BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 2 2009, pages 258–264
doi:10.1093/bioinformatics/btn599

Data and text mining

Bag of Peaks: interpretation of NMR spectrometry

Gavin Brelstaff[1,*], Manuele Bicego[2,†], Nicola Culeddu[3] and Matilde Chessa[4,‡]

[1]Biocomputing, CRS4, 09100 Pula (CA), Sardinia, [2]DEIR, University of Sassari, via Torre Tonda 34, 07100 Sassari, [3]ICB-CNR, 07040 Li Punti, Sassari and [4]Porto Conte Ricerche, Loc. Tramariglio, Alghero, Italy

Received on July 9, 2008; revised on November 12, 2008; accepted on November 16, 2008

Advance Access publication November 18, 2008

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The analysis of high-resolution proton nuclear magnetic resonance (NMR) spectrometry can assist human experts to implicate metabolites expressed by diseased biofluids. Here, we explore an intermediate representation, between spectral trace and classifier, able to furnish a communicative interface between expert and machine. This representation permits equivalent, or better, classification accuracies than either principal component analysis (PCA) or multi-dimensional scaling (MDS). In the training phase, the peaks in each trace are detected and clustered in order to compile a common dictionary, which could be visualized and adjusted by an expert. The dictionary is used to characterize each trace with a fixed-length feature vector, termed Bag of Peaks, ready to be classified with classical supervised methods.

Results: Our small-scale study, concerning Type I diabetes in Sardinian children, provides a preliminary indication of the effectiveness of the Bag of Peaks approach over standard PCA and MDS. Consistently, higher classification accuracies are obtained once a sufficient number of peaks (>10) are included in the dictionary. A large-scale simulation of noisy spectra further confirms this advantage. Finally, suggestions for metabolite-peak loci that may be implicated in the disease are obtained by applying standard feature selection techniques.

Availability: Matlab code to compute the Bag of Peaks representation may be found at http://economia.uniss.it/docenti/bicego/BagOfPeaks/BagOfPeaks.zip

Contact: gjb@crs4.it

1 INTRODUCTION

Distinguishing diseased from healthy subjects presents a recurrent challenge in biomedical data analysis (Lindon *et al.*, 2007), when typically spectrometric samples are available via nuclear magnetic resonance (NMR) or mass spectrometry (MS). By providing an intermediate representation it may be possible to better assist the expert's interpretation of data produced by such devices. This article applies this idea to data acquired by high-resolution proton NMR spectrometry, proposing an intermediate representation based on peaks.

*To whom correspondence should be addressed.

†Present address: Dip. di Informatica, University of Verona, Strada Le Grazie, 15 - 37134 Verona, Italy.

‡Present address: Vincenzo Migaleddu - Imaging s.rl. Viale Caprera n. 3/A, 07100 Sassari, Italy.

NMR spectrometry (Ernst *et al.*, 1990) represents an useful tool for clinical diagnosis and toxicology investigation, since it indicates the metabolic composition of biofluids, such as blood plasma or urine. The standard spectrometer produces a highly detailed single-variable trace over a large range of 'chemical shifts', derived as a digital transformation of the spectrum of radio-frequency free induction decays (FIDs). Depending on the chemical environment of their source nuclei different metabolic species result in different peaks, or groups of peaks. Thus, hundreds of chemical compounds may be revealed by a single act of measurement. Ideally, each peak could be automatically identified with a single species, the area under its curve computed and thus the relative concentration of each species estimated. In practice, peaks often deviate from their true spectral loci due to pH or ionic interactions, or due to minor variations in the preparation or acquisition processes. Therefore, simply locating peaks and looking up the corresponding metabolic species may not be sufficient for the correct interpretation and analysis of the data (even if commercial databases —e.g. Amix— now model variations due to pH changes). In fact, the interpretation may be confounded by non-pH effects, or by the fact that peaks are of finite spread and may overlap. Low amplitude peaks may be entirely lost, whilst more prominent metabolites can still be extracted. Between those two extremes there is a 'grey region', containing partially deformed and shifted peaks. The challenge is therefore to assist the expert in their interpretation, in order to analyse non-obvious metabolites yet not implicated in a particular pathology, or toxin. Indeed, at any peak locus there may be many plausible candidates, and distinguishing between them requires expert knowledge of the biofluid.

2 SYSTEM AND METHODS

2.1 Automated approaches to assist the expert observer

Automated analyses generally assist the expert observer in two distinct ways: (i) by transforming the data in order to visualize meaningful patterns and (ii) by reducing the amount of the data that needs to be subsequently examined. Although these are often combined as a single algorithmic process it can be useful to consider their effectiveness individually. For example, for visualization purposes it is effective to apply chemometric preprocessing techniques in order to calibrate and normalize the signal. This may permit an appropriate visual comparison of features across different NMR traces. On the other hand, dimensionality reduction techniques—such as a principal component analysis (PCA)—may exploit the structure of the dataset in order to project each trace on to a point lying in a low-dimensional feature space (Keun, 2006). In such space, distinctive clusters may emerge. In fact, it

is common practice to plot the traces in a 2D or 3D space (retaining only the first few principal components), in order to visualize and identify meaningful classes (e.g. 'healthy', 'diseased'). Nevertheless, such a complementary scenario between data-reduction and visualization seldom persists in a realistic study, where the classes tend to overlap.

