Feature Selection for Classification with Proteomic Data of Mixed Quality

Elena Marchiori
Department of Computer Science
Vrije Universiteit Amsterdam
The Netherlands
Email: elena@cs.vu.nl

Niels H.H. Heegaard, Mikkel West-Nielsen
Department of Autoimmunology
Statens Serum Institut, Copenhagen
Denmark

Connie R. Jimenez
Department of Biology
Vrije Universiteit Amsterdam
The Netherlands

***Abstract*— In this paper we assess experimentally the performance of two state-of-the-art feature selection methods, called RFE and RELIEF, when used for classifying pattern proteomic samples of mixed quality. The data are generated by spiking human sera to artificially create differentiable sample groups, and by handling samples at different storage temperature. We consider two type of classifiers: support vector machines (SVM) and k-nearest neighbour (kNN). Results of leave-one-out cross validation (LOOCV) experiments indicate that RELIEF selects more stable feature subsets than RFE over the runs, where the selected features are mainly spiked ones. However, RFE outperforms RELIEF in terms of (average LOOCV) accuracy, both when combined with SVM and kNN. Perfect LOOCV accuracy is obtained by RFE combined with 1NN. Almost all the samples that are wrongly classified by the algorithms have high storage temperature. The results of experiments on this data indicate that when samples of mixed quality are analyzed computationally, feature selection of only relevant (spiked) features does not necessarily correspond to highest accuracy of classification.**

I. Introduction

Feature selection (FS) for classification can be formalized as a combinatorial optimization problem, finding the feature set maximizing the quality of the hypothesis learned from these features. FS is viewed as a major bottleneck of supervised learning and data mining [1]–[3]. For the sake of the learning performance, it is highly desirable to discard irrelevant features prior to learning, especially when the number of available features significantly outnumbers the number of samples, as is the case in bioinformatics. Because of its computational intractability, the FS problem has been tackled by means of heuristic algorithms based on various machine learning techniques, e.g., [1], [4], [5].

Biological experiments from laboratory technologies like microarray and proteomic techniques, generate data with a very high number of variables (features), in general much larger than the number of samples. Therefore FS provides a fundamental step in the analysis of such type of data [6]. By selecting only a subset of features, the prediction accuracy can possibly improve and more insight in the nature of the prediction problem can be gained.

Various machine learning and statistical techniques for feature selection have been applied to proteomic data, like [7]–[13], in order to detect potential tumor biomarkers for (early) cancer diagnosis (clinical proteomics). A summary of actual challenges and critical assessment of clinical proteomics can be found, e.g., in [14].

In this paper we consider feature selection for classification with pattern proteomic data consisting of samples of mixed quality. This issue is related to broader questions about reproducibility and validity of results in the discovery-based "omics" research [14], [15]. For instance, in a special session on genomics in a recent issue of *Science* an essay entitled "Getting the noise out of gene arrays" noted that "[t]housands of papers have reported results obtained using gene array ... But are these results reproducible?" [16]. A controversy about reprodicibility and validity of results from pattern proteomic data is ongoing [14], [17]: it has been observed that data and results may change depending on the particular way samples are handled [18], on the instability of the laboratory technology, as well as on the computational analysis techniques applied.

We compare experimentally two state-of-the-art feature selection techniques, RELIEF [19], [20] and RFE [21], and two classification techniques, Support Vector Machines (SVM) [22], [23] and k-nearest neighbours (kNN) [24], on proteomic data of mixed quality. Pattern proteomic profiles generated with MALDI-TOF technology are used. The data is generated from human sera samples, using spiking to artificially create differentiable sample groups, and handling the resulting samples at different storage time. The sample preparation procedure utilized to generate such data is described in detail in [18], where the data has been analyzed using also commercial software.

We consider four algorithms obtained by applying feature selection followed by classification: RELIEF+SVM, RELIEF+kNN, RFE+SVM, RFE+kNN. LOOCV is applied to these algorithms for comparing their performance. Results of the experiments show that RELIEF selects feature subsets that are rather stable over the (leave-one-out) runs, and that contain only relevant (spiked) features. Feature subsets selected by RFE are less stable over the runs and contain also features that are not spiked. Performace of the algorithms in terms of average accuracy varies. An expected result is the superior accuracy of RFE+SVM over RELIEF+SVM, due to the fact that RFE uses a bias from SVM when selecting features. A somewhat suprising result is the superiority of RFE+kNN over

RELIEF+kNN for small values of k. In particular, RFE+1NN achieves perfect LOOCV accuracy, while average accuracy of RELIEF+1NN is 0.9271 (with 0.2614 standard deviation). In general, misclassified samples correspond to samples having high storage temperature, thus substantiating the crucial role of laboratory sample handling for the success of clinical proteomics.

