DISSIMILARITY-BASED METRIC FOR MEDICAL DATA CLASSIFICATION

Agata Manolova

Technical University
8 ave Kliment Ohridski, Sofia 1000, Bulgaria
amanolova@tu-sofia.bg

Abstract

Data mining techniques have been applied to medical services in several areas, including prediction of effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data. In our study we use a dissimilarity-based metric for the classification of different types of medical data for diagnostics such as breast cancer, heart disease, diabetes etc.

Dissimilarity-based pattern recognition offers new possibilities for building classifiers on a distance representation such as kernel methods or the k nearest neighbors (kNN) rule. The goal of this work is to expand the advantageous and rapid adaptive approach to learn only from dissimilarity representations by using the effectiveness of the Support Vector Machines algorithm for real-world classification tasks for medical data. This method can be an alternative approach to the well known methods based on dissimilarity representations and can be as effective as them in terms of accuracy for classification. Practical examples on real medical data show interesting behavior compared to other dissimilarity-based methods.

1. INTRODUCTION

Pattern recognition techniques play a critical role when applied to medical databases by fully automating the process of abnormality detection and thus supporting the development of computer-aided diagnosis (CAD) systems. Often this involves identifying structures such as tumors or lesions, but it can also include monitoring present structures such as the size of the heart in chest X-rays. In most cases, CAD systems are designed to be used for screening purposes, in which large numbers of medical data needs to be examined. They are adopted as an alternative "second opinion" that can assist for example a radiologist in detecting lesions and in making diagnostic decisions. The computerized schemes combine detection and classification of malignancies. The importance of these CAD systems in almost all telemedicine applications is evident and is expected to increase dramatically over the coming years [1], [2].

The goal of this work is to test the advantageous and rapid adaptive approach to learn only from dissimilarity representations by using the effectiveness of the Support Vector Machines algorithm developed by Manolova and Guerin [3] for real-world classification tasks for medical data of different types. This method can be an alternative approach to the known methods based on dissimilarity representations such as Pekalska's dissimilarity classifier [4], Haasdonk's kernel-based SVM classifier [3] and to classic kNN classifier.

The paper is organized as follows: in Section 2 we introduce the motivation of the approach, in Section 3 we describe the theoretical basis of this approach; in Section 4 we provide experimental results on real-life medical data sets. Finally, Section 5 concludes the paper.

2. MOTIVATION OF THE APPROACH

The motivation for the development of a dissimilarity-based classifier is the following: if we assume that "similar" objects can be grouped together to form a class, a "class" is nothing more than a set of these "similar" objects. Based on this idea, it is possible that the notion of proximity (similarity or dissimilarity) is actually more fundamental than that of a feature. Thus, the dissimilarity-based classifiers are a way of defining classifiers between the classes, which are not based on the feature measurements of the individual patterns, but rather on a suitable dissimilarity measure between them. The advantage of this methodology is that since it does not operate on the class-conditional distributions, the accuracy can exceed theoretically the Bayes' error bound. Another salient advantage of such a paradigm is that it does not have to confront the problems associated with feature spaces such as the "curse of dimensionality", and the issue of estimating a large number of parameters.

The distance representation is most commonly used as dissimilarity because is usually the simplest measure. A dissimilarity value expresses a magni-

tude of difference between two objects and becomes zero only when they are identical.

This paper focuses on the incorporation of SVM in to the dissimilarity-based classifier "Shape Coefficient" described in [5], [6]. The Shape Coefficient (Cs) is defined from simple statistics (mean and variance) on the dissimilarity data. The proposed decision rules are based on this Shape Coefficient description and on optimal separating hyper plane with Support Vector Classifier (SVC), using the Cs coefficient as dissimilarity on the input space. This provides a decision rule with a limited number of parameters per class.

3. DESCRIPTION OF THE "SHAPE COEFFICIENT"

Let us consider a two-class classification problem where ω_1 is the first class and ω_2 the second class. Let *N* be a set of objects o_i to be classified, D is the dissimilarity (N×N) table between each object such as: $D = \left[d(o_i, o_j) : 1 \le i, j \le N\right]$. Following [5] and [6], the Shape Coefficient describes the proximity of an object to a given class (for example for ω_1, eq. 1):

$$Cs(o_i, \omega_1) = \frac{\gamma_1[\overline{d^2(o_i, \omega_1)} - I(\omega_1)]^2}{[var(d^2(o_i, \omega_1))]^{\delta_1}} \qquad (1)$$