2.2 Algorithmic approaches that capture expertise

Techniques like PCA appear as an attractive starting point because results may immediately be obtained and visualized. Nevertheless, even if widely applied, PCA-centric techniques may be not the best choice for NMR traces for two reasons. First, their basic assumption scarcely holds: statistical variation is not dominated by additive Gaussian noise in amplitude but rather by unpredictable horizontal, left–right shifts in spectral loci of peak features. Second, and most important, such techniques are problematic (due to their intrinsic *unsupervised* nature): the expert is kept out of the analysis until processing is complete, which could possibly lead to inaccurate or arbitrary decisions. Similar criticisms may apply to other techniques widely employed in this field [see Lindon *et al.* (2001) for a review]: algorithmic refinements of PCA (e.g. Nipals, Press and VariMax), non-linear mappings, hierarchical cluster analysis—where data are visualized as a dendrogram rather than a scatter plot—and others.

Thus, it makes sense to provide the expert with an intermediate representation of data, which permits him to interact with the data at an intermediate stage in the computation. Such intermediate knowledge represents a supplement of the standard training set that is required by any supervised classifier, and contains data samples classified by the supervising expert (typically with the aid of an objective medical examination of chosen patients).

2.3 Intermediate representation based on visible peaks

When examining NMR traces an expert tends to reason on the basis of visible peaks, not troughs or indistinct undulations. Therefore, it seems reasonable to use peaks to define an intermediate representation. Similar considerations have been given in the MS case (Tibshirani *et al.*, 2004), even if their application differs in several significant details.

Our starting observation is that not only large amplitude peaks have a role, but any peak with a well-defined structure—namely any peak having at least one visible flank (i.e. a monotonic fall-off of signal beneath the half-height of the peak). Theory indicates that in ideal spectral isolations each peak should follow a Lorentzian profile; nevertheless, our tests show that, in practice, well-defined peaks can be satisfactorily approximated by fitting the simpler Gaussian function across their visible extent. Moreover, we noted that the theoretical-predicted precise binomial relationships between peak structures (including doublets and triplets) were seldom observed in practice. Thus, it makes sense to model peaks individually before reasoning with them, rather than trying to fit elaborate multi-peak physics-based models directly to the traces data—as some Bayesian approaches might prescribe (Bretthorst *et al.*, 2005). Thus, in our approach the intermediate representation of each trace is based on the set of its well-defined peaks. Once approximated with the Gaussian model, each peak is represented by the following parameters:

- p—the spectral locus of the maximum,
- a—the amplitude at maximum,
- w—the width, estimated from available half-height loci and
- b_l, b_r Boolean indicators of left and right flank existence,

where p and w are measured in parts-per-million (p.p.m).

This representation can be computed in a single pass. This strategy is similar to that adopted by some spectral databases, except that it also codes for peaks lacking one flank. In such a case w is estimated by doubling the interval between the peak locus and the remaining half-height locus. Accepting the shortcoming that such parameters are known with less certainty permits some partial peaks in spectral 'grey region' to continue to be investigated. The spectral energy of each peak (be it partial, or full)

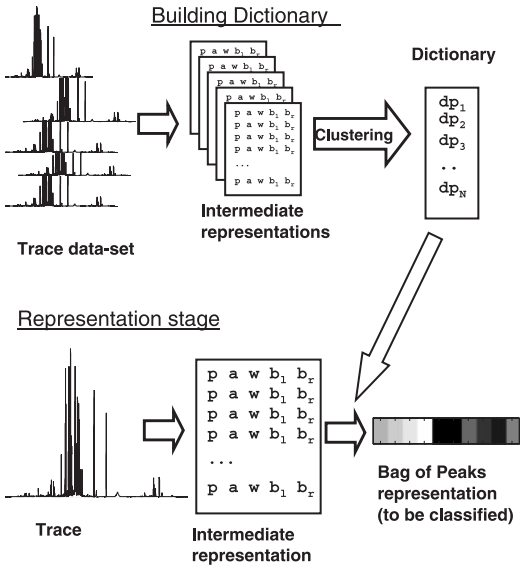

Fig. 1. Computing the Bag of Peaks.

is quantified using the approximation $a.w$, which represents the value of integral beneath the underlying profile, if the peak is indeed a Gaussian. This representation might also be motivated as a form of adaptive sampling, such as 'intelligent bucketing' (Lefebvre *et al.*, 2004).