II. Classification Techniques

We consider two popular classification techniques from the instance based and neural networks learning approach, respectively.

k Nearest Neighbour Classifier

This technique has been thoroughly studied by the machine learning community (see e.g. [24]). In kNN classification, the training set is directly used to classify new samples. Given a sample $\mathbf{x}$, the k training samples which are most similar to $\mathbf{x}$ are selected and the class of the majority of them is assigned to $\mathbf{x}$. The number k of samples is a parameter. The similarity measure between samples that is generally used is the Euclidean distance:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^{n} (\mathbf{x}_1^i - \mathbf{x}_2^i)},$$

where n is the number of features. Despite of its simplicity, kNN is a popular and effective classification technique which has been applied successfully to real world problems (cf., [25]).

Linear Support Vector Machines

In linear SVM binary classification [22], [23] patterns of two classes are linearly separated by means of a maximum margin hyperplane, that is, the hyperplane that maximizes the sum of the distances between the hyperplane and its closest points of each of the two classes (the margin). When the classes are not linearly separable, a variant of SVM, called soft-margin SVM, is used. This SVM variant penalizes misclassification errors and employs a parameter (the soft-margin constant C) to control the cost of misclassification.

Training a linear SVM classifier amounts to solving the following constrained optimization problem:

$$min_{\mathbf{w},b,\xi_k} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m} \xi_i \quad \text{s.t.} \quad \mathbf{w}\cdot\mathbf{x}_i + b \geq 1 - \xi_i$$

with one constraint for each training sample $\mathbf{x}_i$. Usually the dual form of the optimization problem is solved:

$$min_{\alpha_i} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j - \sum_{i=1}^{m}\alpha_i$$

such that $0 \leq \alpha_i \leq C$, $\sum_{i=1}^{m} \alpha_i y_i = 0$. SVM requires $O(m^2)$ storage and $O(m^3)$ to solve.

The resulting decision function $f(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + b$ has weight vector $\mathbf{w} = \sum_{k=1}^{m} \alpha_k y_k \mathbf{x_k}$. Samples $\mathbf{x}_i$ for which $\alpha_i > 0$ are called *support vectors*, since they uniquely define the maximum margin hyperplane.

Maximizing the margin allows one to minimize bounds on generalization error. Because the size of the margin does not depend on the data dimension, SVM are robust with respect to data with high input dimension. However, SVM are sensitive to the presence of (potential) outliers, (cf. [21] for an illustrative example) due to the regularization term for penalizing misclassification (which depends on the choice of C).

III. Feature Selection Techniques

One can distinguish three main approaches for feature selection: wrapper, filter and embedded.

- In the wrapper approach features are selected by taking into account their contribution to the performance of a given type of classifier (e.g., SVM).
- In the filter approach the selection of features is not (directly) biased towards the performance of a specific type of classifier, but is based on an evaluation criterion for quantifying how well feature (subsets) discriminate the two classes.
- Finally, in the embedded approach feature selection is part of the training procedure of a classifier, like in decision trees.

The performance of these approaches depends on the application domain. The wrapper approachis favoured in many works since the selection of the features is directly linked to the performance of a chosen type of classifier. On the other hand, algorithms based on the filter approach, like RELIEF, are in general more efficient and have been successfully applied to real life domains [26]. Techniques based on the embedded approach provide a global approach, where feature selection is a by-product of the training algorithm for classification.

In this work we consider a wrapper technique, called RFE, and a filter technique, called RELIEF.

Recursive Feature Elimination (RFE)

The weights w_i of a linear SVM classifier provide information about feature relevance, where a bigger weight value implies higher feature relevance. In this paper a feature x_i is scored by means of w_i^2, as in the original RFE algorithm [21].

RFE is a recursive algorithm. Each iteration consists of the following two steps. First feature weights, obtained by training a linear SVM on the training set, are used in a scoring function for ranking features as described above. Next, the feature with minimum rank is removed from the data. In this way, a chain of feature subsets of decreasing size is obtained.

In the original RFE algorithm one feature is discarded at each iteration. Other choices are suggested in [21], where at each iteration features with rank lower than a user-given theshold are removed. In this paper we use a simple instance of RFE where at each iteration half of the features are discarded until a user-given number of final features is reached. The pseudo-code of RFE is given below.

```
RFE
%input: training set X, number of features
             to be selected M
%output: subset Selected of M features
Selected = all features;
num_feat = size(Selected);
while num_features < M
train linear classifier with SVM on X
     restricted to Selected;
score features using the squared value of
      the weights of the classifier;
if size(Selected)/2 < M then
num_feat = size(Selected)/2;
else num_feat = size(Selected) - 1;
Selected = num_feat features of Selected
                with highest score;
end;
return Selected;
```

RELIEF

RELIEF [19], [20] is a feature ranking algorithm that assigns a score to features based on how well the features separate training samples from their nearest neighbours from the same and from the opposite class.

