A Comparison on Score Spaces for Expression Microarray Data Classification

Alessandro Perina[1], Pietro Lovato[2], Marco Cristani[2,3], and Manuele Bicego[2]

[1] Microsoft Research, Redmond, USA
[2] University of Verona, Department of Computer Science, Verona, Italy
[3] Italian Institute of Technology, Genoa, Italy

Abstract. In this paper an empirical evaluation of different generative scores for expression microarray data classification is proposed. Score spaces represent a quite recent trend in the machine learning community, taking the best of both generative and discriminative classification paradigms. The scores are extracted from topic models, a class of highly interpretable probabilistic tools whose utility in the microarray classification context has been recently assessed. The experimental evaluation, performed on 3 literature datasets and with 7 score spaces, demonstrates the viability of the proposed scheme and, for the first time, it compares *pros* and *cons* of each space.

1 Introduction

Microarrays represent a widely employed tool in molecular biology and genetics, allowing DNA and/or RNA analysis to be carried out in microminiaturized highly parallel formats. DNA microarray applications are usually directed at gene expression analysis that usually implies to process huge amounts of data. Therefore, fast and robust methodologies are required to face diverse microarray analysis problems such as noise suppression [7], segmentation of spots/background, quantification of the spots, grid matching, clustering or classification [9,17,29,31].

In this paper we focus on this last class of problems, where many approaches have been presented in the literature in the past, each one focusing on different aspects, like computational complexity, effectiveness, interpretability, optimization criterion and others – for a review see e.g. [17,29]. Among others, in recent years some promising techniques were based on a particular class of probabilistic approaches, called topic models, showing optimal and highly interpretable results [2,22,24]. Such probabilistic topic models, the two most famous examples being the Probabilistic Latent Semantic Analysis (PLSA [15]) and the Latent Dirichlet Allocation (LDA [5]), have been imported from the text analysis realm as workhorses in several scientific fields [6,8,33]. Their wide usage is motivated by their simplicity and expressiveness in dealing with very large datasets both in samples and features number. Therefore, they appeared to be a convenient tool for the microarray data analysis problem, and especially in the context of expression microarray classification [2,22]. Nevertheless, not all the potentialities

M. Loog et al. (Eds.): PRIB 2011, LNBI 7036, pp. 202–213, 2011.

of these schemes have been exploited in such context. To overcome this problem, here we make a step forward, by applying a hybrid generative-discriminative paradigm based on the definition of *generative score spaces* [28]. Generative and discriminative classification schemes represent the two main directions for classifying data: each philosophy brings pros and cons with itself and the last research frontier aims at fusing them, following heterogeneous recipes [4,20,19]. In this paper, we adopt the "staged" strategy: the idea is that a generative framework (in this case the PLSA[1]) is instantiated and learned. Then, surrogates of the learning (in the simplest case, likelihood probabilities) are injected as features in a discriminative classifier which is eventually learned. In some cases, this is theoretically proved to rise the purely generative classification performances [16,18,20,30].

In this paper, we show how different strategies to build score spaces lead to diverse classification accuracies, considering different publicly available microarray datasets. Obtained results confirm the goodness of the classification strategies based on topic models in the expression microarray classification context.

2 Methodology

In this section the background concepts regarding topic models and generative embedding are reported. In particular, after introducing the general ideas underlying the PLSA model, we will present it by using the terminology and the notation of the document analysis context. Then we will briefly review how the framework of hybrid generative-discriminative approach can be employed alongside the PLSA, and how it is applied in the microarray classification scenario.

2.1 Probabilistic Latent Semantic Analysis

In Probabilistic Latent Semantic Analysis (PLSA – [15]) the input is a set of D documents, each one containing a set of words taken from a vocabulary of cardinality N. The documents are summarized by an occurrence matrix of size $N \times D$, where $n(w_j, d_i)$ indicates the number of occurrences of the word w_j in the document d_i. 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.*,

$$p(w_j, d_i) = \sum_{k=1}^{Z} p(z_k) \cdot p(w_j|z_k) \cdot p(d_i|z_k) \tag{1}$$

In practice, each k-th topic z_k[2] is a probabilistic co-occurrence of words encoded by the distribution $\beta(w) = p(w|z_k)$, $w = \{w_1,..., w_N\}$, and each document d_i is compactly (usually, $Z < N$) modeled as a probability distribution over the topics,

[1] PLSA is commonly employed as a generative model, even if it is not under a strict formal treatment. See the text for further details.