3 ALGORITHM

3.1 Bag of Peaks approach

Bag denotes, here, an orderless collection of local features. The term has been drawn from the field of linguistics, in which the 'Bag of Words' has established itself as a practical intermediate representation (e.g. Cristianini *et al.*, 2002; Joachims, 1998; Lodhi *et al.*, 2001). Where linguists characterize a document on the basis of the occurrences of a certain set of words in its text, we may characterize a NMR trace on the basis of the occurrences of particular peaks. As with words and documents, we may compile a dictionary of the peaks intended to be representative of those found in a collection of traces. That dictionary, computed off-line, forms the basis by which a representation is obtained, and on which supervised classification is carried out. We call this approach 'Bag of Peaks', following others, working in image or object classification who coined the terms 'Bags of Keypoints' and 'Bags of Features' (e.g. Csurka *et al.*, 2004; Lazebnik *et al.*, 2006; Zhang *et al.*, 2007). As detailed in the following section, the main novelty for NMR traces is that the feature vectors are not populated by histogram values, but by accumulating the integral under each peak. Figure 1 sketches the computational stages involved, as detailed below.

3.2 Building the peak dictionary

A natural way to build the dictionary is by clustering peaks, supplied in a training set selected by the expert. We achieve this on the basis of similarity of peak locus, p, as we describe below. It is well known that no single best technique for clustering exists (Hartigan, 1975; Jain *et al.*, 1999), but their suitability generally depends on type of data, prior knowledge and dimensionality. Here, we adopt a simple yet effective technique, very similar to the Single Linkage

Tree algorithm (Hartigan, 1975). Although somewhat limited with respect to other clustering techniques, this approach permits—as we require—an interpretable interaction by the expert. Later, we show it performs fairly well against other clustering techniques.

In our clustering technique, a single threshold θ controls the process of clustering. The dictionary entries (clusters) emerge as a product of the following method:

(1) Start with a dictionary $\mathcal{D}$ containing a single peak (chosen at random).

(2) For each peak p_i in the training set find, in the dictionary $\mathcal{D}$, its closest neighbour:

$$p_{\min} = \arg\min_{p' \in \mathcal{D}} ||p_i - p'||$$

where $||x||$ is L_1 norm (adopted for simplicity).

(3) Let $d_{\min}$ be the interval between loci p_i and $p_{\min}$, namely $d_{\min} = ||p_i - p_{\min}||$. If $d_{\min} > \theta$ then add a new entry to the dictionary:

$$\mathcal{D} = \mathcal{D} \cup p_i.$$

(4) Otherwise the dictionary entry $p_{\min}$ is updated with the value of p_i, such that it maintains the average value (mean or median) of all the peaks so far in that cluster.

At this point our intermediate representation permits an interaction, revealing all its value. The expert can visually inspect the dictionary and compare it with some of the original traces, or with profiles drawn from a metabolite database. For that purpose, it is useful to project each entry of the dictionary on to a spectral plot. If this plot appears to omit important entries the expert might then vary the value of θ, or change the composition of the training set. Finally, if the expert considers it necessary, he can always edit the dictionary to impose a priori knowledge of the biofluid—e.g. he may decide to split a single entry into two, or to create a new entry nearby. This is analogous to a linguist distinguishing two distinct meanings of a word (like 'like'). The incremental nature of the above method means that, in many circumstances, the dictionary may be rapidly recompiled.

The threshold θ may be chosen by using an assistant algorithm, which employs clustering validation indexes (Jain and Dubes, 1988) to select the appropriate value. The goal is to choose the clustering (i.e. the dictionary), which provides groups which are compact (minimizing scatter within each cluster) and best separated. Different dictionaries are generated in turn by varying θ. The one minimizing the Davies–Bouldin index (Davies and Bouldin, 1979) was chosen. For a given clustering $\mathcal{C}(\theta)$, such index is defined as:

$$DB(\mathcal{C}(\theta)) = \frac{1}{K}\sum_{k=1}^{K} R_k \qquad K \text{ is the number of clusters} \qquad (1)$$

where

$$R_k = \max_{j \neq k}\left\{\frac{S_k + S_j}{d_{ij}}\right\} \qquad (2)$$

S_i is the scatter (or dispersion) *within* the cluster i, and d_{ij} is the scatter (or distance) between cluster i and cluster j.

3.3 Bag of Peaks descriptors

The resultant dictionary serves as the basis to compute the feature vector used to characterize every trace in the dataset. This descriptor, which we call 'Bag of Peaks', is computed for all training samples (in the training phase), and for each testing trace (in the testing phase).

Recall, the original Bag of Words approach aims to characterize a document with a vector that stores the word-count, for each dictionary entry. Here, instead of simply counting the number of peaks corresponding to a dictionary entry, the corresponding intensities are accumulated. This is achieved as follows:

1. Initialize to zero an accumulator value for each dictionary entry: $\forall k, A_k = 0$.
2. For each peak P_i in the trace, find (as before) its nearest dictionary entry dP_k, and add to its accumulator the energy under that peak $A_k = A_k + (a.w)_i$.

