

Fingerprint-Based Classifiers Are Not Generalizable

Brian T. Luke (lukeb@ncifcrf.gov)

As described in detail elsewhere, a [fingerprint-based classifier](#) uses a proteomic pattern, or fingerprint, to determine whether or not an individual has a disease. In general, if their fingerprint is sufficiently similar to a person who has the disease, then this individual is predicted to have the disease also. While this sounds reasonable, problems with the overall flexibility of these classifiers means that they will not be generalizable [[Ran-05a](#), [Ran-05b](#)] to the underlying population. In other words, the accuracy seen for the data so far cannot be presumed for any new samples.

In [an examination](#) of datasets from individuals with benign prostate hyperplasia (BPH) and those with healthy prostates, two different fingerprint-based classifiers were examined. The first was a [decision tree](#) (DT) containing at most seven decision nodes (i.e. used seven features in the classifier) and the second was a [medoid classification algorithm](#) (MCA) similar to the classification procedure used in several studies by the laboratories of Perticoiu and Liotta [[Bro-05](#), [Con-04](#), [Orn-04](#), [Pet-05](#), [Sri-06](#), [Sto-05](#)]. When the DT algorithm was used, a single run produced 2000 unique 6-node decision trees where the average sensitivity and specificity was over 98.7%. MCA identified at least 2000 classifiers based on a 6-feature fingerprint that produced perfect classification results (sensitivity=specificity=100% with no samples receiving an “Undetermined” classification). Since a large number of accurate classifiers can be found for each of these methods, they suffer from a lack of [uniqueness](#) [[Luk-07](#)].

If a new individual’s spectrum is obtained, the only valid option is to determine the individual’s fingerprint for each classifier and determine if they are healthy or have BPH. The probability is high that some classifiers will predict that the individual is healthy, some that they have BPH, and for the MCA classifiers several should state that the individual is “Undetermined”. The reason that the MCA method should yield several “Undetermined” classifications is a set of six features should yield on the order of 10^6 cells, and at most 106 cells are categorized in any specific classifier (at most 52 BPH cells and 54 healthy cells). When this occurs, it is not advisable to allow each classifier to “vote” and use the classification with the largest vote. This is because many of the different fingerprint patterns may be highly similar and if there is an error in this “common region” of the fingerprint then a large number of classifiers may be incorrect. The only recourse is to independently determine if the individual has BPH or a healthy prostate. All classifiers that produced a correct classification can be retained, but those that gave an incorrect classification or an “Undetermined” classification did so because the coverage of the respective fingerprints was incomplete or the fingerprints themselves had an error. For DT classifiers it may be that the cut point of a feature used in a decision node needs to be slightly changed, or for either method a feature in the fingerprint needs to be changed. It may also be that the required fingerprint is incomplete and a new feature needs to be added to specific classifiers. No matter what the result for a particular classifier in the ensemble of available classifiers, the histology of this individual needs to be independently determined, and this individual becomes part of the training set. Its

action in the training set is either to determine which classifiers need to left alone and which need to be augmented or expanded.

While an “adaptive algorithm” may sound useful, after this sample is added to the training set the situation has not changed. The next individual will get various responses from each member of this ensemble of classifiers and will have to be individually examined to determine if they have BPH or a healthy prostate and then added to the training set in exactly the same way. Therefore, the only way to build a fingerprint-based classifier that correctly predicts whether all males in the United States have BPH or healthy prostates, for example, is to test everyone using another procedure and then build a classifier that reproduces the known results. Several years later and individual with a healthy prostate may develop BPH, but there is no guarantee that the classifier or ensemble of classifiers will be able to predict this.

It can be argued that instead of using an ensemble of fingerprint-based classifier that can continually adapt and potentially become more complex, one should find a classifier that minimizes the number of terminal nodes or category-cells. For example, a single node decision tree simply states “yes” or “no” about whether an individual has BPH. The feature used in this single decision node therefore distinguishes the State of the individual, and all individuals in the same State would be grouped together. This feature then distinguishes one State from another, not one individual from another, and is therefore a biomarker-based classifier. Further testing may slightly change the cut-point for this biomarker to make it slightly more accurate, but will not change the biomarker or form of the classifier. Therefore after sufficient testing, the accuracy of this biomarker-based classifier should be generalizable to the underlying population. This means that only biomarker-based classifiers are generalizable.

(Last updated 5/8/07)