Evaluation of Feature Ranking Ensembles for High-Dimensional Biomedical Data: A Case Study

Ludmila I Kuncheva*, Christopher J Smith*, Yasir Syed†‡, Christopher O Phillips§ and Keir E Lewis†‡
*School of Computer Science, Bangor University, Dean Street, Bangor Gwynedd. UK. LL57 1UT
Email: l.i.kuncheva@bangor.ac.uk
†Institute of Life Sciences, College of Medicine, Swansea University, Singleton Park, Swansea, UK. SA2 8PP
‡Respiratory Unit, Prince Philip Hospital, Llanelli, UK. SA14 8QF
§Welsh Centre for Printing and Coating, College of Engineering,
Swansea University, Singleton Park, Swansea, UK. SA2 8PP

***Abstract*—Developing accurate, reliable and easy to use diagnostic tests is based upon identifying a small set of highly discriminative biomarkers. This task can be cast as feature selection within a pattern recognition context. Medical data are usually of the "wide" type where the number of features is substantially larger than the number of instances. With the abundance of feature ranking methods, it is difficult to pick the most suitable one and choose a final consistent feature subset. Ensembles of ranking methods have been recommended for the task but their stability and accuracy have not been evaluated across different ranking methods. Here we present a case study consisting of 429 samples of exhaled air from smokers, 83% of whom suffer from COPD (chronic obstructive pulmonary disease). The task is to identify a discriminative subset of the 1929 volatile organic compounds (VOCs) measured through mass spectrometry. Using Pareto analysis, 16 feature ranking ensembles were evaluated with respect to three criteria: classification accuracy, area under the ROC curve and the stability of the ensemble selection. The t-statistic was rated the best among the 16 feature rankers, outperforming the currently favourite SVM ranker.**

***Keywords*-Feature selection; feature ranking; classifier ensembles, stability index, COPD**

I. Introduction

Medical data sets are often of the "wide" type, characterised by very high dimensionality, small number of instances and sparseness [1]–[3]. This calls for robust feature selection methods stringent experimental protocols and an estimate of the stability of the selection. A typical example of such a data set is used for our case study. The data is organised in a matrix with 429 rows (data points) and 1929 columns (features) as illustrated in Figure 1. The non-zero values, indicated by dots, form just 2% of the data.

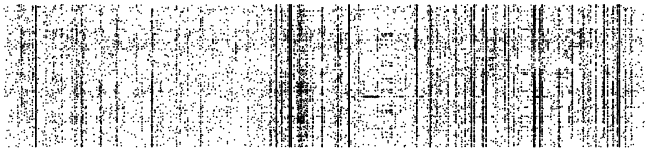

Figure 1. Dimensionality and sparseness of the COPD data set

The task is to select a subset of features (biomarkers) which are easy to obtain and would allow for a robust and accurate diagnosis. Many such methods exist [3], and the challenge is to select the one most suitable for the data at hand. Saeys et al. [3] highlight the potential of ensemble feature selection methods to improve the stability and accuracy of the individual methods.

Classifier ensembles which explicitly or implicitly perform feature selection have been shown to perform well in the bioinformatics context. The Random Subspace (RS) ensemble method [4] builds classifiers on randomly sampled feature subsets. The size of the feature subset and the ensemble size are the two parameters of the method. When the total number of features is tens of thousands, preselection or post-selection have been proposed. Bertoni et al. [5] eliminate redundant and irrelevant features by simple ranking, and subsequently select a feature subset on which the RS ensemble is built, and apply it to gene expression classification. A different approach was proposed by Lai et al. [6], which can be viewed as post-selection. First the feature subsets for the ensemble classifiers are drawn from the whole feature set and then feature selection takes place separately on each feature sample. One drawback of the RS ensemble is that it does not offer an explicit final feature subset. The purpose of the feature selection is to achieve the highest possible classification accuracy and the reduction of the original feature space comes as a by-product.

Ensembles of decision trees trained on bootstrap samples of the data offer a different feature ranking strategy. An example is the Random Forest ensemble (RF) ensemble [7]–[9] with *random tree* classifiers. The difference between a random tree classifier and the conventional tree classifier is in the training. To split the data at a given node of a conventional tree, all features are evaluated in turn and the feature with the best criterion value is selected. In training a random tree, the feature to split a node is selected from a randomly sampled subset of M features. The individual merit of each feature is measured by the classification margin of the ensemble decision – once when the feature is present

and then when the features is replaced with noise. More elaborate feature selection ensembles have been proposed too [10], [11].