The bag of peaks descriptor of a trace j is, therefore, the vector $[A_1, ..., A_K]^T$ (where K is the number of the dictionary entries). Each trace is, therefore, projected in a low (K)-dimensional space (K-dimensional), where the classification task may be performed using any standard technique. The experiment below indicates this space to be highly discriminant, when using simple or complex classifiers.

Note, with this descriptor each trace is represented only by those peaks that occur in the dictionary, and by their respective accumulator values A_k. This thus encodes which important peaks are present and approximates their magnitude.

Finally, it is worth noticing that the accumulation mechanism emphasizes the importance of the choice of the dictionary length: too small a number of entries in the dictionary may imply that many distinct peaks are summarized by a single dictionary entry, which differs from standard NMR analysis (where each peak is analyzed individually). On the other hand, too large a number of entries may lead to a rather poor descriptor, having abandoned the original spirit of the Bag of Words.

4 IMPLEMENTATION

The proposed approach has been tested in a classification task involving 35 Sardinian under 10-year-old children. The goal was to classify the NMR traces derived from their urine samples in two classes (children having or not Type I diabetes). Each sample was analysed by an AVANCE 600MHz spectrometer (Bruker Milan, Italy) at 300K operating at 600.13 MHz in ^{1}H observation mode. To each 400 μl sample aliquot was added 200 μl of sodium phosphate buffer (0.2 M Na_2HPO_4 in H_2O and 0.2 M NaH_2PO_4 in 80:20 H_2O:D_2O, pH 7.4) containing 1 mM sodium trimethylsilyl [2,2,3,3-2H4] propionate (TSP) and 3 mM sodium azide. We used the Eutech Cyberscan 6000 pH meter to measure the pH variations of samples. In order to standardize the experimental conditions the pH of the samples was corrected to a value of 7.4 (DCI or NaOD). Samples were centrifuged at about 1800 *g* for 5 min to eliminate solid debris. NMR acquisition was performed using the first increment of a NOESY sequence with irradiation of the water frequency during the mixing time and relaxation delay, and adopting 128 FIDs, of 64 K data points, over a spectral width of 12376 Hz. All spectra were manually phased; moreover the linear baseline was corrected using the COMET standard (Lindon *et al.*, 2003) routine within TOPSPIN 2.0 (Brucker, Germany) (Schorn, 2002). On receipt of the data, we trimmed each trace to the operating range $[-0.2, 10]$ p.p.m.

Following the procedure described in Section 2.3, we extracted the intermediate peak representation in a single computational pass. Next, a basic Q/A step excluded any trace not having (i) a symmetric profile at the zero p.p.m. locus, or (ii) a peak in the interval [3.035, 3.055] p.p.m. (the typical location of 'Creatine–Creatinine pair' appearing in urine samples). Three of the 35 traces failed the test. The remaining 32 traces were then used in a preliminary evaluation of the Bag of Peaks approach. Each trace was first normalized such that the highest peak reached unit amplitude at maximum. Then, only the parameters of the 10 highest peaks were used to build the dictionary. At this stage we asked our expert to examine the selected peaks and participate in the construction of the dictionary. Their task was to ensure that those metabolite peaks that unmistakably correspond across different traces get assigned to the same dictionary entry. Clearly, the expert will prioritize the peaks for which this matters most. One of the key step in the process of building the dictionary is to set threshold θ, which produces a particular clustering (and dictionary). By working on this parameter the expert may adjust the dictionary to his satisfaction. In our experiment, by a trial and error session our expert arrived at a value of $\theta = 0.005$ which produced a total of 56 dictionary entries. At this point the expert may manually edit the dictionary or employ the assistant algorithm detailed above in order to further refine his choice of θ. In our case, he declined the former and applied the latter to finalize a dictionary of 33 entries ($\theta = 0.02$).

The resulting dictionary served as the basis on which to compute the Bag of Peaks representation of each trace. The suitability of the proposed description was tested in different classification experiments as described below. The first experiment assesses the performance of the method using the whole dataset, comparing them with those of alternative feature extraction techniques. The second identifies the minimum number of peaks needed to discriminate, thus helping in identifying candidate metabolites. Finally, the third experiment simulates a large-scale study comparing the Bag of Peaks approach and the PCA in presence of randomized peak shifts and missing peaks.

4.1 Experiment 1

Four different standard classifiers were used to benchmark the efficacy of the Bag of Peaks representation; the methodology was compared with the standard PCA approach (Jollife, 1986), which is the common baseline choice in the NMR literature (Stoyanova and Brown, 2001), as well as to the *multi-dimensional scaling* (MDS—Cox and Cox, 1994; Kruskal, 1997). MDS is a more sophisticated dimensional reduction technique often used to obtain an appropriate representation of the patterns from the given proximity matrix. It attempts to embed n patterns as points in a d-dimensional space, while keeping the distances between patterns as similar to the input dissimilarity matrix as possible. For a given d, the algorithm minimizes a stress value, which measures the similarity between the given proximity matrix and the inter-point distances of the output pattern matrix. Our implementation relies on the PRTOOLS `mds.m` function (Duin *et al.*, 2004), with pairwise distances between traces being computed using the Euclidean distance.

