ELSEVIER

Kernel method for clustering based on optimal target vector

Leonardo Angelini, Daniele Marinazzo, Mario Pellicoro, Sebastiano Stramaglia *

TIRES-Center of Innovative Technologies for Signal Detection and Processing, Università di Bari, Italy
Dipartimento Interateneo di Fisica, Bari, Italy
Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy

Abstract

We introduce Ising models, suitable for dichotomic clustering, with couplings that are (i) both ferro- and anti-ferromagnetic (ii) depending on the whole data-set and not only on pairs of samples. Couplings are determined exploiting the notion of *optimal target vector*, here introduced, a link between kernel supervised and unsupervised learning. The effectiveness of the method is shown in the case of the well-known iris data-set and in benchmarks of gene expression levels, where it works better than existing methods for dichotomic clustering.
© 2006 Elsevier B.V. All rights reserved.

PACS: 05.10.-a; 87.10.+e

Recent years have been characterized by a dramatic evolution in many fields of life science with the apparition and rapid spread of so-called high-throughout technologies, which generate huge amounts of data to handle various aspects of biological samples or phenomena. The need for efficient methods to represent, analyze and finally make sense out of these data triggered the development of numerous data analysis algorithms, some of which physically motivated: for example in [1] the clustering problem is mapped onto the phase diagram of Potts models, in [2] models of coupled chaotic maps are exploited, while method from quantum mechanics are used in [3]. Among the most effective approaches, kernel methods have quickly gained popularity for problems involving the classification and analysis of high-dimensional or complex data. Well-known supervised kernel algorithms are support vector machines [4], both for regression and classification, and kernel ridge regression [5]. A popular unsupervised kernel method is kernel principal components [6], a nonlinear projection tool which generalizes the classical principal component analysis. In this work we propose use of Ising models for dichotomic clustering, and construct the set of couplings exploiting a simple connection between su-

pervised and unsupervised kernel methods, here presented: the ground state of the Ising model should correspond to the *optimal target vector*, i.e., the target vector leading to the best fit using kernel ridge regression. We compare performances by our method with those from fuzzy c-means [7], the most commonly used method for dichotomic clustering.

In order to describe our prescription for Ising couplings, firstly we recall kernel ridge regression, while referring the reader to [5] or [8] for further technical details. Let us consider a set of ℓ independent, identically distributed data $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$, where $\mathbf{x}_i$ is the n-dimensional vector of input variables and y_i is the scalar output variable. Data are drawn from an unknown probability distribution $p(\mathbf{x}, y)$. The problem of learning consists in providing an estimator $y = f(\mathbf{x})$ out of a class of functions, called *hypothesis space*. In kernel ridge regression, calling $\mathbf{y} = (y_1, y_2, \ldots, y_\ell)^\top$ the vector formed by the ℓ values of the output variable, the estimator is given by

$$y = f(\mathbf{x}) = \sum_{i=1}^{\ell} c_i k(\mathbf{x}_i, \mathbf{x}), \tag{1}$$

where coefficients $\{c_i\}$ are given by

$$\mathbf{c} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y}, \tag{2}$$

* Corresponding author.
   E-mail address: sebastiano.stramaglia@ba.infn.it (S. Stramaglia).

K being the $\ell \times \ell$ matrix with elements $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. λ is called regularization parameter, and its main role is to render well posed the inversion of matrix **K** which in most cases is nearly singular. $k(\cdot, \cdot)$ is a positive definite symmetric function, and Eq. (1) may be seen to correspond to a linear estimator in the feature space

$$\Phi(\mathbf{x}) = \left(\sqrt{\alpha_1}\psi_1(\mathbf{x}), \sqrt{\alpha_2}\psi_2(\mathbf{x}), \dots, \sqrt{\alpha_N}\psi_N(\mathbf{x}), \dots\right),$$

where α_i and ψ_i are the eigenvalues and eigenfunctions of the integral operator with kernel k.