The algorithm constructs iteratively a weight vector, which is initially equal to zero. At each iteration, RELIEF selects one sample, adds to the weight the difference between that sample and its nearest sample from the opposite class (called nearest miss), and subtracts the difference between that sample and its nearest neighbour from the same class (called nearest hit). The iterative process terminates when all training samples have been considered. Subsamplingcan be used to improve efficiency in case of a large training set. The pseudo-code of the RELIEF algorithm used in our experiments is given below.

```
RELIEF
%input: training set X, number of
   features to be selected M
%output: subset Selected of M features
nr_feat = total number of features;
weights = zero vector of size nr_feat;
for all samples exa in training set do
hit(exa) = nearest neighbour of exa
            from same class;
miss(exa) =  nearest neighbour of exa
            from opposite class;
weights =  weights-abs(exa-hit(exa))+
            abs(exa - miss(exa));
end;
Selected = M features with highest
             weights's value;
return Selected;
```

IV. Proteomic Pattern Data

The proteomic pattern data here considered is obtained by surface-enhanced laser desorption/ionization time-of-flight mass spectronomy (MALDI-TOF MS), a recent laboratory technology which offers high-throughput protein profiling. It measures the concentration of low molecular weight peptides in complex mixtures, like serum (cf. e.g. [27]). Because it is relatively inexpensive and noninvasive, it is considered a promising new technology for classifying disease status and for tumor biomarker detection.

SELDI-TOF MS technology produces a graph of the relative abundance of ionized peptides (y-axis) versus their mass-to-charge (m/z) ratios (x-axis). (see Figure 1) The m/z ratios are proportional to the peptide masses, but the technique is not able to identify individual peptides, because different peptides may have the same mass and because of limitations in the m/z resolution. There is no obvious relation between neighbour data points, apart from the fact that they refer to peptides of similar masses and that the resolution is such that the graph should be considered a smoothed version of the true mass density.

Given proteomic profiles from healthy and diseased individuals it is desired to build a classifier for tumor diagnostics and to identify the protein masses that are potentially involved in the disease.

The data considered in this study consists of 96 profiles obtained from human sera of eight adult persons. Spiking has been used to produce the two classes, and 6 different storage temperatures (t=0, 1, 4, 8, 24, 48 hours) have been used to produce data of mixed quality. Each profile contains 22572 m/z points.

The complete procedure for generating such data is described in detail in [18].

V. Results of Experiments

We consider the following four algorithms, obtained by first applying feature selection and next by training a classifier on the training set restricted to the selected features.

1) RFE+SVM: RFE is applied for feature selection and SVM for classification.
2) RFE+kNN: RFE is applied for feature selection and kNN for classification.
3) RELIEF+SVM: RELIEF is applied for feature selection and SVM for classification.
4) RELIEF+kNN: RELIEF is applied for feature selection and kNN for classification.

Because of the small size of the data, LOOCV is used for a comparative analysis of the algorithms (cf., e.g., [28]). At each leave-one-out run, all but one element of the data is used as training set, and the left-out element is used for testing the predictive performance of the resulting classifier.

The number of selected features is set to 10, corresponding to the number of spikes present in the samples. The SVM soft-margin constant C is set to 10, chosen based on results of few runs on one training set (these results indicated that the value of this parameter is not crucial on this data).

A. Comparison of RFE and RELIEF

We compare the sets of features selected by RFE and RELIEF over the runs. RFE selected a total of 37 features

