Exploiting Geometry in Counting Grids

Alessandro Perina[2], Manuele Bicego[1], Umberto Castellani[1], and Vittorio Murino[1,3]

[1] Department of Computer Science, University of Verona, Italy
[2] Microsoft Research Redmond, Washington, USA
[3] Istituto Italiano di Tecnologia (IIT), Genova, Italy

Abstract. In this paper we exploit the use of known information about the geometry structure of a recently proposed generative model, namely Counting Grid (CG) [1] to improve the performance of classification accuracy. Once the generative model is trained, the geometric structure of the model introduces a natural spatial relations among the estimated latent variables. Such relation is generally ignored when standard maximum likelihood approach (or classical hybrid generative-discriminative approach) is employed for classification purpose. In this work, we propose to take into account the geometric relations of the generative model by proposing an ad hoc similarity measure for CG. In particular, the values relative to each point of the grid is spread around its neighborhood by using information coming from the CG training phase. The proposed approach is succesfully applied in two applicative scenarios: expression microarray classification and MRI brain classification. Experiments show a drastic improvement over standard schemes when our approach is employed.

Keywords: generative models, kernels, microarray, MRI.

1 Introduction

In pattern recognition some counting strategies are often introduced, especially when source data is not naturally lying on a vectorial space. A very popular example is the *Bag of Words* approach, where objects are represented as disorganized bags of basic components such as the words of a dictionary. This approach has been succesfully employed in very different applicative domains like computer vision for 2D image or 3D shape retrieval, in bioinformatics for microarray classification, or in medical domain for brain disease detection [2–8]. However, the Bag of Words (BoW) method has some disadvantage since in many situations it looses a lot of important information. For instance, BoW approach does not take into account words relations or co-occurences. To this aim, LDA or pLSA models have been succesfully proposed by showing how inter-relations among words, i.e., *topics* are crucial to improve object encoding [9, 10]. Recently, a new generative model has been proposed, namely Counting Grid (CG) [1] which goes beyond topic-based approach. Indeed, CG exploits not only words co-occurences but also topological relations among words. In particular, with CG an ordering procedure between BoWs is introduced by allowing BoWs to lie in

E. Hancock and M. Pelillo (Eds.): SIMBAD 2013, LNCS 7953, pp. 250–264, 2013.

an n-dimensional grid structure. Such approach has already shown its benefits on document retrieval, 2D scene classification, and microarray expression classification[1, 11]. In all these applications, the classification stage has been computed by standard maximum likelihood scheme, or by employing discriminative classifiers like Support Vector Machine (SVM) with generative kernels, nevertheless without taking into account the peculiar geometry of the model.

In this paper we propose to further exploit the advantage of CGs by studing an ad hoc (dis)similarity measure. We start from the observation that in the CG scenario, the classical classification scheme is based on the grid posterior of a given sample, which is treated as a vector and used for comparison. In such a way, spatial relation between values is lost. Nevertheless, due to the nature of the CG, in the training phase a BoW, or a *count*, is distributed on a local region around a specific point in the grid which is defined by an hidden variable. This leads to a spatial relation among grid points which can be used to improve the classification stage. The idea is to spread the posterior evaluated on a single grid-point around its neighborhood. In this fashion, when two samples are compared, an implicit cross-count evaluation is introduced by avoiding a fully grid alignment constraint. Experiments show that our new (dis-)simmilarity approach leads to a drastic improvement in comparison with standard methods.

The rest of the paper is organized as following. In Section 2 the background on Counting Grids is introduced. Section 3 describes the proposed (dis-)similarity measure for the proposed generative model. Section 4 reports experiments on two applicative domains, namely expression microarray classification and MRI brain disease classification. Finally, conclusions and future work are discussed in Section 5.

2 Background: Counting Grid Model

Data samples are often represented as an unordered bags of features, where each t-th observation is characterized by a vector called *count* vector $\{c_z^t\}$ which contains the number of occurrences of each feature z [12, 9]. For instance, a text document can be described by the number of words occurrences it contains (or an image with the number of occurrences of different visual features it contains). This choice is often motivated by the difficulty or computational efficiency of modeling the known structure of the data.

The counting grid model, recently introduced in [1], is a generative model that extends such representations. The models starts from a common choice in counting data modelling, which implies that the bag of features of a given sample is generated by a latent variable; in the counting grid model, nevertheless it is assumed that a spatial relation between latent variables exists, and can be learnt and used to improve the understanding of the models or to provide rich descriptors for classification. More explicitly, we can unformally say that the generative process of a given bag of features is based on a latent variable but also on some of its spatial neighbours. Formally, the basic counting grid $\pi_{\mathbf{i},z}$ is a

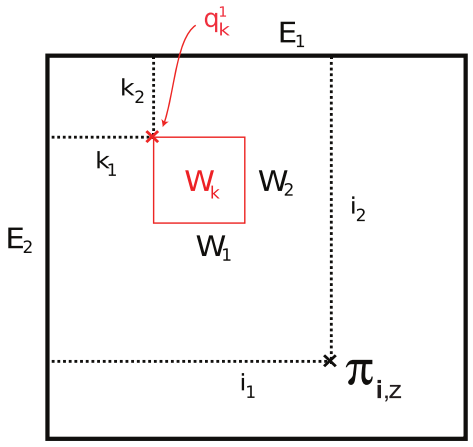

Fig. 1. An example of a counting grid geometry

set of normalized counts of features indexed by z on the 2-dimensional[1] discrete grid indexed by $\mathbf{i} = (i, j)$ where $i \in [1 \dots E_1]$, $j \in [1 \dots E_2]$ and $\mathbf{E} = [E_1, E_2]$ describes the extent of the counting grid. Since π is a grid of distributions, $\sum_z \pi_{\mathbf{i},z} = 1$ everywhere on the grid.

