Expression Microarray data classification using Counting Grids and Fisher Kernel

Alessandro Perina
Istituto Italiano di Tecnologia (IIT) and
Microsoft Research
alessandro.perina@iit.it

Maria Kesa
Tallinn University of Technology
maria.kesa@gmail.com

Manuele Bicego
University of Verona
manuele.bicego@univr.it

***Abstract*—Hybrid generative-discriminative models are useful in biomedical applications– generative modeling extracts interpretable features from raw data, highlighting its properties and increasing classification accuracy when used as input for a discriminative classifier. This raises the question: which generative model should be used for a particular application? In this paper we apply a recently proposed generative model called the Counting Grid to expression microarray data and derive the corresponding Fisher kernel. We justify why this model is particularly well-suited for this application and evaluate classification accuracy on four gene expression data sets, including three tumor data sets and a blood sample data set from schizophrenic patients and healthy controls. We report state of the art results on three of the analyzed data sets and closely match the accuracy from previous work on the other.**

I. Introduction

Microarrays are a widely employed tool in molecular biology and genetics, used to infer expression levels of thousands of genes in different experimental conditions. Among the different computational analyses that can be carried out, an important class is focused on the classification of experiments, namely the discrimination among different samples on the basis of gene expressions. In this context, a class of recent and promising approaches [1], [2], [3], [4] were based on tools typically employed to model documents, called *topic models* [5], [6]. Such approaches start from the parallelism that can be set between the pair word-document and the pair gene-sample. Expression levels are modeled as counts and are used as raw inputs for learning generative models such as topic models, that describe how this "bag-of-features" (simply a collection of counts without specifications of their relationships) arose.

In the context of modeling of general count data, a novel promising approach, called the Counting Grid model, has been recently proposed in [7]. This model starts with a spatial metaphor. The observed expression values in a sample are assumed to be generated independently from probability distributions that differ across samples. To model the generative process, we first create a grid that wraps around itself and connects at the boundaries to form a torus. Every discrete point on this grid is associated with a distribution over genes. To generate the gene expression profile of the sample we move along the grid with a window of fixed size. At a particular position, we average the distributions in the window and generate the gene expression values from that average distribution. The distributions on the grid can be inferred from data using the Expectation Maximization (EM) algorithm.

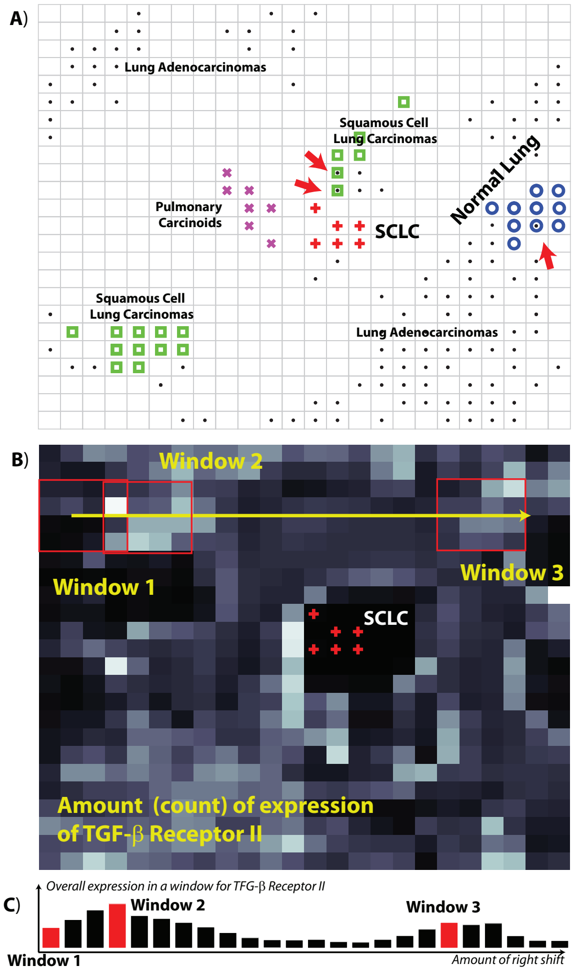

Fig. 1. A) Clustering of the samples of the Lung dataset on the counting grid. In this model samples are mapped in window and here we indicate each sample with a marker placed in correspondence of the upper left corner of its window. Different markers correspond to different types of tumor (Lung Adenocarinoma, Squamous Cell Lung Carcinomas, Small Cell Lung Carcinoma (SCLC), Pulmonary Carninoid and Normal Lung). B) Expression Count for the gene TGF-β Receptor II (e.g., $\pi_{\mathbf{k},z=\text{“}TGF-\beta Receptor I''}$, see Sec. II). C) Variation of the amount of expression of TGF-β Receptor II, while shifting of 1 location, from Window 1 to Window 3. Bars correspond on the sum in of the expression in a window of size $W = 4 \times 4$ (e.g., $h_{\mathbf{k},z=\text{“}TGF-\beta Receptor I''}$, see Sec. II)