Many choices of the kernel function are possible, for example the linear kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}'$ leads to the usual linear estimator. The polynomial kernel of degree p has the form $k(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x} \cdot \mathbf{x}')^p$ (the corresponding features are made of all the powers of $\mathbf{x}$ up to the pth). The Gaussian kernel is $k(\mathbf{x}, \mathbf{x}') = \exp[-(\|\mathbf{x} - \mathbf{x}'\|^2/2a^2)]$ and deals with all the degrees of nonlinearity of $\mathbf{x}$. Specifying the kernel function k one determines the complexity of the function space within which one searches for the estimator, similarly to the effect of specifying the architecture of a neural network, that is number of layers, number of units for each layer, etc. Due to (1) and (2), the predicted output vector $\bar{\mathbf{y}}$, in correspondence of the *true* target vector $\mathbf{y}$, is given by $\bar{\mathbf{y}} = \mathbf{Gy}$, where the symmetric matrix **G** is given by

$$\mathbf{G} = \mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1}. \tag{3}$$

Note that matrix **G** depends only on the distribution of $\{\mathbf{x}\}$ values: **G** embodies information about the structures present in $\{\mathbf{x}\}$ data set. Indeed, the matrix element G_{ij} quantifies how much the target value of the jth point influences the estimate of the target of point i. The training error E can be calculated as follows

$$E = (\mathbf{y} - \mathbf{Gy})^\top(\mathbf{y} - \mathbf{Gy}) = \mathbf{y}^\top \mathbf{H} \mathbf{y},$$

where $\mathbf{H} = \mathbf{I} - 2\mathbf{G} + \mathbf{GG}$ is a symmetric and positive matrix.

It is worth stressing that, given a kernel function, the corresponding features $\phi_\gamma(\mathbf{x})$ are not centered in general. One can show [6] that centering the features ($\phi_\gamma \to \phi_\gamma - \langle \phi_\gamma \rangle$, for all γ) amounts to perform the following transformation on the kernel matrix:

$$\mathbf{K} \to \tilde{\mathbf{K}} = \mathbf{K} - \mathbf{I}_\ell \mathbf{K} - \mathbf{K} \mathbf{I}_\ell + \mathbf{I}_\ell \mathbf{K} \mathbf{I}_\ell,$$

where $(\mathbf{I}_\ell)_{ij} = 1/\ell$, and to work with the centered kernel $\tilde{\mathbf{K}}$. Therefore in the following we will always assume that the kernel matrix **K** has been centered.

Matrix **G** may be seen as a linear convolution filter, mapping target vectors onto the vector of predicted outputs, and carrying information about the structures naturally present in the set of $\{\mathbf{x}\}$ points. In the unsupervised case the data set is made of $\mathbf{x}$ points, $\{\mathbf{x}_i\}_{i=1}^\ell$, and the target vector $\mathbf{y}$ is missing. However we may consider the following question: what is the vector $\mathbf{y}$ such that treating it as the target vector leads to the best fit, i.e., the minimum training error $\mathbf{y}^\top \mathbf{H} \mathbf{y}$? One may expect that this *optimal* target vector would bring information about the structures present in the data. We look for the optimal target vector in the space of binary functions, $\sigma \in \{-1, 1\}^\ell$, the minimizer of
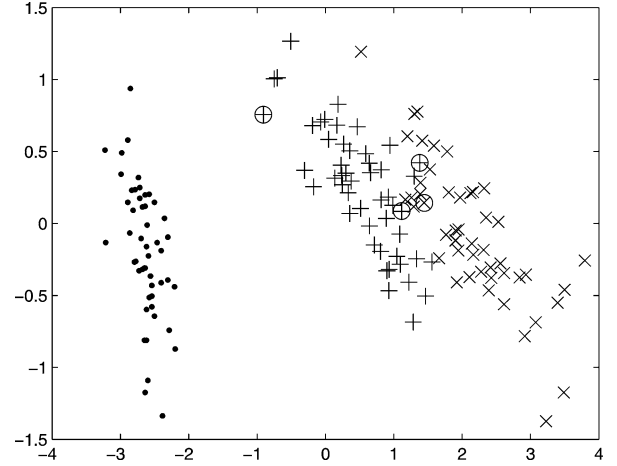

Fig. 1. In the plane of the first and second principal components, the iris data-set is represented. Symbols represent classes: $\cdot$ *setosa*, $+$ *versicolor*, $\times$ *virginica*. The four misclassified points are surrounded by a circle.

