Rogers, S. and Girolami, M. and Krebs, R. and Mischak, H. (2005) Disease classification from capillary electrophoresis: mass spectrometry. *Lecture Notes in Computer Science 3686*:pp. 183-191.

Disease Classification from Capillary Electrophoresis:Mass Spectrometry

Simon Rogers[1], Mark Girolami[1], Ronald Krebs[2], and Harald Mischak[2]

[1] Bioinformatics Research Centre, Department of Computing Science, University of
Glasgow, Glasgow UK,
srogers@dcs.gla.ac.uk,
WWW home page: http://www.dcs.gla.ac.uk/~srogers
[2] Mosaiques Diagnostics and Therapeutics AG,Feodor-Lynen-Str. 21, D-30625
Hannover, Germany

Abstract. We investigate the possibility of using pattern recognition
techniques to classify various disease types using data produced by a new
form of rapid Mass Spectrometry. The data format has several advan-
tages over other high-throughput technologies and as such could become
a useful diagnostic tool. We investigate the binary and multi-class per-
formances obtained using standard classifiers as the number of features is
varied and conclude that there is potential in this technique and suggest
research directions that would improve performance.

1 Introduction

In recent years, microarrays have enabled researchers to measure the expression
of entire genomes simultaneously. Some work has been undertaken to investi-
gate how well classifiers built using microarray data can discriminate between
healthy and diseased samples and samples of differing diseases and disease stages
[1]. However, although this is very interesting from a feature (i.e. gene) selection
perspective, as a general diagnostic tool, it is unlikely to prove useful. There are
several underlying reasons for this. Firstly, the cost of microarray analysis and
the time required to perform the analysis are both currently prohibitive. Sec-
ondly, the mRNA levels measured by a microarray only give a partial picture of
the proteomic activity inside the cell. Finally, samples have to be very localised.
For example, to diagnose a bladder cancer, a sample of bladder tissue would be
required. This is obviously a highly invasive procedure.

In this paper, we consider a new form of biological data (introduced in [2–4])
generated using Mass Spectrometry (MS) and assess whether it has potential as
a diagnosis tool, using various pattern classification techniques. This data can be
obtained very rapidly and inexpensively, suggesting that it may be well suited
for a diagnostic purpose. Also, the data is collected from a urine sample. This
is easily obtained and therefore can potentially be used to diagnose any disease
that will cause a change in the particle content of the urine.

The remainder of the paper is set out as follows. In the next section, we intro-
duce the data generation process. In section 3, we discuss the data pre-filtering

and pre-processing and briefly mention the classification algorithms used. In sections 4 and 5 we present results and conclusions.

2 CE/MS data generation

Recently, a new MS approach has been investigated that couples capillary electrophoresis (CE) directly to MS enabling detailed analysis to be available quickly (< 1 hour) and directly from a suitable (e.g. urine) sample [2, 4, 3]. Traditionally MS has been used to identify individual proteins but typically cannot be performed on a sample consisting of various proteins. Separation techniques exist to isolate individual proteins from such a mixture but these tend to be highly labour intensive and therefore expensive and slow. Here, the CE takes a complex sample of particles (in this case, the particles can be anything that might be found in the urine, not necessarily complete proteins) and by applying a charge differential along the capillary, separates the various particles in time. The output is connected directly to the MS realising a mass profile that evolves with time. This data is then analysed by *Mosaiques Visu* software that detects and outputs intensity values at the unique mass/time peaks (for details, see [4]). The separation in time means that only a small fraction of the particles are applied to the MS at any particular time. If the sample was applied directly to the MS without this stage it would be far more difficult to distinguish between individual particles.

This method has many possible diagnostic advantages over microarrays. Firstly, the analysis is quick and non-invasive. Secondly, in the case of using urine samples, there is the potential to be able to diagnose any diseases that would result in a variation of the products found in urine. However, there are drawbacks to this method. Firstly, the data produced is of a very high dimension ($\sim 30,000$ features) and the number of available samples is relatively small. Secondly, as no real control is imposed on the sample being analysed, it is possible that amongst this high number of features, there will be many due to other, spurious factors.

To date, there has been some research focused on the potential of MS proteomics data as a diagnostic tool, but using serum rather than urine. For example, Lilien *et al* [5] use Principal Components Analysis and a linear discriminant to distinguish between the MS spectra of serum samples from patients with various tumours. Similarly, Wagner *et al* [6] use supervised techniques to try and create protein profiles from MS analysis of serum samples. These approaches are all based on identifying whole, specific proteins whereas CE/MS can detect a much wider range of particles.