The reductions were performed using the same set of traces that had earlier passed the Q/A step. PCA was computed in two ways: first retaining 99.9% of the variance, and second employing the standard Cattell's scree test (Cattell, 1966). MDS was optimized using pseudo-Newton procedure, the dimension of the resulting space was set to the maximum possible, i.e. $N-2$, where N is the number of samples.

Since our N is relatively small from a statistical perspective we carried our evaluations using the well-known cross-validation technique called as *Leave-One-Out* (Theodoridis and Koutroumbas, 1999). This is well suited to small datasets, since it is able to assess the generalization capability of the classifiers (training and testing sets are separated), while maintaining a maximally sized training set. It proceeds as follows: given a dataset of N samples, $N-1$ of them are used to train the classifier, employing the one-left-out for the test. Then, a different sample is left out for testing while training is performed with the remaining $N-1$ samples. This is repeated until all N samples have been left out and tested; the accuracy reported represents the percentage of correctly classified patterns over all samples. The four classifiers used ranged from basic to relatively sophisticated:

- `1-nn`: *Nearest Neighbour* (Fukanaga, 1990). An unknown object is assigned to the same class of the nearest point in the training set (nearest neighbour).
- `k-nn`: *k-nearest neighbour rule* (Fukanaga, 1990). An unknown object is assigned to the most populous class in the k nearest points in the training set. In our case k is estimated on the training set by minimizing the leave-one-out classification error.
- `loglc`: *Logistic Linear Classifier* (Hastie *et al.*, 2001). This models the log-odds (logarithm of the ratio of class posterior probabilities) as linear functions. The weights of the classifier are optimised by maximum likelihood.
- `rbsvm`: *Radial Basis Support Vector Machine* (Schölkopf and Smola, 2002). This is the ν-SVM rule applied to a Gaussian kernel. ν is estimated by the leave-one-out nearest neighbour error on the training set. The scale σ of the Gaussian kernel is determined in a 20-step optimization based on the 5-fold cross-validation error estimation.

All the code was written in Matlab, with the support of the PRTOOLS Matlab toolbox (Duin *et al.*, 2004).

Table 1 shows the results obtained for all four classifiers above using six different input representations: PCA retaining 99.9% variance (29 PCs), PCA controlled by the Cattell Scree test (4 PCs), the MDS (29 dims) and three different versions of our Bag of Peaks representation. In the first the dictionary was generated by the incremental Single Linkage Tree algorithm described in Section 3.2 (resulting in 33 peaks), in the second by K-means (28 peaks) and in the third by the Complete Linkage Tree algorithm (22 peaks) (see

Table 1. Experiment 1: leave-one-out accuracies for the four classifiers and different input representations: PCA, MDS and Bag of Peaks (BOP)

Representation	`1-nn` (%)	`k-nn` (%)	`loglc` (%)	`rbsvm` (%)
PCA (99.9% variance)	84	84	47	84
PCA (Scree test)	87	81	87	81
MDS	84	84	59	84
BOP (Single-Link)	94	94	84	94
BOP (K-means)	97	97	59	97
BOP (Complete-Link)	91	91	72	84

Hartigan (1975); Jain *et al.* (1999) for a review of these methods). In all experiments the number of clusters were computed using the Davies–Bouldin index.[1]

Note: (i) For the Bag of Peaks representations the accuracies obtained are generally higher than those of other representations—with one exception, which occurs when applying the `loglc` classifier. But that classifier might be considered a poor indicator since it performs worst overall. (ii) The 1-NN classifier is the best, or equal best consistently over all six representations. Performance of the other three classifiers varies much with input representation. (iii) Among the three Bag of Peak representations K-means performs slightly better but it is not so useful to us since it does not facilitate user interaction. (iv) To ensure unbiased comparison, in no case was the dictionary edited.

4.2 Experiment 2

Even at low dimensionality the PCA is considered a useful tool, because it ranks its components in order of importance: the first explains most variance, the last the least. A similar characteristic might be expected for the dictionary used by the Bag of Peaks—some entries will be more discriminating than others—but how many dictionary peaks are needed in order to achieve sufficient accuracy? To this end, we extend the comparison made above, performing an analysis over an increasing number of dimensions.

First, we describe the feature selection mechanism by which we choose to rank our dictionary entries. We adopt the forward feature selection scheme (FFSS) (Bishop, 1995) using the classification accuracy of each classifier as the optimality criterion—these kinds of schemes are typically referred to as *wrappers* (Kohavi and John, 1997). The FFSS starts by considering each feature individually and selecting the one producing the optimum value for the criterion. At each successive stage of the algorithm, one additional feature is added to the set, i.e. the one giving rise to the largest increase in the optimality criterion. As a result, a collection of sets of features of growing cardinality are extracted.