[2] Throughout the paper v_k stands for the variable v assuming the value k.

i.e., $p(z|d_i)$, $z = \{z_1,...,z_Z\}$ (note that this formulation, derived from $p(d_i|z)$, provides an immediate interpretation).

The hidden quantities of the model, $p(w|z)$, $p(d|z)$ and $p(z)$, are learnt using Expectation-Maximization (EM) [10], maximizing the model data-loglikelihood $\mathcal{L}$:

$$\mathcal{L} = \prod_{j=1}^{N} \prod_{i=1}^{D} n(w_j, d_i) \cdot \log p(w_j, d_i) \tag{2}$$

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 one can estimates the topic proportion of an unseen document. Here, the learning algorithm is applied by fixing the previously learnt parameters $p(w|z)$ and estimating $p(d|z)$ for the document in hand. For a deeper review of PLSA, see [15].

It is important to note that d is a dummy index into the list of documents in the training set. Thus, d is a multinomial random variable with as many possible values as there are training documents and the model learns the topic mixtures $p(d|z)$ only for those documents on which it is trained. For this reason, PLSA is not a well-defined generative model of documents; there is no natural way to assign probability to a previously unseen document and the procedure just described to estimate $p(d|z)$ is an heuristic [15].

PLSA may be very useful in the expression microarray context, since it may provide powerful and interpretable descriptions of experiments [3,22,24]. In particular there is an analogy between the pairs *word-document* and *gene-sample*: actually it is reasonable to intend the samples as documents and the genes as words. In fact the expression level of a gene in a sample 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, after a proper normalization in order to have positive and integer values.

3 Generative Score-Spaces

Pursuing principled hybrid architectures of discriminative and generative classifiers is currently one of the most interesting, useful, and difficult challenges for Machine Learning. The underlying motivation is the proved complementarity of discriminative and generative estimations: asymptotically (in the number of labeled training examples), classification error of discriminative methods is lower than for generative ones [19]. On the other side, generative counterparts are effective with less, possibly unlabeled, data; further, they provide intuitive mappings among structure of the model and data features. Among these hybrid generative-discriminative methods, "generative score space" approaches grow in the recent years their importance in the literature [6,16,18,20,27,28,30].

Generative score space framework consists of two steps: first, one or a set of generative models are learned from the data; then a score (namely a vector of features) is extracted from it, to be used in a discriminative scenario. The idea is to extract fixed dimensions feature vectors from observations by subsuming the process of data generation, projecting them in highly informative spaces called score spaces. In this way, standard discriminative classifiers such as support vector machines, or logistic regressors are proved to achieve higher performances than a solely generative or discriminative approach.

Using the notation of [27,20], such spaces can be built from data by mapping each observation x to the fixed-length score vector $\varphi^{f}_{\hat{F}}(x)$,

$$\varphi^{f}_{\hat{F}}(x) = \varphi_{\hat{F}} f(\{P_i\,(x|\theta_i))\}), \tag{3}$$

where $P_i\,(x|\theta_i)$ represents the family of generative models learnt from the data, f is the function of the set of probability densities under the different models, and $\hat{F}$ is some operator applied to it. In general, the generative score-space approaches help to distill the relationship between a model parameters θ and the particular data sample.

Generative score-space approaches are strictly linked to generative kernels family, namely kernels which compute similarity between points through a generative model – the most famous example being the Fisher Kernel [16]): Typically, a generative kernel is obtained by defining a similarity measure in the score space, e.g. the inner product.

Score spaces are also called model dependent feature extractors, since they extract features from a generative model. We can divide score spaces in two families: parameters-based and hidden variable-based. Let us review the 7 different score spaces tested in this paper.

3.1 Parameters Based Score Space

These methods derive the features on the basis of differential operations linked to the parameters of the probabilistic model.

The Fisher Score. Fisher kernel [16] was the first example of generative score space. At first, a parameter estimate $\hat{\theta}$ is obtained from training examples. Then, the tangent vector of the data log likelihood $\log p(x|\theta)$ is used as a feature vector. Referring to the notation of [27,20], the score function is the data log likelihood, while the score argument is the gradient.