A given bag of features, represented by counts $\{c_z\}$ is assumed to follow a count distribution found in a patch of the counting grid. In particular, using a window of dimensions $\mathbf{W} = [W_1, W_2]$, 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 patch are averaged to form the histogram $h_{\mathbf{k},z} = \frac{1}{W_1 \cdot W_2} \sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z}$, and finally a set of features in the 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

$$p(\{c_z\}|\mathbf{k}) = \prod_z (h_{\mathbf{k},z})^{c_z} = \frac{1}{W_1 \cdot W_2} \prod_z (\sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z})^{c_z}$$

where with $W_\mathbf{k}$ we indicate the particular window placed at location $\mathbf{k}$ (see Figure 1).

We will refer to E and W respectively as the counting grid and the window size. We will also often refer to the ratio of the CG area and the window area $\kappa = \frac{E_1 \cdot E_2}{W_1 \cdot W_2}$, as the capacity of the model, which can be seen – using a topic models parallelism – as an equivalent number of topics (this is how many nonoverlapping windows can be fit onto the grid). Computing and maximizing the log likelihood of the data turns to be an intractable problem; therefore it is necessary to employ an iterative EM algorithm. The E step aligns all bags of features to grid windows, to match the bags' histograms, inferring the posterior probability $q^t_\mathbf{k}$, the probability that the sample t is generated from the position

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

$\mathbf{k}$, i.e., where each bag maps on the grid. This posterior can be computed as $q^t_{\mathbf{k}} \propto \exp \sum_z c^t_z \cdot \log h_{\mathbf{k},z}$. In the M-step the model parameter, i.e. the counting grid π, is re-estimated. 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 [1].

3 (Dis-)Similarity Measure for CG

Once the training phase is performed, the CG $\pi_{\mathbf{i},z}$ is available and can be used for classification purposes. Given a sample A, represented by counts $\{c^A_z\}$, its posterior $q^A_{\mathbf{k}}$ is computed. In general, the matrix $q^A_{\mathbf{k}}$ can be used in a maximum likelihood scheme or it can be fed in a discriminative classifier such as a Support Vector Machine, after its vectorization, representing a straightforward hybrid generative-discriminative classification approach. When using standard vector-based kernels (like linear kernel), the implicit assumption is that counts are well aligned, so that each count in one sample is only compared to corresponding count in another sample. Here, we exploit cross-count distances by observing that each point in the grid depends by its neighborhood which is defined by $\mathbf{W}$. Indeed, we propose to spread the values $q^A_{\mathbf{k}}$ around a neighborhood region defined by $W_{\mathbf{k}}$. Actually, by construction, the value in a given location $\mathbf{k}$ is computed by using all CG parameters belonging to the subwindow $\mathbf{W}$.

More in details, given two samples A and B, our similarity measure – which we call *Spreading Similarity Measure* is defined by:

$$SSM_S(A, B) = SM(q^A_{\mathbf{k}} * S_{\mathbf{W}}, q^B_{\mathbf{k}} * S_{\mathbf{W}}), \tag{1}$$

where $S_{\mathbf{W}}(\mathbf{x})$ is a box function, of dimension defined by the spreading window $\mathbf{W}$, defined as:

$$S_{\mathbf{W}}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathbf{W} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

and $SM(\cdot, \cdot)$ is any (dis)-similarity measure. In our experiments we evaluate standard inner product[13], histogram intersection [14], and Jensen-Shannon distance [15]. Reasonably, we chose to set the size of the spreading windows as the size of the Counting Grid Window. In the experimental part we make some experiments while varying the dimension of the spreading window, showing that, as expected, our choice is almost always the best choice.

Figure 2 shows the effect of our new (dis)similarity measure. Two posteriors are displayed, each with a peak in a particular zone of the grid. When using a punctual kernel (such as the histogram intersection kernel), which needs aligned grids, we can observe that even if the two peaks are close in the grid the intersection is almost null, and therefore the similarity is null as well (see Figure 2(top)). Conversely, in Figure 2(center) and 2(bottom) the grid intersection, and therefore the similarity, is significative and it increases with the size of the convolution window.

