Applied Artificial Intelligence - Proceedings of the 7th International FLINS Conference,
D. Ruan, P. D'Hondt, P.F. Fantoni, M. De Cock, M. Nechtegael, E.E. Kerre eds.
Genova (Italy), 29-31 August 2006, pp. 557-564

FEATURE SCORING BY MUTUAL INFORMATION FOR CLASSIFICATION OF MASS SPECTRA

C. KRIER[1], D. FRANÇOIS[2], V. WERTZ[2], M. VERLEYSEN[1]

Université catholique de Louvain, Machine Learning Group
[1] DICE, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium
{krier, verleysen}@dice.ucl.ac.be
[2] CESAME Research Center, Av. G. Lemaître 4, B-1348 Louvain-la-Neuve, Belgium
{françois, wertz}@inma.ucl.ac.be

Selecting relevant features in mass spectra analysis is important both for classification and search for causality. In this paper, it is shown how using mutual information can help answering to both objectives, in a model-free nonlinear way. A combination of ranking and forward selection makes it possible to select several feature groups that may lead to similar classification performances, but that may lead to different results when evaluated from an interpretability perspective.

1. Introduction

Mass spectrometry allows identifying chemicals in a substance by their mass and charge. It produces spectra that plot the quantity of chemicals in the substance as a function of their mass to charge ratio (m/z). Typically, several thousands m/z values are considered. Such spectra are said to be high-dimensional.

For illustration purposes, the detection of cancer will be considered. The interesting question for researchers is of course which chemicals are involved in the process and which biomolecules are affected by the disease process. Two objectives are thus sought together: classification performances should be high, and the method should identify which chemicals are affected. Focusing only on features (m/z) that allow building an efficient classification model is not sufficient; indeed several sets of features could lead to similar classification performances, while one set could be of much greater interest for causality interpretability than the other ones.

Another way of identifying relevant features is to examine the statistical dependency between each of them (taken individually) and the class label. While the statistical dependency concept does not make any assumption on the model that is further used for classification, it will discard features that are only relevant in a group, and not individually.

Applied Artificial Intelligence - Proceedings of the 7th International FLINS Conference,
D. Ruan, P. D'Hondt, P.F. Fantoni, M. De Cock, M. Nechtegael, E.E. Kerre eds.
Genova (Italy), 29-31 August 2006, pp. 557-564

In this paper, we suggest to overcome these limitations by using the mutual information measure between features and the class label. Mutual information is a nonparametric, model-free method for scoring a set of features. It can be used to spot all features relevant to the classification, and to identify groups of features that allow building a valid classification model. It is applied to the detection of ovarian cancer through spectra of human serum. The process allows identifying feature sets that can be later assessed from a clinical perspective.

The paper is organized as follows: Section 2 reviews the existing literature, Section 3 introduces the concept of the mutual information. Section 4 proposes some experiments with the Ovarian Cancer dataset and Section 5 concludes.

2. Previous work

Several mass spectrometry classification algorithms have been proposed in the literature [1, 2, 3]; yet only a few studies focus on feature selection.

In a comparative study, Liu [4] considers the Chi-squared test and the t-test, making the assumption that the class populations are normal-distributed. He furthermore uses an entropy-based method, considering the mutual information between pairs of features, and between each feature and the class label. Those methods will however eliminate features relevant only in conjunction with each other.

Petricoin uses a Genetic algorithm [5] prior to probabilistic classification. This allows to find an (sub-)optimal feature subset, but fails at scoring each feature individually. The method may thus find *a* set of features adequate for classification, but not *all* sets that could be of interest. Furthermore, the procedure is model-dependent and prone to convergence issues.

Lilien [6] proposes the use of the Linear Discriminant Analysis, and Back Projection to score the features. The LDA results in a discriminant vector that is then normalized according to the initial features variances. The score associated to each feature is the corresponding element in the normalized discriminant vector. The classification model is thus constructed on all features; therefore, a lot of poorly scored (by the LDA criterion) features may have the same weight in the classification process as a single high-scoring feature. Here again, two features that are relevant only when paired will not be identified as such.

In the following section, we will see that the mutual information allows scoring groups of features, independently from a subsequent classification model, and without making any assumptions about the class sample distributions.

Applied Artificial Intelligence - Proceedings of the 7th International FLINS Conference,
D. Ruan, P. D'Hondt, P.F. Fantoni, M. De Cock, M. Nechtegael, E.E. Kerre eds.
Genova (Italy), 29-31 August 2006, pp. 557-564

3. The mutual information

The mutual information (MI) between two random variables or random vectors measures the "amount of information", i.e. the "loss of uncertainty" that one can bring to the knowledge of the other, and vice versa.