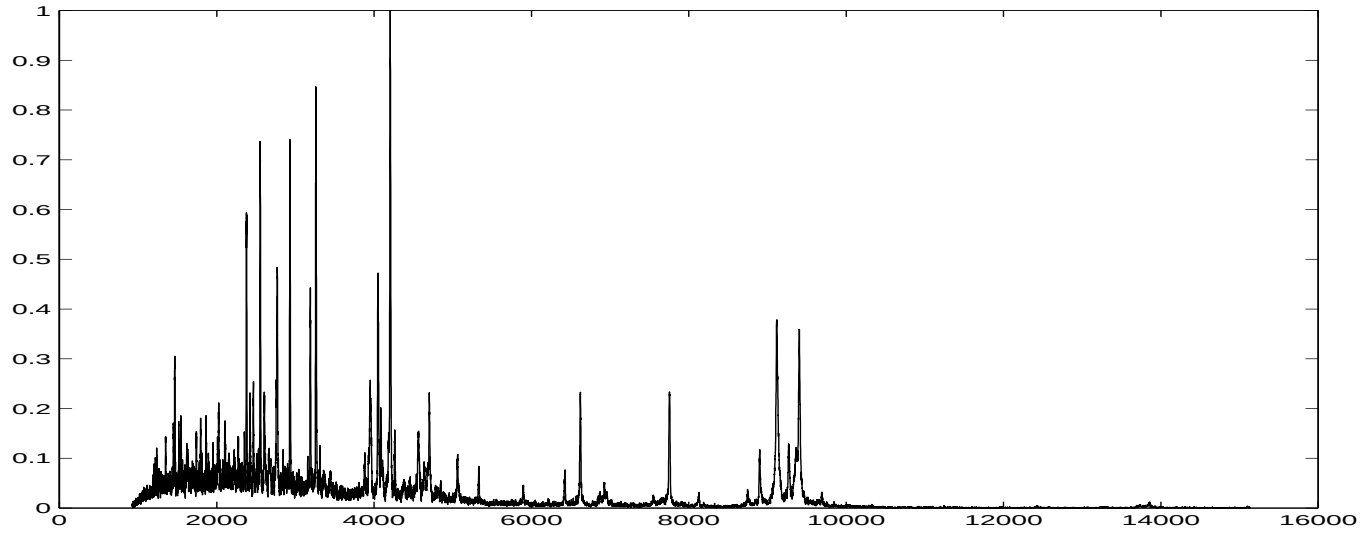

Fig. 1. A MALDI-TOF MS profile: x-axis contains m/z values of peptides and the y-axis their concentration.

over the 96 runs of LOOCV, while RELIEF only 15. Figures 2 and 3 show m/z value (x-axis) and number of occurrences (y-axis) of the selected features. The circles in the figures indicate the position of the spiked peptides. The results indicate that RELIEF is more stable than RFE and selects only features corresponding to spike m/z values, while RFE selects also few features which are not in the neighbourhood of the spikes.

The difference in stability of the two algorithms can be explained by observing that (linear) SVM training is rather sensitive to the samples considered, since the hyperplane generated by SVM is uniquely determined by the support vectors, hence it does change when one support vector is removed. On the other hand, the procedure used by RELIEF to score features is less sensitive to the removal/addition of one sample [24].

B. Comparison of RFE+SVM and RELIEF+SVM

TABLE I

AVERAGE LOOCV SENSITIVITY, SPECIFICITY AND ACCURACY OF SVM WITH RELIEF AND RFE FEATURE SELECTION, RESPECTIVELY. STANDARD DEVIATION IS REPORTED BETWEEN BRACKETS.

	SENSITIVITY	SPECIFICITY	ACCURACY
RELIEF + SVM	0.8542 (0.3567)	0.9583 (0.2019)	0.9062 (0.2930)
RFE + SVM	0.9375 (0.2446)	0.9792 (0.1443)	0.9583 (0.2009)
(RELIEF union RFE) + SVM	0.9375 (0.2446)	1.0000 (0)	0.9688 (0.1749)

We used accuracy, sensitivity and specificity as quality measures for comparing learning algorithms. Other measures, like AUC (Area Under the ROC Curve), can be used. As illustrated in [29], there is good agreement between accuracy and AUC as to the ranking of the performance of the classification algorithms.

When RFE and RELIEF are applied to select features and SVM is trained on the selected feature subsets, we obtain average LOOCV results reported in Table I. Sensitivity and specificity are the fraction of correctly classified spiked and non-spiked samples, respectively. Accuracy is the fraction of samples correctly classified. RELIEF+SVM wrongly classifies a total of 9 samples: 2 normal samples with t=48, and 7 spiked samples, 2 with t=0, 1 with t=1, 2 with t=24 and 2 with t=48. RFE+SVM wrongly classifies a total of 4 samples: 1 normal sample at t=24, and 3 spiked samples, 1 at t=0 and 2 at t=48. Only the latter two samples are wrongly classified by both algorithms.

The results indicate that a significant improvement in sensitivity is obtained by RFE+SVM over RELIEF+SVM, and a non-significant increase in specificity. The superior accuracy of RFE+SVM over RELIEF+SVM is a somewhat expected result, since RFE incorporated SVM into the feature selection process. However, it is not clear why this neat superiority is achieved in terms of sensitivity.

A slight improvement is obtained by taking the union of features selected by RFE and RELIEF as input features for SVM: perfect classification on the set of non-spiked samples is obtained, while sensitivity remains equal to the one of RFE+SVM.

