Expression Microarray Classification using Topic Models

Manuele Bicego*, Pietro Lovato, Barbara Oliboni, Alessandro Perina
Computer Science Department, University of Verona
Strada le Grazie 15, 37134 - Verona, Italy

ABSTRACT

Classification of samples in expression microarray experiments represents a crucial task in bioinformatics and biomedicine. In this paper this scenario is addressed by employing a particular class of statistical approaches, called Topic Models. These models, firstly introduced in the text mining community, permit to extract from a set of objects (typically documents) an interpretable and rich description, based on an intermediate representation called topics (or processes). In this paper the expression microarray classification task is cast into this probabilistic context, providing a parallelism with the text mining domain and an interpretation. Two different topic models are investigated, namely the Probabilistic Latent Semantic Analysis (PLSA) and the Latent Dirichlet Allocation (LDA). An experimental evaluation of the proposed methodologies on three standard datasets confirms their effectiveness, also in comparison with other classification methodologies.

Categories and Subject Descriptors

I. Computing Methodologies [**I.5 Pattern Recognition**]: I.5.1 Models Statistical; J. Computer Applications [**J.3 Life and medical science**]: Biology and genetics

General Terms

Algorithms Experimentation

Keywords

Microarray expression, topic models, latent models, classification, cancer data sets

1. INTRODUCTION

*Corresponding author: email: manuele.bicego@univr.it Tel: +39 045 8027072, Fax: +39 045 8027068. M.B. is also with IIT, Istituto Italiano di Tecnologia, Via Morego, 30, 16163 Genova, Italy.

SAC'10 March 22-26, 2010, Sierre, Switzerland.

The recent wide employment of microarray tools in molecular biology and genetics have produced an enormous amount of data, which has to be processed to infer knowledge. Due to the dimension and complexity of those data, automatic tools coming from Computer Science and Data Analysis research areas have been successfully employed.

Computer Science methodologies may be very useful in the analysis of microarray data: among others, clear examples are tools aiding the microarray probe design, image processing-based techniques for the quantification of the spots, segmentation spot/background, grid matching, noise suppression [5]), methodologies for classification or clustering [18]. In this paper we focus on this last class of problems, and in particular on the classification task. In this context, many approaches have been presented in the literature in the past, each one characterized by different features, like computational complexity, effectiveness, interpretability, optimization criterion and others – for a review see e.g. [14, 17].

This paper provides a further contribution toward this direction. In particular we propose to solve the microarray classification task by employing a particular class of statistical models, typically known as *topic* or *latent models*: the two most famous examples are the Probabilistic Latent Semantic Analysis (PLSA – [13]) and the Latent Dirichlet Allocation (LDA – [3]). These powerful approaches have been introduced in the text understanding community for unsupervised topic discovery in a corpus of documents, in order to correlate the presence of a word in a document to the particular topic discussed in such document – the whole corpus of documents can then be described in terms of these topics. These techniques have also been largely applied in the computer vision community, in order to discover scene classes, by the use of visual topics, from a collection of unlabelled images [4] or to discover groups of geolocated images [7]. One of the main characteristics of this class of approaches is represented by their interpretability. Actually they can model a dataset in terms of hidden topics (or processes), which can reflect underlying and meaningful structures in the problem. Interpretability of techniques and results is going to become a stringent need, especially in bioinformatics: substituting "black box algorithms" (like PCA) with more intuitive representations may help the biologist in interpreting both algorithms and results [6]. This reasoning motivated the definition of a particular instance of the topic models (called Latent Process Decomposition) for clustering expression microarray data [16][1]. In that paper

[1]An optimized training version has been recently proposed

the general training formulation is provided, as well as an experimental evaluation of its ability in discovering meaningful patterns. In any case, no classification experiments have been reported.

