Review

Using support vector machines in diagnoses of urological dysfunctions

David Gil\textsuperscript{a,}\textsuperscript{*}, Magnus Johnsson\textsuperscript{b}

\textsuperscript{a}Computing Technology and Data Processing, University of Alicante, Spain
\textsuperscript{b}Lund University Cognitive Science, Sweden

A B S T R A C T

Urinary incontinence is one of the largest diseases affecting between 10% and 30% of the adult population and an increase is expected in the next decade with rising treatment costs as a consequence. There are many types of urological dysfunctions causing urinary incontinence, which makes cheap and accurate diagnosing an important issue. This paper proposes a support vector machine (SVM) based method for diagnosing urological dysfunctions. 381 registers collected from patients suffering from a variety of urological dysfunctions have been used to ensure the (generalization) performance of the decision support system. Moreover, the robustness of the proposed system is examined by fivefold cross-validation and the results show that the SVM-based method can achieve an average classification accuracy at 84.25%.

1. Introduction

Presently urinary incontinence affects between 10% and 30% of the adult population and it is expected to increase in the next decade with accelerating treatment costs as a consequence (Cortes & Kelleher, 2005; Wein, 2004). This rise in incidence is similar for the male and the female parts of the adult population (Irwin et al., 2006) (see Table 1).

The use of classifier systems in medical diagnosis is increasing gradually. There is no doubt that evaluation of data taken from patients and decisions of experts are the most important factors in diagnosis. However, expert systems and different artificial intelligence techniques for classification have the potential of being good supportive tools for the expert. Classification systems can help in increasing accuracy and reliability of diagnoses and minimizing possible errors, as well as making the diagnoses more time efficient (Akay, 2008).

Some of the related work in the field of the urological diagnosis has been developed basically by means of artificial neural networks (ANNs) (Gil, Soriano, Ruiz, & Montejo, 2007; Gil, Johnsson, Garcia Chamizo, Paya, & Fernandez, 2008). To increase the accuracy and the generalization ability we propose the use of a Support Vector Machine (SVM) based system combined with techniques for dimensionality reduction. In addition to ANNs, the SVM (Cortes & Vapnik, 1995) has also emerged as a powerful tool for classification. SVMs were proposed by Vapnik (1995) and is based on the structured risk minimization (SRM) principle. Hence it tries to minimize an upper bound of the generalization error instead of the empirical error as in the artificial neural networks. Therefore a particular advantage of SVMs over other classifiers is that they can achieve better performance when applied to real world problems (He, Hu, Harrison, Tai, & Pan, 2006). Some classifiers, such as ANNs suffer from the overfitting problem. In the case of the SVM overfitting is unlikely to occur. Overfitting is caused by too much flexibility in the decision boundary.

SVMs are global representatives of the whole set of training points, and there are usually few of them, which gives little flexibility. Thus overfitting is unlikely to occur (Witten & Frank, 2005). SVMs have been successfully applied to a wide variety of applications, e.g. including pattern recognition, biology and financial domains (Hearst, Dumais, Osman, Platt, & Scholkopf, 1998; Hua & Sun, 2001; Huang & Wu, 2006; Shin, Lee, & Kim, 2005; Wu, Huang, & Meng, 2008; Yan, Yan, Han, Wei, & Zhu, 2008).

The remaining part of the paper is organized as follows: first, we give a brief description of some basic SVM concepts. Next we describe the design of our proposal of the SVM-based decision support system with dimensionality reduction and the training of the SVM by the available data. Then we describe our testing of the system and analyze the results. Finally we draw relevant conclusions and suggest future lines of research.

2. Support vector machines

In this section, the basic concept of SVM will be briefly described. More thorough descriptions can be found in Burges (1998), Theodoridis and Koutroumbas (2003), Hsu, Chang, and Lin (2003). A typical two class problem as Fig. 1 shows is similar to the problem of diagnosing urological patients as either ill or healthy.
For a classification problem, it is necessary to first try to estimate a function $f : \mathbb{R}^d \to \{+1\}$ using training data, which are $l$ $N$-dimensional patterns $x_i$ and class labels $y_i$, where

$$(x_1, y_1), \ldots, (x_l, y_l) \in \mathbb{R}^d \times \{+1\}$$  \hspace{1cm} (1)$$

such that $f$ will classify new samples $(x, y)$ correctly.