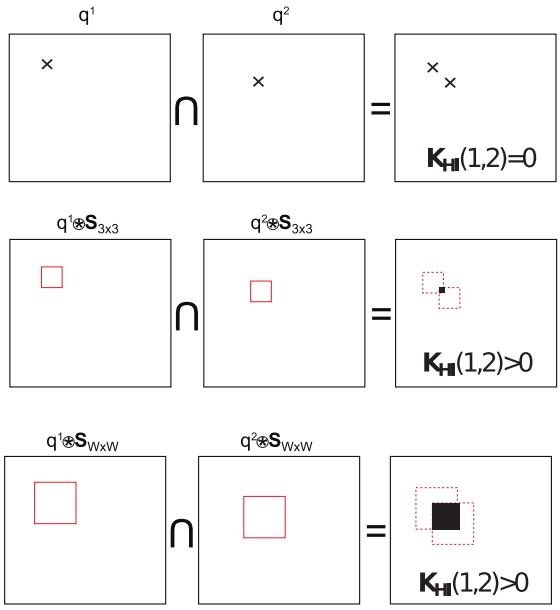

Fig. 2. The spreding effect of using our approach in comparing q_1 and q_2. The histogram intersection ($K_{HI}(\cdot,\cdot)$) is considered as measure $SM(\cdot,\cdot)$. When standard K_{HI} is considered no intersection is observed (top), while using the spreding strategy the similarity between q_1 and q_2 is significative (center), and it increases with the size of $\mathbf{W}$.

As a further note: it is straightforward to show that if $SM(\cdot)$ is a kernel, also our $SSM_S(\cdot,\cdot)$ is a kernel. This may be of great practical importance, since permits to develop a hybrid generative-discriminative scheme where SVM can be used as discriminative classifiers.

4 Experimental Evaluation

In this section the experimental evaluation is presented. In particular, the proposed framework is evaluated within two biomedical applications: cancer classification via the analysis of expression microarray and schizophrenia detection through brain classification using MRI scans.

4.1 Microarray Classfication

In this application, the goal is to analyze gene expression microarray data in order to distinguish between healthy people and people affected by cancer. The starting point is a microarray gene expression matrix, where the element at position (i, j) represents the expression level of the $i-th$ gene in the $j-th$ subject/sample. Methods based on counting values (as CG and topic models)

have been recently and successfully applied in this context (see, *e.g.*, [16, 17, 11]). This is possible if we establish an analogy between a word-document pair and a gene-sample pair; it seems reasonable to interpret samples as documents and genes as words. In this way, the gene expression levels in a sample may interpreted as the word counts in a document. Consequently, we can simply take a gene expression matrix and (of course, after a preprocessing step, for example, to remove possibly negative numbers [16]) interpret it as a count matrix $\mathbf{C}$ from which a CG or a LDA model can be estimated.

The experiments presented in this paper have been performed using two microarray datasets: the ovarian [18] and the colon [19] datasets, whose characteristics are summarized in table 1.

Table 1. Summary of the employed microarray datasets

Dataset Name	n. of genes	n. of samples	n. of classes	citation
1. Ovarian cancer	1513	53	2	[18]
2. Colon cancer	2000	62	2	[19]

4.2 Brain Classification

In this application the main goal is to distinguish between healthy and schizophrenic people through the classification of MRI brain scans.

Data Set. The study population used in this work consists of 42 patients (21 male, 21 female) who were being treated for schizophrenia and 40 controls (19 male, 21 female) with no DSM-IV axis I disorders and had no psychiatric disorders among first-degree relatives. Diagnoses for schizophrenia were corroborated by the clinical consensus of two psychiatrists. T1 weighted structural MRI scans were acquired with a 1.5 Tesla machine and to minimize biases and head motion, restraining foam pads were used. The original image size is 384x512x144; these images are then rotated and realigned to a resolution of 256x256x192. After this alignment, they were segmented into specific brain regions called Regions of Interest (ROIs) manually by experts following a specific protocol for each ROI [20]. In this work, we use three ROIs from the two hemispheres of the brain summing upto a total of six different brain regions: Dorsolateral prefrontal cortex (*ldlpfc* and *rdlpfc*), Entorhinal Cortex (*lec* and *rec*), and Thalamus (*lthal* and *rthal*) which are found to be impaired in schizophrenic patients.

Preprocessing. After the alignment and ROI tracing, DARTEL [21] tools within SPM software [22] was used to pre-process the data. Initially, images are segmented into grey and white matter in *Native* and *DARTEL imported* spaces. The DARTEL imported images have lower resolution than the original images but are used to spatially align to standard MNI atlas. In the second step, DARTEL template generation is applied which creates an average template from the input data while simultaneously aligning white and grey matter. In this step, the

flowfields of the registration are also computed which will be used to segment the MNI space normalized images into ROIs. In the final step, the DARTEL template is used to spatially normalize all images into standard MNI space. In this way, smoothed (12 mm Gaussian), and Jacobian scaled grey matter images are constructed which is general practice in neuroimaging applications.

Feature Extraction. The images at the end of the preprocessing pipeline are the intensity probability maps which are then used to construct the features for our classification experiments. Since we already have ROI segmented source images, using the flow fields computed in the second step of preprocessing we create the intensity maps for every subject and ROI instead of extracting a single set of features for the whole brain. Since the ROIs have different bounding boxes, the sizes of these images are not the same for all subjects. By applying thresholding at 0.2 level, we compute histograms of probability maps for every subject and ROI. Number of bins in each histogram is chosen to be 40 which showed the best performance in our experiments. As a result, we have a data set of six different ROIs, 82 subjects with a counting vector of size 40 which we apply our classification pipeline.