Here we adopt a simple and intuitive ensemble procedure for feature ranking suitable for biomedical data [1], [3], [12]; we call this procedure the Ensemble Feature Ranker (EFR). Individual feature ranking is applied to bootstrap samples from the data, and the ranks for each feature are averaged to obtain one final ensemble ranking. The EFR based on the Support Vector Machine classifier (SVM) was found to be both more stable and more accurate than the individual SVM ranker [1].

Stability of the selected feature set is important in the light of the typically small number of observations. Stability of a variety of *individual* feature ranking methods has been evaluated in the literature [13]. However, apart from the SVM feature ranker, the *ensemble* stability of these methods has not been discussed. Furthermore, stability and accuracy must be evaluated together rather than individually [14], [15], which suggests using Pareto optimality for identifying the non-dominated EFRs.

This paper compares 16 feature ranking methods in the EFR context using both accuracy and stability. The task is to select volatile organic compounds (VOCs) for diagnosing chronic obstructive pulmonary disease (COPD) from a chemical analysis of the exhaled air.

II. Feature ranking methods (criteria)

We consider classification into two classes – positive (COPD) and negative (healthy). If the feature of interest x is continuous-valued, the distributions for the two classes can be estimated and compared. The larger the difference, the better the feature. Given the relatively small sample size, the typical choices are the t-test and the Mann-Whitney U test. They both provide an instant feature ranking based on the respective test statistic.

A continuous-valued x can be thresholded to give a positive or a negative label to the object. This allows for using the area under the Receiver Operating Characteristic (ROC) curve as another measure of quality of the features. If x is binary, many more ranking criteria can be added. A wealth of binary ranking criteria have been studied for text classification [16]. Following Forman's rationale, we also experiment with binary features. Using binary values will reduce sensitivity of the feature but it will also reduce the noise, which, on balance, may prove to be beneficial.

The optimal threshold for splitting a feature is derived from the training data, as proposed by Altidor et al. [13]. In a training data set of N objects, assuming that all values of x are different, there are N possible split points. Ideally, all N values should be checked, and the most favourable threshold should be taken forward to define the binarisation. However, the computational cost of this operation may be prohibitive, hence we chose 10 uniformly spread threshold points for each feature.

A split of x at a given threshold T leads to a binary feature x' and the following contingency table with *proportions* calculated from the training data

	$x'=1$	$x'=0$
positive	a (true positive)	b (false negative)
negative	c (false positive)	d (true negative)

$$a+b+c+d=1. \tag{1}$$

Here we do not assume that larger values of x correspond to the positive class. Therefore the labels 1 and 0 can be assigned to x' in reverse order, giving a mirror table to (1)

	$x'=1$	$x'=0$
positive	b (true positive)	a (false negative)
negative	d (false positive)	c (true negative)

$$a+b+c+d=1. \tag{2}$$

When calculating the value of a feature, the better of the two assignments was taken forward.

Following reference sources [13], [16] and [17], and adding SVM ranking and the Random ranking as benchmark methods, we formed a collection of 16 ranking methods (criteria) as detailed below.

A. Individual criteria

1) Accuracy: $Acc = a + d$.

2) Probability ratio: $PR = \frac{a(c+d)}{c(a+b)}$.

3) Odds ratio: $Odds = \frac{ad}{bc}$.

4) Power: $Pow = \left(\frac{d}{c+d}\right)^k - \left(\frac{b}{a+b}\right)^k$, where k is a parameter. Forman recommends a value of $k = 5$.

5) GM measure: , the geometric mean of sensitivity and specificity

$$GM = \sqrt{sensitivity \times specificity}$$
$$= \sqrt{\frac{ad}{(a+b)(c+d)}}.$$

This measure bypasses the possible imbalance of the class prevalence.