In this paper we propose to investigate the use of this class of techniques in the expression microarray *classification* scenario; in particular we adapt the PLSA and LDA models to the microarray case, also providing a possible interpretation; moreover we customize the technique proposed in [16] in order to deal with the classification task. We show the suitability of PLSA and LDA in two expression microarray classification tasks. Since the goal of this paper is to investigate the representation power of the topic models, here we employ the simple K-Nearest Neighbor rule [10] on the topic model-based representations. The methodologies have also been compared with baseline methods (as PCA [10]) and with similar results proposed in other papers, with really encouraging performances.

The rest of the paper is organized as follows: section 2 contains a general description of the PLSA and the LDA techniques, together with their adaptation and interpretation in the expression microarray case. The experimental evaluation is reported in section 3. Finally, in section 4 conclusions are drawn and future perspectives are envisaged.

2. TOPIC MODELS

Topic models were introduced in the linguistic scenario, in order to describe and model documents. The basic idea underlying these methods is that each document is characterized by the presence of one or more topics (e.g. sport, finance, politics), which may induce the presence of some particular words. From a probabilistic point of view, the document may be seen as a mixture of topics, each one providing a probability distribution over words. A topic model represents a generative model for documents, since a simple probabilistic procedure permits to specify how documents are generated. In particular, a new document may be generated in the following way: first choose a distribution over topics; then, for each word in that document, randomly select a topic according to its distribution, and draw a word from that topic. It is possible to invert the process, in order to infer the set of topics that were responsible for generating a collection of documents.

The representation of documents and words with topic models has one clear advantage: each topic is individually interpretable, providing a probability distribution over words that picks out a coherent cluster of correlated terms. This may be really advantageous in the expression microarray context, since the final goal is to provide knowledge about biological systems, and provide possible hidden correlations.

A variety of probabilistic topic models have been used to analyze the content of documents and the meaning of words. These models all use the same fundamental idea – that a document is a mixture of topics – but make slightly different statistical assumptions. In the following we will briefly review the mathematics of the two topic models employed in this paper, namely PLSA and LDA.

2.1 PLSA

In the Probabilistic Latent Semantic Analysis (PLSA – [13]) the input is a dataset of N documents $\{d_i\}$, i=1,..., N, each one containing a set of words. Before applying PLSA, the dataset is summarized by a co-occurrence matrix of size $M \times N$, where the entry $<w_j, d_i>$ indicates the number of occurrences of the word w_j in the document d_i, also called $n(w_j, d_i)$. Each document d_i has n_i words. The presence of a word w_j in the document d_i is mediated by a latent *topic* variable, $z \in Z = \{z_1,..., z_Z\}$, also called *aspect* class, *i.e.*,

$$P(w_j, d_i) = \sum_{k=1}^{Z} P(w_j|z_k)P(z_k|d_i)P(d_i). \quad (1)$$

In practice, the topic z_k is a probabilistic co-occurrence of words encoded by the distribution $P(w|z_k)$, $w = \{w_1,..., w_M\}$, and each document d_i is compactly (usually, $Z < M$) modeled as a probability distribution over the topics, *i.e.*, $P(z|d_i)$, $z = \{z_1,..., z_Z\}$; $P(d_i)$ accounts for varying number of words.

The hidden distributions of the model, $P(w|z)$ and $P(z|d)$, are learnt using Expectation-Maximization (EM) [8], maximizing the model data-likelihood L:

$$L = \prod_{i=1}^{N} \prod_{j=1}^{M} P(w_j, d_i)^{n(w_j, d_i)} \quad (2)$$

The E-step computes the posterior over the topics, $P(z|w, d)$, and the M-step updates the hidden distributions. Once the model has been learnt, the most used inference, also called recognition inference, estimates the topic distribution of a novel document. Here, the learning algorithm is applied fixing the previously learnt distribution $P(w|z)$ and estimating $P(z|d)$ for the query document. For a deeper review of PLSA, see [13].