3 Method

3.1 Data

The data set we shall use consists of analysis of 632 samples that come from one of 22 separate classes from individuals with various renal diseases, cancers

and diabetes as well as samples from healthy individuals. We have performed binary classification with a variety of algorithms on a large number of pairs of classes from this set, however, in this investigation, we will concentrate on a group of five classes - Bladder Cancer (BLA - 47 instances), Renal Cancer (REN - 25 instances), Prostate Cancer (PCA - 8 instances), Benign Prostate (PB - 12 instances) and Healthy (NK - 41 instances). The total number of features is 28378. The inclusion of benign prostate samples is interesting as clinical differentiation between individuals with prostate cancer and those with a benign growth is challenging and the two different conditions require vastly different treatment.

3.2 Feature pre-filtering

This particular form of data has several important characteristics. Firstly, although the total number of possible features is very high ($\sim 30,000$), in each sample, only a small proportion of these values are non-zero, indicating that this particular particle was not present or, and this distinction may be important, not detected. Therefore, if we call our N (samples) $\times M$ (features) dataset $\mathbf{X}$, the vast majority of the x_{ij} values are zeros. Secondly, those values that are present take values over a very large range (see figure 1(a) (top)). To overcome this second problem, we have adopted a log transform[3]. Figure 1(a) shows the binary classification performances for a wide range of pairwise comparisons and algorithms with and without this transform. We can see that in the vast majority of problems the log transform improves performance (all points below the $y = x$ line).

The first problem is not quite so straightforward to address. As an initial step, we perform a simple pre-filtering. For a given classification problem (i.e. 2 or more classes), we only keep features that appear (i.e. are non-zero) in at least $\rho\%$ of the data samples for one or more of the classes. Note that we do not force the feature to be present $\rho\%$ of the time across *all* of the classes as it is possible that both presence and absence of a feature as well as presence with varying magnitude could be indicative of changing condition. We will investigate the effect of varying ρ in more detail later.

3.3 Classifiers

In this investigation we limit ourselves to two main classes of classifiers. Naive Bayes classifiers (NB) and Support Vector Machines (SVM's). This is by no means a complete list but serves as a reasonable starting point. Due to limitations of space, a description of these algorithms is omitted, readers are referred to [7, 8] for more details. When using a NB classifier, it is necessary to determine the parametric form of the density function that will be used for each feature. Here, we have considered the following four (defining the data matrix $\mathbf{X}$ as before, and

---

[3] specifically, $\log(x_{ij} + 1)$, where the additive term ensures that our zero values remain at zero
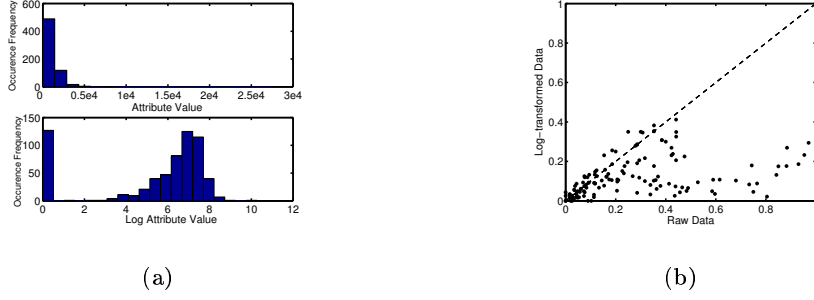
(a)   (b)

Fig. 1. Log transformation example (left, note that due to the wide range of values, many bars on the top histogram are too small to be visible) and binary classification errors with and without log transform

indexing the individual features with $m = 1 \ldots M$ and the classes $c = 1 \ldots C$ and defining the $M \times 1$ data vector $\mathbf{x}$)

