Signal Discrimination Using a Support Vector Machine for Genetic Syndrome Diagnosis

Amit David & Boaz Lerner*
Pattern Analysis and Machine Learning Lab
Department of Electrical & Computer Engineering
Ben-Gurion University, Beer-Sheva, Israel
davidami@ee.bgu.ac.il; boaz@ee.bgu.ac.il

Abstract

In this study, a support vector machine (SVM) classifies real world data of cytogenetic signals measured from fluorescence in-situ hybridization (FISH) images in order to diagnose genetic syndromes. The study implements the SVM structural risk minimization concept in searching for the optimal setting of the classifier kernel and parameters. We propose thresholding the distance of tested patterns from the SVM separating hyperplane as a way of rejecting a percentage of the miss-classified patterns thereby allowing reduction of the expected risk. Results show accurate performance of the SVM in classifying FISH signals in comparison to other state-of-the-art machine learning classifiers, indicating the potential of an SVM-based genetic diagnosis system.

1. FISH image analysis and signal representation

In recent years, fluorescence *in-situ* hybridization (FISH) has emerged as one of the most significant new developments in the analysis of human chromosomes. FISH offers numerous advantages compared with conventional cytogenetic techniques since it allows detection of numerical chromosome abnormalities during normal cell interphase. An important application of FISH is dot counting, i.e. the enumeration of signals (dots) within the nuclei, as the dots in the image represent the inspected chromosomes. Dot counting is used for diagnosing numerical chromosomal aberrations in, e.g., haematopoietic neoplasia, solid tumors and prenatal diagnosis [1]. However, a major limitation of the FISH technique for dot counting is the need to examine large numbers of cells. This is required for an accurate estimation of the distribution of chromosomes over cell population, especially in applications involving a relatively low frequency of abnormal cells. As visual evaluation by a trained cytogeneticist of large numbers of cells and enumeration of hybridization signals is expensive and time consuming, FISH analysis for dot counting can be expedited by automating the procedure.

A neural network (NN) has recently been proposed [2] discriminating between real signals and artifacts resulting from noise and out-of-focus images, thus enabling FISH dot counting. Aiming at clinical diagnosis we are interested here in improving the NN classification accuracy by studying a support vector machine (SVM) classifying signals of two genetic syndromes. The SVM classifier, having superior generalization capability and reputation of a highly accurate paradigm, is described in Section 2. The SVM-based experimental framework and its results in classifying FISH signals are respectively given in Sections 3 and 4, before concluding in Section 5.

2. Support vector classifiers

2.1. Structural risk minimization

In statistical learning theory [3], we bound the difference between the expected risk, $R(\alpha)$ (mean error rate measured on the test set) and the empirical risk, $R_{emp}(\alpha)$ (mean error rate measured on the training set) when both sets are assumed to be generated from the same underlying probability distribution $P(\mathbf{x},y)$. The empirical risk is calculated by $R_{emp}(\boldsymbol{\alpha}) = \frac{1}{2l}\sum_{i=1}^{l}\left|y_i - f(\mathbf{x}_i,\boldsymbol{\alpha})\right|$ where l is the size of the training set, $\boldsymbol{\alpha}$ are the model parameters and $f(\mathbf{x}_i,\boldsymbol{\alpha})$ is the classifier output for a training vector $\mathbf{x}_i$ having a corresponding label $y_i \in \{-1,1\}$. The expected risk for an unseen test vector $\mathbf{x}$ is $R(\boldsymbol{\alpha}) = \int \frac{1}{2}\left|y - f(\mathbf{x},\boldsymbol{\alpha})\right| dP(\mathbf{x},y)$. Vapnik [3] showed that with probability $1-\eta$ for some $0\le\eta\le1$, a bound on the expected risk could be calculated as $R(\boldsymbol{\alpha}) \le R_{emp}(\boldsymbol{\alpha}) + \sqrt{\left[h\left(\log\left(2l/h\right)+1\right)-\log\left(\eta/4\right)\right]l^{-1}}$ trading between tight and reliable bounds for η close to 1 and 0, respectively. The parameter h is a non-negative integer called the Vapnik Chervonenkis (VC) dimension defined