2.2 LDA

Latent Dirichlet Allocation (LDA - [3]) represents an extension of the PLSA. It is based on the same concepts of PLSA, namely words, documents and topics. The differences stem in the fact that the PLSA model does not make any assumptions about how the mixture weights are generated, making it difficult to generalize the model to new documents. On the contrary LDA extends this model by introducing a Dirichlet prior on the mixture weights, permitting a true generative model for the whole corpus of documents. The formulation becomes more formal and elegant, for more details see [11]. In this work we employ the LDA adaptation given in [16], which was originally designed for clustering expression microarray. The optimization of the likelihood function came out to be intractable, so that variational inference was employed to estimate the parameters of the model (through an EM approach). The mathematical formulation of this model is rather complex, and is out of the scope of this paper – interested readers are referred to the original paper [16].

2.3 Topic Models and Classification of Expression Microarray

As deducible from the previous sections, topic models may be very useful in the expression microarray context, since they may provide powerful and interpretable descriptions of experiments. In particular there is an analogy between the pairs word-document and gene-sample: actually it seems reasonable to intend the samples as documents and the genes as words. In fact each sample is characterized by a vector of genes expressions: the expression level of a gene in a sample

in [19].

may be easily interpreted as the count of words in a document (the higher the number the more present/expressed the word/gene is). In our case, therefore, we can consider the expression matrix as the count matrix $<w_j, d_i>$ of topic models. However we are forced to normalize and preprocess the matrix in order to have positive and integer values. Here we adopted a very simple technique, namely shifting and rounding the matrix. In some cases, moreover, a proper matrix scaling had to be applied in order to obtain suitable topic models. It is worth noting that gene expression is subject to complex co-regulation mechanisms, and there are aspects of this interdependence that cannot be captured with words co-occurrence. Nevertheless, we will show later that our methods may work properly even if disregarding this biological aspect.

Topic models have been originally introduced for clustering sets of documents: given the dataset, models are trained and analyzed in order to find clusters. In the classification scenario, however, we should describe how to carry out the training and the testing phases. The training phase, in both PLSA and LDA, is carried out by first learning the topic models on the training set. Then a set of features is extracted from each document (through the learned model); the transformed training set is then used to train a classifier. In the testing phase, the same feature extraction process is applied to the test document, resulting in a feature vector which is then classified using the trained classifier. In our work we adopted a similar but not identical feature extraction process in the PLSA and in the LDA cases. In particular, for PLSA, we employed the scheme proposed in [4] for natural scene categorization: given a sample d_i, we evaluate the topic probability $p(z|d_i)$, this representing a valid signature for the sample. In the LDA case, in a similar way, we used as feature vector the variational Dirichlet parameters obtained after feeding the sample to the trained generative model (called γ_{ik} in [16] – see the paper for more details).

In the experimental part we employed a very simple but effective classifier, namely the K-Nearest Neighbor (using the Euclidean distance – more complex distances may be used). Since we are interested in the description power of the proposed representations, we avoided the use of complex classifiers. In any case, we will see in the experimental part that such simple classifier is able to produce really competitive results.

3. EXPERIMENTAL EVALUATION

In this section the experimental evaluation is presented. In particular, we describe the employed datasets and the experimental protocols and parameters; then we provide results, discussion and comparative evaluations.

3.1 Experimental details and results

The suitability of the two topic models described in the previous section have been tested in two cancer classification tasks (classification of Leukemia and of Colon cancer), involving three well known datasets: the *Leukemia2* dataset [2], the *Leukemia1* dataset [12] and the *Colon Cancer* dataset [1]. The *Leukemia2* dataset contains the expressions of 72 samples (organized in 3 classes) with 11225 genes (following [17] all the genes with 'absent' calls in all samples were excluded from the analysis). The *Leukemia1* dataset contains 72 samples (2 classes) and 7129 genes. Finally, the *Colon cancer* dataset contains 62 samples (2 classes) and 2000 genes (which are a selection of the original 6500 genes, as in [1]).