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.311

The position of a sample in the grid (the upper-left corner of the window) assumes the role of a latent variable. The mathematical details are presented in the Section II. This model is particularly well suited for data that exhibits smooth variation between samples. Expression values are biologically constrained to lie within certain bounds by purifying selection [8] and variation in only a few expression values can cause a pathology. This specific property of the data is captured well by the counting grid model. This is illustrated in Fig. 1 where we embedded tumour samples [9] on a counting grid. It is apparent that samples with different labels cluster and small shifts on the grid can cause a change in class. In the same figure (panel B), we show the level of expression in each location of the grid for the "TGF-β Receptor II" gene which is known to be related to lung tumors [9]. Samples are generated taking a window and starting from "Window 1" and shifting to the right. The overall expression in a window rises to reach a peak in "Window 2",then smoothly decreases and finally rises again in "Window 3". The smooth variation of the expression is illustrated in the panel C. The grid also shows that small-cell lung carcinomas cluster in a region, where 'TGF-β Receptor II" expression is very low. Loss of expression of this gene is a well-known characteristic of the disease [10].

In previous works, counting grids have been employed for gene selection [11] and classification [12], by explpoiting the mapping of each sample on the grid. However, there are still margins for improvements, particularly in classification. In this paper we propose and investigate the effect of a sophisticated generative kernel over the counting grid. Generative kernels represent the most striking example of the hybrid generative-discriminative approaches, a recent trend in pattern recognition aimed at combining the best of both generative and discriminative paradigms [13], [14], [6]. When using a generative kernel, the generative model is learned from the data and used to derive a kernel between objects, to be used with a discriminative classifier (e.g. SVM). The use of these kernels for microarray data has proven to drastically improve the performances of classification techniques based on topic models [3], in comparison to standard classification rules. The main goal of this paper is to investigate if such kernels can be beneficial also when derived for richer generative models, like the counting grid. In particular here we derive and employ the Fisher Kernel [13], which represents the first and the most employed generative kernel and turned out to be very robust in [3]. The Fisher Kernel between two samples can be computed as the inner product between the so called Fisher Scores, which are the first derivatives of the log-likelihood with respect to the parameters of the generative model, evaluated in the two given samples. In the case of the counting grid, the Fisher kernel is intractable, because its log likelihood is intractable. However, following the same trick used for the latent Dirichlet allocation [15], we are able to extract the Fisher scores from the free energy (its derivatives with respect to the parameters are functions of the variational parameters). This is a contribution of this paper, since the Fisher Kernel for the counting grid has never been computed or tested.

The rest of the paper is organized as follows: Section II reviews the basic mathematics of the counting grid Model, used to derive the formulation of the Fisher Kernel, presented in Section III. Such section also contains a critical comparison with the kernel proposed in [12]. Section IV discusses our empirical evaluation of the method and Section V concludes the paper.

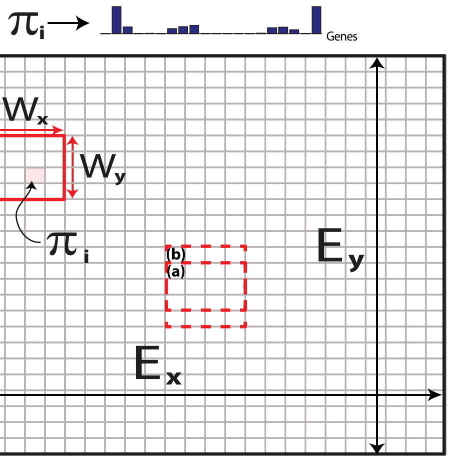

Fig. 2. An example of counting grid geometry.

II. The Counting Grid model

In Pattern Recognition, data samples are often represented as bags of features without particular order; each t-th observation is characterized by a vector – often called count vector $\{c_z^t\}$ – containing the number of occurrences of each feature z [16]. For example, a text document may be described by the number of occurrences of the different words it contains (or an image with the number of occurrences of different visual features it contains). This choice is often motivated by the difficulty of and computational problems associated with modeling the full structure of the data. It has been shown in [1], [2], [3], [4] that the bag-of-features representation is well-suited also for microarray data, providing interpretable and descriptive signatures. Each sample can be seen as an independent observation; the gene expression value is then interpreted as the "count" of that gene in the sample: the higher the expression level, the "more present" the gene is in such experiment.