the training error $\sigma^\top \mathbf{H} \sigma$ thus naturally provides a partition of points in two classes. The minimizer is the ground state of an Ising model (see, e.g., [9]) with long range symmetric couplings **J** given by

$$J_{ij} = 4G_{ij} - 2\sum_{k=1}^\ell G_{ik}G_{kj},$$

for $i \neq j$, and $J_{ii} = 0$. We note that, unlike the Potts model introduced in [1], here the couplings are both ferro- and antiferromagnetic. Therefore there is room for frustration and multiple minima. Many algorithms can be used to find an estimate of the ground state of Ising models, for example simulated annealing [10]. Here we use the mean field annealing method [11], which iteratively solves the mean field equations for the local magnetization vector $\mathbf{m} = \langle \sigma \rangle$,

$$\mathbf{m} = \tanh(\beta \mathbf{J} \mathbf{m}), \tag{4}$$

while decreasing the temperature (increasing β). The starting temperature β_{cr} may be chosen as the inverse of the maximum eigenvalue of matrix **J**. Note that $\mathbf{m} \in \{-1, 1\}^\ell$ in the limit of large β.

Let us now discuss the application to iris data-set, published by Fisher [12], where the sepal length, sepal width, petal length, and petal width are measured in millimeters on fifty iris specimens from each of three species, *iris setosa*, *iris versicolor*, and *iris virginica*. Firstly we process all the 150 points, using a linear kernel and $\lambda = 1$: we find two clusters, one is made of 51 points (50 belonging to *iris setosa*), the other is made of the remaining 99 points. Then, we process the 99 points with linear kernel and the same value of λ, obtaining two clusters of 48 points (47 belonging to *iris versicolor*) and 51 points (49 belonging to *iris virginica*). Globally four points are misclassified, with efficiency of classification 0.973. In Fig. 1 we show the four misclassified points. Results are quite insensitive to λ: the same classification is obtained varying $\lambda \in [0.05, 100]$. Using a Gaussian kernel we obtain even better efficiency: using $a = 3$ only two points are misclassified, with efficiency 0.987.
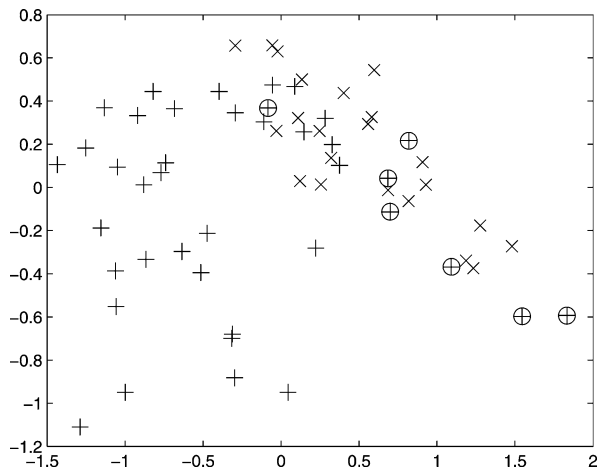
Fig. 2. The colon data set, + normal and × tumor, is represented in the plane of the first and second principal components (extracted over the 100 most discriminating genes). Misclassified points are surrounded by a circle.

Fig. 3. The leukemia data set, + AML and × ALL, is represented in the plane of the first and second principal components (extracted over the 500 most discriminating genes). Misclassified points are surrounded by a circle.

We now compare our results with those from fuzzy c-means; applying twice dichotomic fuzzy c-means, as in the case of the proposed method, we obtain 0.847 efficiency. Application of fuzzy c-means, with three centers, leads to 0.893 efficiency. Hence in both cases the proposed algorithm outperforms fuzzy c-means.