6) F1 measure: is the harmonic mean of recall and precision, often used in document retrieval

$$F1 = \frac{2 \times recall \times precision}{recall + precision} = \frac{a}{2a+b+c}.$$

7) Gini index: is a measure used in constructing decision trees. It measures the reduction of impurity associated with the split of x defined by the contingency table (1). The impurity of the data is

$$1 - (a+b)^2 - (a+c)^2 = 2(a+b)(c+d).$$

The impurity of the set for which $x' = 1$ is $\frac{2ac}{(a+c)^2}$. It must be weighted by $(a+c)$, which is the probability of $x' = 1$. Including the impurity for $x' = 0$, the overall reduction of impurity is

$$Gini = 2(a+b)(c+d) - \frac{2ac}{(a+c)} - \frac{2bd}{(b+d)}.$$

8) Mutual Information (MI): is one of the most advocated feature selection criteria. Brown et al. [18] offer a valuable review and propose a general framework, which accommodates most of the past MI criteria. For the 2-by-2 contingency table (1),

$$\begin{aligned} & MI(x;\ \text{class}) \\ = \ & a \log \frac{a}{(a+b)(a+c)} + b \log \frac{b}{(a+b)(b+d)} \\ + \ & c \log \frac{c}{(a+c)(c+d)} + d \log \frac{d}{(b+d)(c+d)}. \end{aligned}$$

9) Chi square:: Calculate first the table with the expected proportions assuming independence between the class class label and feature x'

$$\begin{aligned} e \ &= \ \{e(i,j)\} \qquad (3) \\ &= \begin{array}{lll} & x'=1 & x'=0 \\ \hline \text{positive} & (a+b)(a+c) & (a+b)(b+d) \\ \text{negative} & (a+c)(c+d) & (b+d)(c+d) \end{array}. \end{aligned}$$

The χ^2 statistic is calculated as

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(x'(i,j) - e(i,j))^2}{e(i,j)}.$$

Larger values of χ^2 indicate a better feature.

10) Bi-Normal separation::

$$BNS = \left| \Phi^{-1}\left(\frac{a}{a+b}\right) - \Phi^{-1}\left(\frac{c}{c+d}\right) \right|,$$

where Φ^{-1} is the inverse of the cumulative probability function of the Normal distribution. This ranking criterion was highly recommended by Forman [16].

11) Kolmogorov-Smirnov:

$$KS = \left| \frac{a}{a+b} - \frac{c}{c+d} \right|,$$

Criteria 1-11 require an exhaustive run through the K splits of x to find the optimal threshold T. Formally, if the criterion is denoted by $C(T)$ for threshold T, the value which is used to rank the features is

$$C^* = \max_T \{C(T)\}.$$

The next 3 measures treat x as a continuous-valued variable, and estimate its worth in a single calculation.

12) t-test: . The Student t-test statistic has been used extensively in fMRI data analysis for ranking the voxels and determining statistically significant relationships. We will use the statistic as a ranking criterion (without the test), assuming unequal variances of the two classes. Denote by $m_{(+)}$ and $m_{(-)}$ the means of x for the positive and the negative class respectively, calculated from the training data. Denote by $s_{(+)}$ and $s_{(-)}$ the respective unbiased estimates of the standard deviations. Since we are interested in the magnitude of the difference and not its sign,

$$t = \left| \frac{m_{(+)} - m_{(-)}}{\sqrt{\frac{s_{(+)}}{n_{(+)}} + \frac{s_{(-)}}{n_{(-)}}}} \right|,$$

where $n_{(+)}$ and $n_{(-)}$ are the class counts.

13) Mann-Whitney U test: or Wilcoxon rank-sum test is a non-parametric test for version of the test. We add this statistic to the collection because it is less affected by outliers compared to the t-test. To calculate the U statistics arrange x in ascending order and calculate the ranks. Sum the ranks for the two classes separately to get $R_{(+)}$ and $R_{(-)}$. Then

$$U = \min \left\{ R_{(+)} - \frac{n_{(+)}(n_{(+)}+1)}{2}, \right.$$

$$\left. R_{(-)} - \frac{n_{(-)}(n_{(-)}+1)}{2} \right\}.$$

14) Area under the ROC curve (AUC): has been one of the preferred criteria for evaluating the quality of classification algorithms. We will use it here as a feature ranking criterion, acknowledging the recently published study which warns against over-trusting AUC for small sample sizes [19]. The feature values are arranged in increasing order, and a threshold is set between every pair of values. A classifier is associated with each threshold, whose sensitivity and specificity determine a point on the ROC curve. The AUC is approximated from these discrete points.