4.3 Experimental Details

The experimental evaluation is aimed at validating the proposed approach. In particular, we start by assessing the baseline CG results, without any spreading operation, using the ovarian dataset. Then we evaluate the impact of the proposed approach. Third, we investigate the impact of the dimension of the spreading window. Finally, we show some more results on the colon microarray experiment and on the Brain MRI classification task.

For all the experiments the following protocol has been adopted:

- Since, as a base level, we are mostly interested in the quality of unsupervised learning of the distributions over the samples, the whole dataset has been used to train a CG (of course labels are ignored in this phase), in a transductive way [13, 4]. As explained in the methodological section, here we employed bidimensional squared Counting Grid models (in principle, also higher dimensional/not squared grids can be used, see [1]). Two parameters should be set when learning the Couting Grid: the dimension of the Grid $\mathbf{E}$ and the dimension of the Window $\mathbf{W}$. Here we performed a large scope analysis, reporting results for many different configurations, with $\mathbf{E}$ ranging from $[10, 10]$ to $[90, 90]$, and $\mathbf{W}$ ranging from $[4, 4]$ to $[19, 19]$[2]. An interesting parameter which can be used to summarize the dimension of a Counting Grid is the capacity κ, which, as explained in the methodological section, represents the ratio between the dimension of the grid and the dimension of the window, and can be seen as the number of topics in the standard topic models.

[2] Of course only valid configurations were retained – e.g. $\mathbf{E} = [10, 10], \mathbf{W} = [15, 15]$ is not a valid configuration.

- In order to avoid to get stuck in local optima during the learning procedure (given the initialization, E-M converges to the nearest local optima), we repeated the training 10 times, starting from random initialization, retaining the model with the highest training likelihood.
- Given the model, an hybrid generative-discriminative approach is used to perform classification. In particular, for every pair of samples A, B, represented by counts $\{c_z^A\}$, $\{c_z^B\}$, we computed its posterior $q_{\mathbf{k}}^A$, $q_{\mathbf{k}}^B$ given the learned counting grid, comparing them with a kernel, employed to perform a discriminative classification via Support Vector Machines. In all experiments the parameter C of the SVM was set, after some preliminary evaluations, to 10000.
- In all experiments, classification accuracy has been computed using Leave-One-Out Cross validation, as typically done with these small size problems.
- In all the experiments we also computed and reported the performances of the Latent Dirichlet Allocation (LDA - [23]), the most famous topic model, whose usefulness has been already shown in these contexts [17, 16, 24]. LDA can straightforwardly be considered as a counting grid where the Window Size is equal to 1, since there are no interactions between latent variables (i.e. topics). For classification, the same hybrid generative-discriminative approach explained before is used. In this case, given a pair of samples A, B, the posterior Dirichlet parameters have been computed through the learned LDA model and compared via a kernel, to be used in a SVM classification scenario. Given the parallelism between the concept of the capacity of the Counting Grids and the number of topics, we performed an experiment with LDA for every capacity value experimented for our approach.

Similarity Measures and Kernels Concerning the similarity measures / kernels to be adopted in our hybrid generative-discriminative scheme, different options can been used. Given the modularity of our proposed scheme, we can straightforwardly apply the same kernel $S(\cdot, \cdot)$ with and without performing the spreading via the convolution. This will permit us to directly investigate the impact of the spreading operation. In particolar, we experimented three different options:

1. *Linear Kernel.* This is the standard inner product between the representations of the two objects, namely

$$K^{\mathrm{LI}}(q_{\mathbf{k}}^A, q_{\mathbf{k}}^B) = q_{\mathbf{k}}^A \cdot q_{\mathbf{k}}^B \tag{3}$$

2. *Jensen Shannon Kernel.* This represents a standard and well known Information Theoretic Kernel, namely a kernel based on probability measures. These kernels have been shown very effective in classification problems involving text, images, and other types of data [25–27]. Very recently, moreover, they have been found to be very suitable in hybrid generative discriminative scenarios [28]. Given two posterior probabilities $q_{\mathbf{k}}^A$ and $q_{\mathbf{k}}^B$, representing two objects, the Jensen-Shannon kernel is defined as

$$K^{\mathrm{JS}}(q_{\mathbf{k}}^A, q_{\mathbf{k}}^B) = \ln(2) - JS(q_{\mathbf{k}}^A, q_{\mathbf{k}}^B), \tag{4}$$

with $JS(q^A_{\mathbf{k}}, q^B_{\mathbf{k}})$ being the Jensen-Shannon divergence

$$JS(q^A_{\mathbf{k}}, q^B_{\mathbf{k}}) = H\left(\frac{q^A_{\mathbf{k}} + q^B_{\mathbf{k}}}{2}\right) - \frac{H(q^A_{\mathbf{k}}) + H(q^B_{\mathbf{k}})}{2}, \tag{5}$$

where $H(p)$ is the usual Shannon entropy.