The fisher score for the PLSA model has been introduced in [14], starting from the asymmetric formulation of PLSA. In this case, the log-probability of a document d_i is defined by

$$l(d_i) = \frac{\log p(d_i, w)}{\sum_m n(d_i, w_m)} = \sum_{j=1}^{N} \hat{p}(w_j|d_i) \log \sum_{k=1}^{Z} p(w_j|z_k) \cdot p(d_i|z_k) \cdot p(z_k), \tag{4}$$

where $\hat{p}(w_j|d_i) \equiv n(d_i, w_j)/\sum_m n(d_i, w_m)$ and where $l(d_i)$ represents the probability of all the word occurrences in d_i normalized by document length.

Differentiating Eq. 4 with respect to $p(z)$ and $p(w|z)$, the PLSA model parameters, we can compute the score. In formulae:

$$\frac{\partial l(d_i)}{\partial p(w_r|z_t)} = n(d_i, w_t) \cdot \frac{p(d_i|z_t) \cdot p(z_t)}{\sum_k p(w_r|z_k) \cdot p(d_i|z_k) \cdot p(z_k)} \tag{5}$$

$$\frac{\partial l(d_i)}{\partial p(z_t)} = \sum_{j=1}^{W} n(d_i, w_j) \cdot \frac{p(d_i|z_t) \cdot p(w_j|z_t)}{\sum_k p(z_k) \cdot p(w_j|z_k) \cdot p(d_i|z_k)} \tag{6}$$

As visible from Eq. 5-6, the samples are mapped in a space of dimension $W \times Z + Z$. The fisher kernel is defined as the inner product in this space. We will refer to it as FSH.

TOP Kernel Scores. Top Kernel and the tangent vector of posterior log odds score space were introduced in [30]. One of the aim of the paper was to introduce a performance measure for score spaces. They considered the estimation error of the posterior probability by a logistic regressor and they derived the TOP kernel in order to maximize the performance.
Whereas the Fisher score is calculated from the marginal log-likelihood, TOP kernel is derived from Tangent vectors Of Posterior log-odds. Therefore the two score spaces have the same score function (i.e., the gradient) but different score argument, which, for TOP kernel $f(p(x|\theta)) = \log p(c = +1|x, \theta) - logp(c = -1|x, \theta)$ where, c is the class label. We will refer to it as TOP.

Log Likelihood Ratio Score Space. The loglikelihood ratio score space is introduced in [28]. Its dimensions are similar to the Fisher score, except that the procedure is repeated for each class: a model per class is learnt θ_c and the gradient is applied to each class-loglikelihood $\log p(x|\theta_c)$. The dimensionality of the resulting space is C-times the dimensionality of the original Fisher score. We will refer to it as LLR.

3.2 Random Variable Based Methods

These methods, starting from considerations in [20], seek to derive feature maps on the basis of the log likelihood function of a model, focusing on the random variables rather than on the parameters in their derivation (as done in the parameter-based score spaces).

Free Energy Score Space. In the Free Energy Score Space [20], the score function is the free energy while the score argument is its unique decomposition in addends that composes it[3]. Free energy is a popular score function representing

[3] This is true once a family for the posterior distribution is given. See the original paper for details.

a lower bound of the negative log-likelihood of the visible variables used in the variational learning. For PLSA it is defined by the following equation:

$$\begin{aligned}\mathcal{F}(d_i) = &\sum_w n(d_i, w) \cdot \sum_z p(z|d, w) \cdot \log p(z|d, w) \\ &- \sum_w n(d_i, w) \cdot \sum_z p(z|d, w) \cdot \log p(d, w|z) \cdot p(z)\end{aligned} \tag{7}$$

where the first term represents the entropy of the posterior distribution and the second term is the cross-entropy. For further details on the free energy and on variational learning see [12], on the PLSA's free energy see [15].

As visible in Eq. 7 both terms are composed of $Z \times N$ addends $\{f_j\}_{j=1}^{Z \times N}$, and their sum is equal to the free energy. In generative classification, a test data is assigned to the class which gives the lower free energy (i.e., higher loglikelihood). The idea of FESS is to decompose the free energy of each class in its addends, i.e., $\mathcal{F}(d_i)^c = \sum_j \{f_{j,c}\}$ and to add a discriminative layer by estimating a set of weights $\{w_{j,c}\}$ through a discriminative method.