The counting grid model, recently introduced in [7], is a generative model for such representations. Formally, the basic counting grid $\pi_{\mathbf{k},z}$ is a set of normalized counts of features indexed by z on the 2-dimensional[1] discrete grid indexed by $\mathbf{k} = (x, y)$ where $x \in [1 \dots E_x]$, $y \in [1 \dots E_y]$ and $\mathbf{E} = E_x \times E_y$ describes the extent of the counting grid. Since π is a grid of distributions, $\sum_z \pi_{\mathbf{k},z} = 1$ everywhere on the grid (see Fig. 2 for an illustration of the geometry and Fig. 1B for where we show π for a particular feature z.).

A given bag-of-features, represented by counts (expression levels) $\{c_z\}$ is assumed to follow a count distribution found in a window of the counting grid. In particular, using a window of dimensions $\mathbf{W} = W_x \times W_y$, each bag can be generated by first selecting a position $\mathbf{k}$ on the grid and then by placing the window in the grid such that $\mathbf{k}$ is its *upper left corner*. Then, all counts in this window are averaged to form the histogram $h_{\mathbf{k},z} = \frac{1}{W_x \cdot W_y} \sum_{\ell \in W_{\mathbf{k}}} \pi_{\ell,z}$, and finally a set of features in the

[1] N-dimensional in general, here we focus on 2 dimensions.

bag is generated. In other words, the position of the window $\mathbf{k}$ in the grid is a latent variable given which the probability of the bag of features $\{c_z\}$ is

$$p(\{c_z\}|\mathbf{k}) = \prod_z h_{\mathbf{k},z}^{c_z} = \prod_z (\frac{1}{W_x \cdot W_y} \cdot \sum_{\ell \in W_{\mathbf{k}}} \pi_{\ell,\mathbf{z}})^{c_z} \qquad (1)$$

where with $W_{\mathbf{k}}$ we indicate the particular window placed at location $\mathbf{k}$ (see Fig. 2). We will also often refer to the ratio between the counting grid area and the window area $\kappa = \frac{E_x \cdot E_y}{W_x \cdot W_y}$, as the capacity of the model.

To learn a counting grid, we need to maximize the log likelihood of the data:

$$\log P = \sum_t \log \Big(\sum_{\mathbf{k}} \cdot \prod_z h_{\mathbf{k},z}^{c_z^t} \Big) \qquad (2)$$

The sum over the latent variables $\mathbf{k}$ makes it difficult to perform assignment to the latent variables while also estimating the model parameters. The problem is solved by employing a variational EM procedure [17], which iteratively learns the model, by minimizing a bound B on $\log P$ by alternating the E and M-step. B is often referred to as the free energy of the model and it is equal to

$$\begin{aligned} \log P \geq B = & - \sum_t \sum_{\mathbf{k}} q_{\mathbf{k}}^t \cdot \log q_{\mathbf{k}}^t + \qquad (3) \\ & + \sum_t \sum_{\mathbf{k}} q_{\mathbf{k}}^t \cdot \sum_z c_z^t \cdot \log \sum_{\ell \in W_{\mathbf{k}^t}} \pi_{\ell,z} \end{aligned}$$

where $q_{\mathbf{k}^t} = P(\mathbf{k}|t)$ is the variational distribution over the latent mapping onto the counting grid of the t-th sample. Each of these variational distributions can be varied to maximize the bound: the E step aligns all bags of features to grid windows, to match the bags' histograms. In the M-step we re-estimate the counting grid π so that the histogram matches are even better. To avoid severe local minima it is important to consider the counting grid as a torus, and perform all windowing operation accordingly. For details on the learning algorithm and on its efficiency see [7].

III. Fisher Kernel from Counting Grids

In the last years, hybrid generative discriminative paradigms, and in particular generative score spaces, have been proposed for classification. They consist 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 for every object through the learned model(s), to be used as features for a discriminative classifier. 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. Among the score spaces the most famous is the Fisher kernel [13].

The Fisher kernel makes use of the the first derivative of the log-likelihood with respect to its parameters $U_t = \bigtriangledown_\theta \log P(x^t|\theta)$. In our case the samples are gene expressions $x^t = c_z^t$ and the parameter is the counting grid $\theta = \{\pi\}$. U_t is called the *Fisher score* and it evaluates the effect of the sample on the model parameters. The Fisher kernel between the t-th and the s-th sample is then defined as the inner product between their corresponding Fisher scores, $K(t,s) = U_t^T \cdot U_s$.

