Concerns for Any Dataset or Classifier

Brian T. Luke (lukeb@ncifcrf.gov)

Bias, Chance and Generalizability

Ransohoff [Ran-05a, Ran-05b] has presented three factors that must be explored in any classification study; *bias*, *chance* and *generalizability*. A dataset contains a bias if there is some factor other than the presence of absence of the disease that distinguishes individuals from each group. For example, if all individuals in the disease state are being given a particular drug, there is no way to determine if the change in a feature value is due to the disease or the drug. There is no way to remove this bias, and such situations should be excluded in the initial study design. It is also important that samples from diseased and healthy individuals are not collected at different hospitals or clinics. If it is a multi-institutional study, standardized procedures for collection, storage, processing and transportation must be used [Ban-05]. It is also possible to introduce bias during the production of SELDI spectra. For example, if the volume of sera was significantly larger for healthy than diseased individuals, only healthy samples may have enough serum to be used to test different chip surfaces, reagents, and other protocols. This would mean that the healthy samples may be thawed and re-frozen more times than the diseased samples, and this could change the nature of their serum proteome [Mit-05].

Bias may also be present if individuals in the disease state may be significantly older than those in the healthy state. Many diseases are more prevalent in older individuals and it may be very difficult to find age-matched patients who are disease free or are not on a regular drug treatment. If a random collection of age-matched individuals without signs of the particular disease state are taken to be the healthy category, it is likely that this category will be composed of a number of states due to other diseases or drug responses. Markers separating each of these “healthy” states from the disease state would have to be found. Finding all required biomarkers would be very difficult within a single set of features (i.e. a single microarray or mass spectral study). In addition, if the number of individuals in a particular healthy state was small, the significance of any biomarker may be suspect (see below). For this case, the affect of age can be examined. If there is no correlation between the feature value and the age of the individual in either the disease or healthy state, one can conclude that age is not the source of the difference in feature values [Hab-06].

A published example of a study with an underlying bias is the high-resolution mass spectra of women with and without ovarian cancer [Con-04]. Since the goal of this study was to test a QA/QC procedure, all healthy samples were run first, followed by all cancerous samples. They purposely did not recalibrate the machine between runs and noticed a decrease in the total ion current of the spectra. Therefore, one can never be sure if the resulting classifier distinguished between healthy and cancerous samples, or simply determined the order in which the samples were run. This bias [Bag-05] cannot be removed from the analysis, and has lead to the conclusion that the samples should be randomized when they are examined. A recent investigation [Hon-05] has shown that there is no systematic variability in the spectra between plates, chips or spots on which

the samples were assayed, so the only requirement is a randomization of the order in which the samples are processed.

In contrast to bias, which relates to the dataset, chance is a factor that must be examined in the classifier. If the available individuals in the disease and healthy states are divided into a training set and a testing set, it is theoretically possible to construct one or more classifiers using the training set that can accurately classify the individuals in the testing set without using a marker that depends upon the presence of the disease. Such a classifier is a chance fit to the available data, and we have shown that accurate results can be obtained for certain classifiers without any disease-specific markers being present in the set of available features [Luk-07]. Therefore, simply constructing a good classifier is not sufficient to demonstrate the presence of a disease-specific marker.

It is often assumed that if a classifier is able to accurately classify both a training set and a testing set of data, then this classifier can be used for all individuals in the population from which these individuals were taken. In other words, any classifier that accurately classifies a sufficient sample from a population should be generalizable to the entire population. We assert that this assumption may be true only if the classifier is strictly composed of disease-specific markers. Any classifier that is a chance fit to the available data, or is a fingerprint-based classifier, will not be generalizable to the entire population. This is further explained in the nest section.

Coverage, Uniqueness and Significance

A classifier constructed from state-specific markers is only generalizable to the underlying population to the extent that the coverage of known marker values accurately reflects the full range of intensities. For example, if the state-specific marker in Figure 1 is used by itself to construct a classifier, it contains the basic assumption that the distribution of marker values in the population is well represented by the available samples. If many diseased individuals in the population had values that were substantially below the range shown in this figure, the generalizability of this classifier would be greatly diminished. If too few samples are used to construct the classifier, the probability that there is an adequate coverage of the marker values is reduced.