For PLSA this results in a space of dimension equal to $C \times 2 \times Z \times W$; we will refer to this score space FESS L_3.

In [20] the authors point out that, if the dimensionality is too high, some of the sums can be carried out to reduce the dimensionality of the score vector before learning the weights. The choice of the addend to optimize is intuitive but guided by the particular application. In our case, as previously done in [18, 21], we perform the sums over the word indices, optimizing the topics contribute. The resulting score space has dimension equal to $C \times 2 \times Z$; we will refer to this score space FESS L_2.

Posterior Divergence. Posterior Divergence score space is described in [18]. Like FESS it takes into account how well a sample fits the model (crossentropy terms in FESS) and how uncertain the fitting is (entropy terms in FESS, Eq. 7) but it also assesses the change in model parameters brought on by the input sample, i.e. how much a sample affects the model. These three measures are not simply stacked together, but they are derived from the incremental EM algorithm which, in the E-step only looks at one or few selected samples to update the model in each iteration. Details on posterior divergence score vector for PLSA and on its relationships with FESS case can be found in [18]. We will refer to this score space as PD.

Classifying with the Mixture of Topics of a Document. Very recently, PLSA has been used as dimensionality reduction method in several fields, like computer vision, bioinformatics and medicine [6, 2, 8]. The idea is to learn a PLSA model to capture the co-occurrence between visual words [6, 8], or gene expressions [2], which represent the (usually) high-dimensional data description; co-occurrences are captured by the topics. Subsequently, the classification is performed using the topic distribution of a document as its descriptor.

Since we are extracting features from a generative model, we are defining a score space which is the (Z-1)-dimensional simplex. In this case, the score argument f, a function of the generative model, is the topic distribution $p(z|d)$ (using Bayes' formula, one can easily derive $p(z|d)$ starting from $p(d|z)$), while the score function is the identity. We will refer to this score space as TPM, or citing [2], the first work in the context of microarray classification that used this technique.

Summary. Summarizing, here we propose to face the expression microarray classification task by first learning a PLSA model, then extracting a score space, and finally classifying the samples in this new space, using a discriminative classifier (e.g. a Support Vector Machine).

4 Experimental Evaluation

The suitability of the proposed classification schemas has been tested using three different well-known datasets, briefly summarized in Tab. 1. The whole description of each dataset may be found in the reported reference.

Table 1. Summary of the employed dataset

Dataset Name	n. of genes	n. of samples	n. of classes	citation	BIC
1. Colon cancer	2000	62	2	[1]	6
2. Ovarian cancer	1513	53	2	[11]	4
3. DLBCL	6285	77	2	[26]	4

As in many expression microarray analysis, a beneficial effect may be obtained by selecting a sub group of genes, using a prior belief that genes varying little across samples are less likely to be interesting. Hence, we decided to perform the experiments by retaining the top 500 genes ranked by decreasing variance, as done also in [24].

A crucial issue arising when learning a topic model is to decide beforehand the number of topics. Here we employed the well-known Bayesian Information Criterion [25], which penalizes the likelihood with a penalty term which depends on the number of free parameters of the model – in such way, larger models which do not lead to a substantial increase in the likelihood are discouraged. In the PLSA model, the free parameters are $(D-1)\cdot Z+(N-1)\cdot Z+(Z-1)$, where Z and N refers to the number of topics and the number of words respectively. The penalization term is then given by

$$Pen. = \frac{1}{2} \cdot ((D-1) \cdot Z + (N-1) \cdot Z + (Z-1)) \cdot \log \sum_{j=1}^{N} \sum_{i=1}^{D} n(d_i, w_j) \qquad (8)$$

The best number of topic is found by searching for the maximum of the penalized likelihood, varying the number of topics from 2 to 50. The optimal number of topic for each dataset is shown in Tab. 1, column BIC.