Given this classification problem the SVM classifier, as described by Vapnik (1995), Guyon, Boser, and Vapnik (1992), Cortes and Vapnik (1995), satisfies the following conditions:

$$\begin{cases} 
\langle w', \phi(x) \rangle + b > +1 & \text{if } y_i = +1, \\
\langle w', \phi(x) \rangle + b \leq -1 & \text{if } y_i = -1,
\end{cases}$$  \hspace{1cm} (2)$$

which is equivalent to

$$y_i[\langle w', \phi(x) \rangle + b] > 1, \quad i = 1, 2, \ldots, l.$$  \hspace{1cm} (3)$$

Here training vectors $x_i$ are mapped into a higher-dimensional space by the function $\phi$. The equations of (2) construct a hyperplane $\langle w', \phi(x) \rangle + b = 0$ in this higher-dimensional space that discriminates between the two classes shown in Fig. 1. Each of the two half-spaces defined by this hyperplane corresponds to one class, $H_1$ for $y_i = +1$ and $H_2$ for $y_i = -1$. Therefore the SVM classifier corresponds to decision functions:

$$y(x) = \text{sign}(\langle w', \phi(x) \rangle + b)$$  \hspace{1cm} (4)$$

Thus the SVM finds a linear separating hyperplane with the maximal margin in this higher-dimensional space. The margin of a linear classifier is the minimal distance of any training point to the hyperplane which is the distance between the dotted lines $H_1$ and $H_2$ and the solid line showed in Fig. 1. The points $x$ which lie on the solid line satisfy $\langle w', \phi(x) \rangle + b = 0$, where $w'$ is normal to the hyperplane, $|b|/|w'|$ is the perpendicular distance from the hyperplane to the origin, and $|w'|$ is the Euclidean norm of $w$. $1/|w'|$ is the shortest distance from the separating hyperplane to the closest positive (negative) example. Therefore, the margin of a separating hyperplane will be $1/|w'| + 1/|w'|$. To calculate the optimal separating plane is equivalent to maximizing the separation margin or distance between the two dotted lines $H_1$ and $H_2$.

It has to be considered that $H_1 : \langle w', \phi(x) \rangle + b = +1$ and $H_2 : \langle w', \phi(x) \rangle + b = -1$ are parallel since they have the same normal $w'$ and perpendicular distance from the origin, $|1 - b|/|w'|$ for $H_1$ and $|1 - b|/|w'|$ for $H_2$, and that no training points fall between them. Thus we expect the solution for a typical two dimensional problem to have the form shown in Fig. 1. Those training points which gives equality in (3) are lying on one of the

![Fig. 1. The mapping between input space and feature space in a two class problem with the SVM. Mapping the training data nonlinearly into a higher-dimensional feature space via function $\phi$. $H_1$ and $H_2$ are parallel since they have the same normal $w'$ and perpendicular distance from the origin, $|1 - b|/|w'|$, and that no training points fall between them. The support vectors are the gray triangles and circles respectively located on $H_1$ and $H_2$. The distance from $w'$ to these support vectors is $1/|w'|$ and the margin is simply $2/|w'|$.](image-url)
hyperplanes $H_1$ and $H_2$ are called support vectors and they are indicated in Fig. 1 by means of a gray color.

In practice, a separating hyperplane may not exist, i.e. if a high noise level causes a large overlap of the classes. To allow for the possibility of examples violating the edges of the margin (when perfect separation is not possible) (Martin, Fowlkes, & Malik, 2003), one introduces slack variables (Bennett & Mangasarian, 1992; Cortes & Vapnik, 1995; Scholkopf, Smola, Williamson, & Bartlett, 2000; Vapnik, 1995)

$$\xi_i \geq 0, \quad i = 1, 2, \ldots, l.$$  \hspace{1cm} (5)

in order to relax the constraints to

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, l.$$ \hspace{1cm} (6)

A classifier which generalizes well is then found by controlling both the classifier capacity (via $w$) and the sum of the slacks $\sum_i \xi_i$. The latter is done as it can be shown to provide an upper bound on the number of training errors which leads to a convex optimization problem. One possible realization, called C-SVC, of a soft margin classifier is minimizing the objective function

$$\min_{w,b,\xi} J(w, b, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$ \hspace{1cm} (7)

subject to the constraints (5) and (6). The constant $C > 0$ determines the trade-off between the flatness of a function and the amount up to which deviations are tolerated from the actually obtained targets $y_i$ for all the training data. The positive real constant $C$ is a tuning parameter in the algorithm. Therefore $C \sum_{i=1}^{l} \xi_i$ minimizes the misclassification error.