3.1. *Definition of the mutual information*

The concept of uncertainty of a random variable is expressed by its entropy [7]. Although the notion of entropy has first been developed for discrete variables, it can be extended to continuous variables rather easily. The entropy *H(Y)* of a random variable *Y* with probability density function (pdf) μ_Y is defined by

$$H(Y) = -\int \mu_Y(y) \log \mu_Y(y) dy \,. \tag{1}$$

The entropy of a random variable or vector *Y* when the value of some other random variable *X* is known is the conditional entropy:

$$H(Y \mid X) = -\int \mu_X(x) \int \mu_Y(y \mid X = x) \log \mu_Y(y \mid X = x) dydx \,. \tag{2}$$

The mutual information is the difference between the entropy of a variable and the conditional entropy *I(X,Y) = H(Y) – H(Y|X).* The mutual information can be expressed as the Kullback–Leibler divergence between the joint distribution $\mu_{X,Y}$ of the variables and the product of the marginal distributions μ_X and μ_Y :

$$MI(X,Y) = \int \mu_{X,Y}(x,y) \log \frac{\mu_{X,Y}(x,y)}{\mu_X(x)\mu_Y(y)} dxdy \,. \tag{3}$$

When *X* and *Y* are independent, the mutual information is zero; the higher the dependency between variables *X* and *Y*, the higher is their mutual information. Contrarily to the correlation, the mutual information measures any relationship between variables, and not only linear relations. In the above equations, *X* and *Y* can be random vectors instead of random variables. If Y is a binary class label, definition (3) holds. Its extension to multi-class problem is not obvious though, as an adequate class labeling has to be provided.

3.2. *Estimation*

Equations (1) to (3) are not applicable as such, as the pdf are not known in practice. The estimation of the mutual information given finite samples is thus a problem of density estimation. Density estimation can be achieved in several ways [8], for instance with histograms, kernels, B-splines, or Nearest

Applied Artificial Intelligence - Proceedings of the 7th International FLINS Conference,
D. Ruan, P. D'Hondt, P.F. Fantoni, M. De Cock, M. Nechtegael, E.E. Kerre eds.
Genova (Italy), 29-31 August 2006, pp. 557-564

Neighbours. The latter has the major advantage to be reasonably efficient for the estimation of a multivariate density (when a random vector is involved), while the other ones suffer more dramatically from the ‘curse of dimensionality’ (the required number of samples needed for the estimation grows exponentially with the dimension of the random vector).

There exists an extensive literature on density-based entropy estimators [9, 10]; recently, they have been extended to the Mutual Information by Kraskov et al [11]. The latter estimator is used in the experimental part of this paper.

3.3. *Using the mutual information for feature scoring/selection*

A high mutual information between a feature X and the class label Y thus means that feature X is relevant, regardless of the classification algorithm. However, the mutual information can be used in several ways to select (sets of) features.

First, the mutual information scores can be estimated between each feature individually (m/z) and the class label. The highest scores correspond to features that are most relevant in discriminating between the two classes. In contrast, the features with a mutual information near zero are statistically independent from the class label. The drawback of this method is that features that are relevant together but useless individually cannot be accurately spotted.

Secondly, the mutual information can be used to search for the optimal feature subset (which may or may not be the subset of optimal features) in a forward manner: the feature with the highest mutual information with the class label is chosen first. Then, pairs of features containing the already selected one and any remaining one are built. The mutual information between each of these pairs and the class label are measured; the second chosen feature is the one contained in the pair with the highest mutual information score. The procedure is then iterated until the adequate number of features has been reached. Although this procedure, which is greedy in the sense that the choice of a feature is never questioned afterwards, can lead to a sub-optimal feature subset, it performs most often efficiently, and definitely better than the previous option.

While the second procedure is good at identifying the most relevant subset, it will probably not select all features that could be relevant for the problem, as redundancy between features is avoided. Both procedures have advantages; therefore, in order to identify all features relevant as well individually as in conjunction with others, they are merged into a single one, inspired from [12]:

1. N features are selected by individual mutual information
2. M features are selected by the forward procedure
3. All 2^{N+M} possible feature subsets are constructed and their mutual information with the class label is estimated.