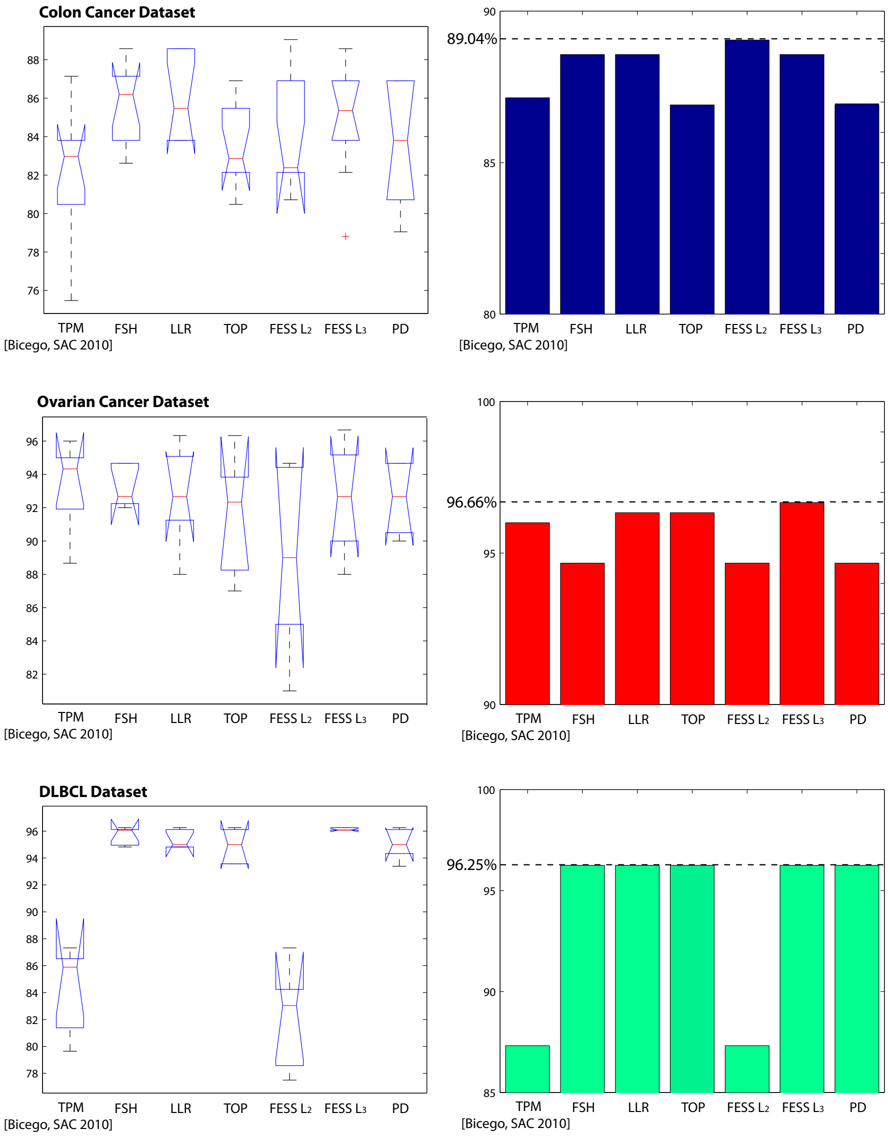

Fig. 1. Microarray classification results. TPM stands for the method present in [2], FSH is the fisher score space, LLR is the loglikelihood ration score space, TOP is the tangent of posterior log-odds score space, FESS L_2 and FESS L_3 are two complexities of the free energy score space while PD is the posterior divergence score space. See the text for details. Please print in color.

The errors have been found using a cross validation scheme: in particular the subdivision in training and testing set is carried out using 10-fold crossvalidation. Differently from [2] we learned the generative models only with the training data; this is necessary since LLR, TOP, FESS and PD require to learn a model per class, and we cannot use labels at training time. In order to have a significant results, we repeated each test 10 times and the results of this 10 repetitions have been validated through the standard anova variance test [23]. Finally, as final classifier we used a support vector machine with linear kernel; as in [16,30], the similarity between two datapoints is defined as the inner product of their scores. Before computing the kernels, the scores are normalized to have zero mean and unit variance; the constants to perform the normalization are computed with the training set and applied to each test sample.

Results are shown in Fig. 1, where each row of the figure describes a dataset. The graph on the left is a boxplot useful to assess the statistical significance of the results. The red bar is the median accuracy of the 10 repetitions, the edges of the blue box are the 25^{th} and 75^{th} percentiles, while the whiskers (black dotted bars) extend to the most extreme data points not considered outliers. Outliers are plotted individually with a red cross. Two medians are significantly different at the 5% significance level if their intervals (boxes) do not overlap.

The bar graphs on the right represent the accuracy obtained in correspondence of maximum loglikelihood value of the generative model. The best result among all the score spaces considered is textually reported on the left of the figure.