Experiment 1 was thus repeated by employing sets of increasing dimensionality (retaining the first PCA component, the first two and so on—the same for MDS and our Bag of Peaks). Again, we compute the leave-one-out accuracies using the same four classifiers as before. The accuracies of all classifiers are averaged and reported in Figure 2.

Note that when the dictionary is composed of more than 10 peaks, the accuracy for the Bag of Peaks grows to be significantly better than for the PCA and the MDS. Before that it is relatively poor: i.e. in order to discriminate the algorithm needs a descriptor of sufficient dimension. Again no expert editing was permitted.

4.3 Experiment 3

Even if a proper preprocessing helps in avoiding errors due to signal distortion, other types of variations may be present in the signal. First, the precise trace location of any metabolite peak is typically confounded by a random left–right shift despite corrective controls for pH. Our approach, that makes provision for such variation, ought to out-perform PCA-based approaches that are designed to best combat Gaussian amplitude noise. Second, it may be possible for peaks to appear or disappear in response to uncontrolled aspects of diet that modify the composition of biofluid. Since we directly encode peaks, this may perturb our method more than the others. Clearly, a proper demonstration of these two conditions requires a large statistical study beyond the scope of this article. In the anticipation of that study we have simulated two larger datasets that exhibit both variations described above: random shifts and missing peaks. We bootstrapped the original training set—in a way similar to that done in phylogeny (Felsenstein, 1985)—by generating additional traces obtained by randomly injecting these two kinds of variation. In the first set, for any trace, N_S random peaks are selected for shifting, taking care of minimizing the discontinuities. The value of each (randomly left or right) shift is equal to γw, with γ a positive scalar parameter. In the second set, for any trace, N_R random peaks are selected and removed. Figures 3 and 4 show two examples of noisy spectra (only a part of the spectrum is displayed).

[1]For K-means, this number varies over different runs, depending on the initialization of the algorithm. The most frequent result employed in these tests is 28.

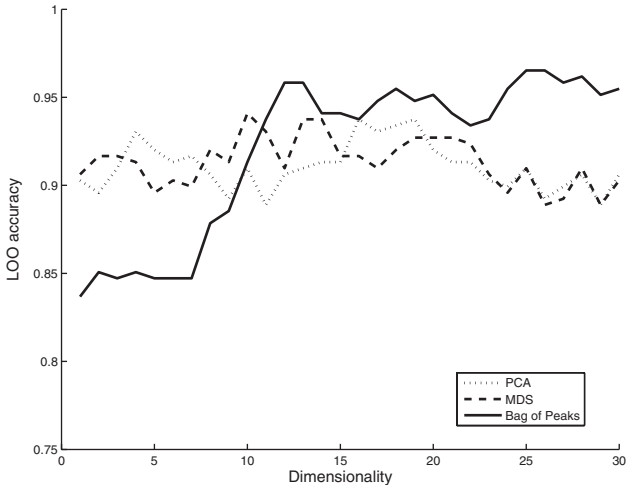

Fig. 2. Averaged accuracies of PCA, MDS and Bag of peaks representations. The accuracy, at a given dimensionality, is the average over the four classifiers.

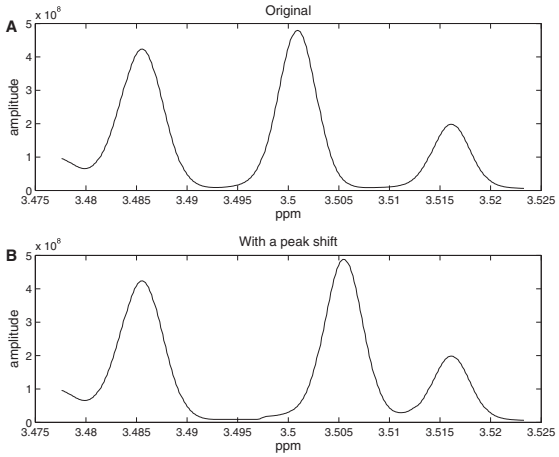

Fig. 3. (**A**) Original spectrum; (**B**) spectrum with a shifted peak: the shift-level γ is 1.

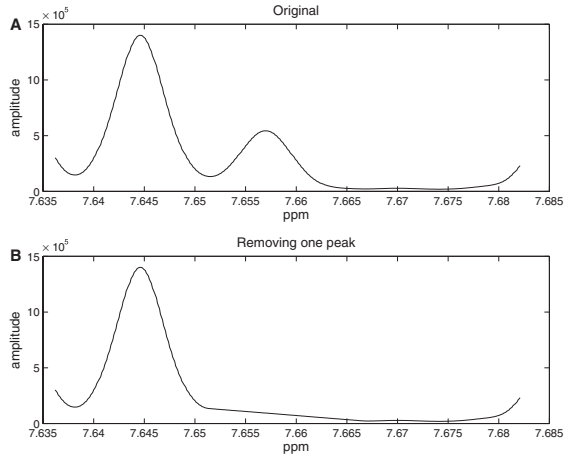

Fig. 4. (**A**) Original spectrum; (**B**) Spectrum with a missing peak.