where $\overline{d(o_i, \omega_1)^2}$ is the empirical average of the dissimilarity between object o_i and all the observations in class ω_1, $var(d(o_i, \omega_1))$ is the empirical variance, and I (ω_1) is the class inertia computed as the empirical mean of all the squared dissimilarities between objects in class ω_1. The numerator deals with the "position" of the observation o_i relatively the class center. The denominator interpretation is more complex, taking into account the "structure" (orientation, shape, intrinsic dimension…) of the observations distribution in the class. Then the parameters γ_1 and δ_1 are learning parameters to best fit this data structure. The equation for $Cs(o_i\ \omega_2)$ with the class ω_2 is equivalent to (1) and has two fitting parameters γ_2 and δ_2. The decision rule for a two-class classification problem for an object o_i is given then by the following equation:

$$Cs(o_i, \omega_1) \underset{2}{\overset{1}{<}} Cs(o_1, \omega_2) \qquad (2)$$

3.1. Decision rule using SVC optimization

The quantities $Cs(o_i\ \omega_1)$ and $Cs(o_i\ \omega_2)$ being positive, we can transform (2) using the logarithmic function as follows:

$$\begin{aligned} &\log\left(\frac{\gamma_1}{\gamma_2}\right) + 2\log(\overline{d^2(o_i, \omega_1)} - I(\omega_1)) - \\ &- 2\log(\overline{d^2(o_i, \omega_2)} - I(\omega_2)) - \\ &- \delta_1 \log(var(d^2(o_i, \omega_1))) + \\ &+ \delta_2 \log(var(d^2(o_i, \omega_2))) \underset{2}{\overset{1}{<}} 0 \end{aligned} \qquad (3)$$

This is in fact, a linear decision rule in a 4-dimensional input space. Following (3), we can represent each object o_i using a vector x_i with 4 features, $x_i = [x_{i1}\ x_{i2}\ x_{i3}\ x_{i4}]^T$:

$$\begin{aligned} x_{i1} &= 2\log(\overline{d^2(o_i, \omega_1)} - I(\omega_1)) \\ x_{i2} &= -2\log(\overline{d^2(o_i, \omega_2)} - I(\omega_2)) \\ x_{i3} &= -\log(var(d^2(o_i, \omega_1))) \\ x_{i4} &= \log(var(d^2(o_i, \omega_2))) \end{aligned} \qquad (4)$$

So now, the decision rule (3) becomes:

$$\beta^T x_i + \beta_0 \underset{2}{\overset{1}{<}} 0 \qquad (5)$$

with $\beta = [1\ 1\ \delta_1\ \delta_2]^T$ be the normal to the optimal separating hyper plane and $\beta_0 = \log\left(\gamma_1 / \gamma_2\right)$ be the bias from the hyper plane to the origin. Labeling the objects with the auxiliary variables per class, such as $y_i = -1 \text{ for } o_i \in \omega_1$ and $y_i = 1 \text{ for } o_i \in \omega_2$, we have the following classical linear decision rule:

$$y_i = sign(\beta^T x_i + \beta_0) \quad (6)$$

This is the standard decision rule for SVC. Here, the difference is the vector β normal to the optimal hyper plane: it is constraint to have the same two first components: $\beta_1 = \beta_2$. Thus finding the optimal hyper plane when the 2 classes are inseparable consists of this optimization problem solved by using the Lagrange multipliers [BUR98]:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i-1}^{N}\zeta_i$$
$$\text{subject to } y_i(\beta^T x_i + \beta_0) \geq 1-\zeta_i, \quad (7)$$
$$\zeta_i \geq 0, i = 0,..., N$$

where ζ_i are the slack variables, associated with all the objects. If the object o_i is classified in the wrong class then $\zeta_i > 1$. The parameter C corresponds to the penalty for errors and it is chosen by the user. In order to introduce the constraints on the β vector, we consider the observations x_i into two orthogonal subspaces such as:

$x_i = [x_i' \; x_i'']^T$, $x_i' = [x_{i1} \; x_{i2}]^T$, $x_i'' = [x_{i3} \; x_{i4}]^T$ an d also:

$$\beta = [\beta' \; \beta'']^T,$$
$$\beta' = \|\beta'\|.[1\,1]^T / \sqrt{2}, \quad (8)$$
$$\beta'' = [\beta_3 \; \beta_4]^T$$