By looking at the figures and examining the results, the following observations may be extracted:

Colon Cancer dataset : Fisher, Loglikelihood Ratio and FESS L_3 are statistically better than the method presented in [2], while all the other methods are clearly better but without statistical significance.

Ovarian Cancer dataset : all the methods seem to be equivalent even if, once again, the best result is obtained with FESS.

DLBCL dataset : it is clear that [2] and FESS L_2 perform significantly worse than all the other score space, which in turn do not differ much.

To better understand the differences between the considered score spaces, we varied the number of topics between 2 and 20, to see how they are robust to this value. Mean accuracies (over 10 repetitions) for Fisher, FESS L_3, PD and TPM are shown in Fig. 2. Score spaces based on the parameters are clearly less sensitive to this value[4], while for TPM and FESS the results obtained for different Z's are statistically different (t-test, significance level 5%). Despite being a score space based on random variables like TPM and FESS, PD looks very robust (see also the very small variance among the repetitions) to Z. This is not surprising since PD is composed by entropy and crossentropy terms (as FESS), but also it has a set of extra terms that assess the change in model parameters brought on by the

[4] We have found that for this test the results of TOP are nearly identical to FSH.

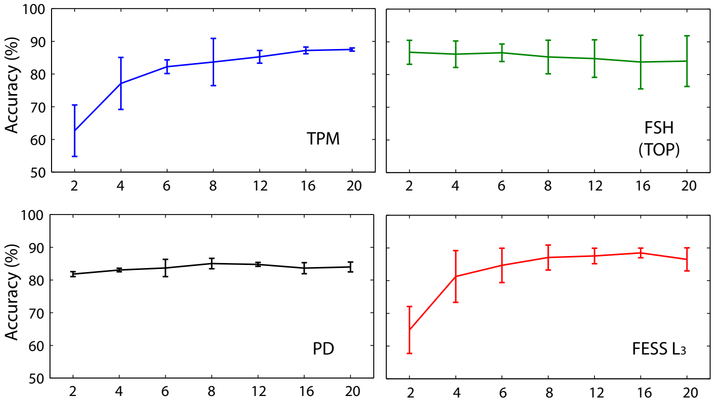

Fig. 2. Robustness of the score space to the number of topics (Z) for the Colon Cancer dataset. On the x-axis we show the number of topic, on the y-axis the accuracy. The four lines are the mean accuracy, while the vertical bars represent the variance computed across the 10 repetitions.

input sample d, which is characteristic of parameter based methods. This extra set permits to inherit the peculiar robustness to changes in Z of the parameters based method.

As a last test we tried transductive learning, namely learning a single model for all the available data, not using labels [32, 13] – on Fisher, FESS L_2, FESS L_3 and [2], to assess if this has some influence on the accuracy. For each fold and for each repetition, we learned a *single* model using all the data. Subsequently we used the training labels to train the discriminative classifier. Transductive learning has the problem that it requires to learn a model each time a test sample is available.

We performed anova test considering the three datasets together as different factors, with the following null hypothesis: "*Transductive and non-Transductive learning do not differ*". The hypothesis is confirmed with p-values respectively of 0.8473, 0.094 and 0.8683. This means that FESS is the less robust even we cannot claim that the difference of the results is statistically significant.

5 Conclusions

In this paper different generative score spaces have been evaluated, with the aim of classifying expression microarray data. Such score spaces are built on the PLSA generative model, a probabilistic tool whose usefulness in this context has been already assessed. Experimental results confirm the viability of the proposed hybrid schemes, also in comparison with the state of the art. In particular, all the score spaces introduced here outperform the previously published frameworks on microarray classification [2, 22]. FESS reached the best classification results even if the variance across repetitions or changes in Z was sensibly higher than the other score spaces. Fisher, TOP and LLR performed similarly and they are sufficiently robust. PD presented performances slightly inferior to the other methods but it has shown the best robustness to changes in number of topics and multiple restarts. Finally, the accuracies reported here can be further improved using more complex kernels, like done in [22].

Acknowledgements. The authors want to thank Xiong Li for providing the code for the score space based on the posterior divergence.