15) SVM: The support vector machine classifier SVM has proven its worth for high-dimensional data [20]. SVM is particularly suited to wide data because it scales linearly along the feature dimension while tolerating the small sample size by ensuring large classification margins. The linear-kernel SVM can be used as a feature ranking algorithm. A feature's relevance is measured by the absolute value of the weight for this feature in the linear discriminant function of the trained SVM. This feature ranking method can be thought of as pseudo-multivariate and falls into the category of *embedded* methods [3]. The features are ranked by their worth if the whole feature set is used with the SVM but this does not mean that if the top k features were cut off, they will make a good subset. The Recursive Feature Elimination (RFE) algorithm is an addition to the SVM feature selector, which brings it a step closer to a true multivariate procedure [20]. Starting with an SVM on the entire feature set, a fraction of the features with the lowest weights is dropped. A new SVM is trained with the remaining features, and subsequently reduced in the same way. The procedure stops when the set of the desired cardinality is reached. While SVM-RFE has been found to be extremely useful for wide data such as functional magnetic resonance imaging (fMRI) data [17], it was discovered that the RFE step is not always needed [1], [2], [12]. This may happen when the features are loosely related, and a single SVM captures adequately their relationship. In such problems, it can be expected that other single-pass ranking algorithms will fare well too.

16) Random: Features are arranged in a random order.

B. Ensemble Feature Ranker (EFR)

The ensemble consists of L feature rankers, each one built on a bootstrap sample from the data. In this study, ensemble size of $L = 30$ was deemed sufficient for a stable overall ranking. An interesting alternative to the ensemble approach with the SVM ranker is proposed by Han and Yu [21]. The authors suggest that stable feature sets are obtained in a single ranking, if the instances are weighted by their respective margins.

C. Ensemble Feature Ranker (EFR)

The ensemble consists of L feature rankers, each one built on a bootstrap sample from the data. In this study, ensemble size of $L = 30$ was deemed sufficient for a stable overall ranking. An interesting alternative to the ensemble approach with the SVM ranker is proposed by Han and Yu [21]. The authors suggest that stable feature sets are obtained in a single ranking, if the instances are weighted by their respective margins.

III. Evaluation criteria

A. Classification accuracy.

The EFR is used to produce a ranking of the features. The ensemble classification accuracy is measured on feature subsets of specified cardinalities. For cardinality k, an SVM classifier is trained with the top k features and tested on the testing part of the data.

B. AUC.

The area under the ROC curve is estimated through the same protocol as the classification accuracy. The SVM classifier produces a label output by comparing the discriminant function value with 0. However, for the calculation of the AUC we need a continuous-valued output that can be thresholded, therefore we took the discriminant function directly, prior to inferring the class label.

C. Stability of feature ranking.

A stability index for measuring the agreement between feature rankings has been proposed by Kuncheva [14]. It is based on the concept of consistency between feature subsets. Suppose that there are two feature rankings of T features, R_A and R_B. Let A be the set of the top k features according to R_A and B be the set of the top k features according to R_B. The Consistency Index between A and B is

$$I_C(A,B) = \frac{r - \frac{k^2}{T}}{k - \frac{k^2}{T}} = \frac{rT - k^2}{k(T-k)}, \tag{4}$$

where r is the number of common features in sets A and B. The maximum value of the index, $I_C(A,B) = 1$, is achieved when A and B contain the same features (note that A and B are sets, and there is no order of the features within). The minimum value of the index is bound from below by -1. The limit value is attained for $k = \frac{T}{2}$ and $r = 0$. It should be mentioned that $I_C(A,B)$ is not defined for $k = 0$ and $k = T$. These are the trivial cases where either no feature is selected or all features are selected. They are not interesting from the point of view of comparing feature subsets, so for completeness we can assume $I_C(A,B) = 0$ for both cases. Finally, $I_C(A,B)$ will assume values close to zero for independently drawn A and B because r is expected to be around $\frac{k^2}{T}$.

A stability index for L sets, $S_1, S_2, \ldots, S_L$, all of cardinality k, coming from different rankings, is the average pairwise consistency

$$INDEX(k) = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} I_C(S_i(k), S_j(k)). \tag{5}$$

The consistency index is used here to determine to what extent the feature rankings differ across ensembles.

D. Pareto optimality.

Consider an evaluation of the 16 EFR rankers using the three criteria. The Pareto frontier is a set which contains only the non-dominated methods. A method A is said to be non-dominated if there is no other method B which is no worse than A on all criteria and is strictly better on at least one criterion.

IV. The experimental study

A. Data

The data set in Figure 1 comes from a clinical investigation of Chronic Obstructive Pulmonary Disease (COPD). COPD is a global health problem predicted to become the third leading cause of death worldwide by 2020 [22]. Breath analysis offers a non-invasive means to study biochemical processes in the body. Since Pauling's initial description of around 200 volatile organic compounds (VOCs) in exhaled breath over 30 years ago [23], considerable progress has been made in various trapping, detecting and analytical techniques. This study is a step towards in the quest to identify a key VOCs as markers of COPD which will help to develop smaller bed-side detectors.