Now we consider application on gene expression data sets. Firstly the colon cancer data set of [13], consisting in 40 tumor and 22 normal colon tissues samples, each described by 2000 gene expression levels; data are available on the Kent Ridge Bio-medical Data Set Repository [14]. We are not going to face, in this place, the important task of feature selection which is fundamental in the analysis of gene expression data. The following preprocessing is used to normalize data. First, for each gene, expression levels are rescaled so as to get unit mean over tissues. Then, for each tissue, expression levels are linearly transformed to have zero mean (over genes) and unit variance. Application of our method, using all the genes and a linear kernel ($\lambda = 1$), leads to 12 misclassified points with an efficiency of 0.806 (the efficiency from fuzzy c-means is 0.532). Then we select the 100 most discriminant genes, using a supervised step where nonparametric Wilcoxon test is used to asses the capability of each gene in discriminating tumor and normal tissues. After ranking genes by Wilcoxon test, we apply our method (still with linear kernel and $\lambda = 1$) using only 100 attributes corresponding to the first 100 genes. The output is depicted in Fig. 2, seven points are misclassified with efficiency 0.887 (fuzzy c-means efficiency is 0.613). Next we consider the leukemia data set of [15], consisting of samples of tissues of bone marrow samples, 47 affected by acute myeloid leukemia (AML) and 25 by acute lymphoblastic leukemia (ALL), 7129 attributes. Data are normalized as in the previous example. Using all genes, and linear kernel, leads to a poor performance with efficiency 0.569 (fuzzy c-means shows 0.555 efficiency in this case). Ranking genes by Wilcoxon test (assessing the capability of each gene in discriminating AML and ALL tissues) and selecting the best 500 genes, leads to better performances: our method with a linear kernel misclassifies 8 points, with effi-
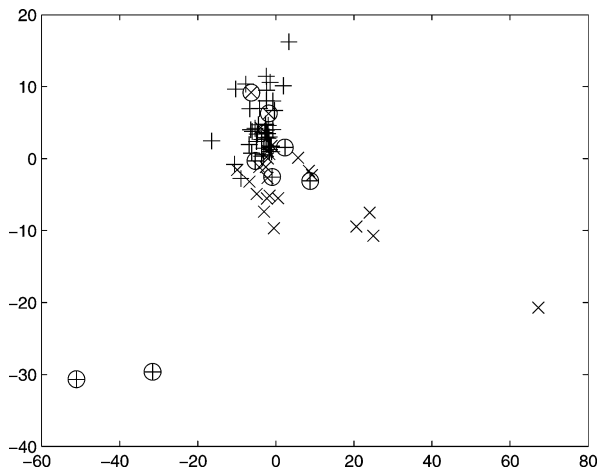
ciency 0.889, see Fig. 3 (the classification from fuzzy c-means has an efficiency 0.722). Our results are quite insensitive to the choice of λ.

We now summarize our method. Dichotomic clustering is described as the problem of evaluating the ground state of Ising models, with couplings values determined in the frame of kernel methods for supervised learning. The effectiveness of the proposed approach is shown here on real examples in terms of high efficiency of classification, much higher than those from fuzzy c-means. As the output depends very weakly on the regularization parameter λ, only the choice of the kernel matters. In this Letter we mostly deal with the simplest kernel, the linear one: the problem of kernel selection, common to any kernel method, will be considered elsewhere.

Some remarks are in order. Firstly, it is worth stressing that some information about structures can also be recovered from diagonal elements G_{ii}. In the leave-one-out scheme [4] a data point i is removed from the data set and the model is trained using the remaining $\ell - 1$ points: let us denote $\tilde{y}_i$ the target value thus predicted, in correspondence of $\mathbf{x}_i$. It can be shown that the leave-one-out error $\tilde{y}_i - y_i$ and the training error $\bar{y}_i - y_i$, obtained using the whole data set, satisfy

$$\tilde{y}_i - y_i = \frac{\bar{y}_i - y_i}{1 - G_{ii}}. \tag{5}$$

This formula shows that the closer G_{ii} to one, the farther the leave-one-out predicted value from those obtained using also point i in the training stage. If point i is in a dense region of the feature space then one may expect that removing this point from the data-set would not change much the estimate since it can be well "interpolated" by neighboring points. Therefore points in low density regions of the feature space are characterized by diagonal values G_{ii} close to one, while G_{ii} is close to zero for points $\mathbf{x}_i$ in dense regions. In Fig. 4 we depict $\{G_{ii}\}$ for the case of iris.