For both cases, from each original spectrum five simulated spectra were generated (obtaining 160 entries in total). These spectra were then classified using the best performing classifier in Experiment 1: `1-nn`. Note, in order to assess the generalization capability of the representations in these noisy cases, all training phases were computed on the original dataset. Indeed, the training phase includes both the computation of the representation (the PCA space and the Peak Dictionary[2]) and the classifiers training. Testing was then performed with the simulated sets: patterns were projected in the PCA and Bag of Peak spaces and classifiers then evaluated. This results in appropriately disjoint training/testing sets. Random fluctuations in reported accuracies were reduced by simulating the datasets 30 times using different random peak selections, and averaging the results.

Results obtained for different values of the noising process parameters are shown in Table 2. The Bag of Peaks is always more accurate than the PCA when peaks are shifted (set 1), while there is no significant difference between the two methods when peaks are removed (set 2). In set 1, the difference is greatest when 100 or more peaks are shifted by a random magnitude between 0.5 and 1.5 peak-widths—where accuracy improves by around 10%.

5 DISCUSSION

Although we have promoted the Bag of Peaks approach on the basis of its interpretability and potential for intermediate expert interaction, here we have presented results produced in a fully automated manner, without manual editing of the dictionary or setting threshold. Even in that absence, the three experiments above show that the Bag of Peaks fares well with respect to the other algorithms to which it has been compared.

In the first experiment, concerning our small, but real, dataset of NMR traces, we have shown that the accuracy of the Bag of Peaks is consistently better than both the PCA and MDS approaches.

[2]In order to have a fair comparison, MDS is omitted from the analysis, since its projection space is computed each time it is used—and not once for all as in the PCA or Bag of Peaks cases.

Table 2. Simulated datasets with 160 traces each

Set 1					
	$\gamma = 0.1$ (%)	$\gamma = 0.5$ (%)	$\gamma = 1$ (%)	$\gamma = 1.5$ (%)	$\gamma = 2$ (%)
$N_S = 1$	+0.08	+0.00	−0.02	+0.19	+0.06
$N_S = 10$	+0.06	+0.65	+0.40	+0.71	+0.77
$N_S = 50$	+1.35	+2.54	+2.71	+2.52	+3.19
$N_S = 100$	+2.42	+4.75	+5.10	+6.15	+5.77
$N_S = 200$	+4.27	+10.10	+9.58	+10.44	+5.88
$N_S = 300$	+7.73	+13.98	+12.08	+9.13	+3.69
Set 2					
$N_R = 1$ (%)	$N_R = 5$ (%)	$N_R = 10$ (%)	$N_R = 15$ (%)	$N_R = 20$ (%)	N_R=25 (%)
+0.062	+0.23	−0.08	+0.16	−0.04	+0.02

Advantage percentage accuracy obtained by the Bag of Peaks approach over the PCA approach (the number of retained components was computed using the Scree Test). A positive value indicates that Bag of Peaks has better accuracy than PCA. (Set 1) Simulation with peaks shifting: N_S represents the number of random peaks shifted in each trace, whereas γ scales the overall amplitude of the random shifts; (Set 2) simulation with peaks removing: N_R represents the number of random peaks removed in each trace.

Moreover, in the second experiment we have shown how to rank, in the dictionary, the peaks that most influence classification using the FFSS. This has a direct practical application: identifying metabolites needed to discriminate a particular disease can become simply a process of selecting peaks in the dictionary—perhaps with the aid of a metabolite database. In our application regarding Type I diabetes in Sardinian children it turned out that the most discriminating peak loci occur around (in p.p.m.): 1.46, 3.50, 3.26 (Arginine triplet), 3.04 and 3.05 (Creatine Creatinine pair) and peaks in the range [3.24:3.29]. Here, the error tolerances on individual loci are bounded by the magnitude of the parameter $\theta/2 = 0.01$ p.p.m.

Finally, paucity of available traces is a condition that effects many pilot studies other than ours. Approaches, like the Bag of Peaks, that facilitate a pragmatic degree of interpretation of the small datasets collected for budget-limited studies ought to help instruct follow-up studies. In our experimental session, we tried to enlarge the scope of the analysis by generating two large sets of 160 artificial traces, obtained by simulating random peak shifts and peak loss in original traces. With these datasets, we were able to illustrate (Table 2) further the advantage of the Bag of Peaks over the standard PCA method.

