Artificial Intelligence in Medicine 54 (2012) 103–114

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Improved modeling of clinical data with kernel methods

Anneleen Daemen[a,*], Dirk Timmerman[b], Thierry Van den Bosch[b], Cecilia Bottomley[c], Emma Kirk[d], Caroline Van Holsbeke[b,e], Lil Valentin[f], Tom Bourne[b,g], Bart De Moor[a]

[a] Department of Electrical Engineering, Katholieke Universiteit Leuven, 3001 Leuven, Belgium
[b] Department of Obstetrics and Gynecology, University Hospitals Leuven, Katholieke Universiteit Leuven, 3000 Leuven, Belgium
[c] Department of Obstetrics and Gynaecology, St. George's Hospital, St. George's University of London, London SW17 0RE, UK
[d] Early Pregnancy and Gynecological Unit, St. George's Hospital, St. George's University of London, London SW17 0RE, UK
[e] Hospital Oost-Limburg, 3600 Genk, Belgium
[f] Malmö University Hospital, Lund University, SE 20502 Malmö, Sweden
[g] Hammersmith Hospital, Imperial College London, London W12 0NN, UK

ARTICLE INFO

Article history:
Received 27 November 2009
Received in revised form 22 October 2011
Accepted 7 November 2011

Keywords:
Machine learning
Support vector machine
Kernel function
Biostatistics
Clinical data representation
Clinical decision support system
Gynecology
Breast cancer

ABSTRACT

Objective: Despite the rise of high-throughput technologies, clinical data such as age, gender and medical history guide clinical management for most diseases and examinations. To improve clinical management, available patient information should be fully exploited. This requires appropriate modeling of relevant parameters.

Methods: When kernel methods are used, traditional kernel functions such as the linear kernel are often applied to the set of clinical parameters. These kernel functions, however, have their disadvantages due to the specific characteristics of clinical data, being a mix of variable types with each variable its own range. We propose a new kernel function specifically adapted to the characteristics of clinical data.

Results: The clinical kernel function provides a better representation of patients' similarity by equalizing the influence of all variables and taking into account the range r of the variables. Moreover, it is robust with respect to changes in r. Incorporated in a least squares support vector machine, the new kernel function results in significantly improved diagnosis, prognosis and prediction of therapy response. This is illustrated on four clinical data sets within gynecology, with an average increase in test area under the ROC curve (AUC) of 0.023, 0.021, 0.122 and 0.019, respectively. Moreover, when combining clinical parameters and expression data in three case studies on breast cancer, results improved overall with use of the new kernel function and when considering both data types in a weighted fashion, with a larger weight assigned to the clinical parameters. The increase in AUC with respect to a standard kernel function and/or unweighted data combination was maximum 0.127, 0.042 and 0.118 for the three case studies.

Conclusion: For clinical data consisting of variables of different types, the proposed kernel function – which takes into account the type and range of each variable – has shown to be a better alternative for linear and non-linear classification problems.

1. Introduction

During an examination, patient-specific information such as age, menopausal status and medical history is registered. Histopathological parameters such as tumor size, lymph node status and relapse rate, and ultrasound data such as endometrium thickness are often registered as well, with the set of clinical parameters characterizing a patient depending on the investigated disease. Such parameters or combinations thereof have been evaluated as prognostic indicators (for example [1,2]). Because clinicians prefer interpretable decision support systems, clinical management for diagnosis and prognosis and decisions concerning therapy response are for most of the diseases and examinations fully based on clinical and pathological indicators.

Besides clinical data, high-throughput technology – and especially microarray technology – has considerably advanced basic biological science and the entire field of cancer taxonomy, biomarker development and identification of prognostic and predictive markers [3–5]. In numerous studies, multiple high-throughput data sources were collected and simultaneously

[*] Corresponding author at: Department of Cancer & DNA Damage Responses, Life Sciences Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, 94720 Berkeley, CA, USA. Tel.: +1 510 486 5202.
E-mail address: anneleen.daemen@gmail.com (A. Daemen).

0933-3657/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2011.11.001

studied while omitting clinical parameters. High-throughput data, however, are in general much more difficult and expensive to collect while clinical parameters are routinely measured by clinicians. The latter have been used by clinicians for decades and should be included in the investigation, moreover because a critical study on the prediction of breast cancer outcome has suggested that clinical markers and profiles obtained from high-throughput technologies have similar power for prognosis [6].

Advanced mathematical models can aid clinical decision support. In many previous studies [7–10], the support vector machine (SVM) [11] was used for this purpose. Several disadvantages, however, occur when applying the SVM directly to clinical data, due to the heterogeneous nature of clinical data compared to high-throughput data sources. The influence of each variable on patients' similarity will be proportional to its range, thereby enlarging the influence of irrelevant continuous variables and diminishing the contribution of important discrete variables. As it has been shown that better results can be obtained by adapting the kernel function to the structure of the data and defining a kernel function per domain [12], a distinction is made between continuous variables, ordinal variables with an intrinsic ordering but often lacking equal distance between two consecutive categories, and nominal variables without any ordering.