C. Comparison of RFE+kNN and RELIEF+kNN

We analyze experimentally the LOOCV predictive performance of the three classifiers obtained by feature selection followed by kNN. Figures 4, 5 and 6 show plots of average

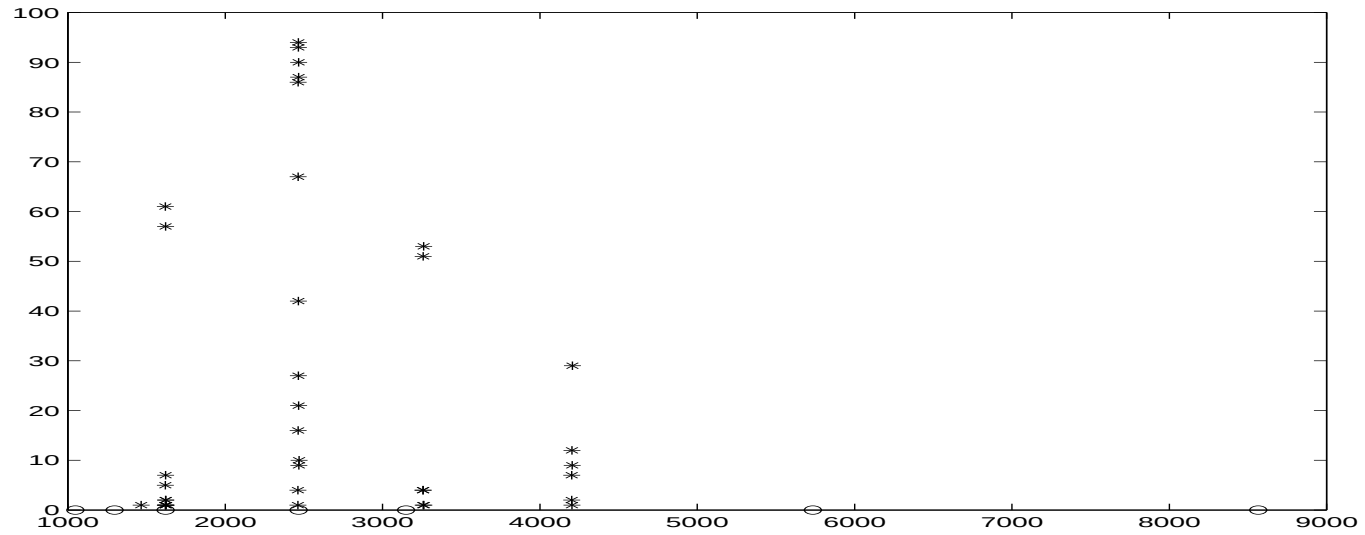

Fig. 2. Features selected by RFE: the x-axis contains the m/z values and y-axis the number of times a m/z value is selected over all the LOOCV runs. The circles on the x-axis correspond to the m/z values of spiked peptides.

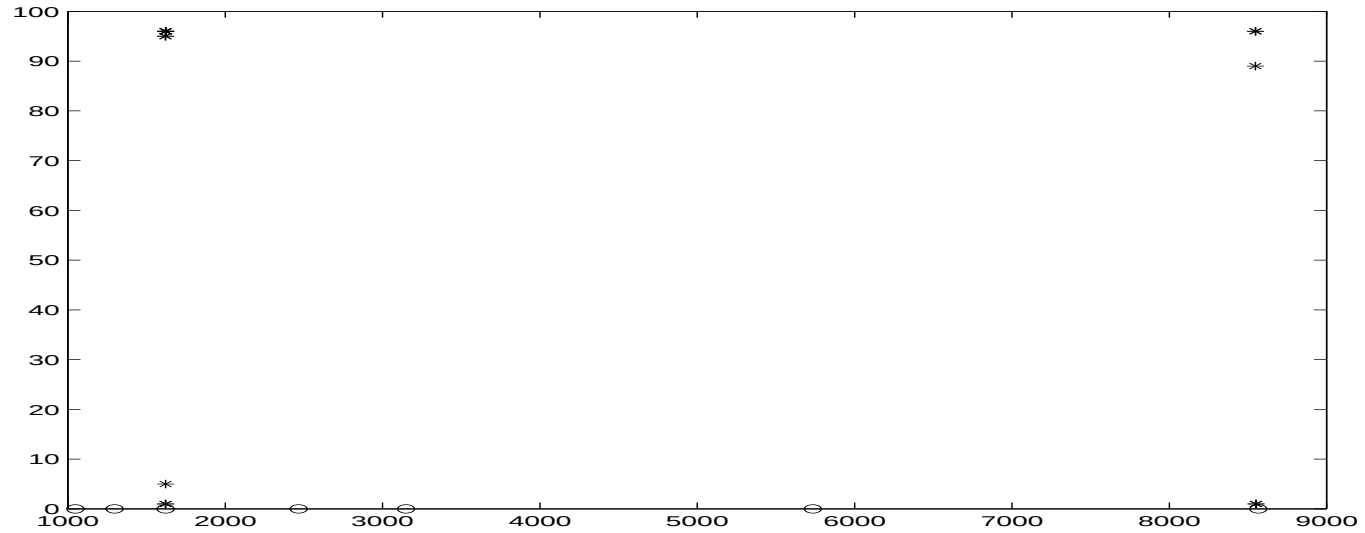

Fig. 3. Features selected by RELIEF: the x-axis contains the m/z values and y-axis the number of times a m/z value is selected over all the LOOCV runs. The circles on the x-axis correspond to the m/z values of spiked peptides.