* Corresponding author

as the largest number of patterns that can be separated (shattered) in all possible ways using the family of functions $f(\alpha)$ implemented by the classifier. Thus, the VC dimension is a measure of the capacity (complexity) of the class of functions, in relation to the available amount of data. Bounding the risk manifests a well known pattern recognition trade off where a highly complex hypothesis is able to discriminate between much more complex data structures, but at the cost of increasing the difference between the empirical and the expected risk, i.e., the VC confidence, in a phenomenon usually referred to as "over fitting". To find an optimal hypothesis reducing the VC confidence, the concept of structural risk minimization is introduced. It establishes a nested set of hypothesis spaces in which each nested space has a smaller VC dimension than that of its outer space. The concept builds upon finding the hypothesis space that minimizes the bound on the risk. This is accomplished by training a machine for each space to minimize the empirical risk. The machine having the minimal sum of empirical risk and VC confidence is then being selected.

2.2. Maximal margin classifier

The basic principle of SVM is the determination and selection of the optimal hyperplane (hypothesis) yielding the maximum margin of separation between the classes [3]. Figure 1 shows separation of a two-dimensional two-class separable classification example by optimal, insuring a maximum margin, and non-optimal hyperplanes.

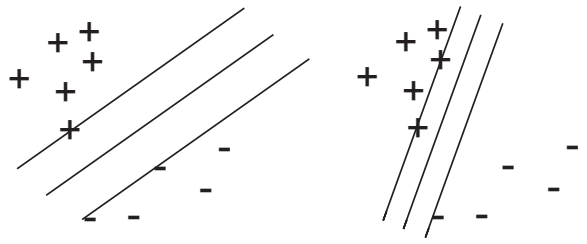

Figure 1. Two separating hyperplanes for a separable case – an optimal (left) and non-optimal (right)

A large margin leads to an error bound depending on an "effective" VC dimension smaller than the VC dimension, thus providing a classifier having higher generalization capability [4]. Training an SVM is a quadratic optimization problem, in which $L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$ is maximized w.r.t $\alpha_i \geq 0,\ \sum_{i=1}^{l} \alpha_i y_i = 0,\ i = 1, l$. Each α_i is a Lagrange multiplayer that when kept non-negative ensures correct classification of $\mathbf{x}_i$. We derive the maximal margin classifier $f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i \cdot (\mathbf{x} \cdot \mathbf{x}_i) + b \right)$, where b is computed from the Karush-Kuhn-Tucker conditions $\alpha_i \cdot \left[y_i ((\mathbf{x}_i \cdot \mathbf{w}) + b) - 1 \right] = 0,\ i = 1, l$, for any i for which $\alpha_i \neq 0$ or averaged over all of these points, and $\mathbf{w}$ is normal to the hyperplane. Because the problem is quadratic, the optimization process always results in a global maximum. During training the SVM finds those vectors defining the maximal margin, thus called support vectors. These support vectors are the only vectors necessary for shaping the decision boundary and they cannot be overlooked during classifier training. Correspondingly, their Lagrange multipliers are the only coefficients which do not vanish during training.

For the non-separable case no particular separating hyperplane can be found as some of the α_i's get increasingly large values. Two main solutions exist. First, we penalize errors of the classifier by a user-defined parameter C, thereby bounding $0 \leq \alpha_i \leq C$. Second, we construct a non-linear SVM projecting the data onto a higher (even infinite) dimensional space in which the data can be linearly separated by a maximal margin classifier. We may implement the projection using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ leading to the same training procedure and solution as for the linear case independently of feature space dimension. The kernel replaces the dot product between patterns modifying L_D, as well as the decision boundary to $f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i \cdot K(\mathbf{x}, \mathbf{x}_i) + b \right)$. This is a nonlinear optimal separating hypothesis.

2.3. Multi class SVM

The SVM is a binary classifier which can be extended by fusing several of its kind into a multi class classifier. In this study, we fuse classifier decisions using the error correction output codes (ECOC) approach [5]. In the ECOC approach several binary classifiers are trained (the exact number is a function of the number of classes), each of them aimed at separating a different combination of classes. Comparing classifier results to an optimal output code representing each of the classes makes the final decision. The correct class is the closest, in the Hamming distance sense, to that achieved by the classifiers.