Applied Artificial Intelligence - Proceedings of the 7th International FLINS Conference,
D. Ruan, P. D'Hondt, P.F. Fantoni, M. De Cock, M. Nechtegael, E.E. Kerre eds.
Genova (Italy), 29-31 August 2006, pp. 557-564

The subset with the highest mutual information can be chosen for classification purposes; however, all other subsets associated to a high value of the mutual information with the class label can be considered as relevant for the problem too. In this way, several subsets of features can be identified, hopefully allowing spotting all features relevant to the problem, either individually or in conjunction with others. The subsets can thus be ranked and further application-dependent investigations performed. The values of N and M should be chosen as high as possible, while keeping the 2^{N+M} MI estimations tractable. Despite the fact that the complexity of the method is proportional to the square of the number of features in the worst case, the average number of computations is linear with the number of features. In practice, the computation of all MI values does not exceed a few tenth of minutes on a standard computer if, as an example, $N+M$ is limited to 7. Furthermore, the whole variables selection process can be performed off-line, as it does not need to be repeated to classify a new sample.

4. Example

The method is illustrated on an ovarian cancer dataset from the Clinical Proteomics Program of the U.S. National Cancer Institute [13]. The spectra result from SELDI-TOF experiments. The healthy samples come from women showing risks of cancer from a clinical perspective, while the positive cancer samples come from women with various tumors types and severity (see [5] for details.) To get a tractable number of feature subsets to assess, three features were chosen by the forward selection method; then, four other ones were chosen among the highest scored features not already selected. To assess the relevance of the selected features, a linear classification is performed on a test set, as in [6].

4.1. *Results*

Figure 1 shows the mutual information score for each m/z value. The vertical lines indicate which features were chosen by the forward strategy and not by the ranking procedure (in this case, the m/z ratios 2.7921 and 24.2851). Few features have really high mutual information scores. Note that negative values are obviously the result of the estimates bias (without consequence on the ranking) and variance (that gives an idea of the estimator accuracy).

The final set of selected features is given in Table 1, with the corresponding m/z values. Feature 1679 has the highest mutual information with the class label. The features selected by the ranking method are obviously highly correlated; nevertheless, we will see that they are not totally equivalent for classification.

Applied Artificial Intelligence - Proceedings of the 7th International FLINS Conference,
D. Ruan, P. D'Hondt, P.F. Fantoni, M. De Cock, M. Nechtegael, E.E. Kerre eds.
Genova (Italy), 29-31 August 2006, pp. 557-564

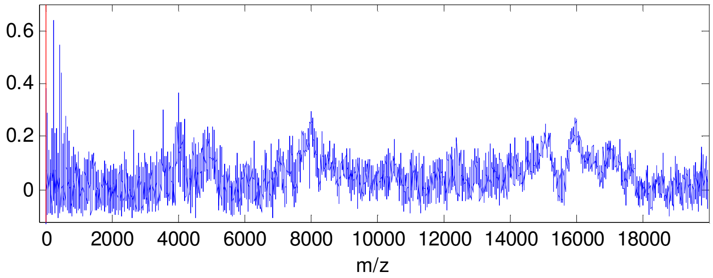

Figure 1. Mutual information for each m/z feature.

Table 1. The seven selected features, along with their corresponding m/z values. An O in regard of the name of a method indicates that the feature was selected by the method.

Feature	*181*	*530*	*1678*	*1679*	*1680*	*1681*	*1682*
m/z	*2.7921*	*24.2851*	*244.6604*	*244.9525*	*245.2447*	*245.5370*	*245.8296*
Forward	O	O		O			
Ranking			O	O	O	O	O

The mutual information of each possible of the 128 feature subsets is given in Figure 2, along with the performances of a linear classifier built using that subset. The feature groups are ordered by increasing mutual information. Table 2 presents some of those feature groups. Figure 2 confirms that the classification performances are highly correlated to the mutual information (0.9041).

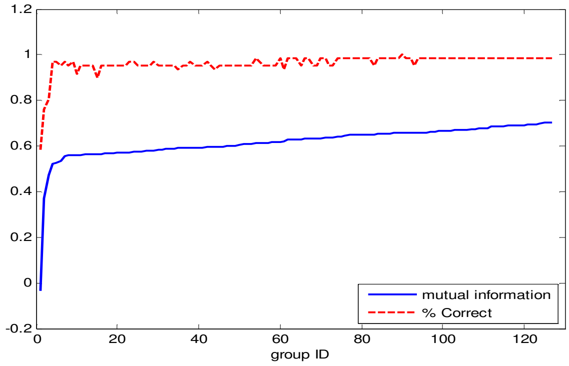

Figure 2. The mutual information and classification performances for a linear classifier built on every possible subset of the selected features.