accuracy, sensitivity and specificity of the three resulting kNN classifiers, for value of k (number of nearest neighbours) equal to 1,3,5,10,20,30,40,50. The results of experiments indicate that RFE+kNN is superior to RELIEF+kNN. For all the values of k considered, RFE+kNN obtains higher sensitivity than RELIEF+kNN, and higher specificity for values of k smaller than 40. In particular, perfect classification is achieved by RFE+1NN. The plots show that taking the union of features selected by the two algorithms affects positively specificity, which becomes equal to 1, while sensitivity remains inferior to the one of RFE+kNN.

VI. Conclusion

We have considered the issue of feature selection for classification with pattern proteomic data of mixed quality. We analyzed computationally data from a previous study about sample preparation in mass spectrometry proteomics. Two state-of-the-art feature selection techniques, RFE and RELIEF, and two popular classification techniques, kNN and SVM, have been applied to the data.

A comparative analysis of these techniques have been performed, in terms of stability, intepretability, and quality of results. The results of LOOCV experiments indicate that RFE selects features that yield better predictive performance both when SVM and kNN is used as classifier. On the other hand,

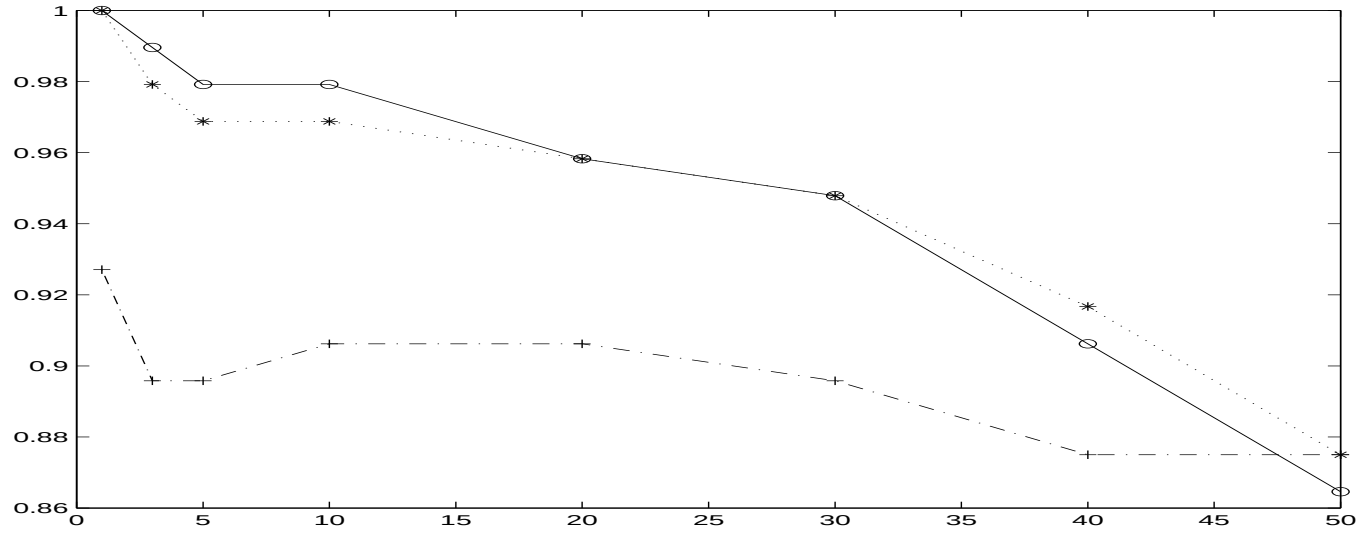

Fig. 4. average accuracy (y-axis) of LOOCV with k-NN for different values of k (x-axis): data restricted to m/z values selected by RELIEF (dashdotted line), RFE (solid line), and the union of them (dotted line).