Fig. 1 shows the basic idea of SVM, which is to map the data into some other dot product space (called the feature space) $F$ via a nonlinear map

$$\Phi: \mathbb{R}^n \rightarrow F$$ \hspace{1cm} (8)

and perform the above linear algorithm in $F$. As can be seen, this only requires the evaluation of dot products.

$$K(x,x') = \langle \Phi(x) \cdot \Phi(x') \rangle$$ \hspace{1cm} (9)

Clearly, if $F$ is high-dimensional, this approach is not feasible and we have to find a computationally cheaper way. The key observation then is to substitute it for a simple kernel $K$ that can be evaluated efficiently.

Kernel substitution is a method for using a linear classifier algorithm to solve a nonlinear problem by mapping the original nonlinear observations into a higher-dimensional space (as it is indicated in Fig. 1), where the linear classifier is subsequently used; this makes a linear classification in the new space equivalent to nonlinear classification in the original space.

That transformation is done using Mercer’s theorem (Williamson, Smola, & Scholkopf, 2001), which states that any continuous, symmetric, positive semi-definite kernel function $K(x, x')$ can be expressed as a dot product in a high-dimensional space.

In that point the tools to construct nonlinear classifiers are defined. We substitute $\Phi(x_i)$ for each training example $x_i$ and perform the optimal hyperplane algorithm in $F$. Because we are using kernels, we will thus end up with nonlinear decision function of the form

$$y(x) = \text{sign} \left[ \sum_{i=1}^{l} y_i K(x, x_i) + b \right].$$ \hspace{1cm} (10)

where $K(x, x')$ is called the Kernel function, whose value is equal to the inner product of two vectors $x$ and $x'$ in the feature space; namely, $K(x, x') = \phi(x') \cdot \phi(x')$. Through the Kernel function, all computations are performed in the low-dimensional input space. Any function that satisfies the Mercers condition (Courant & Hilbert, 1953) can be used as the Kernel function.

There are many possibilities to define a function to be used as a Kernel. However, typical examples of kernels used in SVM, which have been successfully applied to a wide variety of applications, are linear, polynomials, radial basic functions and hyperbolic tangent:

Linear kernel: $k(x, x') = x \cdot x'$ \hspace{1cm} (11)

Polynomial kernel: $k(x, x') = (x \cdot x' + c)^d$ \hspace{1cm} (12)

RBF kernel: $k(x, x') = \exp \left( -\frac{||x-x'||^2}{\sigma} \right)$ \hspace{1cm} (13)

Sigmoid kernel: $k(x, x') = \tanh(\gamma x \cdot x' + c)$ \hspace{1cm} (14)

In this study, radial basis functions kernel (RBF) function has been adopted because we believe that it is a suitable choice for our problem. The RBF kernel nonlinearly maps samples into a higher-dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear.

Furthermore, the linear kernel is a special case of RBF as (Keerthi & Lin, 2003) shows that the linear kernel with a penalty parameter $C$ has the same performance as the RBF kernel with some parameters ($C', \gamma$). In addition, the sigmoid kernel behaves like RBF for certain parameters (Lin & Lin, 2003). Moreover, the number of hyperparameters influences the complexity of model selection. The polynomial kernel has more hyperparameters than the RBF kernel. Finally, the RBF kernel has less numerical difficulties. One key point is $0 < K_p \leq 1$ in contrast to polynomial kernels of which kernel values may go to infinity ($\gamma x_i^2 + r < 1$) or zero ($\gamma x_i^2 + r > 1$) while the degree is large.

3. Experimentation

3.1. Urological data

The input data in the system starts when a patient reports to a physician. Then, a large number of information to be considered

