shifted down along the vertical axis, as shown in Figure 2.

The ASM for \( c \) feature subsets is obtained by averaging the pairwise similarity for all pairs of feature subsets, and is given by:

\[
ASM = \frac{2}{c(c-1)} \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} S_A(s_i, s_j) .
\]

The ASM (Equation 9) differs from the USM (Equation 2) in that the similarity measure \( S_A \) used in ASM is adjusted for the commonality of a pair of feature subsets that may be obtained purely by chance while the similarity measure \( S_j \) used in USM is unadjusted for the effect of chance.

**Application to a Biomedical Dataset**

We applied the adjusted and the unadjusted stability measures, ASM and USM, to a proteomic dataset obtained from the University of Pittsburgh Cancer Institute. This dataset contains 240 samples, a total of 70 features (protein probes) and a binary class variable (cancer vs. healthy). We evaluated three different classifier based feature selection methods that included Support Vector Machines (SVM) with a linear kernel which has been shown to perform well on high-dimensional biomedical datasets, Logistic Regression (LR) and Naïve Bayes (NB). For the experiments, we utilized the classifiers implemented in the Waikato Environment for Knowledge Acquisition (WEKA version 3.6) set to their default parameter values.

For each classifier, we performed greedy forward search for feature selection, starting with an empty set of features and adding a single feature in each iteration that most improved the overall classifier performance. The search was terminated when none of the remaining candidate features was unable to improve the accuracy of the classifier.

We performed 10 runs of 10-fold stratified cross-validation, and computed the mean stability measures and the mean classifier accuracies from a total of 100 folds. While bootstrap sampling simulates the sampling distribution underlying the data better, we opted to perform \( N \)-fold stratified cross-validation since it is more commonly used for evaluating classifier performance.
Results

Figure 3 compares the adjusted and the unadjusted stability measures, ASM and USM, for the three classifier based feature selection methods, namely, SVM, LR and NB. Based on the USM, NB had better stability than both SVM and LR. However, the ASM shows that when adjusted for chance NB has worse stability than the other two classifiers. This indicates that the good performance of NB on the unadjusted measure is mainly due to its larger feature subsets (average of 54 features) as compared to LR (average of 32 features) and SVM (average of 38 features). On examination of the feature subsets, NB indeed generated subsets of widely varying cardinalities ranging from 0 to 70. Figure 3 also shows that at larger feature subsets, the effect of the correction incorporated by ASM is bigger as seen in the difference between ASM and USM for NB, and at smaller feature subsets sizes the two stability measures generate similar values as can be seen in the case of SVM and LR.

The average accuracies of the three classifiers learned from the optimal subset of features selected by each classifier based feature selection method are reported in Table 1. Overall, SVM had the highest average accuracy and NB had the lowest with LR in between. Stability and accuracy are not necessarily related; though in our example SVM simultaneously attains the highest accuracy and stability.

Discussion

Typically, classifier based feature selection methods are evaluated and compared on the quality of the subset of the features selected, such as, the ability of the feature subset to predict a target variable of interest. However, an additional measure to evaluate the stability of feature selection methods is needed, especially in tasks such as biomarker discovery. Such a measure should have a low value when there are large variations in the sizes of the feature subsets; lower stability values may indicate either that the

<table>
<thead>
<tr>
<th>Classifier</th>
<th>LR</th>
<th>SVM</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Accuracy</td>
<td>76.84%</td>
<td>87.24%</td>
<td>73.18%</td>
</tr>
<tr>
<td>SEM</td>
<td>3.68</td>
<td>2.04</td>
<td>2.89</td>
</tr>
</tbody>
</table>

Table 1. Average accuracy with standard error of mean (SEM) obtained from 10 x 10-fold cross-validation for the three classifiers. The classifiers were learned from the best feature subset chosen by the wrapper method.
feature selection method is not robust or that the data contains many correlated, redundant or noisy features. Of note, measures that evaluate the quality of the selected features are unrelated to the measures of stability. Specifically, a stability measure provides no information on the quality of the selected features and a quality measure such as classifier accuracy provides no information on the stability of the selected features. In the domain of biomarker discovery these measures provide complementary information to domain experts.

We have developed a stability measure called the ASM that can be applied to feature selection and feature ranking methods. A positive value on this measure indicates that the feature selection method is more stable than random feature selection and a negative value indicates that it is more unstable than random feature selection. When compared to an unadjusted stability measure, we demonstrated on an example dataset, that the ASM is a more robust measure. Thus, the ASM has the advantage that it can be applied directly.

The ASM can be used as a secondary measure for selecting between classifier based and other feature selection methods that have similar performance on a quality measure such as classifier accuracy. Some attempts at combining stability and classifier performance have been described in the literature. For example, David et al.9 describe a weighted combination of stability and classification performance for selecting reliable gene signatures for microarray classification. Classification and feature selection methods that are able to jointly optimize stability and quality may be more useful than those that only optimize one or the other measure.

**Future Work**

In future work, we plan to evaluate the ASM extensively on several biomedical datasets to compare the stability of the commonly used feature selection methods. We also plan to develop a measure that combines ASM with a quality measure such as classification accuracy for evaluating feature subsets.

**Acknowledgements**

This research was funded by the National Library of Medicine (T15-LM007059) and the National Institute of General Medical Sciences (GM 071951). We thank Dr. William Bigbee who heads the Molecular Biomarkers Group at the University of Pittsburgh Cancer Institute for the lung cancer dataset.

**References**