3. *Histogram Intersection Kernel.* This Kernel, initially designed to compare histograms, can be safely used also in case of multinomials (as the Counting Grid posteriors), which are simply normalized Histograms. Given two object representations $q^A_{\mathbf{k}}$ and $q^B_{\mathbf{k}}$, the kernel is defined as [29]

$$K^{\,\mathrm{HI}}(q^A_{\mathbf{k}}, q^B_{\mathbf{k}}) = \sum_k \min(q^A_k, q^B_k) \tag{6}$$

A further note: by looking at the formulation of our proposed dissimilarity measure, some similarities with the diffusion distance [30] can be found. Actually, in both cases, the value of every particular point is spread/diffused in its neighborhood. It seems therefore interesting to compare our approach with this distance[3], applied on the original model posteriors. More in detail, the distance between two representations $q^A_{\mathbf{k}}$, $q^B_{\mathbf{k}}$ is defined as a temperature field $T(\mathbf{k}, t)$ with $T(\mathbf{k}, 0) = q^A_{\mathbf{k}} - q^B_{\mathbf{k}}$. Using the heat diffusion equation

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial \mathbf{k}^2}$$

which has a unique solution

$$T(\mathbf{k}, t) = T(\mathbf{k}, 0) * \phi(\mathbf{k}, t)$$

where

$$\phi(\mathbf{k}, t) = \frac{1}{(2\phi)^{1/2} t} \exp -\frac{\mathbf{k}^2}{2t^2},$$

we can compute the distance D as:

$$D(q^A_{\mathbf{k}}, q^B_{\mathbf{k}}) = \int_0^r \eta(|T(\mathbf{k}, t)|) dt$$

where $\eta(\cdot, \cdot)$ is a norm which measures how $T(\mathbf{k}, t)$ differs from 0. Given this distance, we can obtain a kernel following the extended gaussian kernels recipe [31]:

$$K^{\,\mathrm{DD}}(q^A_{\mathbf{k}}, q^B_{\mathbf{k}}) = e^{-\rho D(q^A_{\mathbf{k}}, q^B_{\mathbf{k}})} \tag{7}$$

In our experiments, following the suggestion given in [32], the scale parameter ρ has been set to the average diffusion distance between all pairs of objects in the training set.

[3] The code has been taken from the author's home page:
`http://www.ist.temple.edu/~hbling/code_data.htm`