COPD patients were recruited through hospital and primary care registers. Smoking status was validated using exhaled carbon monoxide (CO). Batches of tubes with exhaled air were loaded into an Ultra unit (Markes International, Llantrisant, UK) for automated processing via thermal desorption and gas chromatography - mass spectrometry (GC/MS).

In result, 2075 VOCs were recorded for each sample. There were 429 samples taken from smokers or ex-smokers, of which 357 (83.22%) came from COPD patients and the remaining 72 came from healthy controls. Of the 2075 VOCs, 146 had zero values for all samples, leaving 1929 VOCs to select from.

B. Experimental protocol

We carried out 30 runs where the data was split into 90% training and 10% testing parts. Consider one such split. The following calculations were done entirely on the training part. Thirty bootstrap samples were taken to train the ensemble feature rankers. For methods 1-11, each feature was binarised by checking 10 thresholds equally spaced along the range of its values. The EFR rankings were subsequently calculated for methods 1-15 and a random permutation of the features was added as method 16.

C. Results and discussion

Figures 2, 3 and 4 show the resultant curves for three feature ranking methods. The grey lines show the 16 methods, and the black line shows the chosen method. The three methods selected for display are: t-test (12, found to be the best in this study), SVM (15, baseline recommended and used elsewhere) and Random (16, chance).

If there was no knowledge of the prevalence of the classes and a random class label was picked for each object, the classification accuracy would be close to 0.5. Assuming that we know the prevalent class within the population of interest (e.g., the proportion of COPD among the patients with pulmonary complaints in a surgery), the trivial classifier would label all objects in the prevalent class. Our sample was imbalanced in that the proportion of COPD was 83.3%, which was an experimenter's choice and was not intended to be a true reflection of the prevalence of the disease. In this case a classifier would be valuable if the accuracy is significantly higher than the largest prior in the data. This value is plotted in Figure 2 with a dashed line. T-test was carried out to compare the average accuracy across the 30 splits with the largest prior. The subset sizes that were significantly better ($p < 0.05$) are marked with red dots on the curves. The accuracy curves for all 16 measures barely exceed the largest prior mark for small number of features and then plumet quickly to chance level. One possible reason for this behaviour is that the classes are very difficult to separate, and given the "wide" propetry of the data, the SVM classifier overfits the training data for larger feature subsets. Note that even the random feature ranking manages to outperform the largest prior (red circles for subsets of 2 and 7 features), suggesting that all features contain some albeit small amount of discriminatory information.

The blue circles in Figure 3 indicate statistically significant improvement on chance, also estimated through t-test ($p < 0.05$). The T-method is the clear winner among the 16 feature rankers on the AUC criterion while the SVM method recommended in the literature has a mediocre performance. Finally, Figure 4 reveals an interesting pattern of stability. As expected, the Random method gives a flat line at 0 stability. The SVM ranker shows consistent high stability for medium set sizes but low stability for small sizes. Note that the accuracy and AUC of SVM are best for small subsets but these subsets appear to be unstable. Conversely, the T ranker shows high stability for small feature subsets, which is exactly where it holds the high accuracy and the high AUC.

Pareto frontiers were determined using the three criteria: Classification accuracy, AUC and Stability, separately for each of the 38 feature set sizes. Figure 5 shows the number of times each method was in the Pareto frontier out of the 38 calculations. The T-method is positioned first with 34/38 while the current favourite SVM is 6th with 14/38.

Finally, to recommend a subset size, Figure 6 shows a zoomed view of three selection criteria: t-test (12, found to be the best in this study), SVM (15, baseline recommended and used elsewhere) and Random (16, chance). It can be seen that feature sets of size about 50 offer high accuracy, AUC and stability.