<table>
<thead>
<tr>
<th>Field</th>
<th>Values normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>General information</td>
<td>Age (0, 1)</td>
</tr>
<tr>
<td>Neurological physical examination</td>
<td>Perineal and perianal sensitivity (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Anal tone (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Voluntary control of the anal sphincter (0, 1)</td>
</tr>
<tr>
<td>Free Flowsometry</td>
<td>Volume of urine (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Post void residual (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Maximum flow rate (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Average flow rate (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Micturition time (0, 1)</td>
</tr>
<tr>
<td>Cystometry</td>
<td>Bladder storage (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Detrusor pressure filling (0, 1)</td>
</tr>
<tr>
<td></td>
<td>First sensation of bladder filling (0, 1)</td>
</tr>
<tr>
<td>Test pressure/flow</td>
<td>Detrusor contraction (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Abdominal pressure (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Volume of urine in micturition (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Post void residual (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Maximum pressure detrusor (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Maximum flow rate (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Average flow rate (0, 1)</td>
</tr>
<tr>
<td></td>
<td>Micturition time (0, 1)</td>
</tr>
</tbody>
</table>
during the diagnosis will be saved in a database. In this study, an exhaustive urological exploration with 20 different measurements has been carried out by using 381 patients with dysfunctions in the lower urinary tract (LUT). The 20 input variables (Table 2) that are essential to the diagnosis of the LUT diseases of interest are extracted from the urological database. The Table 2 helps us to understand the dimension of the problem to deal with (different types of data, ranges and incomplete fields). These variables can be divided into five classes:

(a) General information about a patient, in our case only the age, normally the gender is also included but we use only female patients (the urological service is specialized in female incontinence).
(b) Neurological physical examination (three features in total).
(c) Free flowmetry analysis (five features in total).
(d) Cystometry (three features in total).
(e) Test pressure/flow (eight features in total).

Moreover, the 20 input variables (categorized into three groups or classes) are prearranged using the following structures:

(a) Numerical variables such as age, volume of urine and micturition time are normalized onto the interval (0, 1). For instance, the patients ages may span from 0 to 100 years old. However, in practice the oldest is 86, and thereby the age of a 86-year-old patient can be normalized to the value of 86/86 = 1 whereas a 63-year-old patient can be normalized to the value of 63/86 = 0.73.
(b) The variables with only two independent attributes are prearranged with binary values (0, 1). We only have the representation of anal tone where 1 means normal and 0 denotes relaxed.
(c) The variables with three independent attributes, such as perineal and perianal sensitivity, detrusor contraction or abdominal pressure are prearranged using the ternary values (1, 0, 1). For example, the abdominal pressure will take −1 for the absence, 0 represents weak and 1 represents normal.

3.2. Dimensionality reduction

Research indicates that application of data dimensionality reduction as a pre-step to the classification procedure does improve the classification accuracy (Fu & Wang, 2003; Raymer, Punch, Goodman, Kuhn, & Jain, 2000). Furthermore, feature selection can also provide a better understanding of the underlying process that generated the data (Guyon & Elisseeff, 2003). It is beneficial, and also demanded in many cases, to limit the number of input features when building classification systems in order to have a good predictive and less computationally intensive model (Zhang, 2000). In the area of medical diagnosis, a small feature subset means lower test and diagnostic costs. There are several approaches to go about the dimensionality reduction process. The wrapper and the filter approach are two of the most important. These select the most significant attributes (Li, Wu, Liu, & Tang, 2004). The wrapper approach uses the selected data mining algorithm in its search for the attribute subsets (Kohavi & John, 1997) while in the filter approach, undesirable attributes are filtered out of the data before classification begins (Duch, 2006). Feature wrappers use induction learning in order to evaluate a feature subset whereas feature filters use a heuristic in order to do the same task. Feature filters are known to be faster than feature wrappers. However, since wrappers use the performance function as the evaluation function, they are known to outperform filters in terms of accuracy. Therefore, the decision between filter or wrapper selection depends on the problem. If data analysis needs to be performed in high-dimensional space, selecting the filter approach seems to be a better alternative in order to avoid computational cost. However, if accuracy is important and computational cost is manageable, then wrappers should be utilized (Uncu & Türkşen, 2007). The wrapper approach uses a predetermined learning algorithm along with a statistical re-sampling technique (usually cross-validation) to determine a feature (variable) subset that results in the most accurate model. Wrapper methods integrate feature selection with the machine learning process. The wrapper approach includes a target classifier as a black box for performance evaluation. In other words, a computation-intensive evaluator is performed many times on candidate feature subsets to choose relevant features. When wrappers are included in the training process, they are often called as embedded methods. The wrapper model requires one predetermined learning algorithm and uses the performance of the learning algorithm to evaluate and determine which features are selected. An algorithm commonly used by the wrapper is a genetic algorithm (GA), which is employed to search optimal feature subsets under the assessment of the classifiers predictive accuracy (Kim, 2004; Min, Lee, & Han, 2006).