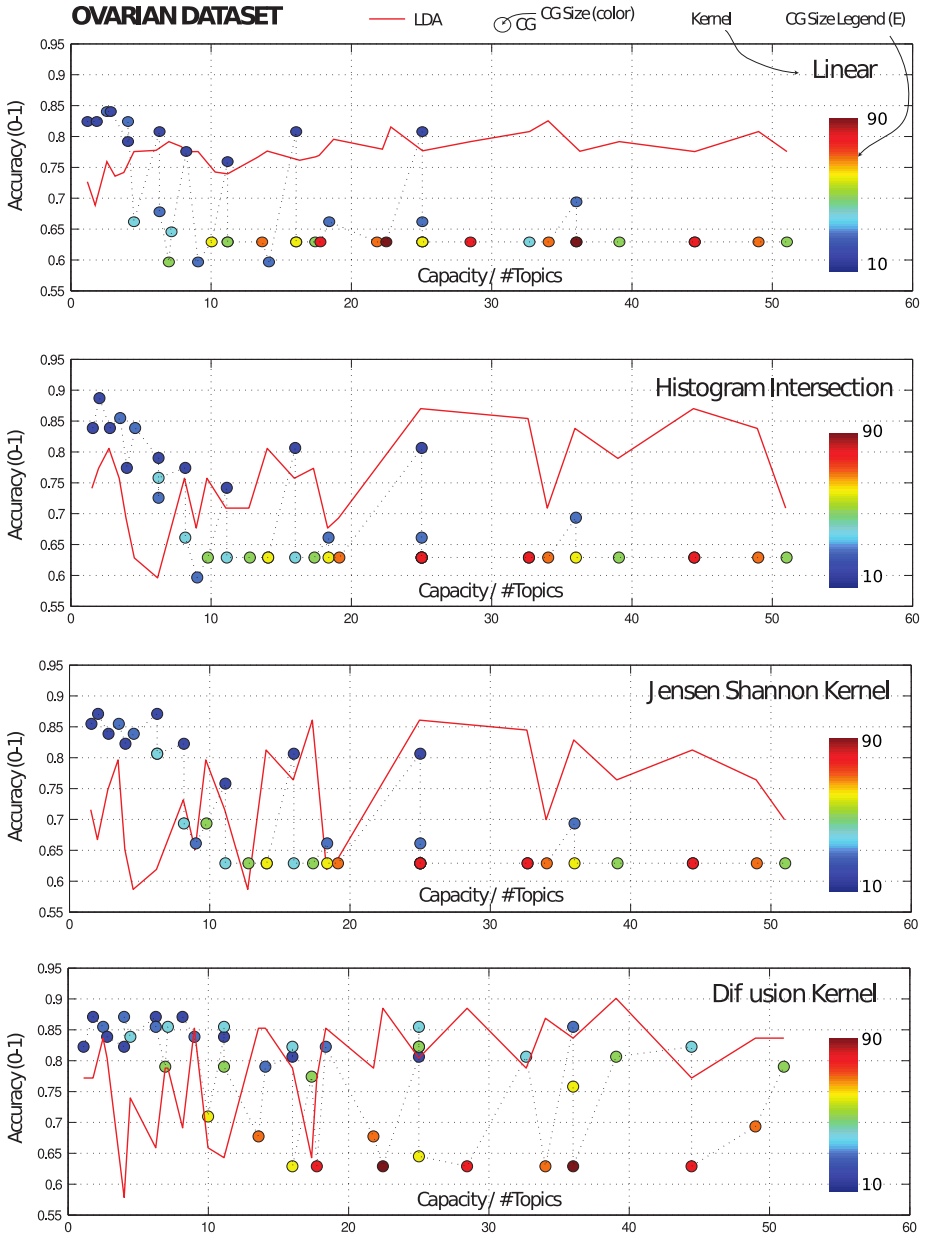

Fig. 3. Baseline results

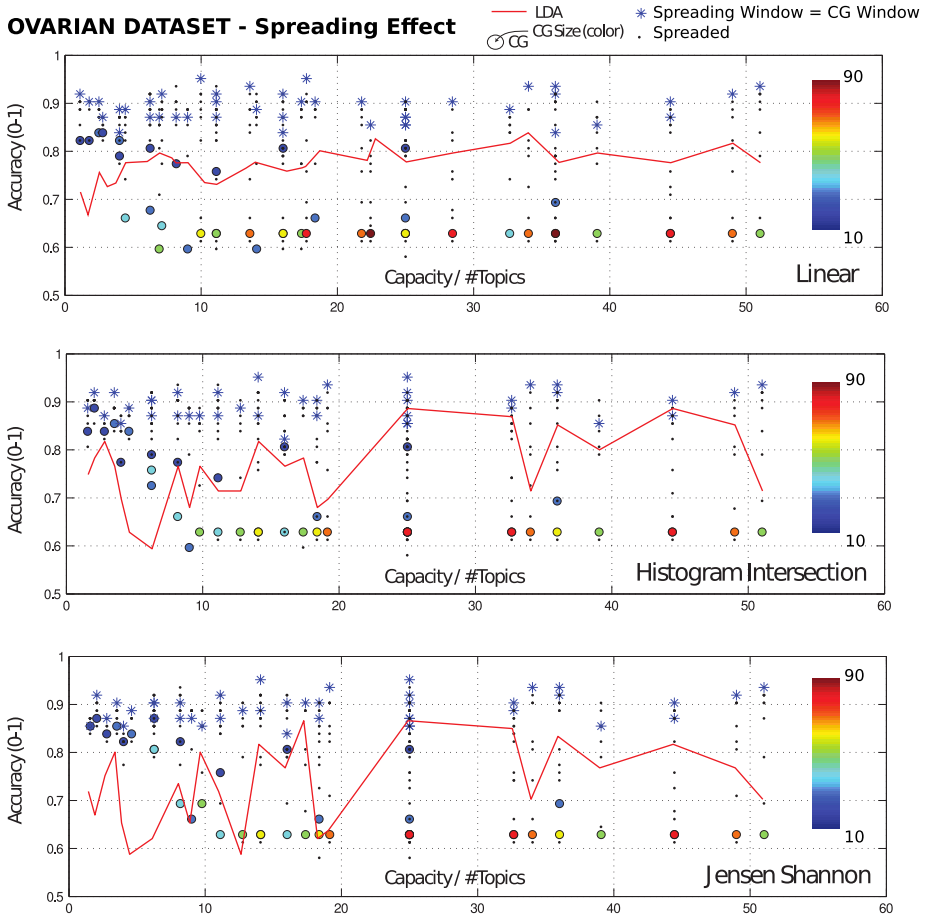

Fig. 4. Results obtained with the proposed spreading operation

4.4 Results

Results are presented in figures 3, 4, 5 and 6. More in detail, in Figure 3 the performances of the original Counting Grids scheme, without any spreading operation, are presented for the different kernels. In particolar, on the x-axis we have the different model size (different capacities), whereas in the y-axis we reported the accuracy. The solid line represents the performances of the LDA. The dimension E of the counting grid is represented by the color. From this figure we can infer that Counting grids are better than the LDA model only for small capacities, whereas for larger capacities the simpler LDA model is preferrable. Moreover it can be noted that the diffusion distance-based kernel represents the best choice (especially for LDA), confirming the intuition that diffusing the values of the posterior represents a good idea. This is more evident by looking at Figure 4, where we plot also the results with the proposed approach (marked with