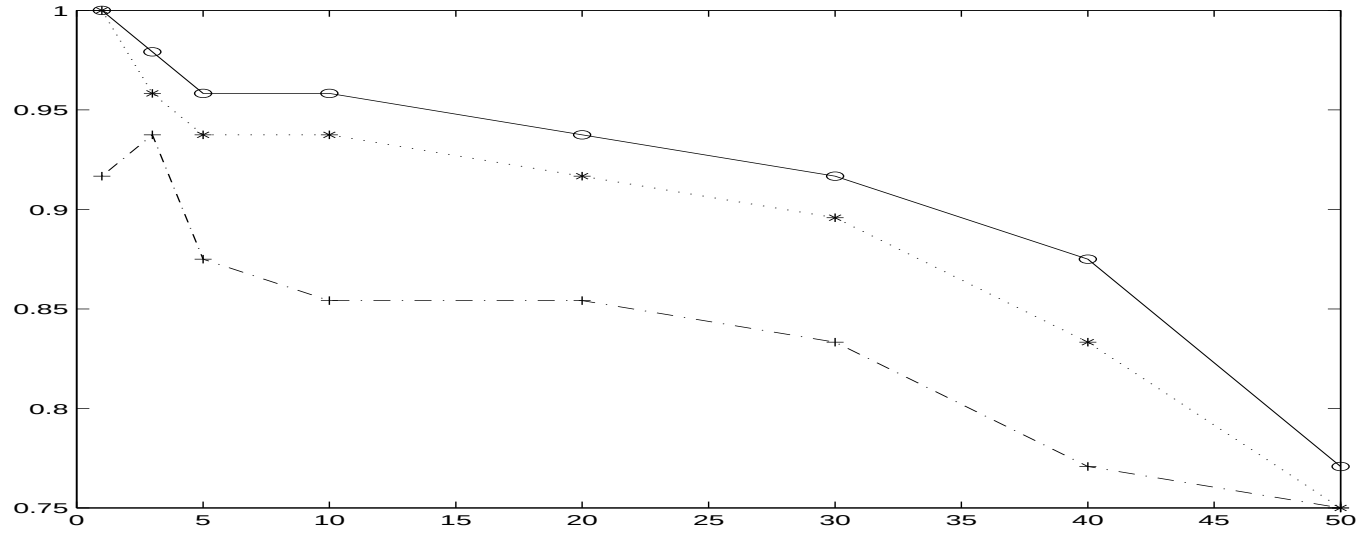

Fig. 5. average sensitivity (y-axis) of LOOCV with k-NN for different values of k (x-axis): data restricted to m/z values selected by RELIEF (dashdotted line), RFE (solid line), and the union of them (dotted line).

features selected by RELIEF are more stable over the leave-one-out runs, and consist of relevant (that is spiked) features. While features selected by RFE are more diverse over the runs and contain also features that do not correspond to spiked peptides.

In general, the results seem to indicate that better predictive performance does not necessarily correspond to stability of the method and intepretability of results. This observation fits in the ongoing discussion about results in the discovery-based "omics' research: better understanding of results does not seem necessarily to imply better prediction, and better prediction seems possible without complete understanding of the results.

REFERENCES

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Machine Learning*, vol. 3, pp. 1157–1182, 2003, special Issue on variable and feature selection.

[2] H. Lie and e. H. Motoda, *Feature Extraction, Construction and Selection: a Data Mining Perspective*. International Series in Engineering and Computer Science. Kluwer, 1998.

[3] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *International Conference on Machine Learning*, 1994, pp. 121–129.

[4] I.-S. Oh, J.-S. Lee, and B.-R. Moon, "Local search-embedded genetic algorithms for feature selection," in *16 th International Conference on Pattern Recognition (ICPR'02)*. IEEE Press, 2002.

[5] M. Raymer, W. Punch, E. Goodman, L. Kuhn, and A. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 2, pp. 164–171, 2000.