Compared to the experimental results where 90% of the data was used for the training, this final set is obtained from larger data and is expected to be better. However, we refrained from running an additional cross-validation experiment with the final selection. Such an experiment would be qualified as "peeking" in that the data used to select the features is reused to test the subset, be it in a cross-validation manner. It is important to highlight this point here because, in spite of the multiple well argued warnings [3], [24], optimistically biased results are often reported through

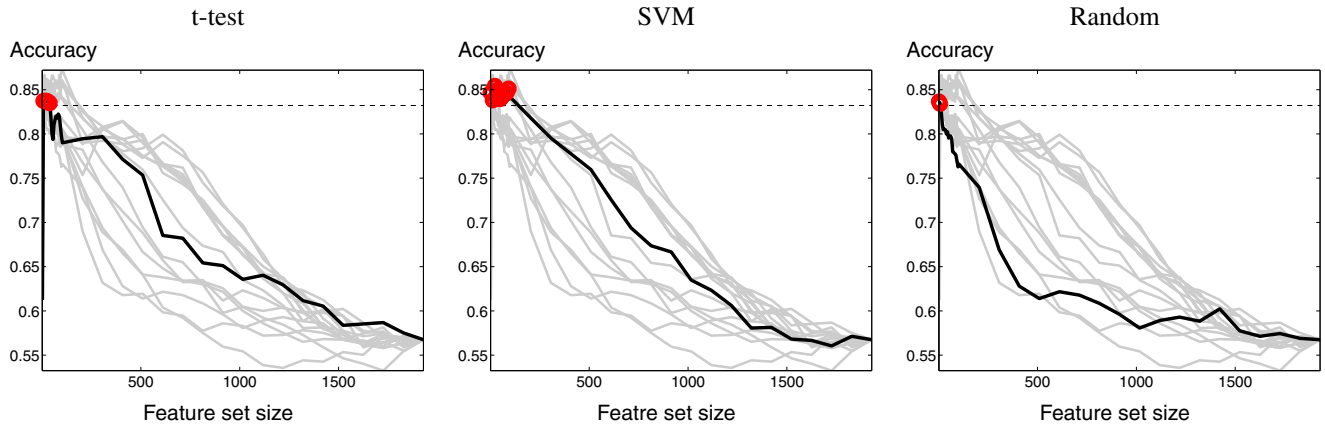

Figure 2. Classification accuracy of the ensemble feature ranker (EFR) as a function of subset size. The dashed line is the prevalence of the larger class. The red circles indicate statistically significant improvement on the trivial classifier which labels all objects in the majority class.

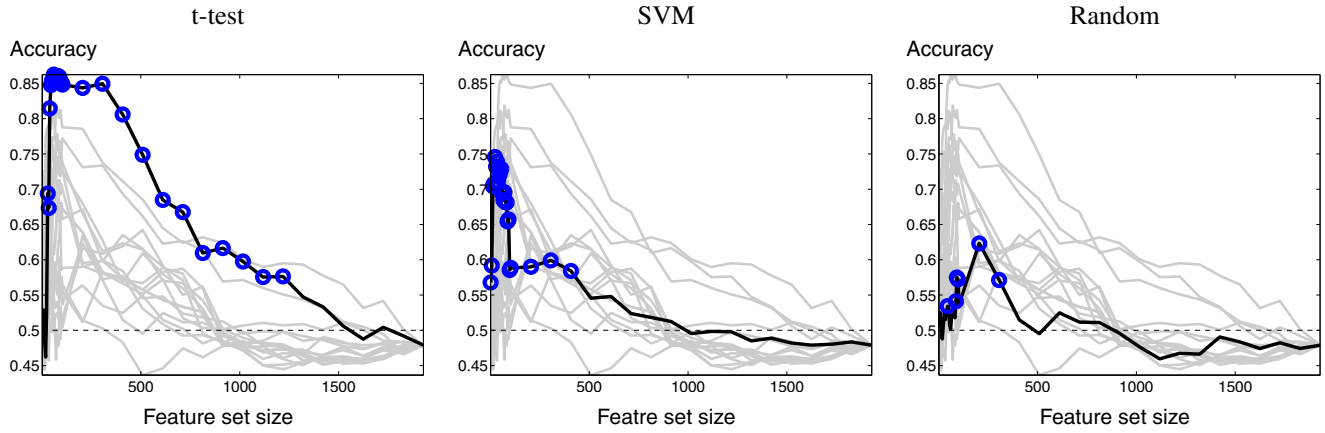

Figure 3. AUC of the ensemble feature ranker (EFR) as a function of subset size. The dashed line at 0.5 corresponds to AUC obtained with random labelling. The blue circles indicate statistically significant improvement on chance.