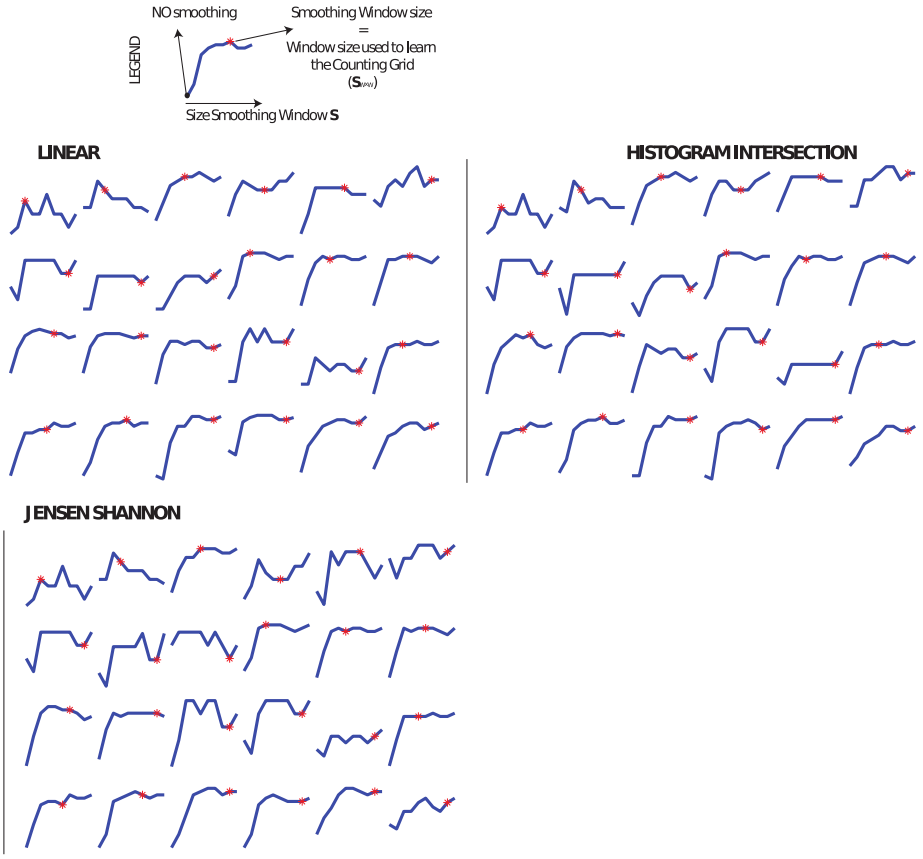

Fig. 5. Analysis of the impact of the dimension of the spreading window

an asterisk), for three of the four kernels – we excluded the diffusion distance-based kernel since already possessing the property of spreading the values. In this case, results with the Counting Grids always outperform the corresponding LDA, making the choice of the capacity less crucial. In that figure, moreover, we also plotted the different accuracies obtained by varying the dimension (from 2 to 10) of the spreading window (marked with a dot). From this figure, it is evident that selecting as the size of the spreading window the size of the counting grid window almost always represents the best choice, as expected. This can be confirmed with the analysis plotted in Figure 5, where for some configurations of the Counting Grid the accuracy for different values of the spreading window is plotted. Also in this case, the asterisk indicates the CG window size, which is almost everywhere among the best values.

Finally, with the same visualization scheme of figure 4, in figure 6 we plot results for the MRI Brain dataset and for the colon cancer microarray dataset. Also in these cases it is evident the gain obtained by the spreading approach.

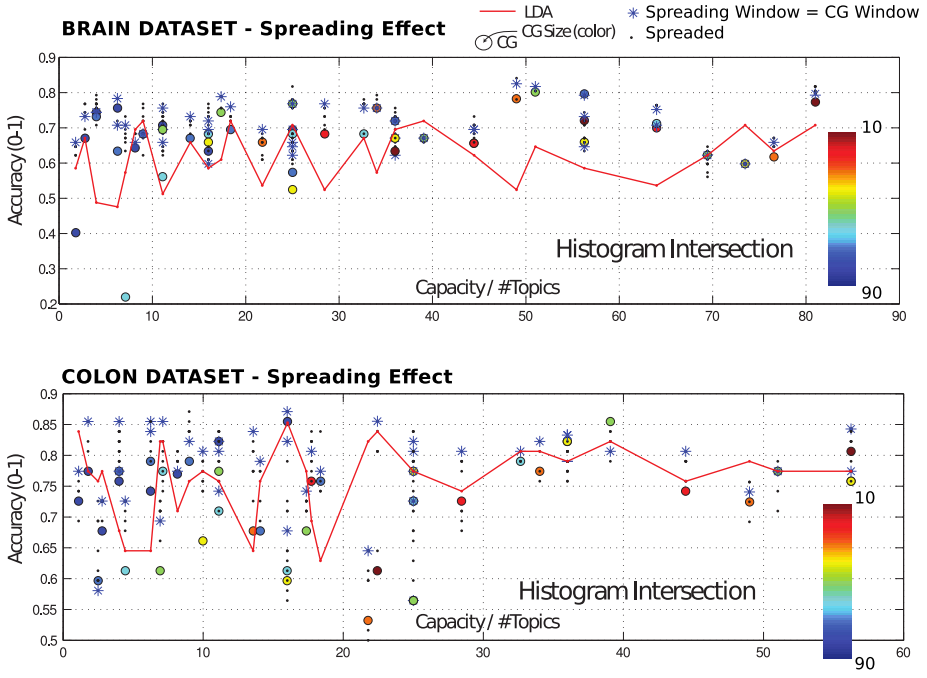

Fig. 6. Results on other datasets

5 Conclusions

In this paper a new approach to compare data represented by counts is introduced. Starting from the recently proposed CGs, we show how the classification perfomance can increase by carefully taking into account of information coming from the generative learning procedure. The proposed Spreading Similarity Mesure leads to a drastic improvement in comparison with standard approaches as shown on different applicative scenarios. In particular, our SSM approach outperfoms diffusion distance which is known to well dealing with cross-count contraints.