6 CONCLUSION

In this article, a novel approach for NMR spectra analysis has been proposed, in which a fixed length descriptor, based on peaks, was used to characterize a single trace. Such scheme, which we called Bag of Peaks, represents an interpretable and intermediate descriptor of NMR traces, allowing the interaction of the expert. Nevertheless, in the experimental evaluations presented in the article, we preferred to neglect that advantage, in order to demonstrate its viability with respect to standard automated methods. Having established that viability, we intend in subsequent articles to demonstrate what additional advantage may be obtained by the interactive participation of human experts, in the way already motivated. Here, we have

presented results produced in a fully automated manner—without manual editing of the dictionary or setting threshold. Even in that absence, the three experiments above show that the Bag of Peaks can and does perform better than standard PCA—and the MDS-based approaches. Not only can it produce more accurate classifications, it also delivers practical suggestions for metabolite peak loci that may be implicated in the disease under study: here Type I diabetes in Sardinian children.

ACKNOWLEDGEMENTS

We acknowledge Sergio Uzzau and the facilities provided by Porto Conte Ricerche, Alghero in Sardinia. We thank E. Grosso, A. Lagorio, M. Gessa and M. Cadoni for helpful discussions. We also thank the anonymous reviewers for helpful suggestions.

Conflict of Interest: none declared.

REFERENCES

Bishop,C. (1995) *Neural Network for Pattern Recognition*. Clarendon Press, Oxford.

Bretthorst,G.L. *et al*. (2005) Exponential parameter estimation (in NMR) using Bayesian probability theory. *Concepts Magn. Reson. A*, **27**, 55–62.

Cattell,R. (1966) The screen test for the number of factors. *Multivariate Behav. Res.*, **1**, 245–276.

Cox,T. and Cox,M. (1994) *Multidimensional Scaling*. Chapman and Hall, London.

Cristianini,N. *et al*. (2002) Latent semantic kernels. *J. Intell. Inf. Syst.*, **18**, 127–152.

Csurka,G. *et al*. (2004) Visual categorization with bags of keypoints. In *Proceedings of the Workshop Pattern Recognition and Machine Learning in Computer Vision*. Grenoble, France.

Davies,D. and Bouldin,D. (1979) A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 224–227.

Duin,R. *et al*. (2004) Prtools4, a matlab toolbox for pattern recognition. Delft University of Technology.

Ernst,R. *et al*. (1990) *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Clarendon Press, Oxford, England.

Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Fukanaga,K. (1990) *Introduction to Statistical Pattern Recognition*. 2nd edn. Academic press, San Diego.

Hartigan,J. (1975) *Clustering Algorithms*. John Wiley & Sons, Oxford, England.

Hastie,T. *et al*. (2001) *The Elements of Statistical Learning*. Springer Verlag.

Jain,A. and Dubes,R. (1988) *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ, USA.

Jain,A. *et al*. (1999) Data clustering: a review. *ACM Comput. Surv.*, **31**, 264–323.

Joachims,T. (1998) Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the European Conf. Machine Learning*, Springer Verlag, Heidelberg, Germany, pp. 137–142.

Jollife,I.T. (1986) *Principal Component Analysis*. Springer Verlag, New York.

Keun,H.C. (2006) Metabonic modeling of drug toxicity. *J. Pharmacol. Ther.*, **109**, 92–106.

Kohavi,R. and John,G. (1997) Wrappers for feature subset selection. *Artif. Intell.*, **97**, 273–324.

Kruskal,J. (1997) Multidimensional scaling and other methods for discovering structure. In *Statistical Methods for Digital Computers*, John Wiley & Sons, New York, pp. 296–339.

Lazebnik,S. *et al*. (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2. Association for Computing Machinery, NY, USA, pp. 2169–2178.

Lefebvre,B. *et al*. (2004) Intelligent bucketing for metabonomics - part 1. In *Metabolic Profiling: Pathways in Discovery*. ACD Labs, Toronto, Ontario, Canada.

Lindon,J. *et al.* (2003) Contemporary issues in toxicology - the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicol. Appl. Pharmacol.*, **187**, 137–146.

Lindon,J.C. *et al*. (2001) Pattern recognition methods and applications in biomedical magnetic resonance. *Prog. Nucl. Magn. Reson. Spectrosc.*, **39**, 1–40.

Lindon,J. *et al*. (2007) *The Handbook of Metabonomics and Metabolomics*. Elsevier, Amsterdam, Holland.

Lodhi,H. *et al*. (2001) Text classification using string kernels. In *Advances in Neural Information Processing Systems*, Vol. 13. MIT Press, Cambridge, MA, USA.

Schölkopf,B. and Smola,A. (2002) *Learning with Kernels*. MIT Press, Cambridge, MA.

Schorn,C. (2002) *NMR Spectroscopy: Data Acquisition*. Wiley-VCH Verlag GmbH & Co., Weinheim, Germany.

Stoyanova,R. and Brown,T. (2001) NMR spectral quantitation by principal component analysis. *NMR Biomed.*, **14**, 271–277.

Theodoridis,S. and Koutroumbas,K. (1999) *Pattern Recognition*. Academic Press, New York, USA.

Tibshirani,R. *et al*. (2004) Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, **20**, 3034–3044.

Zhang,J. *et al*. (2007) Local features and kernels for classifcation of texture and object categories: a comprehensive study. *Int. J. Comput. Vis.*, **73**, 213–238.