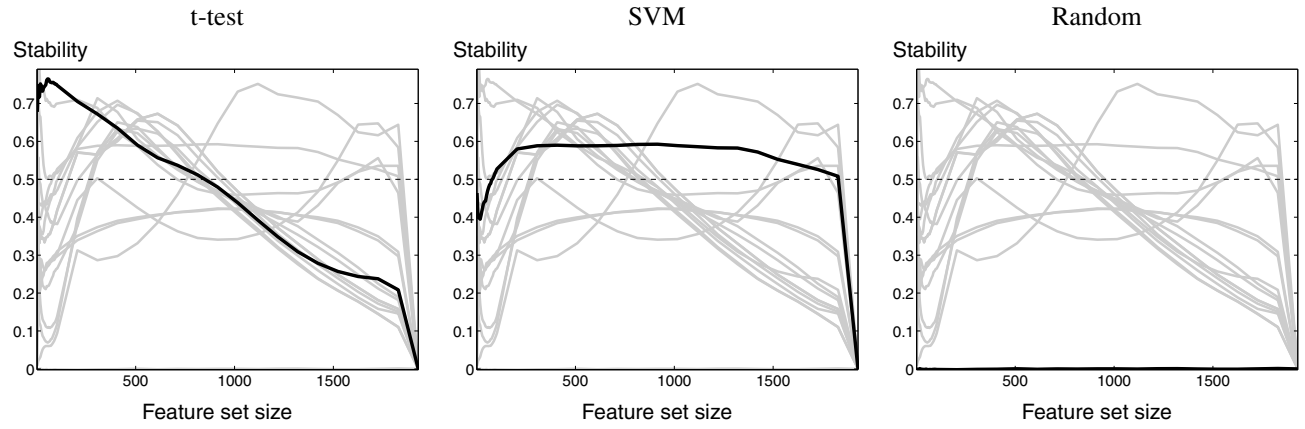

Figure 4. Stability of the ensemble feature ranker (EFR) as a function of subset size. Independent rankings (random rankings) are expected to have stability of 0. The dashed line at 0.5 indicates ”reasonably good” stability above chance.

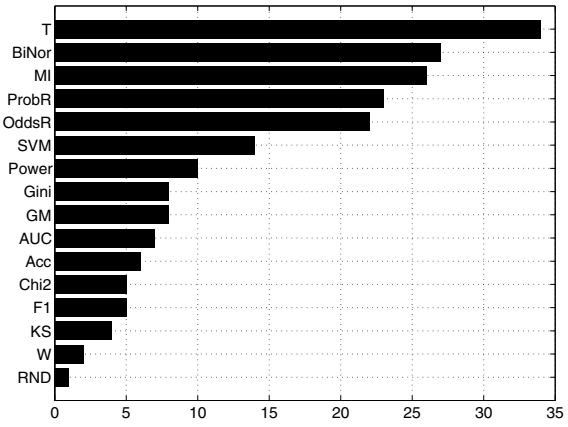

Figure 5. Number of times each method was among the Pareto frontier out of the 38 calculations for different feature subset sizes.

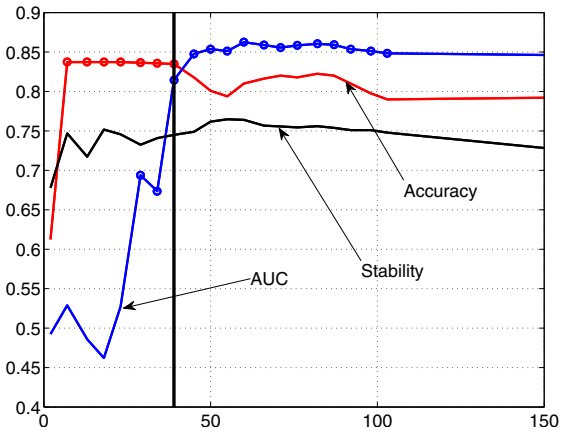

Figure 6. Zoom on the t-test FRE curves

flawed experimental protocols. An estimate of classification accuracy, AUC and stability of the set of the top VOCs subset can be read from Figure 6 at the position of the solid vertical line (at 39). The accuracy for this VOC size was 83.49%, and the area under the ROC curve was 0.8144, both found to be significantly better than chance. The stability of the ensemble selections was 0.7450, which again indicates good match.

V. Conclusion

We examined feature ranking ensembles with 16 feature ranking methods for selecting volatile organic compounds (VOCs) for diagnosing COPD in smokers. The results from the Pareto analysis singled out the T-test statistic as the most accurate and stable method. In our experiment, the T-method outperformed the SVM, the currently favourite ranking method. Our experimental results suggest that, while all VOCs may contain some discriminatory information, a reliable classification of COPD from exhaled air is not an easy task. The classification accuracy was not high enough to allow us to recommend a diagnostic method based on VOCs. Further investigations are needed starting with developing a better technology for measuring the VOCs. Larger data sets should be collected and the prevalence of COPD should be either pre-estimated or reflected in the sampling.