Acknowledgements. Authors would like to thank Dr. A. Ulas for the help in the preprocessing of the Brain MRI dataset.

References

1. Jojic, N., Perina, A.: Multidimensional counting grids: Inferring word order from disordered bags of words. In: Uncertainty in Artificial Intelligence (2011)
2. Rogers, S., Girolami, M., Campbell, C., Breitling, R.: The latent process decomposition of cdna microarray datasets. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2005)

3. Bicego, M., Lovato, P., Oliboni, B., Perina, A.: Expression microarray classification using topic models. In: SAC, pp. 1516–1520 (2010)
4. Perina, A., Lovato, P., Cristani, M., Bicego, M.: A comparison on score spaces for expression microarray data classification. In: Loog, M., Wessels, L., Reinders, M.J.T., de Ridder, D. (eds.) PRIB 2011. LNCS, vol. 7036, pp. 202–213. Springer, Heidelberg (2011)
5. Cruska, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision, pp. 1–22 (2004)
6. Toldo, R., Castellani, U., Fusiello, A.: The bag of words approach for retrieval and categorization of 3D objects. The Visual Computer 26(10), 1257–1268 (2010)
7. Brelstaff, G., Bicego, M., Culeddu, N., Chessa, M.: Bag of peaks: interpretation of nmr spectrometry. Bioinformatics 25, 258–264 (2009)
8. Castellani, U., Rossato, E., Murino, V., Bellani, M., Rambaldelli, G., Perlini, C., Tomelleri, L., Tansella, M., Brambilla, P.: Classification of schizophrenia using feature-based morphometry. Journal of Neural Transmission 119, 395–404 (2012)
9. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
10. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42(1-2), 177–196 (2001)
11. Lovato, P., Bicego, M., Cristani, M., Jojic, N., Perina, A.: Feature selection using counting grids: application to microarray data. In: Gimel'farb, G., Hancock, E., Imiya, A., Kuijper, A., Kudo, M., Omachi, S., Windeatt, T., Yamada, K. (eds.) SSPR&SPR 2012. LNCS, vol. 7626, pp. 629–637. Springer, Heidelberg (2012)
12. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
13. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
14. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision 7(1), 11–32 (1991)
15. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory 37(1), 145–151 (1991)
16. Bicego, M., Lovato, P., Perina, A., Fasoli, M., Delledonne, M., Pezzotti, M., Polverari, A., Murino, V.: Investigating topic models' capabilities in expression microarray data classification. IEEE/ACM Trans. Comput. Biology Bioinform. 9(6), 1831–1836 (2012)
17. Perina, A., Lovato, P., Murino, V., Bicego, M.: Biologically-aware latent Dirichlet allocation (balda) for the classification of expression microarray. In: Dijkstra, T.M.H., Tsivtsivadze, E., Marchiori, E., Heskes, T. (eds.) PRIB 2010. LNCS, vol. 6282, pp. 230–241. Springer, Heidelberg (2010)
18. 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)
19. 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, 6745–6750 (1999)
20. Baiano, M., Perlini, C., Rambaldelli, G., Cerini, R., Dusi, N., Bellani, M., Spezzapria, G., Versace, A., Balestrieri, M., Mucelli, R.P., Tansella, M., Brambilla, P.: Decreased entorhinal cortex volumes in schizophrenia. Schizophrenia Research 102(1-3), 171–180 (2008)

21. Ashburner, J.: A fast diffeomorphic image registration algorithm. Neuroimage 38(1), 95–113 (2007)
22. Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W. (eds.): Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press (2007)
23. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
24. 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, Part II. LNCS, vol. 6362, pp. 177–184. Springer, Heidelberg (2010)
25. Martins, A., Smith, N., Xing, E., Aguiar, P., Figueiredo, M.: Nonextensive information theoretic kernels on measures. Journal of Machine Learning Research 10, 935–975 (2009)
26. Cuturi, M., Fukumizu, K., Vert, J.P.: Semigroup kernels on measures. Journal of Machine Learning Research 6, 1169–1198 (2005)
27. Jebara, T., Kondor, R., Howard, A.: Probability product kernels. Journal of Machine Learning Research 5, 819–844 (2004)
28. Bicego, M., Ulas, A., Castellani, U., Perina, A., Murino, V., Martins, A., Aguiar, P., Figueiredo, M.: Combining information theoretic kernels with generative embeddings for classification. Neurocomputing 101, 161–169 (2013)
29. Odone, F., Barla, A., Verri, A.: Building kernels from binary strings for image matching. IEEE Transactions on Image Processing 14(2), 169–180 (2005)
30. Ling, H., Okada, K.: Diffusion distance for histogram comparison. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006, vol. 1, pp. 246–253 (2006)
31. Chapelle, O., Haner, P., Vapnik, V.: Support vector machines for histogram-based image classifcation. IEEE Transactions on Neural Networks 10(5), 1055–1064 (1999)
32. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision 73(2), 213–238 (2007)