- **Gaussian:** Each class is defined by an M dimensional Gaussian distribution with diagonal covariance. i.e. $p(\mathbf{x}|c, \mu_c, \sigma_c) = \prod_{m=1}^{M} \mathcal{N}(x^m | \mu_c^m, \sigma_c^m)$
- **Binomial:** The data is transformed to a binary representation (i.e. value present (non-zero) or absent (zero). Each class is then defined by an M dimensional vector of probabilities $\mathbf{p}_c$, where $p_c^m = P(x_m = 1|c)$. i.e. $p(\mathbf{x}|c, \mu_c, \sigma_c) = \prod_{m=1}^{M} (p_c^m)^{x_m} (1 - p_c^m)^{1-x_m}$
- **Multinomial:** As binomial but with > 2 possible states. Each class is now defined by M vectors of state probabilities.
- **Exact-zero Gaussian:** This density is intended to capture more accurately the characteristics of the particular problem. For each class, we now have an M dimensional vector of probabilities as in the binomial case. If the value is non-zero we then assume it can be modelled by a Gaussian. Therefore, defining the indicator variable t_m which is 1 iff x_m is non-zero, $p(\mathbf{x}|c, p_c, \mu_c, \sigma_c) = \prod_{m=1}^{M} (p_c^m \mathcal{N}(x^m | \mu_c^m, \sigma_c^m))^{t_m} (1 - p_c^m)^{1-t_m}$.

In all cases, we define the prior distributions for each class to be the proportion of training instances from that class and use the standard maximum likelihood solutions for the parameter values.

As an alternative to the probabilistic, Naive Bayes classifiers, we consider the SVM. To use an SVM, a kernel function must be defined. We use a linear kernel (a simple dot product in the input space) after normalising each feature to have zero mean and unit variance. It may be the case that there are other more suitable kernels that could be used however, due to the high dimensionality, this is a reasonable starting point. In addition to this, we set the margin parameter C to infinity (i.e. a hard-margin).

Multi-class SVM Classifiers Naive Bayes classifiers are naturally multi-class. SVM's however can not be naturally generalised to the multi-class setting. However, various tree based heuristics can be used to split the problem down into a set of binary decisions. We experiment with two of these here

- **Directed Acyclic Graph (DAG):** In a C class problem, the DAG SVM [9] formulates the problem as a tree with $(C(C-1)/2)$ nodes. At each node, an SVM is trained between two of the classes in the problem. When testing a point, we start by assuming that the point could belong to any of the C classes. It then moves through the tree and at each SVM, one class is removed from the possible solution until only one class remains. For example, in a 3 class problem, we might train our first classifier on class 1 versus class 3. When testing, if the test point is classified as 1, we remove 3 from the list of possible solutions and move on to the classifier between 1 and 2.
- **Divide-by-2 (DB2):** The DB2 [10] classifier operates by repeatedly splitting the C class problem into binary problems. For example, in a four class problem, the first classifier might split the data into the meta-classes (1,2) and (3,4). If a test point is classified as belonging to the first class, it is then applied to a classifier between 1 and 2 etc.

In either of these systems, the particular form of SVM can vary between nodes. Presently, we have kept them all the same (linear kernel, $C = \infty$) but employing different ones for different classifications is an obvious next step. This is particularly promising for the DB2 model where it may be sensible to have different classifiers built from different features at different levels in the hierarchy. We will discuss this further below.

4 Results

4.1 Binary Classification

Initially, we have investigated the pairwise classification performance between relevant pairs of classes in the dataset. This has been performed for many pairwise combinations but due to space limitations we will only consider those belonging to the cancer subset here. Table 1 shows the results for our five binary classifiers (SVM and 4 different NB). Each value is the best leave-one-out (LOO) performance obtained when varying ρ, the feature filtering threshold. In some cases the best performance was obtained for several different values of ρ. In these cases, we have shown the minimum and maximum ρ values. We can see from the table that generally, the performance is reasonably good with low values LOO error. The highest errors obtained are for the classification between Prostate Cancer and benign Prostate with a minimum of 10% LOO error (= 2 data points). This is to be expected, partly due to the difficulty of the problem and partly due to the fact that there is such a small number of samples in each class (8 and 12 in PCA and PB respectively). We also note that no-one classifier out-performs the others although the best performance can generally be found

from an SVM or Naive Bayes with Gaussian or binomial densities. This is especially interesting as it suggests that in some cases, the magnitude of a value (if it is non-zero) does not improve performance whereas in other cases it does. The relatively poor performance of the exact-zero system seems to suggest that it is not necessary to use both presence and absence information and magnitude information at once. It is worth mentioning that the exact-zero mixture requires the fitting of considerably more parameters than the individual Gaussian and Binomial models and it will be interesting to see if this error rate can be improved as more data becomes available. The results are promising and suggest that discrimination is possible using CE/MS data. However, further validation will be acquired through a planned blind test.