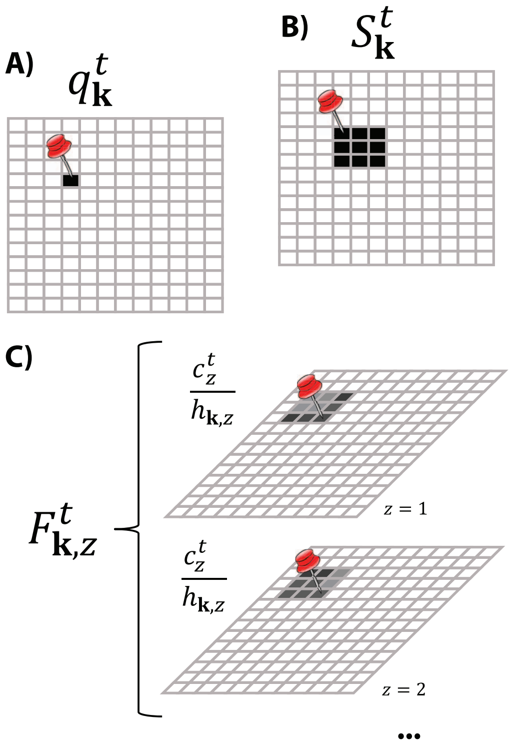

Fig. 3. A) The variational posterior $q_{\mathbf{k}}^t$ that defines the mapping on the counting grid. $q_{\mathbf{k}}$'s are generally peaky (either 0 or 1) by construction as they are discrete variables. B) The score used in [12]. As in the experiments we considered a window of size $\mathbf{W} = 3 \times 3$. C) Fisher score illustration. In A,B and C we highlighted the mapping position of the window using a "pin".

In the case of counting grid, the Fisher kernel is intractable to compute for the very same reason its log likelihood (Eq. 4) is. However, as for latent Dirichlet allocation [15], we can extract the Fisher scores from the free energy B and its derivatives with respect to the parameters $\pi_{\mathbf{k},z}$ are functions of the variational distributions q_ℓ. In formulae:

$$\begin{aligned} U_t &= [\ldots, \frac{\partial B}{\partial \pi_{\mathbf{k},z}}, \ldots] \quad \forall \ \mathbf{k}, z \qquad (4) \\ &= [\ldots, \frac{\partial}{\partial \pi_{\mathbf{k},z}} \Big(\sum_\ell q_\ell^t \cdot \sum_z c_z^t \cdot \log \big(\sum_{\mathbf{i} \in W_\ell} \pi_{\mathbf{i},z} \big) \Big), \ldots] \end{aligned}$$

where we ignored the first term of Eq. 4 as it does not depend on π. Since most of the terms in the partial derivative do not depend on the current choice for $\mathbf{k}$ or z, Eq. 4 simplifies as follows

$$\begin{aligned} \frac{\partial B}{\partial \pi_{\mathbf{k},z}} &= \frac{\partial}{\partial \pi_{\mathbf{k},z}} \Big(\sum_\ell q_\ell^t \cdot \sum_z c_z^t \cdot \log \big(\sum_{\mathbf{i} \in W_\ell} \pi_{\mathbf{i},z} \big) \Big) \\ &= \frac{\partial}{\partial \pi_{\mathbf{k},z}} \Big(\sum_{\ell | \mathbf{k} \in W_\ell} q_\ell^t \cdot c_z^t \cdot \log \big(\sum_{\mathbf{i} \in W_\ell} \pi_{\mathbf{i},z} \big) \Big) \\ &= \sum_{\ell | \mathbf{k} \in W_\ell} q_\ell^t \cdot c_z^t \cdot \frac{1}{\sum_{\mathbf{i} \in W_\ell} \pi_{\mathbf{i},z}} \end{aligned}$$

$$= \quad c_z^t \cdot \sum_{\ell | \mathbf{k} \in W_\ell} \frac{q_\ell^t}{h_{\ell,z}} = F_{\mathbf{k},z}^t \tag{5}$$

The concatenation of all the partial derivatives of Eq. 5 comprises the Fisher score for of sample c_z^t from which we can compute the Fisher kernel that measures the similarity between two microarray experiments. The Fisher score $F_{\mathbf{k},z}^t$ has dimensionality $Z \times E_x \times E_y$ as it depends on the grid locations $\mathbf{k}$ and on the genes z.