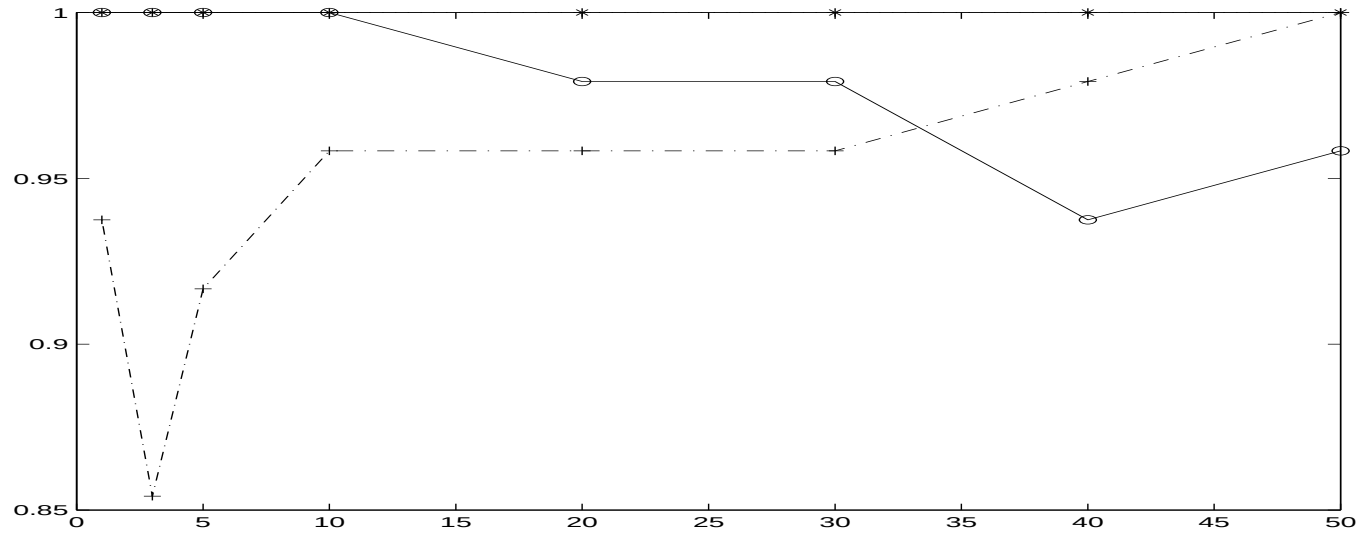

Fig. 6. average specificity (y-axis) of LOOCV with k-NN for different values of k (x-axis): data restricted to m/z values selected by RELIEF (dashdotted line), RFE (solid line), and the union of them (dotted line).

[6] E. Xing, "Feature selection in microarray analysis," in *A Practical Approach to Microarray Data Analysis*. Kluwer Academic, 2003.

[7] J. Li, Z. Zhang, J. Rosenzweig, Y. Wang, and D. Chan, "Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer," *Clinical Chemistry*, vol. 48, no. 8, pp. 1296–1304, 2002.

[8] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome Informatics*, vol. 13, pp. 51–60, 2002.

[9] Z. W. et al., "Detection of cancer-specific markers amid massive mass spectral data," *PNAS*, vol. 100, no. 25, pp. 14 666–14 671, 2003.

[10] K. Jong, E. Marchiori, M. Sebag, and A. van der Vaart, "Feature selection in proteomic pattern data with support vector machines," in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2004.

[11] P. E. et al., "Serum proteomic patterns for detection of prostate cancer," *Journal of the National Cancer Institute*, vol. 94, no. 20, pp. 1576–1578, 2002.

[12] ——, "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, no. 9306, pp. 572–7, 2002.

[13] Q. Y. et al., "Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients," *Clin. Chem*, vol. 48, no. 10, pp. 1835–43, 2002.

[14] D. Ransohoff, "Lessons from controversy: Ovarian cancer screening and serum proteomics," *Journal of the National Cancer Institute*, vol. 97, pp. 315–319, 2005.

[15] M. S., K. S., and H. C., "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488–92, 2005.

[16] E. Marshall, "Getting the noise out of gene arrays," *Science*, vol. 306, pp. 630–631, 2004, issue 5696.

[17] E. Diamandis, "Analysis of serum proteomic patterns for early cancer diagnosis: Drawing attention to potential problems," *Journal of the National Cancer Institute*, vol. 96, no. 5, pp. 353–356, 2004.

[18] M. West-Nielsen, E. Hogdall, E. Marchiori, C. Hogdall, C. Schou, and N. Heegaard, "Sample handling for mass spectrometric proteomic investigations of human sera," *Analytical Chemistry*, 2005.

[19] L. A. Rendell and K. Kira, "A practical approach to feature selection," in *International Conference on machine learning*, 1992, pp. 249–256.

[20] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Tenth National Conference on artificial intelligence*, 1992, pp. 129–134.

[21] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, 2002.

[22] V. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.

[23] N. Cristianini and J. Shawe-Taylor, *Support Vector machines*. Cambridge Press, 2000.

[24] T. Cover and P. Hart, "Nearest neighbour pattern classification," *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, 1967.

[25] T. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[26] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, 2003, pp. 856–863.

[27] H. I. et al., "SELDI-TOF MS for diagnostic proteomics," *Anal. Chem.*, vol. 75, no. 7, pp. 148A–155A, 2003.

[28] T. Evgeniou, M. Pontil, and A. Elisseeff, "Leave one out error, stability, and generalization of voting combinations of classifiers," *Mach. Learn.*, vol. 55, no. 1, pp. 71–97, 2004.

[29] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 6, pp. 1145–1159, 1997.