The classification strategies described in the previous section have been applied to these datasets, by varying the number of topics of the topic models from 3 to 50 (step 2). However, the experimental evaluation showed that this parameter is not particularly crucial: in most experiments classification results did not vary too much when varying it. Classification errors have been computed using 10-fold cross validation (with 40 repetitions), using the K-NN classifier (K has been found with cross validation on the training set). As explained before, since we are interested in the description power of the proposed representations we avoided the use of complex classifiers. In order to augment the statistical significance of the results, PLSA and LDA trainings have been repeated 4 times (and results averaged). We used the PLSA implementation of J. Verbeek [2], whereas the LDA was adapted from the version developed by the authors of [16][3]. The presented approaches have been directly compared with a standard approach, which performs a Principal Component Analysis reduction [10] on the whole dataset, retaining a number of components able to explain the 99% of the variance.

As in many expression microarray analyses, a beneficial effect may be obtained by selecting a sub group of genes, in order to limit the dimensionality of the problem and to reduce the possible redundancy present in the dataset. Here we decided to perform the experiments described above by repeatedly giving in input different expression matrices, each one obtained by retaining a particular quantity of relevant genes. Relevance of genes may be measured using different methodologies, ranging from the simple variance up to complicate statistics. Here we employed a quite recent and promising technique, called Minimum-Redundancy Maximum-Relevance feature selection [9, 15][4]. We defined different datasets using a growing number of genes (in a logarithmic scale between 4 and 1024) and performed the experiments.

Results for *Leukemia2*, *Leukemia1* and *Colon cancer* datasets are displayed in Table 1.

3.2 Discussion and comparative evaluations

From Table 1 it is evident that the proposed approaches perform rather accurately in these experiments. In particular, the application of a proper gene selection is beneficial for both techniques, up to a certain level: if too few genes are retained, the co-occurrence matrices are not able to properly describe the problem (reasonably, a document with only 4 different words may be very difficult to characterize in terms of topics). Results are also competitive if compared to results proposed in [17] on the same datasets, obtained with different classifiers (in particular different versions of Support Vector Machines – see the paper for all details), for clarity reported in Table 2. Even if an absolute comparison may not be carried out (in [17] a slight different testing protocol has been used, gene selection is not performed everywhere), it is evident that our description performs comparatively well.

[2]http://lear.inrialpes.fr/ verbeek/software.
[3]http://www.enm.bris.ac.uk/lpd/.
[4]http://www.mathworks.com/matlabcentral/fileexchange/14916.

Leukemia2

N. of genes	PLSA	LDA	PCA
4	14.97% (29)	13.93% (5)	14.11%
8	5.65% (3)	6.77% (5)	10.71%
16	6.18% (27)	3.87% (5)	4.69%
32	4.98% (7)	0.9% (7)	**2.85%**
64	**2.89%** (3)	0.02% (5)	2.93%
128	3.26% (21)	**0%** (9)	2.97%
256	4.07% (5)	0.25% (5)	4.11%
512	4.03% (5)	1.18% (7)	5.61%
1024	3.78% (13)	9.66% (7)	5.32%
11225	**7.22%** (49)	17.19% (7)	9.9%

Leukemia1

N. of genes	PLSA	LDA	PCA
4	5.56% (19)	7.24% (35)	6.62%
8	5.13% (41)	6.69% (27)	8.73%
16	3.87% (9)	4.68% (7)	6.89%
32	3.40% (11)	1.68% (27)	4.50%
64	2.25% (9)	0.88% (35)	5.32%
128	**1.82%** (13)	**0.82%** (37)	6.16%
256	1.95% (7)	2.06% (13)	**3.26%**
512	2.62% (19)	1.52% (13)	6.55%
1024	2.93% (35)	2.26% (13)	4.50%
7129	**9.05%** (43)	14.92% (33)	11.68%