Thanks to this integration with the training process wrapper is a beneficial approach considering that we immediately see the result of the reduction of the feature space being translated into the corresponding classification with our subset of features selection. Figs. 2 and 3 illustrate both methods. Recently, some researchers proposed a hybrid approach for feature selection (Das, 2001; Xing, Jordan, & Karp, 2001). The basic idea is to select a small number of variable subsets based on variable correlation and then determine the best subset by explicitly evaluating the loss function.

In this work the wrapper approach has been selected to decide the importance degree of every input data. In particular, an SVM classifier which evaluates the worth of every attribute. Then, the results in Table 3 indicates this order of importance of each of them.

3.3. Architecture of the SVM

Each selected attribute of the urological data studied in the previous section constituted the input nodes of the SVM-based structure shown in Fig. 4, which shows the architecture of SVM.

The method to evaluate the efficiency of the SVM system is to obtain some measures as classification accuracy, sensitivity, specificity, positive predictive value, negative predictive value and confusion matrix. A confusion matrix (Kohavi & Provost, 1998) contains information about actual and predicted classifications done by a classification system. Table 4 shows the confusion matrix for a two class classifier. Classification accuracy, sensitivity,
can be defined by using the elements of the confusion matrix as specificity, positive predictive value and negative predictive value. Table 4 defines the confusion matrix with the value for every measure.

### Table 4
Definition of confusion matrix with the value for every measure.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True positive (TP) = 302</td>
<td>False positive (FP) = 37</td>
<td>False negative (FN) = 23</td>
</tr>
<tr>
<td>Negative</td>
<td>False positive (FP) = 37</td>
<td>True negative (TN) = 19</td>
<td></td>
</tr>
</tbody>
</table>

![Image: Architecture of the SVM](image)

In this paper we have evaluated the performance of a classifier constructed by means of the SVM method when applied to the diagnosis of urological dysfunctions. The SVM were trained with data from a database with registers of patients with urological dysfunctions. The experiment starts with a preprocessing of the uro-dynamical measures from every patient. This preprocessing includes missing data treatment and normalization process. After that, data are provided to the SVM which determines whether there is a dysfunction of the LUT or not. Data preprocessing is the stage in which the most time of the experimentation is spent. However, it is worth it because the accuracy of the system will increase notably approximately of 15%.

In this work we obtained valuable comments from the urologists after using the system. They remarked its advantages to give a more precise diagnosis and, therefore, to save time and money to health sector institutions. They also pointed out that the goal of the system is to imitate the human expert in the sense of being able to generalize by using the experience. The system can also help them to detect similarities between fields or urodynamical samples.

The accuracy of the SVM has been tested and shows a degree of certainty of 84.25% which is slightly higher than other supervised methods as the multilayer perceptron (82.9%) and quite higher than the unsupervised ones (around 75%).

The next step is to distinguish and classify the most common dysfunctions.

In order to develop a system which is able to classify different types of dysfunctions, an improved class distribution is needed. In other words, all sets of data of every class should have (approximately) the same amount of number in order to obtain a high accuracy in the diagnosis. In the experiment section, the problem regarding the skewed output class distribution was described (the number of healthy patients was considerably higher than the one of ill patients). When this distribution is not proportional, the system tends to give a higher weight to the majority class.

In this work we obtained valuable comments from the urologists after using the system. They remarked its advantages to give a more precise diagnosis and, therefore, to save time and money to health sector institutions. They also pointed out that the goal of the system is to imitate the human expert in the sense of being able to generalize by using the experience. The system can also help them to detect similarities between fields or urodynamical samples.

The accuracy of the SVM has been tested and show a degree of certainty of 84.25% which is slightly higher than other supervised methods as the multilayer perceptron (82.9%) and quite higher than the unsupervised ones (around 75%).

The next step is to distinguish and classify the most common dysfunctions.

In order to develop a system which is able to classify different types of dysfunctions, an improved class distribution is needed. In other words, all sets of data of every class should have (approximately) the same amount of number in order to obtain a high accuracy in the diagnosis. In the experiment section, the problem regarding the skewed output class distribution was described (the number of healthy patients was considerably higher than the one of ill patients). When this distribution is not proportional, the system tends to give a higher weight to the majority class.
Acknowledgments

We want to express our acknowledgments to the urologists of the Hospital of San Juan (Alicante-Spain), who have made it possible to reach a better understanding of the different types of urological dysfunctions. Moreover, the data used in the development of this system is the result of several years of this collaboration.

References


Lin, H. T., & Lin, C. J. (2003). A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Taipei: Department of Computer Science and Information Engineering, National Taiwan University.