The optimization problem is then transformed such as:

$$\min_{\|\beta'\|,\beta'',\beta_0} \frac{1}{2}\|\beta'\|^2 + \frac{1}{2}\|\beta''\|^2 + C\sum_{i-1}^{N}\zeta_i$$
$$\text{subject to } y_i(\|\beta'\|u_i' + \beta''^T.x_i'' + \beta_0) \geq 1-\zeta_i, \quad (9)$$
$$\zeta_i \geq 0, i = 0,..., N$$

with u_i' the scalar product such as:

$$u_i' = \left\langle [1 \quad 1]^T, x_i' \right\rangle / \sqrt{2} \quad (10)$$

4. EXPERIMENTAL RESULTS

All the experiments are done using SVM^{Light}, an implementation of Support Vector Machines in C by Thorsten Joachims (http://svmlight.joachims.org) and Matlab. We have made source modifications in order to implement the supplementary constraints on the β vector.

The medical datasets come from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). The information of the medical dataset used in our experiments is gathered in Table 1.

The information is collected from real world patients or microbiological laboratories and consists of mixed types of data: continuous, dichotomous and categorical variables (ex. age, sex, chest pain type, blood sugar, heart condition, proteins, blood pressure etc.). There are also datasets with missing values (ex. the Heart dataset). For the multi class SVC optimization procedure we use "one versus all" method.

Table 1. Medical Datasets used in the experiments

Data	Class	Class sizes	Dissimilarity
Heart	2	139/164	Gower's distance
Proteins	5	72/72/ 39/30/13	Evolutionary Distance
Cat Cortex	4	18/10/ 18/19	Determined by Expert

Table 2 summarizes the results for the average classification error for these datasets with the classifier "Shape Coefficient" and the classifiers 1-NN (Nearest Neighbor), K-NN (K Nearest Neighbors), the SVM with 3 different kernels (linear, polynomial and Gaussian) from [3] and [4].

Table 2. Average classification error [in %] for the medical datasets in LOO

Data	Heart	Proteins	Cat Cortex
1-NN	26.8	1.66	5
K-NN	22.6	1.66	3.84
Cs	22.6	1.11	3.46
SVM	21.5	0.89	3.09

5. CONCLUSION

We have proposed a new way of optimizing the parameters of the proximity index "Shape Coefficient". It used the SVM decision rules which allow us to find the optimal solution for our classification problem. With only two parameters per class, the model for class description is compact and parsimonious. The model is flexible, effective and fast in different classification tasks as already proven in [5] and [6]. The result of the comparison with the K-NN and 1-NN shows better results for the classification error. The Cs with SVC optimization procedure is a global method with adjustable parameters according to the properties of the class so it performs better then the K-NN rule in case of (1-NN or 3-NN). The good performance the SVM with linear kernel on proteins and cat-cortex data is a hint on the linear separability of these two datasets. The result is confirmed by the Cs classifier. Indeed, the polynomial and Gaussian kernel improve the results of the linear kernel for most datasets. The Gaussian ker-

nel even slightly outperforms the polynomial in most cases so in Table 2 only the best results are shown.

The results with the real-world medical datasets encourage us to propose this metric as a good alternative to other dissimilarity-based classifiers for this kind of tasks – assisting the medical personnel to take decisions about the condition of a patient for example. Because the metric uses only 2 parameters per class and a linear kernel, data classification is very fast (0.07 seconds in SVM^{Light} for 200 points).

6. ACKNOWLEDGEMENTS

This work was financed by a project grant of the National Fund for Scientific Research of the Bulgarian Ministry of Education and Science by the contract VU-I-305.

References

[1] Anke Meyer-Base, *Pattern Recognition for Medical Imaging*, Elsevier Academic Press, 2004.

[2] G. Dougherty, *Digital Image Processing for Medical Applications*, Cambridge University Press, 2009.

[3] B. Haasdonk, C. Balhmann, "Learning with Distance Substitution Kernels", *Pattern Recognition -Proc.* of the 26th DAGM Symposium, Tubingen, Germany, August/September 2004.

[4] R.P.W. Duin, E. Pekalska, *Object representation, sample size and dataset complexity*, Springer-Verlag, pp. 25-58, 2006.

[5] Manolova, A. Guerin-Dugue, C*assification of dissimilarity data with a new flexible Mahalanobis-like metric*. Pattern Anal. Appl. 11(3-4): 337-351 (2008), Springer Link.

[6] Manolova, A. Guerin-Gugue, "Dissimilarity-based metric for data classification using Support Vector Classifiers", RSFC, Grenoble, France, 2009, pp. 37-41.