2.4. Rejection

Since most of the classification errors occur near the separating hypothesis where classes overlap, rejecting data vectors located at this region can reduce the overall

risk. However, part of the rejected vectors may have been classified correctly, thus the rejection reduces both the error rate and correct classification rate. To implement classification, the SVM computes the distance of each data vector from the separating hypothesis, thus rejection can be performed by thresholding this distance.

3. The empirical study

3.1. Methodology

We have studied and applied an SVM in classifying real and artifact FISH signals of two genetic syndromes. Over 3,000 signals were segmented from 400 FISH images and represented by twelve features of size, shape, intensity and color [2]. Structural risk minimization for the SVM was explored empirically by employing different families of functions, characterized by their kernels, and different settings of classifier parameters. The settings of parameters were evaluated by training and validating SVMs having different settings on a partition of the data including 70% of the signal patterns using a CV5 experiment. The setting corresponding to the SVM achieving the minimum risk on the validation set on average was employed to train an SVM on that partition and then to test it on the other partition of the data containing the remaining 30% of the patterns. Measured on the test set, the SVM expected risk was compared to that of other machine learning state-of-the-art classifiers.

3.2. SVM Kernels

In exploring the non-linear projection best suitable for the FISH data, we have employed three different kernels. The linear kernel composes of a dot product between data vectors $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$ leading to a separating hypothesis which is a hyperplane. The only parameter that should be set for this kernel is the cost parameter *C*, and values checked are in the range [1,10000]. A more complicated kernel is a polynomial of degree *D* in the data $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^D$ leading to a polynomial of degree *D* separating hypothesis. *D* is determined independently of the dimensionality of the data. Settings of the kernel are selected from the range [0.5,10] and [1,10000] for *D* and *C*, respectively. The third kernel used is a symmetric hyper Gaussian, called sometimes a radial basis function, $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right\}$. The separating hypothesis is a linear combination of symmetric hyper Gaussians which their widths are controlled by changing the value of the parameter σ, in our case in the range [0.5,1000]. *C* is set as previously.

3.3. Optimization tools

Training an SVM is mainly solving a quadratic optimization problem. For large databases it is computationally demanding, causing slow convergence of the optimization process. Approaches to reduce the computational demands divide the problem into smaller problems, solving each one separately and combining the results into a complete solution. We examined here two of these approaches. The first approach was a MATLAB wrapper [6] to SVM*light* [7] and the second a routine [6] utilizing the MATLAB optimization toolbox. The first approach was found superior both in the classification error it achieved and in the speed of convergence.

4. Results

Binary classification We first tested the SVM in classifying the FISH data as real signals or artifacts. For each of the evaluated kernels, we optimized the SVM settings as described before. Table 1 depicts the SVM risk for the optimal settings for each of the kernels.

Table 1. The SVM risk for FISH signal classification using different kernels and corresponding optimal settings

Kernel	Kernel parameter	*C*	Error Rate (%)
Linear	-	10	15.45
Polynomial	*D*=3	7	13.14
Gaussian	σ=5	100	12.82

Four-Class classification Extending previous experiment, we employed a 7-bit ECOC algorithm for the SVM to discriminate signals as real or artifact of two genetic syndromes indicated in the images by red and green signals. The four-class SVM used the settings of the optimal binary classifier, i.e., Gaussian kernel and its parameters achieving a risk of 15.36%. Table 2 gives the confusion matrix of the resulted SVM.

Table 2. A confusion matrix for the FISH four-class classification problem

SVM results / True Class	Real Red	Artifact Red	Real Green	Artifact Green
Real Red	88.82	11.18	0.00	0.00
Artifact Red	13.20	84.52	0.00	2.28
Real Green	1.89	0.00	81.13	16.98
Artifact Green	0.42	4.18	12.55	82.85

Rejection By raising the threshold on the distance between an erroneous classification and the corresponding separating hyperplane we increased the number of rejected patterns leading to a reduced risk though also reduced accuracy. Figure 2 demonstrates this trade-off for binary classification and threshold changes in the [0,1] range where most errors occur.