Relationship with other methodologies

In this section we highlight differences and relationships with [12], where the authors exploited the geometric reasoning of the counting grid to define a generative kernel. Without loss of generality we will assume a peaky variational posterior $q_\mathbf{k}$ equal to either 0 or 1 as shown in Fig. 3A. By construction, each point in the grid depends by its neighborhood, defined by $\mathbf{W}$ and in [12] each sample is described by its mapping window $S_\mathbf{k}^t$ as illustrated in Fig. 3B. The intuition is that samples whose windows intersect have similar overall genes expression and therefore may be similar. One of the biggest problem of this approach is that if two samples are mapped in the same point, they will have identical signature and when they belong to different classes it is impossible to disambiguate.

The Fisher kernel proposed here builds upon similar geometric reasoning and it is illustrated in Fig. 3C. Its only non-zero dimensions $F_{\mathbf{k},z}^t \neq 0$ are the ones whose window $\mathbf{W}_\ell$ contains the mapping position $\mathbf{k}$ of the t-th sample ("$\ell | \mathbf{k} \in W_\ell$" in Eq. 5) therefore, again, only samples mapped close on the grid may have non-zero similarity. The influence window is shifted by $-\mathbf{W}$ wrt [12] (see Fig. 3B and C), however this does not affect the similarity relation as the grid is a torus (i.e. it has wraparound). The second difference with [12] is that the Fisher kernel also explicitly takes into account the gene expression level c_z^t while the dependence of [12] on the expression is only implicit: gene expressions are used to compute the mapping, but not in the kernel. Finally, the counting grid is a well defined generative model and its Fisher kernel is theoretically better than the maximum likelihood classification based on the same model [13]. This nice property clearly does not hold for [12].

Relationship with raw expression classification

It is also easy to prove that the raw gene expression classification is a special case of our kernel. In fact when $\mathbf{E} = \mathbf{W}$, we have that $\sum_\mathbf{k} F_{\mathbf{k},z}^t = \alpha_z \cdot c_z^t$, where α_z is a constant that only depends on the particular gene z. This is a nice property means that model selection strategies can take into account the raw data classification.

IV. Experimental results

We tested our approach using four different well-known datasets, briefly summarized in Tab. I: in particular we employed three tumor data sets and a blood sample data set from schizophrenic patients and healthy controls. The whole description of each dataset may be fo und in the reported reference.

TABLE I. Datasets considered in our study

Dataset Name	*Protocol*	*# classes*	*# samples*	*Citation*
Lung	5-Folds	5	203	[9]
Prostate	LOO	2	102	[18]
Brain	4-Folds	5	90	[19]
Schizophrenia	3-Folds	2	202	[20]

TABLE II. Comparison between score spaces

Method	Score Space	Dimensionality	Kernel considered
CG Fisher	$\nabla_\theta \log P(c_z^t \| \theta_{CG})$	$\|\mathbf{E}\| \times Z$	linear
[12]	Map. Window (Fig. 3B)	$\|\mathbf{E}\|$	linear / HI
[2]	$p(topic \| c_z^t)$	$\|\mathbf{E}\| / \|\mathbf{W}\|$	linear
[3]	$\nabla_\theta \log P(c_z^t \| \theta_{PLSA})$	$\|\mathbf{E}\| / \|\mathbf{W}\| \times Z$	linear

As in [1], [3] we filtered the genes by variance and retained the top 500 genes, using a prior belief that genes varying little across samples are less likely to be interesting. We compared our method with previous work on topic models [2], [3], on counting grids [12] and with the same baseline. In [2] a pLSA model [16] is learnt from the data and each sample is described by its topic proportions. In [3] the kernel is derived from the pLSA model [16] yielding a significant improvement upon [2]. Finally, in [12], the gene expressions are mapped on a counting grid and their mapping window is used as a signature. All these methods belong to the family of generative score spaces and they are summarized in Tab. II in terms of score space, dimensionality and kernel employed. We considered counting grids of various sizes $\mathbf{E} = [6 \times 6, 9 \times 9, \ldots, 30 \times 30]$ and we set the window size as $\mathbf{W} = 3 \times 3$. As in previous CG literature [7], [12], we acknowledge that the capacity of the model $\kappa = |\mathbf{E}|/|\mathbf{W}|$, which measures how many independent windows can fit into the grid, is roughly equivalent to the number of pLSA topics and we used this parallelism to compare the accuracies.