The experimental protocol proposed here is applicable to any feature selection problem. It is interesting to examine the ranking methods on different types of wide data, e.g., mass spectroscopy, gene expression and fMRI.

References

[1] 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 (Oxford, England)*, vol. 26, no. 3, pp. 392–3988, 2010.

[2] P. Geurts, M. Fillet, D. de Seny, M.-A. Meuwis, M. Malaise, M.-P. Merville, and L. Wehenkel, "Proteomic mass spectra classification using decision tree based ensemble methods," *Bioinformatics (Oxford, England)*, vol. 21, no. 14, pp. 3138–3145, 2005.

[3] Y. Saeys, I. n. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics (Oxford, England)*, vol. 23, no. 19, pp. 2507–2517, 2007.

[4] T. K. Ho, "The random space method for constructing decision forests," vol. 20, no. 8, pp. 832–844, 1998.

[5] A. Bertoni, R. Folgieri, and G. Valentini, "Feature selection combined with random subspace ensemble for gene expression based diagnosis of malignancies," in *Biological and Artificial Intelligence Environments*. Springer, 2005, pp. 29–36.

[6] C. Lai, M. J. Reinders, and L. Wessels, "Random subspace method for multivariate feature selection," *Pattern Recognition Letters*, vol. 27, no. 10, pp. 1067–1076, 2006.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[8] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[9] P. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products," *Chemometrics and Intelligent Laboratory Systems*, vol. 83, no. 2, pp. 83–90, 2006.

[10] B. Liu, Q. Cui, T. Jiang, and S. Ma, "A combinational feature selection and ensemble neural network method for classification of gene expression data," *BMC Bioinformatics*, vol. 5, no. 1, p. 136, 2004.

[11] B. Jin, A. Strasburger, S. Laken, F. A. Kozel, K. Johnson, M. George, and X. Lu, "Feature selection for fMRI-based deception detection," *BMC Bioinformatics*, vol. 10, no. Suppl 9, p. S15, 2009.

[12] S. Van Landeghem, T. Abeel, Y. Saeys, and Y. Van de Peer, "Discriminative and informative features for biomolecular text mining with ensemble feature selection," *Bioinformatics (Oxford, England)*, vol. 26, no. 18, pp. i554–i560, 2010.

[13] W. Altidor, T. Khoshgoftaar, and A. Napolitano, "A noise-based stability evaluation of threshold-based feature selection techniques," in *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*. IEEE, 2011, pp. 240–245.

[14] L. Kuncheva, "A stability index for feature selection," in *Proc. IASTED, Artificial Intelligence and Applications*, Innsbruck, Austria, 2007, pp. 390–395.

[15] P. Krizek, J. Kittler, and V. Hlavac, "Improving stability of feature selection methods," in *Proceedings of the 12th international conference on Computer analysis of images and patterns*, ser. CAIP'07, vol. 4673. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 929–936.

[16] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[17] F. De Martino, G. Valente, N. Staeren, J. Ashburner, and R. G. a E. Formisano, "Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns," *NeuroImage*, vol. 43, no. 1, pp. 44–58, 2008.

[18] G. Brown, A. Pocock, M. Zhao, and M. Lujan, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[19] B. Hanczar, J. Hua, C. Sima, J. Weinstein, M. Bittner, and E. R. Dougherty, "Small-sample precision of ROC-related estimates," *Bioinformatics (Oxford, England)*, vol. 26, no. 6, pp. 822–830, 2010.

[20] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Bioinformatics (Oxford, England)*, vol. 22, no. 19, pp. 2348–2355, 2006.

[21] Y. Han and L. Yu, "A variance reduction framework for stable feature selection," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 206–215.

[22] World Health Organisation, "World health statistics," 2008, [cited 2011 13 August].

[23] L. Pauling, A. B. Robinson, R. T. R, and P. Cary, "Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography," *Proceedings of the National Academy of Sciences USA*, vol. 68, pp. 2374–2376, 1971.

[24] P. Smialowski, D. Frishman, and S. Kramer, "Pitfalls of supervised feature selection," *Bioinformatics*, vol. 26, no. 3, pp. 440–443, 2010.