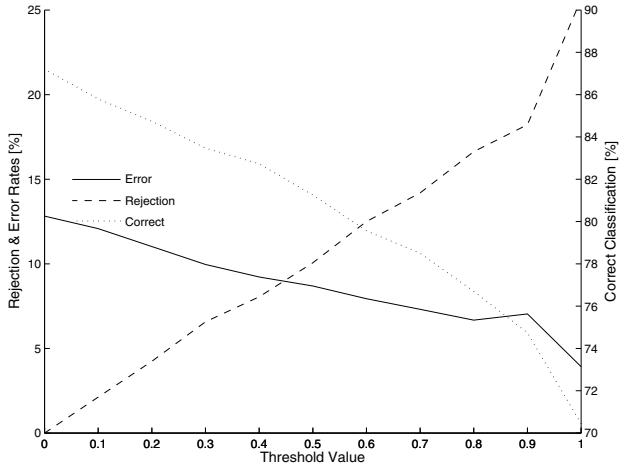

Figure 2. Rejection implemented in the SVM

Comparison to other machine learning classifiers A comparison of the optimal SVM to other state-of-the-art classifiers for the FISH two-class classification problem is provided in Table 3. The SVM is the most accurate classifier except of the Bayesian neural network.

Table 3. A comparison of the optimal SVM to other machine learning classifiers optimized to the domain

Classifier	Error Rate (%)
Binary Gaussian SVM	12.82
7-Nearest-Neighbor	13.24
Neural Network [8]	13.6
Bayesian Neural Network [8]	11.8
Naïve Bayesian Classifier [8]	17.0
Linear Classifier [8]	15.9

5. Discussion

We have suggested an SVM for the discrimination of FISH signals as real or artifacts. The Gaussian kernel optimized to the domain yielded an SVM having a more accurate decision boundary than those associated with the linear and polynomial kernels. Thresholding the distance of classification errors from the hyperplanes established using the support vectors provided a mechanism of trading off between making erroneous decisions and rejecting a fraction of the data, accompanied by the opportunity to further increase precision hierarchically. This mechanism is rarely employed in the SVM literature.

Accurately discriminating FISH signals, the SVM classifier is a major contribution to genetic syndrome diagnosis. The classifier outperforms most of other machine learning classifiers and is inferior only to the Bayesian neural network, the latter having pronounced computational requirements. As the superiority of the Bayesian neural network is attributed to the exploitation of a priori knowledge additionally to that acquired from the data, we are interested in ways to combine these information sources within the SVM framework in order to improve the classifier performance.

Acknowledgment

This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben Gurion University, Beer-Sheva, Israel.

6. References

[1] H.J. Tanke, R.J. Florijn, J. Wiegant, A.K. Raap, and J. Vrolijk. “CCD microscopy and image analysis of cells and chromosomes stained by fluorescence in situ hybridization”, *Histochemical Journal*, 27, 4-14, 1995

[2] B. Lerner, W.F. Clocksin, S. Dhanjal, M.A. Hult'en and C. M. Bishop, “Feature representation and signal classification in fluorescence in-situ hybridization image analysis”, *IEEE Trans. on Systems, Man and Cybernetics A,* 31, 655-665, 2001.

[3] Vapnik, V. *The nature of statistical learning theory*, Springer-Verlag, New York, 1995.

[4] Cristianini, N. and J. Shawe-Taylor, *An introduction to support vector machines*, Cambridge Press, 2000.

[5] T.G. Dietterich and G. Bakiri, “Solving multiclass learning problems via error-correcting output codes”, *Journal of Artificial Intelligence Research*, 2, 263-286, 1995

[6] A. Schwaighofer, “SVM toolbox for Matlab”, 2002, http://www.cis.tugraz.at/igi/aschwaig/software.html

[7] T. Joachims, “SVM-Light support vector machine”, 1999, http://svmlight.joachims.org/

[8] B. Lerner and N.D. Lawrence, “A comparison of state-of-the-art classification techniques with application to cytogenetics”, *Neural Computing & Applications*, 10, 39-47, 2001.