Our subdivision in training and testing set is carried out using the dataset author's protocol which is reported in Tab. I. As in [3], [12] we firstly learned the generative models (pLSA or counting grid) using all the data (not using the labels, therefore following a transductive learning approach [21]) and we described each sample with the appropriate score. Then we learned a support vector machine (SVM) using the training data and we classified the test data. To highlight the contribution of the modeling, we considered the simple linear kernel which defines the similarity between two data points as the inner product of their scores. For [12] we also considered the histogram intersection which directly measures the extent of window shared between samples. We considered values for SVM's cost parameter C values among $2^{-1.1}, 2^{-0.2}, \ldots, 2^{5.1}$ and (for each svm classification) we reported the best accuracy. The mean accuracies are shown in Fig. 4: the Fisher kernel extracted from counting grids outperforms [3], [2], [12] the other tested approaches on all the datasets and provides the most consistent results. Interestingly, it was also found to be less sensitive to the model complexity and the value of the cost parameter of the SVM did not affect the result. These two properties are very important in this context, as microarray datasets are usually small and cross-evaluation strategies to pick a optimal parameter generally work poorly. In Tab. III we report the mean, across all the complexities, variance aross the choices for C for each dataset and for each method: [2], [12], [3] reported higher variances, making their accuracies in Fig. 4

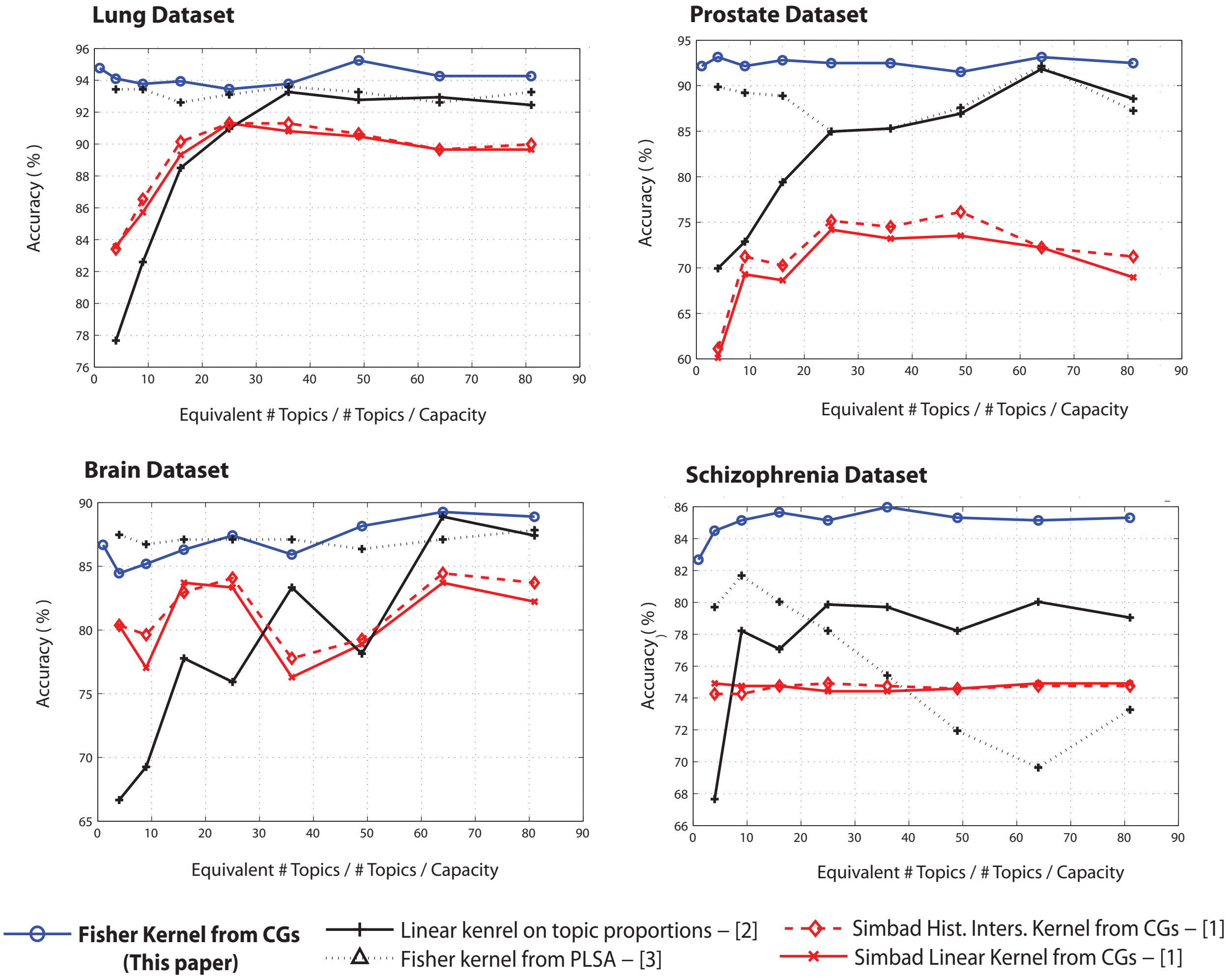

Fig. 4. Classification Results (Mean accuracy over 3 repetitions)