Colon Cancer

N. of genes	PLSA	LDA	PCA
4	13.30% (21)	18.70% (9)	17.05%
8	13.28% (45)	12.26% (15)	15.26%
16	12.82% (47)	12.04% (13)	15.84%
32	10.43% (47)	**11.43%** (31)	16.97%
64	**9.81%** (7)	12.84% (9)	14.70%
128	10.09% (13)	14.22% (35)	13.09%
256	10.15% (5)	12.21% (39)	**10.38%**
512	11.05% (7)	15.57% (35)	14.12%
1024	11.08% (13)	14.60% (15)	15.00%
2000	**11.47%** (9)	23.25% (13)	20.68%

Table 1: Classification errors of the proposed approaches for different datasets and for different number of retained genes. For PLSA and LDA, only the best results over the different number of topics has been reported (between brackets). In bold the best result for each technique.

In particular we were able to get an almost perfect accuracy on the datasets while using a very simple classification technique, the KNN: this means that the LDA space is really discriminative, as can be inferred by looking at the Figure 1, where the three classes are displayed in the 3-topic LDA space. Concerning the *Colon Cancer* dataset it may be noticed that in this case there is a beneficial effect in using PLSA, whereas LDA is not able to properly capture the underlying model, performing slightly worse than the baseline. In any case, the obtained results are really competitive, if compared with those published in [14] – and briefly reported in Table 3. The results are obtained with different methods of gene selection and using a wide variety of classifiers, spanning from simple linear discriminant analysis to complex support vector machines. In the table only the best classifier for each gene selection scheme has been reported – see the original paper for detailed description of the techniques. Also in this case comparison may not be absolute, since a slightly different cross validation procedure has been employed, but a clear idea of the comparative effectiveness of the presented methodologies may be obtained. It is worthwhile to notice that all our results have been obtained with the simple K-NN, while some improvements may be obtained by using more sophisticated machines. In order to show that, we perform on this dataset a further experiment employing Support Vector Machines (with radial basis function kernel – parameters have been set again with cross validation on the training set). For topic models, we chose the configuration leading to the best result with K-NN, namely PLSA with 64 genes and 7 topics. With this configuration we were able to reduce the classification error from 9.81% to 6.52% – this confirming the suitability of the proposed techniques.

A final comment concerns the interpretability of the re-

Method	*Leukemia2*	*Leukemia1*
SVM version 1	2.50%	2.68%
SVM version 2	2.68%	4.11%
SVM version 3	3.93%	4.11%
SVM version 4	2.50%	4.11%
SVM version 5	2.50%	4.11%
KNN	16.43%	12.86%
Neural Networks	23.39%	8.97%
Probabilistic Neural Networks	15.00%	16.79%

Table 2: Other results on the *Leukemia1* and *Leukemia2* dataset respectively, obtained from [17].

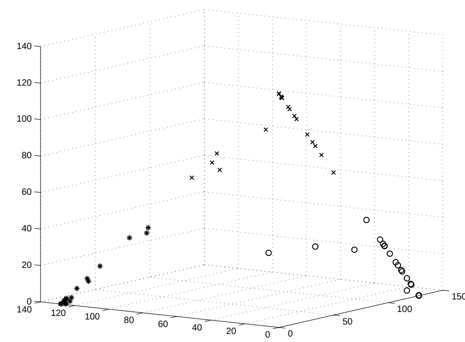

Figure 1: Resulting space after the application of the LDA (3 topics, 128 genes selected by MRMR.) on the *Leukemia2* dataset (3 classes, 24 samples each).

Gene selection	Best Method	Error
BSS/WSS	SVM-Rad	14%
Rank-based	Diagonal Linear Discr. An.	14%
Soft-thresholding	SVM-Rbf	12%

Table 3: Other results on the *Colon Cancer* dataset, obtained from [14].