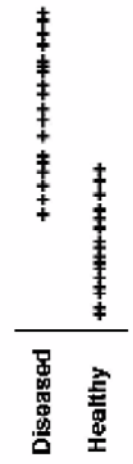

Figure 1

For fingerprint-based classifiers, the issue of coverage is much larger. Here there is the assumption that all fingerprint patterns present in the testing set have a sufficiently similar pattern in the training set; otherwise the testing sample cannot be classified. In other words, each member of the testing set must have a "match" in the training set. This means that an *a priori* division of the individuals into training and testing sets may cause certain fingerprint patterns to have insufficient coverage. A simple example of this is represented by the decision tree in Figure 2a. Assuming that the entire dataset is composed of 60 diseased and 60 healthy individuals, the intensity of Feature 1 splits the dataset into two groups; 40 diseased and 20 healthy individuals if the intensity of this feature is below Cut-1 and 20 diseased and

40 healthy individuals if its intensity is above Cut-1. The left branch is further divided using Feature 2 into a diseased node (D1) that contains 38 diseased and 3 healthy individuals and a healthy node (H1) that contains 2 diseased and 17 healthy individuals. The right branch is divided using Feature 3 into a healthy (H2) and a diseased (D2) terminal node.

Overall this decision tree would yield a sensitivity and a specificity of 90%, but the general procedure is to divide the data into a training set and a testing set and construct the classifier using only the training data. If one-third of the data was removed to form the testing data, the situation in Figure 2b could be produced. In this example, 16 of the 20 healthy samples happened to come from H1 and 16 of 20 diseased samples from D1. This training distribution would make the use of Feature 2 unnecessary and may result in different features being used at each node. If only Features 1 and 3 were used, the training set would have a sensitivity of 90% and a specificity of 82.5%, while the testing data would have a sensitivity of 100% but a specificity of only 20%. The basic reason for this large change in sensitivity is that the fingerprint needed to describe the healthy subjects in Group H1 is no longer present in the training data.

Though this is never done, studies that construct fingerprint-based classifiers would have to examine all possible divisions of the individuals into training and testing sets in order to find those fingerprints that had sufficient coverage in the training set for the testing set samples. Even if such divisions and fingerprints are found, one still has to assume that the coverage of the fingerprint patterns in the training set is sufficient to classify all samples in the underlying population.

Uniqueness is related to the number of classifiers of a given form that accurately classify all available data. For a given set of features, a very small number of unique state-specific markers are generally found. For a given form of the classifier, such as a five-nearest neighbor classifier, using more than two markers generally does not improve the quality of the classifier and all that results is a one- or two-feature classifier. In a study using a published set of mass spectral peaks [Ada-02] we have shown that a five-feature medoid based classifier similar to that used by the groups of Petricoin and Liotta [Bro-05, Con-04, Orn-04, Pet-05, Sri-06, Sto-05] was able to accurately classify all available data for thousands of sets of five features [Luk-07]. These classifiers spanned a wide range of divisions between training and testing sets, but all performed very well on both sets. Similarly, a large number of decision trees that used different sets of features at up to seven decision nodes were able to accurately classify all of the available data. Therefore, for both forms of fingerprint-based classifiers, there was definitely not a unique or small set of accurate classifiers.

Significance is related to how well a classifier of the same form is able to classify data that does not contain a marker. This is done by using the same set of data, but permuting the category labels of the individuals. As expected, a search of a large number of classifiers was able to identify features that were regularly used, but the quality of these putative markers, and any classifier built using them, performed very badly on the training data. This means that if the dataset does not contain a state-specific marker,

none will be found and the results will not be good. In contrast, many classifiers were found using both the medoid based classification algorithm and a single decision tree that classified all available samples to a high quality [Luk-07]. To the authors' knowledge, no study that presented a fingerprint-based classifier has ever exhaustively searched the original dataset or this dataset with permuted categories to examine the uniqueness or significance of their classifier.

Since fingerprint-based classifiers have problems with coverage, uniqueness and significance, it is possible to show that classifiers of this type are not generalizable to the underlying population, and that generalizability is only possible if the classifier is constructed using only state-specific markers.

Figure 2: (a) Hypothetical decision tree using all available data and (b) the corresponding tree when one-third of the samples are removed as testing data.