TABLE III. Mean Variance $\hat{\sigma}$

	Lung [9]	Prostate [18]	Brain [19]	Schizoph. [22]
[2]	4.9	1.3	3.7	1.2
[12]	0.4	0.1	0.5	0.6
[3]	1.5	1.3	3.5	0.7
CG Fisher	0	0	0	0

TABLE IV. Comparison the state of the art

	Lung [9]	Prostate [18]	Brain [19]	Schizoph. [22]
CG Fisher	**95,1%**	96,3%	**90,0%**	**88,1 %**
Raw Data	94,1%	92,1%	86,5%	82,7%
SoA	93,8%	**98,2%**	86,5%	87,6%

somewhat optimistic.

To highlight the benefit of generative modeling, we compared in Tab. IV with the linear kernel on the raw expressions ("Raw Data" row) and with the state-of-the-art (results are taken from [2]). Despite the fact that we did not use any discriminative gene selection strategy (as done in [2]), our best results are very close to the state of the art on prostate dataset, and we set it on schizophrenia, brain and lung datasets.

As final test, we evaluated the effect of $\mathbf{W}$ on the Fisher kernel. It is known that CGs are not sensitive to the window size, however this holds for the generative model (maximum likelihood classification) and not necessarily for a kernel extracted from it. In this last experiment we considered windows $\mathbf{W} = 2 \times 2, 4 \times 4, 6 \times 6, 5 \times 2$. For every window choice, we varied the grid size $\mathbf{E}$ roughly keeping the same model capacity considered in the previous test[2]. We performed the classification experiment using the usual procedure. In Fig. 5A we show the effect of the $\mathbf{W} = 5 \times 2$ on the schizophrenia dataset [22]. By using a rectangular window, a shift of the window in one dimension, in this case x, provides a different degree of variation in the gene expression, in fact 50% of the window's content changes. This did not influence the results, especially for larger grids where the algorithm has more space to lay down the samples. In Fig. 5B we compared the other windows choices: despite $\mathbf{W} = 3 \times 3$ is clearly the best choice, all the results are satisfactory and all ouperform [3], [2], [12]. The lower performances of $\mathbf{W} = 2 \times 2$ can be

[2] A CG of complexity $\mathbf{W} = 3 \times 3 \;-\; \mathbf{E} = 6 \times 6$ has the same capacity of a CG $\mathbf{W} = 4 \times 4 \;-\; \mathbf{E} = 8 \times 8$

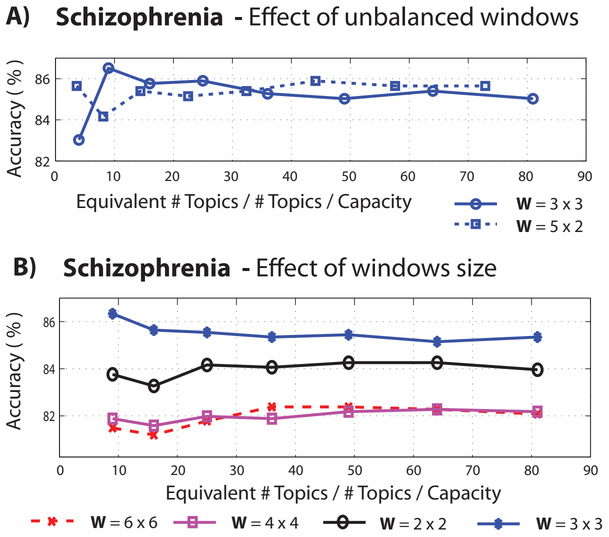

Fig. 5. Classification Results (Mean accuracy over 3 repetitions)

explained by a lack of (overlapping) modeling power while for the larger windows the reason is clearly over-training. To keep the capacity κ fixed, larger windows require larger grids and thus the Fisher score (Eq. 5) has higher dimensionality. Tests on the other dataset yielded to similar results.