The scale of the input data was already known to influence model performance. A rough distinction according to variable type was incorporated in LS-SVMlab, a Matlab/C toolbox containing a variety of techniques and algorithms for the least squares support vector machine (LS-SVM) with applications in classification and non-linear regression [13]. Binary variables were re-scaled to $\{-1, 1\}$ while continuous variables were normalized, avoiding attributes in larger numeric ranges to dominate those in smaller ranges. Other variables, however, were kept unchanged, thereby not distinguishing ordinal from nominal variables.

We will propose an alternative kernel function specifically developed for clinical data, which does not suffer from the ambiguity of data preprocessing by equally taking into account all variables. First, we will show the improvement obtained with this alternative kernel function when applied to four clinical data sets within gynecology. Secondly, the advantage of this kernel function will be illustrated for the combination of clinical and microarray data in three case studies on breast cancer.

2. Methods

2.1. Kernel methods and least squares support vector machine

Kernel methods are a powerful class of algorithms for pattern analysis. They work in a high dimensional feature space to which data x is mapped from the original input space with the function $\Phi(x)$ [14,15]. The kernel function $k(x^i, x^j)$ efficiently computes the inner product $\langle \Phi(x^i), \Phi(x^j) \rangle$ between all pairs of data items x^i and x^j in the feature space, resulting in the $N \times N$ kernel matrix K with N the number of data items. Any symmetric, positive semi-definite function is a valid kernel function, resulting in many possible kernels. However, no formal proof of optimality exists for the use of one kernel function above an other. The functions that are most frequently employed in classification problems are the linear kernel $x^{i^T} x^j$, the polynomial kernel $(x^{i^T} x^j + \tau)^d$ with – as kernel parameters – the intercept constant $\tau \in \mathbb{R}^+$ and degree $d \in \mathbb{N}$, and the radial basis function $\exp(-||x^i - x^j||_2^2/\sigma^2)$ with $\sigma \in \mathbb{R}^+$ representing the width of a Gaussian distribution centered on the data points. The polynomial kernel corresponds to a feature space spanned by all products of d variables at the most. This kernel results in a quadratic separating surface in the input space for $d = 2$, and it represents the cubic kernel for $d = 3$. More complex kernel functions have been proposed as well, such as graph and wavelet kernels [16,17]. In this paper, the linear kernel function is compared with a newly introduced kernel function for clinical data, referred to as the *clinical kernel function* (see Section 2.3).

A kernel algorithm for supervised classification is the LS-SVM, a simplified version of the SVM [11] and developed by Suykens et al. [18,19]. Given is a training set for classification $\{x^i, y_i\}_{i=1}^N$ of N samples with feature vectors $x^j \in \mathbb{R}^p$ and binary output labels $y_i \in \{-1, +1\}$. The aim of supervised classification is to train a function $f(x) = y$ that correctly classifies unseen samples $\{x, y\}$. Data points x^i with $f(x^i) \geq 0$ are assigned the label +1, data points with $f(x^i) < 0$ the label -1. A non-linear function of the form $f(x) = w^T \Phi(x) + b$, with w representing the normal vector on the decision hyperplane $w^T \Phi(x) + b = 0$ and variable b the bias term, can be obtained with the following constrained optimization problem for the LS-SVM:

$$\min_{w,b,e} \left(\frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} \zeta_i e_i^2 \right) \text{ subject to } y_i[w^T \Phi(x^i) + b]$$

$$= 1 - e_i \quad i = 1 \ldots N \quad \text{with } \zeta_i = \begin{cases} \dfrac{N}{2N_P} & \text{if } y_i = +1 \\ \dfrac{N}{2N_N} & \text{if } y_i = -1 \end{cases},$$

and N_P and N_N representing the number of positive and negative samples, respectively.

The regularization parameter γ represents the trade-off between maximization of the distance between samples of the two considered classes (that is, $2/||w||_2$) and minimization of the squared error contribution. Regularization by keeping γ small allows tackling the problem of overfitting by enforcing low complexity and good generalizability while tolerating misclassifications in case of overlapping distributions. Because in many two-class problems data sets are skewed in favor of one class with $N_P \gg N_N$ or $N_N \gg N_P$, we used an adapted version of the LS-SVM in which a different factor ζ_i is assigned to positive and negative samples [20]. In this way, the contribution of false negative and false positive errors to the objective function is balanced.

In dual space, the equivalent problem of this optimization problem is a system of linear equations in function of the number of samples [18,19]. All experiments and calculations in this study were therefore performed in dual space, using Matlab 7.0.0 for Windows.

2.2. Kernel-based integration of multiple data sets

The representation of any data set with a real-valued kernel matrix, independent of the nature or complexity of the data to be analyzed, makes kernel methods ideally positioned for heterogeneous data integration. In [21], Daemen and colleagues investigated whether clinical and microarray data can be efficiently combined. In most microarray studies on cancer, the focus is on the microarray analysis while clinical data are not modeled in the same manner. When integrating both heterogeneous data sources, advantage can be taken from the strength of both data sources. This approach has been improved and extended towards the inclusion of multiple high-throughput data sources [22]. Three ways to simultaneously learn from multiple data sources were discussed, differing in the stage of the model building process at which integration occurs and referred to as early, intermediate and late integration [21]. With early integration, the microarray and clinical data sets would be concatenated before model building. Due to the huge amount of genes, clinical variables would need to be very significant before being selected. The late integration approach in which the two