Classes	Class Sizes	SVM	NB Gauss	NB Bin	NB Mult	Exact Zero
PCA v PB	8 v 12	15.00 (90)	15.00 (95)	10.00 (90)	20.00 (55→ 95)	25.00 (15 → 85)
PCA v NK	8 v 41	0.00 (85)	4.08 (25→95)	0.00 (95)	12.24 (50→85)	6.12 (40→95)
PB v NK	12 v 41	1.80 (65)	0.00 (10→20)	5.66 (30→85)	11.32 (55→75)	7.55 (70)
REN v NK	25 v 41	1.52 (75)	1.52 (20)	6.06 (15)	7.58 (35→50)	6.06 (10→65)
BLA v NK	47 v 41	2.30 (65)	4.55 (15→20)	7.95 (5)	6.82 (45→80)	3.41 (10→15)

Table 1. Binary LOO performances (errors are percentages and the value in brackets is value for ρ for this particular level of performance)

4.2 Multi-class Classification

Binary classifications are interesting but are limited from a diagnostic point of view. One of the possible benefits from CE/MS data from urine samples is that any number of different diseases could be identified. Therefore, we turn our attention to multi-class schemes. As discussed above, the various Naive Bayes classifiers can be naturally expanded to a multi-class scenario. For the SVM's, we have used two tree based approaches, DAG and DB2. For DB2, it is necessary to define the hierarchy - i.e. how we want to perform the successive partitions of the C classes to create a series of binary problems. In this example, we have decided on a hierarchy that is sensible from a clinical point of view. The hierarchy is shown in figure 2(a). At the top level, we split NK from everything else (i.e. healthy versus unhealthy). If the point is classified as unhealthy, we perform a further split into (PCA, PB) v (BLA, REN) and then perform a standard binary classification on whichever of these pairs is chosen. The best results (again, as ρ is varied) can be seen in table 2. A plot of the number of features retained against ρ can be seen in figure 2(b). In this case, we see that the two SVM schemes out-perform the various Naive Bayes classifiers. This may be due to the fact that the two SVM schemes do a series of more simple binary classifications, rather than one more complicated multi-class one (as is the case with the Naive Bayes). The DB2 SVM defines a hierarchy over the possible diseases and so is able to classify at varying levels of abstraction. For example, at the top level of the tree (i.e. simply classifying between healthy and unhealthy) there is just one misclassification (a NK) and at the next level (ignoring the 1 wrong NK from the previous classification), there are no errors, suggesting that the errors all occur in the most specific, lowest level. This shows the power of a possible hierarchical

approach - currently we are unable to reliably classify between PCA and PB but it appears that we can reliably classify that something is either PCA or PB from other classes. The ability to visualise the decreasing certainty as we move down the hierarchy is a great bonus to such an approach; something that is lacking from a flat structure such as a simple Naive Bayes classifier.
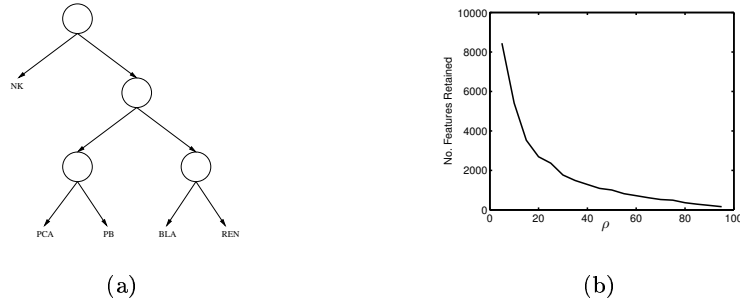

(a)   (b)

Fig. 2. Example hierarchy (left) and number of features retained in multi-class problem as ρ is increased (right)

	SVM(DAG)	SVM(DB2)	NB Gauss	NB Bin	NB Mult	Exact Zero
All (%)	6.02	6.77	14.29	19.55	24.81	16.45
ρ (No.Feat)	10 (5411)	25 (2366)	40 (1287)	25 (2366)	60 (720)	60 (720)
PCA	0 / 0	37.5 / 3	37.5 / 3	25.0 / 2	100 / 8	50.0 / 4
PB	33.3 / 4	16.7 / 2	41.67 / 5	41.67 / 5	91.7 / 11	66.7 / 8
REN	12.0 / 3	8.0 / 2	8.0 / 2	16.0 / 4	32.0 / 8	20.0 / 5
BLA	2.1 / 1	0.0 / 0	6.4 / 3	12.8 / 6	2.0 / 1	2.0 / 1
NK	0.0 / 0	4.9 / 1	14.6 / 6	22.0 / 9	12.2 / 5	9.8 / 4

Table 2. Multi-Class LOO errors. Top line shows overall percentage, second line shows ρ with the actual number of features in brackets, lower rows give the percentage of errors in each particular class / absolute number of errors in each class.

5   Conclusions and Future work

In this paper, we have described CE/MS a new rapid, high-throughput form of proteomic data and have performed several simple experiments to try to give some indication of the diagnostic capabilities of the data. The data has several advantages over other similar data formats such as microarray data. It can be produced very rapidly in a non-invasive manner (normally through analysis of a urine sample) and has the potential to be able to diagnose many diseases. However, like microarray data, it is noisy and the number of features is far greater than the number of collected samples - this latter problem is likely to improve as the data generation process is a fraction of the cost of a microarray experiment and obtaining samples for analysis is much more straightforward.

Results presented suggest that pattern recognition techniques combined with CE/MS data has potential as a diagnostic tool. In these basic experiments low

LOO errors were observed with only a very basic choice of classifiers and very crude feature pre-filtering. Although results have been presented for only 5 of the 22 available classes, the same general level of performance is observed across other subsets that have been investigated. As might be expected, multi-class performances are worse than binary performances but the performance of the two tree-based SVM approaches is promising. Particularly, the DB2 SVM enables us to classify in a hierarchical manner, revealing where the errors are made and giving a more useful diagnosis. Such an approach isn't limited to SVM classifiers - any particular classifier could be used at each node and it would be expected that by tuning classifiers to the different hierarchical problems performance could be considerably improved. i.e. by extracting relevant features at each level. This is something for future investigation. The only feature selection considered here has involved the initial pre-filtering step. Examining the results, we see that the value of ρ for best performance varies dramatically. This suggests that for some problems, there are a small number of features that are consistently varied for the different diseases, whereas for others, we are obtaining useful information from features that are very rarely present. In these latter cases, it should be remembered that it is possible that several different features could correspond to the same particle that has undergone some small change. This suggests that performance may be improved by combining features or developing more applicable kernel functions, possibly including the mass and time information available for each detected peak. It may also be beneficial to include extra meta-data in the decision making process. This could be clinical history or more general observations (e.g. the gender of the individual). This would be particularly interesting in the hierarchical classifier as different meta-data could be incorporated at each level.

Of the multi-class methods investigated, the hierarchical SVM methods look to be the most promising and it is likely that the performances considered presented here could be improved by careful selection of the classifier at each level. This would involve more careful selection of kernels and a more rigorous feature selection stage.

Finally, the diseases that have been investigated so far have all been chosen due to the fact that they are very likely to produce a change in the urine profile. It would be interesting in future work to investigate whether or not such techniques could be used to diagnose diseases without such an obvious effect or in other testing circumstances.

References

1. Alizadeh, A., Eisen, M., Davis, R., et al.: Different types of diffuse large b-cell lymphoma identified by gene expressing profiling. Nature **403** (2000) 503–511
2. Kolch, W., Neususs, C., Pelzing, M., Mischak, H.: Capillary electrophoresis: Mass spectrometry as a powerful tool in clinical diagnosis and biomarker discovery. Mass Spectrometry Reviews (2005 (in press))

3. Kaiser, T., Wittke, S., Just, I., et al.: Capillary electrophoresis coupled to mass spectrometer for automated and robust polypeptide determination in body fluids for clinical use. Electrophoresis **25** (2004) 2044–2055

4. Weissinger, E., Wittke, S., Kaiser, T., et al.: Proteomic patterns established with capillary electrophoresis and mass spectrometry for diagnostic purposes. Kidney International **65** (2004) 2426–2434

5. Lilien, R.H., Farid, H., Donald, R.: Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. Journal of Computational Biology **10** (2003) 925–946

6. Wagner, R., Naik, D., Pothen, A., et al.: Computational protein biomarker prediction: a case study for prostate cancer. BMC Bioinformatics **5** (2004)

7. Mitchell, T.: Machine Learning. McGraw-Hill (1997)

8. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)

9. Platt, J., Cristianini, N., Shawe-Taylor, J.: Large margin DAG's for multiclass classification. Advances in Neural Information Processing Systems **12** (2000) 547–553

10. Vural, V., Dy, J.: A hierarchical method for multi-class support vector machines. In: Proceesings of the 21st International Conference on Machine Learning. (2004)