V. Discussion

The paper discusses the use of counting grids to model microarray expression data and derives the Fisher kernel, building a successful classification framework. As the kernel proposed in [12], the Fisher kernel exploits the clustering of the samples on the counting grid but it also *1)* takes explicitly into account the individual expressions of each gene and *2)* inherits theoretical properties of [13]. We have also shown that raw expression classification is a special case of our framework. Our results indicate that the CG is a better fit than topic models for this type of data. Our method compares favorably with the state-of-the art on several datasets and very importantly it proved to be insensitive to SVM parameters and to counting grid complexity. We have also investigated the use of higher dimensional grids and more complex kernels like rbf, polynomial, histogram intersection and χ^2 but they did not provide enough improvement to justify the additional parameters to set (e.g., SVM's γ). This is explainable as the Fisher kernel is defined as the dot product of the scores.

Finally, while this paper only focuses on classification, future work will be devoted to exploiting the expressive power of counting grids to understand the variation of gene expression values in health and disease. Regions of gene expression profiles that characterize a disease can, in fact, be highlighted on the grid. For example, in Fig. 1B the small-cell lung carcinomas (see the red crosses) are mapped in an area where the gene is not expressed. This confirms the findings by [9], [10]. By focusing on the border regions between different classes, we can compute which genes vary most in the direction of the transitions and obtain biomarkers for diseases.

References

[1] S. Rogers, M. Girolami, C. Campbell, and R. Breitling, "The latent process decomposition of cdna microarray datasets," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2005.

[2] M. Bicego, P. Lovato, A. Perina, M. Fasoli, M. Delledonne, M. Pezzotti, A. Polverari, and V. Murino, "Investigating topic models' capabilities in expression microarray data classification," *IEEE/ACM Trans. Comput. Biology Bioinform.*, vol. 9, no. 6, pp. 1831–1836, 2012.

[3] A. Perina, P. Lovato, M. Cristani, and M. Bicego, "A comparison on score spaces for expression microarray data classification," in *PRIB*, 2011, pp. 202–213.

[4] M. Bicego, P. Lovato, B. Oliboni, and A. Perina, "Expression microarray classification using topic models," in *SAC*, 2010, pp. 1516–1520.

[5] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, apr 2012.

[6] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic, "Free energy score spaces: Using generative information in discriminative classifiers." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1249–1262, 2012.

[7] N. Jojic and A. Perina, "Multidimensional counting grids: Inferring word order from disordered bags of words," in *Uncertainty in Artificial Intelligence*, 2011.

[8] I. Jordan, L. Marino-Ramirez, and E. Koonin, "Evolutionary significance of gene expression divergence," *Gene*, vol. 345, no. 1, pp. 119 – 126, 2005.

[9] A. Bhattacherjee, W. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, and M. G. et al., "Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses," *Proc. Natl Acad. Sci*, vol. 98, pp. 13 790–13 795, 2001.

[10] R. Jonge, L. Garrigue-Antar, V. Vellucci, and V. Reiss, "Frequent inactivation of the transforming growth factor beta type ii receptor in small-cell lung carcinoma cells," *Oncology research*, vol. 9, no. 2, pp. 89–98, 1997.

[11] P. Lovato, M. Bicego, M. Cristani, N. Jojic, and A. Perina, "Feature selection using counting grids: application to microarray data," in *Proc. Int. Workshop on Statistical Techniques in Pattern Recognition (SPR2012)*, 2012.

[12] A. Perina, U. Castellani, M. Bicego, , and V. Murino, "Exploiting geometry in counting grids," in *Proc. Int. Work. on Similarity-Based Pattern Recognition*, 2013, pp. 250–264.

[13] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *NIPS*, 1999.

[14] J. Lasserre, C. Bishop, and T. Minka, "Principled hybrids of generative and discriminative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York, 2006.

[15] G. Chandalia and M. Beal, "Using fisher kernels from topic models for dimensionality reduction," in *NIPS Workshop*, 2009.

[16] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1-2, pp. 177–196, 2001.

[17] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[18] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, and J. R. et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 98, pp. 203–209, 2002.

[19] S. Pomeroy and P. e. a. Tamayo, "Prediction of central nervous system embryonal tumour outcome based on gene expression." *Nature*, vol. 415, no. 6870, pp. 436–42, 2002.

[20] S. de Jong et al., "A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes," *PLoS ONE*, vol. 7, no. 6, p. e39498, 06 2012.

[21] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[22] M. Takahashi et al., "Diagnostic classification of schizophrenia by neural network analysis of blood-based gene expression signatures," *Schizophrenia Research*, vol. 119, no. 13, pp. 210 – 218, 2010.