Secondly, one may look for the optimal target vector $\mathbf{y}$ in $\mathbb{R}^\ell$; to avoid the trivial solution $\mathbf{y} = \mathbf{0}$, one may constrain $\mathbf{y}$ to have unit norm, $\mathbf{y}^\top \mathbf{y} = 1$. This problem then becomes equivalent to
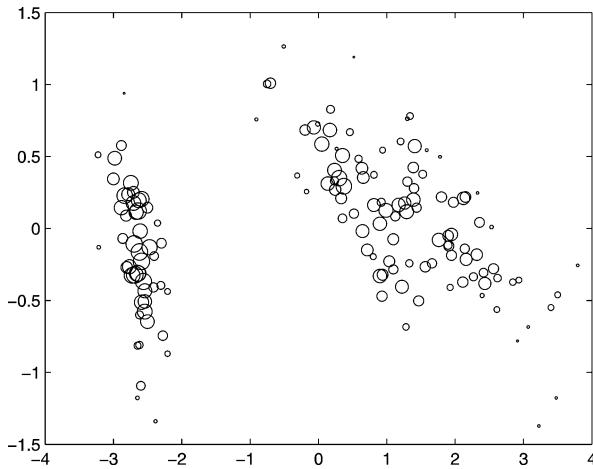
Fig. 4. The 150 points of iris data-set are depicted (first and second principal components). The radius of the circle around each point is proportional to $1 - G_{ii}$, the matrix $\mathbf{G}$ being evaluated using a Gaussian kernel with $\sigma = 0.5$ and $\lambda = 1$.

find the normalized eigenvector of $\mathbf{H}$ with the smallest eigenvalue. On the other hand, matrix $\mathbf{H}$ is a function of matrix $\mathbf{K}$: hence it has the same eigenvectors of $\mathbf{K}$ while the corresponding eigenvalues are related by a monotonically decreasing function. Therefore, optimal target functions in $\mathbb{R}^\ell$ coincide with the kernel principal components: this shows that the method introduced in [6] may be motivated also as the search for the optimal target functions.

Finally, as already shown in the iris example, we remark that multiclass clustering may be obtained by repeated applica-

tion of the proposed dichotomic method. The notion of optimal target vector is an interesting bridge between supervised and unsupervised learning: here we have considered application of this notion for dichotomic clustering. In our opinion it will carry to the developments of other effective algorithms for the analysis of complex data.

References

[1] M. Blatt, S. Wiseman, E. Domany, Phys. Rev. Lett. 76 (1996) 3251;
G. Getz, E. Levine, E. Domany, Proc. Natl. Acad. Sci. 97 (2000) 12079.
[2] L. Angelini, F. De Carlo, C. Marangi, M. Pellicoro, S. Stramaglia, Phys. Rev. Lett. 85 (2000) 554;
L. Angelini, L. Nitti, M. Pellicoro, S. Stramaglia, Phys. Lett. A 285 (2001) 279.
[3] D. Horn, A. Gottlieb, Phys. Rev. Lett. 88 (2002) 018702.
[4] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[5] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge Univ. Press, Cambridge, 2004.
[6] B. Scholkopf, A. Smola, K.-R. Muller, Neural Comput. 10 (1998) 1299.
[7] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York, 1981.
[8] N. Ancona, R. Maestri, D. Marinazzo, L. Nitti, M. Pellicoro, G.D. Pinna, S. Stramaglia, Physiol. Meas. 26 (2005) 363.
[9] L.P. Kadanoff, Statistical Physics: Statics, Dynamics and Renormalization, World Scientific, Singapore, 2000.
[10] S. Kirkpatrick, C.D. Gelatt Jr., M.P. Vecchi, Science 220 (1983) 671.
[11] J.J. Hopfield, D.W. Tank, Biol. Cybern. 52 (1985) 141;
C. Peterson, B. Soderberg, Int. J. Neural Systems 1 (1989) 3.
[12] R.A. Fisher, Ann. Eugenics 7 (1936) 179.
[13] U. Alon, et al., Proc. Natl. Acad. Sci. 96 (1999) 6745.
[14] http://sdmc.lit.org.sg/GEDatasets/Datasets.html.
[15] T.R. Golub, et al., Science 286 (1999) 531.