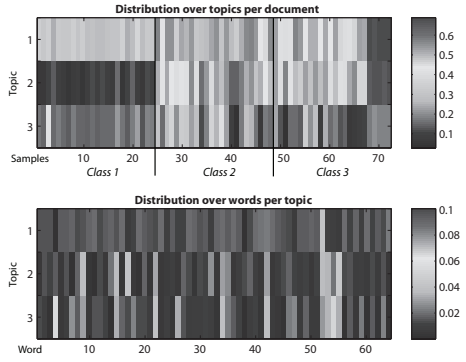

Figure 2: Probabilities of the PLSA (3 topics, 64 genes selected by MRMR) on the *Leukemia2* dataset. The top plot shows $p(z|d)$, the bottom $p(w|z)$. From the top one we can observe that the third topic highly characterizes the first class. In the distribution displayed below, we can observe that there are some words (like the fourth) peculiar for that topic: since words represent genes, this information may be exploited to understand the biological process.

sults. In particular, as described in the previous section, samples correspond to documents and genes to words. Therefore, some interesting properties of the genes may be obtained by visualizing the probability of the documents given the topics and that of the words given the topics. An example is reported in Fig. 2, where $p(z|d)$ and $p(w|z)$ are displayed. Since there is a particular set of topics related to one class (in particular the second topic), we may consider the most important words induced by such topic as the most crucial genes in that particular class.

4. CONCLUSIONS

In this paper we investigated the suitability of topic models in the expression microarray classification problem. In particular, we investigated the Probabilistic Latent Semantic Analysis (PLSA) and the Latent Dirichlet Allocation (LDA), giving their interpretation and adaptation in the peculiar applicative scenario. An experimental evaluation of the proposed methodologies on three standard datasets confirms the effectiveness of the proposed techniques, also in comparison with other classification methodologies.

5. REFERENCES

[1] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.*, 96(12):6745–6750, 1999.

[2] S. Armstrong, J. Staunton, L. Silverman, R. Pieters, M. den Boer, M. Minden, S. Sallan, E. Lander, T. Golub, and S. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41–47, 2002.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. of Mach. Learn. Res.*, 3:993–1022, 2003.

[4] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via PLSA. In *Proc. of European Conference on Computer Vision*, volume 4, pages 517–530, 2006.

[5] N. Brändle, H. Bischof, and H. Lapp. Robust DNA microarray image analysis. *Machine Vision and Applications*, 15:11–28, 2003.

[6] G. Brelstaff, M. Bicego, N. Culeddu, and M. Chessa. Bag of peaks: interpretation of nmr spectrometry. *Bioinformatics*, 25(2):258–264, 2009.

[7] M. Cristani, A. Perina, U. Castellani, and V. Murino. Geo-located image analysis using latent representations. In *Proc. Conf. Computer Vision and Pattern Recognition, 2008*, pages 1–8, 2008.

[8] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977.

[9] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. In *Proc. of IEEE Computer Society Bioinformatics Conference*, pages 523–529, 2003.

[10] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, second edition, 2001.

[11] M. Girolami and A. Kabán. On an equivalence between plsi and lda. In *Proc. of ACM SIGIR conf. on Research and development in informaion retrieval*, pages 433–434, 2003.

[12] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999.

[13] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196, 2001.

[14] J. Lee, J. Lee, M. Park, and S. Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869–885, 2005.

[15] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(8):1226–Ű1238, 2005.

[16] S. Rogers, M. Girolami, C. Campbell, and R. Breitling. The latent process decomposition of cdna microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(2):143–156, 2005.

[17] A. Statnikov, C. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005.

[18] F. Valafar. Pattern recognition techniques in microarray data analysis: A survey. *Annals of the New York Academy of Sciences*, 980:41–64, 2002.

[19] Y. Ying, P. li, and C. Campbell. A marginalized variational bayesian approach to the analysis of array data. *BMC Proceedings*, 2(Suppl 4):S7, 2008.