References

1. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A.: 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. Bicego, M., Lovato, P., Oliboni, B., Perina, A.: Expression microarray classification using topic models. In: ACM SAC - Bioinformatics track (2010)
3. Bicego, M., Lovato, P., Ferrarini, A., Delledonne, M.: Biclustering of expression microarray data with topic models. In: Proc. of International Conference on Pattern Recognition (2010)
4. Bishop, C., Lasserre, J.: Generative or discriminative? getting the best of both worlds. Bayesian Statistics 8, 3–24 (2007)
5. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
6. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
7. Brändle, N., Bischof, H., Lapp, H.: Robust DNA microarray image analysis. Machine Vision and Applications 15, 11–28 (2003)
8. Castellani, U., Perina, A., Murino, V., Bellani, M., Rambaldelli, G., Tansella, M., Brambilla, P.: Brain morphometry by probabilistic latent semantic analysis. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6362, pp. 177–184. Springer, Heidelberg (2010)
9. de Souto, M., Costa, I., de Araujo, D., Ludermir, T., Schliep, A.: Clustering cancer gene expression data: A comparative study. BMC Bioinformatics 9 (2008)
10. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39, 1–38 (1977)
11. Dhanasekaran, S., Barrette, T., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K., Rubin, M., Chinnaiya, A.: Delineation of prognostic biomarkers in prostate cancer. Nature 412(6849), 822–826 (2001)
12. Frey, B.J., Jojic, N.: A comparison of algorithms for inference and learning in probabilistic graphical models. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005)
13. Gammerman, A., Vovk, V., Vapnik, V.: Learning by transduction. In: Proc. of Uncertainty in Artificial Intelligence (1998)
14. Hofmann, T.: Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In: Adv. in Neural Information Processing Systems (1999)

15. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42, 177–196 (2001)
16. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Adv. in Neural Information Processing Systems (1998)
17. Lee, J., Lee, J., Park, M., Song, S.: An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis 48(4), 869–885 (2005)
18. Li, X., Lee, T.S., Liu, Y.: Hybrid generative-discriminative classification using posterior divergence. In: Proc. of Conference on Computer Vision and Pattern Recognition (2011)
19. Ng, A., Jordan, M.: On discriminative vs generative classifiers: A comparison of logistic regression and naive Bayes. In: Adv. in Neural Information Processing Systems (2002)
20. Perina, A., Cristani, M., Castellani, U., Murino, V., Jojic, N.: Free energy score space. In: Adv. in Neural Information Processing Systems (2009)
21. Perina, A., Cristani, M., Castellani, U., Murino, V., Jojic, N.: An hybrid generativediscriminative framework based on free energy terms. In: Proc. of the International Conference on Computer Vision (2009)
22. Perina, A., Lovato, P., Murino, V., Bicego, M.: Biologically-aware latent dirichlet allocation (balda) for the classification of expression microarray. Proc. of Pattern Recognition in Bioinformatics (2010)
23. Rao, C.R.: Diversity: Its Measurement, Decomposition, Apportionment and Analysis. Sankhy: The Indian Journal of Statistics, Series A 44(1), 1–22 (1982)
24. Rogers, S., Girolami, M., Campbell, C., Breitling, R.: The latent process decomposition of cdna microarray data sets. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2(2), 143–156 (2005)
25. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics 6, 461–464 (1978)
26. Shipp, M., Ross, K.: Diffuse large b-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. Nature Medicine 8, 68–74 (2002)
27. Smith, N., Gales, M.: Speech recognition using svms. In: Adv. in Neural Information Processing Systems (2002)
28. Smith, N.D., Gales, M.J.F.: Using SVMs to Classify Variable Length Speech Patterns. Tech. rep., Cambridge University Engineering Dept. (2002)
29. Statnikov, A., Aliferis, C., Tsamardinos, I., Hardin, D., Levy, S.: A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21(5), 631–643 (2005)
30. Tsuda, K., Kawanabe, M., Rotsch, G., Sonnenburg, S., Mueller, K.R.: A new discriminative kernel from probabilistic models. In: Neural Computation. MIT Press (2001)
31. Valafar, F.: Pattern recognition techniques in microarray data analysis: A survey. Annals of the New York Academy of Sciences 980, 41–64 (2002)
32. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1995)
33. Xing, D., Girolami, M.: Employing latent dirichlet allocation for fraud detection in telecommunications. Pattern Recogn. Lett. 28, 1727–1734 (2007)