4.2. *Discussion*

From the analysis of Table 2, it appears that:

- The group of features achieving the best classification is not the group of most individually relevant features nor it is the group identified by the

Applied Artificial Intelligence - Proceedings of the 7th International FLINS Conference,
D. Ruan, P. D'Hondt, P.F. Fantoni, M. De Cock, M. Nechtegael, E.E. Kerre eds.
Genova (Italy), 29-31 August 2006, pp. 557-564

Table 2. Some values from Figure 2.

Group ID	Feature group	Mutual information	% Correct classification
126	181	0.3694	74.86
127	530	-0.0349	58.00
118	1678	0.3694	75.86
113	530 ; 1678	0.5644	90.00
38	181 ; 530 ; 1678	0.6571	100.00
1	181; 1678; 1679; 1680; 1682	0.7026	98.43
33	181; 530; 1678; 1679; 1680; 1682; 1681	0.6585	98.43
59	1678; 1679; 1680; 1681; 1682	0.6347	95.31
34	181; 530; 1680	0.6583	98.43
35	1678 ; 1679	0.6581	95.23

forward procedure. Group 38 is the best group and contains only one of the highest-ranked features. Furthermore, the group discovered by the forward procedure achieves less good classification performances; this is because the choice of the first feature was never questioned. Using a forward or ranking procedure alone does not lead to the optimal feature subset.

- Some individually less relevant features help building more accurate classifiers than if using individually relevant features only. Features 530 (Group 127 – low MI) 1678 together reach 90% of correct classification (Group 113), while feature 1678 (Group 118) classifies only 76% of the samples correctly. It can thus be assumed that feature 530 is involved in the process. Only ranking features may prevent from spotting relevant ones.
- There are groups of different features that achieve very similar results. For example Groups 34 and 35 share no variable, although their classification performances (95 and 98 %) are rather close. Simply relying on the best feature subset according to the mutual information or to some model-based algorithm does not allow recovering all features involved in the process.
- The performances in classification reached by the method are similar to the results obtained by Lilen with LDA and Back Projection [6].
- In this problem, it appears that only features with low m/z ratio are relevant.

5. Conclusion

This paper shows that using the mutual information between features and class label in mass spectra analysis help choosing relevant feature sets. The method based on the combination of feature ranking and forward selection, and using mutual information on sets of features rather than individually, makes it possible to rank feature subsets. Then, an application-driven procedure can be used to assess the (clinical in this example) relevance of the feature sets, starting from the highest-ranked ones by the proposed procedure.

Acknowledgments

M. Verleysen is Research Director of the Belgian FNRS. C. Krier and D. François are funded by a Belgian FRIA grant. Parts of this research results from the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The scientific responsibility rests with its authors.

References

1. BL. Adam et al., Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, *Cancer Research*, **62**, 3609 (2002).
2. G. Ball et al., An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics*, **18,** 395 (2002).
3. H. Zhou et al., Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry, *Nature Biotechnology*, **20,** 512 (2002).
4. 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* **13,** 51 (2002)
5. E. Petricoin III et al., Use of proteomic patterns in serum to identify ovarian cancer, *The Lancet*, **359** 572 (2002)
6. R. H. Lilien, H. Farin, B. R. Donald, Probabilistic Disease Classification of Expression-Dependent Proteomic data from Mass Spectrometry of Human Serum, *Journal of Computational Biology* **10(6)**, 925 (2003).
7. C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, *University of Illinois Press*, Urbana, IL, 1949.
8. D.W. Scott, Multivariable Density Estimation: Theory, Practice, and Visualization, *Wiley,* New-York, 1992.
9. R.L. Dobrushin, A simplified method of experimental evaluation of the entropy of stationary sequence*, Theory Prob. Appl.* **3(4)**, 462 (1958).
10. O. Vasicek, Test for normality based on sample entropy, *J. Royal Statist. Soc*. **B38**, 54 (1976).
11. A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, *Phys. Rev*. **E69:**066138 (2004).
12. F. Rossi, A. Lendasse, D. François, V. Wertz, M. Verleysen, Mutual information for the selection of relevant variables in spectrometric nonlinear modeling, *Chemometrics & Intelligent Lab. Systems*, **80**, 215-226 (2006).
13. http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp