



UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

38050 Povo – Trento (Italy), Via Sommarive 14
<http://www.dit.unitn.it>

VALIDATION OF CFS CLASSIFICATION WITH DIFFERENT
DATA SOURCES

Marco Bassetti*, Massimiliano Bernabe’*, Manuel Borile*, Cesare
Desilvestro*, Tarcisio Fedrizzi*, Alessandra Giordani*, Roberto
Larcher*, Alida Palmisano*, Angelo Salteri*, Stefano Schivo*, Nicola
Segata*, Linda Tambosi*, Roberto Valentini*, Periklis Andritsos^o, Paolo
Fontana^s, Andrea Malossini^o, Enrico Blanzieri^o

February 2006

Technical Report # DIT-06-004

Validation of CFS classification with different data sources

Marco Bassetti*, Massimiliano Bernabe’*, Manuel Borile*, Cesare Desilvestro*, Tarcisio Fedrizzi*, Alessandra Giordani*, Roberto Larcher*, Alida Palmisano*, Angelo Salteri*, Stefano Schivo*, Nicola Segata*, Linda Tambosi*, Roberto Valentini*, Periklis Andritsos^o, Paolo Fontana^s, Andrea Malossini^o, Enrico Blanzieri^o

(*) University of Trento. (°) University of Trento. Department of Information and Communication Technology (°) Istituto Agrario di San Michele all’Adige. Correspondent Author: Enrico Blanzieri, Department of Information and Communication Technology via Sommarive 14, 38059 - Trento Italy. blanzier@dit.unitn.it.

Summary. The difference between patients with CFS patient and healthy ones could, in principle, be detected by examining a variety of data. We systematically used the CAMDA 2006 available data sets in order to assess the patients’ discrimination using supervised and unsupervised techniques. Our results suggest that data sets that are predictive are the clinical as well as the microarray data sets. On the other hand, our analysis of the proteomics data suggests that subjects with diseases different from CFS could be among the healthy ones. Finally, we indicate a set of genes extracted from the microarray data and validate then with an automatic comparison with Gene Ontology information. A set of these genes with high GO proximity may contribute to CFS.

Introduction. Chronic Fatigue Syndrome (CFS) is clinically defined through symptoms and disabilities that proved to be elusive when it was attempted to characterize them in terms of analysis-based diagnosis and etiology. The data sets provided for the CAMDA 2006 contest contained clinical data, SNPs, blood data, proteomics and microarray genetic expression data. The integration between different sources has been explored in [Whistler et al, 2003], however not to the extension required by the CAMDA contest, in the sense that the authors addressed the problem of characterization of subgroups of CFS. [Vernon et al, 2002] reported the identification of a set of genes that could play the role of biomarkers. However, as the authors stated, limitation in the number of samples and no independent validation influenced the validity of the results. Therefore, the problem of isolating biomarkers for the syndrome remains open.

In this paper, we present our attempt to use the data available for the CAMDA contest in order to validate the diagnosis related to healthy patients and patients suffering from CFS. Our working hypothesis is, that

given the elusiveness of the syndrome, the CFS labeling of patient records can possibly be wrong. Hence, in order to increase the probability of detecting relevant biomarkers, we would like to assess the inherent ability of each data set to produce a consistent classification. Depending on the data, we expect a difference in the way individuals are classified. In particular, we expect an optimal classification when using clinical data since CFS is defined in terms of symptoms. From the literature, it seems that we cannot expect blood data to be very useful in order to predict the class. Moreover, we do not have any indication of the properties of the SNPs and proteomics data. After analyzing the aforementioned data sets, we validate our results using information derived from the Gene Ontology, which provides an automatic way of assessing the biological consistency of the discovered genes.

The activities whose results are reported here were partially carried out as assignments during the course of Data Mining for Computer Science students of the University of Trento. In the following sections, we report our analysis and results for each one of the data sets. We also provide an interpretation of our results.

Clinical Data. Using this data set we tried to analyze the real meaning of the column headers. Then, based on these results, we classified the data set using a Support Vector Machine (SVM) classifier. In our first attempt to classify the data, we used a multi-class SVM classifier (M-SVM) and all the features. The accuracy of the result was very low and, thus, we decided to remove some redundant features (e.g. columns “Intake classic”, “DOB”, “Exclusion”) and some instances (rows regarding patients excluded for medical or psychiatric reasons, and those having null-values in key-columns). However, the accuracy obtained using the M-SVM on this new data set was still low (25% marked as “Incorrectly classified”).

A possible explanation for this low accuracy is that the data were classified using 15 different class labels. In order to reduce the number of classes, we discarded all the instances not labeled as CFS or NF, reducing the data set to 58 instances (15 of which are CFS). Classification using binary-SVM produced a significantly higher accuracy (1.72% “Incorrectly classified”, see Table 1).

Then we applied two filtering algorithms (see [Malossini, 2006]) in order to detect and remove possible incorrectly labeled instances. The two filtering algorithms we used (*CL-stability* and *LOOE-sensitivity*) identified only one such suspect instance. This sample was removed from the data (for convenience, we name this last data set with *57instancesCFS-NFonly*).

A	B	←Classified as
15	0	A - CFS
1	42	B - NF

Table 1 Confusion Matrix on cleaned clinical data

Classification of the *57instancesCFS-NFonly* data set using binary-SVM resulted in complete accuracy (no instances were marked as “Incorrectly classified”).

SNPs Data. Our objective in the experiment was to find a reasonably correct relationship between the presence of particular SNPs and the state of health of a patient. From the given data sets named *CAMDA_SNP description* and *CAMDA-SNP-Genotype data 4-14-05*, we extracted a subset of data, ignoring patients labeled as ISF. Moreover excluded any patients without a corresponding record in the clinical data. The resulting patients were 111 females and 22 males.

After the preprocessing of the data set we imposed several thresholds of correctness in order to distinguish between good and bad results:

- >90%: we find a reasonable possible cause of CFS;
- 75-90%: a good result but not so good to derive conclusions;
- 40-75%: Not so bad, but not sufficient to derive conclusions;
- <40%: very bad.

We used the functions of the Weka Data Mining tool [Weka] for attribute selection in order to assess if one or more attributes could predict the status of the patient. In particular we consider the individual predictive ability (cfsubset eval), the computation of Chi-squared statistics (chisquaredsubseval), and a classifier based procedure (ClassifierSubseval). The Chi-squared test showed that none of the single attributes was sufficient to predict the disease of a patient. The methods CFS Subset and Classifier Subset gave us a reasonably correct result, as a subset of the features seemed to be predictive. Those features were:

POMC_3227244,	TH_243542,	MAOA_878819,
MAOB_15959461,	TPH2_8376042,	COMT_2539273,
COMT_3274705,	NR3C1_11837659,	NR3C1_11159943,
NR3C1_1046361,	NR3C1_1046360,	5HTT_7911132,
5HTT_7911143,	CRHR1_7450777,	CRHR2_15872871,
CRHR2_15960586,	HTR2A_8695278;	

Trying to cluster or classify the data set with this features, however, led to reduced correctness.

In our experiments, we used several clustering techniques: K-means, Expectation-Maximization (EM) and Cobweb. The results obtained after clustering were: K-means grouped 50.38% of the patients correctly, Cobweb provided the worst results with 3% of them grouped correctly and, finally, EM gave us a correctness of 35.3%. Hence, the results obtained through clustering methods were totally inconclusive.

The methods used for classification are: SVM, NBTree and Nearest Neighbor. SVM with Leave-one out validation classified only 54.14% of the patients correctly, and NBTree only 49.62%. Nearest Neighbor, with 5 neighbors, gave us a better correctness of 62.41%.

With our work on SNPs we did not obtain any results that could increase our knowledge on CFS nor give evidence that SNPs information could discriminate the classes.

Blood Data. The study of CAMDA’s supplied data about blood analysis was carried out in a number of steps. First of all, the data of each of the 34 “classical” blood exams was graphically visualized, dividing the subjects according to their gender and different state of their health, obtained by the two columns named “Intake Classific” and “Empiric” of the clinical data set. The progress of each graph was essentially the same for all exams: there were values out of the normal ranges with similar proportion in each group; so it was not possible to point out a trend that could help to assign a patient to a particular category.

The second part of the study, focused on the classification and clustering of the blood analysis data. First of all, we decided to exclude from the next analysis all patients whose “Empiric” label contained the string “Med” because, as explained in the CAMDA document [Camda, 2006 prot], those patients had some other medical reasons that could suggest a wrong diagnosis.

Therefore, the final group of analyzed patients contained 191 people and, because of the various values that the “Empiric” column could assume, we decided to classify the patients according to the following criteria:

1. healthy/ill: the patients with label “NF” were classified as healthy, and the other as ill;
2. NF/ISF/CFS: the patients were classified as healthy if the value of the column “Empiric” contains the substring “NF”. The same approach was used for ISF and CFS;
3. 9 distinct classes: the patients were classified according to the exact value of the associated column “Empiric” in the original clinical table

Using Weka [WEKA] for the analysis, we chose 2 classification methods: the Nearest-Neighbour and the Support Vector Machine (SVM), both using 10-fold crossing validation and leave-one-out. Besides, we tried to use different kernels for SVM (polynomial with exponents 2 and 3, and RBF with different parameters) but we always obtained bad classification results: in general the percentage of correctly classified ill patients was around 70%, while the percentage of correctly classified healthy patients was always below 20%.

Then, we tried to cluster the data using k-means, specifying 2, 3 and 9 clusters, according to the categories identified before. However the percentage of wrongly clustered instances ranged from 50 to 77% in this case as well. Hence, our task was to support an argument about the uselessness of blood data in CFS diagnosis, using numerical results. Such an argument may also be found in [Reeves, 2003], [Fukuda, 1994].

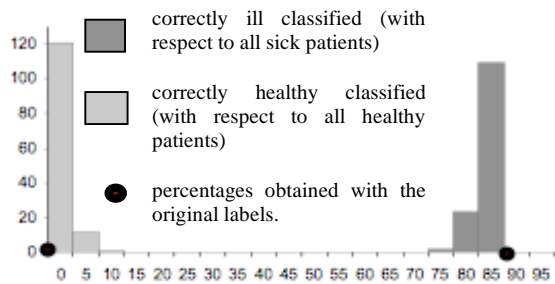


Figure 1. Accuracies for ill and healthy patients from the permuted datasets

In order to support it, we did some random permutations of the patients' "healthy/ill" label and we executed the classification described above once again. Figure 1 depicts the results obtained by SVM classification (the results of other classification methods are similar).

As we can see, the percentages of the original CAMDA labels were very similar to the ones obtained with random labels: so we can conclude that all "classical" blood analysis data are not able to distinguish healthy people from the patients that actually have CFS.

Proteomics Data. Proteomics research seeks to gain a better understanding of the role of proteins and gene function in the biology of a certain disease. The aim of our analysis was to identify serum biomarkers in CFS using Surface-Enhanced Laser Desorption/Ionization Time of Flight Mass Spectrometry (SELDI-TOF). To understand the molecular basis of CFS, we applied the hypothesis that different molecular patterns could be identified in samples from subjects with CFS compared to Quality Control subjects (QC).

The proteomics data sets given by CAMDA were very spread. Fractionated serum of 63 samples ran in duplicate was spotted into ProteinChips, profiled under several analytical conditions (IMAC, H50, CM10LS, CM10HS) and then read with both high and low energy laser.

According to the proteomic hypothesis we looked for the best fraction/ProteinChip combination that allowed us to discover biomarkers, *i.e.* peaks of protein intensities in CFS spectra that were not present in control serum. Given the spectra of samples coming from the same CAMDA classification group, we obtained an average spectrum for every condition, fraction and laser intensity of each group. One of the most relevant biomarkers discovered, comparing CFS pattern with QC pattern, was found with the combination fraction6/H50, and is shown in Figure 2A. In that figure, we can also notice that those patterns, supposed to be different, are on the contrary very close to each other, as if CFS couldn't be well diagnosed using mass spectrometry.

To validate this result, we provided a comparative analysis of the 7000 spectra given, based on the distance of spectra of 63 samples from QC. The method yielded, for every patient, the value of her

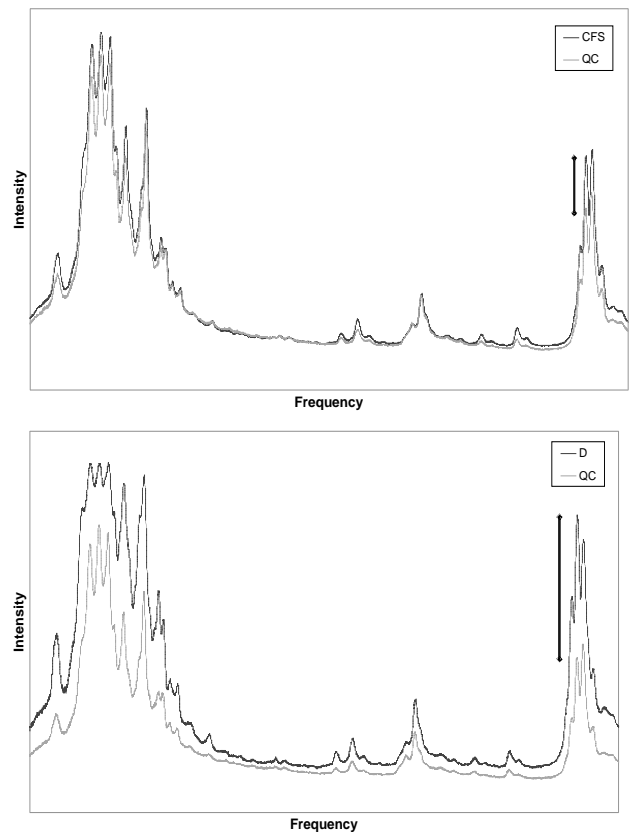


Figure 2: Spectra from the 6th fraction serum read with high laser and condition H50. Single biomarkers were found at frequency $\approx 12\text{kHz}$ (spectra plotted from frequency 8kHz to 13kHz). **Top:** Two patterns identified in samples from subjects with CFS and in QC subjects, according to CAMDA classification. **Bottom:** QC pattern and the average spectrum of diseased subjects (D) according to proteomic clustering.

distance to the quality control. Given all the distances, we produced an ordered list of all patients from the closest to the furthest from the QC. The ordering corresponded to a classification from the healthiest patient to the most diseased one. Euclidean distance and Pearson correlation were calculated between every sample spectrum and the relative QC spectrum for each condition and laser energy. The normalization of those distances, between 0 and 1, showed an obvious opposite correlation between the values of the Euclidean distance and the values of the Pearson correlation (and thus similarity) measure. For that reason we merged the two ordered lists, obtaining for each set of spectra, the average distance of Euclidean's and inverted Pearson values. We obtained average distances for different energy lasers and conditions, and since values were very similar, it seemed natural to cluster all patients according to those average distances. Using k-means with $k=3$ we obtained three clusters: the cluster H contains patients more similar to the QC (supposed to be healthy), the cluster D with distances close to 1 contains patients considered as

diseased since they were very far from QC, and an intermediate cluster M, which contains all the other

H	M	D	
8	21	2	NF
5	4	2	ISF
13	6	2	CFS

Table 2: Confusion matrix of assignment to clusters.

patients. Using three clusters, the predictive power of this analysis was very low: as shown in the confusion matrix of Table 2, proteomics predicted only 22% of the assignments to each classification group given by the CAMDA clinic data.

Given the previous clustering we build, as explained before, D patterns as the average spectra of our diseased patients. Comparing QC patterns with the D patterns we discovered a lots of biomarkers, as well as biomarkers reported comparing it with CFS pattern. In particular, for the same fraction/ProteinChip combination used before, we can see (Figure 2B) that D and QC patterns are very different and the previously reported biomarker is even more evident (the intensity is triplicate).

In other words using mass spectrometry we can identify few biomarkers that characterize clusters given by CAMDA. Moreover, through our distance based method we provided another classification whose results reported the same peaks and identified new ones. Those other peaks could not necessary identify biomarkers of CFS or CFS-like patient populations, as there are many biological factors that influence the role of proteins. Anyhow, as a consequence of our findings we can say that molecular patterns can be identified in samples from subjects with CFS compared to control subjects, but they are less significant than other patterns that could be identified from other groupings of same subjects.

Microarray Data. Analyzing the gene expression data set we noticed that the values were very noisy; in fact, even in the same chip, the values of the same gene spotted in spatially different places were very unstable. In this scenario, the normalization process was very important; we started the normalization by aligning the medians of the sample values with the scaling operation. The scaling factor applied to all the values for a specific sample i was calculated in the following way:

$$sca_factor_i = \frac{med(med(s_1), med(s_2), \dots, med(s_{177}))}{med(s_i)}$$

where med is the median.

In order to normalize the distribution of the samples, we applied the classical *Quantile Normalization* to the logarithms of the scaled values; we call the resulting data set *Logarithmic Normalized Values* (LNV) in order to distinguish it from another data set, called *Cubic Root Normalized Values* (RNV), resulting from the same process but with the cubic root operation

instead of the logarithm. Since the clustering of the complete data set (as expected) did not give any interesting results, in this context the goal was to identify genes with statistically significant changes in expressions between the samples labeled as CFS (fatigued – 89 samples) and the samples labeled as NF (non fatigued - 41); in this phase we ignored the ISF samples in order to improve sensibility. We applied two statistical methods to the two data sets (NV and RNV): conventional *t-tests* and *Significance Analysis of Microarrays* (SAM) [Tusher, 2001]. We obtained 4 analyses, but varying the parameters of the methods (alpha for *t-test* and delta for SAM) we obtained a set of possible results; from this set we selected the most precise results, obtaining 9 consistent sets of genes. A set of genes is considered precise enough if it respected the following two conditions:

1. The analysis parameters must be restrictive enough (the alpha parameter for the *t-test* must be under a threshold Th_{α} and the delta parameter for SAM must generate a FDR median value under another threshold Th_{δ}).
2. The genes provide high unsupervised classification accuracy. The confusion matrix between the CFS and NF labeled samples and the two clusters of samples obtained applying the (2 class) k-means clustering on the set of genes, must respect the following constraint:

$$CFS_{rate} \times NF_{rate} > Th_{precision}$$

where CFS_{rate} is the percentage of the right classified CFS samples and NF_{rate} is the percentage of the correctly classified NF samples. We were aware that there existed clustering algorithms more precise than k-means, but the goal of this clustering was a relative comparison between the analyses, not an absolute measurement.

The results of four analyses (one for each main group) are available on the web as additional material¹.

We chose to give each gene isolated by one of the nine best analyses, a value which represented the confidence level that we had on it. This value for a gene is simply calculated as the percentage of consistent analyses in which it appeared; the table of the best 30 genes is shown in Table 3.

We also tried to apply the same method on the *57instancesCFS-NFonly*, finding worse values of accuracy on the confusion matrices obtained by the *k-means* clustering on the resulting sets of genes; in particular, the best analysis on *57instancesCFS-NFonly* produced an accuracy value of 71% against a mean accuracy on the original set of 77%.

In conclusion, we can assume that the set of genes shown in Table 3 are the most likely to be in relation with CFS.

Integration with GO information. In order to verify and test the biological plausibility of the genes that were isolated as differential expressed in the CFS

¹http://dit.unitn.it/~blanzier/TN2camda06add_data.htm

GENES	DETECTED IN:
AK075162 XM_087606	8 analyses out of 9
NM_014149 BC001439	7 analyses out of 9
BC035807 AF172066 AF151022 AF100928	
AF449187 BC007072 AF492830	6 analyses out of 9
NM_006189 BC004166 BC002462	
NM_006278 S76825 AK022571	5 analyses out of 9
NM_002280 BC022270 AK024524	
AK095113 NM_004364	
AB002380 NM_005263	
AF113616 D37827 NM_000570	
AF142099 NM_015846 D14665	4 analyses out of 9
AF075430 NM_003608	
NM_001256 AK000759	
BC025394 BC012070 AF035933	
AB083606	

Table 3 Genes differentially expressed in microarray data

patient and in the healthy ones we assessed their relation with GO structure. We took the list of genes with their amino-acid sequence and we executed the Blast of these sequences against Uniprot database (updated 9 december 2005). Considering only biological processes, for each alignment of a gene, we assigned its weight to all its pertinent GO nodes (updated 10 january 2006), "bringing up" the same weight to ancestors (a node has a weight corresponding to the sum of its own weight plus all the weights of its children). Then we computed the information content of each node dividing the number of its descendents by the total number of nodes in the graph and calculating the natural logarithm of the resulting value. For each gene, the output were the GO with the maximum value of information content multiplied for the weight. In this way we choose the nodes that better characterize the query sequence in term of scoring and information associated to the GO nodes. We applied this technique to the list of genes in Table 3 and we obtained subgroups of them with respect to GO graph: the relation between GO terms is evaluated with Lin's formula [Lin 1998]. The subgroups are:

- 1) AF151022 AF492830 AK000759 AK075162 BC002462 BC004166 BC007072 BC022270 BC035807 D37827 NM_004364 NM_006278 NM_014149 GO:0008152
- 2) AF040958 GO:0005975
- 3) AF356527 GO:0050789
- 4) AF374726 GO:0007165
- 5) AK093494 GO:0008152
- 6) BC015761 GO:0050896
- 7) XM_087062 GO:0008152.

These 13 sequences of the first subgroup are annotated with the general GO term metabolism 0008152 and so it is difficult inferring in which type of metabolic pathway they are involved. Probably this information, coupled with microarray expression data, is informative about some transcription regulation pathway in which these genes are involved as emerged by a more accurate analysis of blast results.

Conclusions. The clinical data after cleaning and data selection were able to correctly classify the distinction between CFS and NF. This is not surprising given that the diagnosis is based on clinical information. We cleaned the data in a way that left us with a completely consistent and classifiable data set. SNPs and Blood data sets were not effective in classification. Proteomics data analysis detects patterns less evident in CFS patients of the one detected in other groups. Analysis of gene differentially expressed in microarray data identify a set of genes. There is a consistency between this genes in terms of proximity in Gene Ontology under the general term metabolism. The indication that some metabolic pathway could be involved in CFS is consistent with the literature.

References

- [Camda2006 prot] wichita_clinical_irb_protocol.doc
<http://www.camda.duke.edu/camda06/datasets/>
[CAMDA 2006] [://www.camda.duke.edu/camda06](http://www.camda.duke.edu/camda06)
[Fukuda, 1994] Keiji Fukuda, Stephen E.Straus, Ian Hickie, Michael C.Sharpe, James G.Dobbins, Anthony Komaroff, International Chronic Fatigue Syndrome Study Group. The Chronic Fatigue Syndrome: A Comprehensive Approach to Its Definition and Study. *Annals of Internal Medicine* 1994 Dec 15;121(12):953-9
[Lin, 1998] Dekang Lin. An Information-Theoretic Definition of Similarity. Proc. 15th International Conf. on Machine Learning, 1998.
[Malossini, 2006] Andrea Malossini, Enrico Blanzieri and Raymond T. Ng. Detecting Uncertainty in Microarray Data Labeling by Data Perturbation. Submitted.
[Reeves, 2003] William C Reeves, Andrew Lloyd, Suzanne D Vernon, Nancy Klimas, Leonard A Jason, Gijs Bleijenbergh, Birgitta Evengard, Peter D White, Rosane Nisenbaum, Elizabeth R Unger and the International Chronic Fatigue Syndrome Study Group. Identification of ambiguities in the 1994 chronic fatigue syndrome research case definition and recommendations for resolution. *BMC Health Services Research* 3:25, 2003.
[Tusher, 2001] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response *PNAS*, Apr 2001; 98: 5116 – 5121.
[Vernon, 2002] Suzanne D. Vernon, Elizabeth R. Unger, Irina M. Dimulescu, Mangalathu Rajeevan, William C. Reeves. Utility of the blood for gene expression profiling and biomarker discovery in chronic fatigue syndrome. *Disease Markers* 4 (18) 2002 193-199.
[WEKA] <http://www.cs.waikato.ac.nz/ml/weka/>
[Whistler, 2003] Whistler T, Unger ER, Nisenbaum R, Vernon SD. Integration of gene expression, clinical, and epidemiologic data to characterize Chronic Fatigue Syndrome. *J Transl Med.* 2003;1:10.