16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

BREAST CANCER DIAGNOSIS FROM FINE-NEEDLE ASPIRATION USING SUPERVISED COMPACT HYPERSPHERES AND ESTABLISHMENT OF CONFIDENCE OF MALIGNANCY

Tingting Mu, Asoke K. Nandi

Department of Electrical Engineering and Electronics, The University of Liverpool
Brownlow Hill, L69 3GJ, Liverpool, UK
phone: + (44) 151 794 4525, fax: + (44) 151 794 4540, email: {tingting.mu,a.nandi}@liv.ac.uk
web: www.liv.ac.uk

ABSTRACT

An automatic classification methodology is implemented to analyze breast masses from fine-needle aspiration, including feature selection, diagnostic decision, and computation of confidence of malignancy for each mass. Feature selection is performed using a genetic algorithm based on a measure of alignment. A kernel-based classifier is employed for diagnostic decision, which learns two supervised compact hyperspheres (SCHs), each encompassing the most training patterns from one class, while also the least training patterns from the other class. Instead of only providing a binary diagnostic decision of "malignant" or "benign", we assign a confidence of malignancy to each mass for the first time, by calculating probabilities of being benign and malignant. Compared with several well-known classifiers, the SCH-based classifier provides the highest training accuracy of 100% and test accuracy of 99%.

1. INTRODUCTION

Breast cancer is the most common form of cancer and the second most common cause of cancer deaths among women in the world; the disease affects approximately 10% of all women at some stage of their life in the Western world [1]. Breast cancer may be detected via a careful study of clinical history, physical examination, and imaging with either mammography or ultrasound. However, definitive diagnosis of a breast mass can only be established through fine-needle aspiration (FNA) biopsy, core needle biopsy, or excisional biopsy. Among these methods, FNA is the easiest and fastest method of obtaining a breast biopsy, and is effective for women who have fluid-filled cysts. Research works on the Wisconsin Diagnosis Breast Cancer (WDBC) data grew out of the desire to diagnose breast masses accurately based solely on FNA [2, 3, 4]. To improve the accuracy and efficiency of the detection of breast cancer, a number of research projects are focusing on developing methods for computer-aided diagnosis (CAD) of breast cancer from FNA, including works on image analysis and computational intelligence [2, 3, 5, 6, 7].

The minimum enclosing ball (MEB) problem was first proposed by Sylvester [8] in computational geometry, which seeks the smallest ball that contains all the samples in a training set, and can be solved by several traditional algorithms [9, 10], and the core vector machine [11] with low computational complexity. The MEB problem can be easily shown to be equivalent to the hard-margin support vector data description (SVDD) [12], which seeks a hypersphere containing all the training samples from the target class for one-class classification. Tax and Duin [12, 13] also proposed the soft-margin SVDD by involving the 1-norm of slack variables to allow the possibility of outliers in the training set. In this work, we extend the soft-margin SVDD to breast cancer diagnosis by involving the 2-norm of slack variables and employing the negative samples outside the target class. Two smallest hyperspheres are learned, named as supervised compact hyperspheres (SCHs), each encompassing the maximum possible number of training patterns from one class and the least possible number of training patterns from the other class. Kernel functions are employed to incorporate the nonlinearity.

As an additional consideration beyond pattern classification, given the question "Are we equally confident about the classification of all masses?", the answer is, generally, negative. Different input samples may have different levels of the output value, although they have the same predicted label. In such a situation, we aim to explore methods to assign a measure of the confidence of malignancy to an individual mass based on its corresponding output value, instead of only providing a binary diagnostic decision of "malignant" or "benign".

Built on the success of the previous works by Wolberg et al. [2, 3], the following methodology is employed to diagnose breast masses based solely on FNA: (1) A genetic algorithm (GA) is employed to reduce the dimensionality of features, based on alignment of the inner-product matrix with the target function [14]. (2) The SCH-based classifier is applied to detect malignant tumors. (3) Probabilities of being benign and malignant are computed for each breast mass based on the output values obtained by different classifiers.

2. SCH-BASED CLASSIFICATION METHOD

A set of l labeled training samples $z = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l} \in (R^n \times Y)$ is considered, where R^n is the n-dimensional real feature space with a binary label space $Y = \{1, -1\}$, and $y_i \in Y$ is the label assigned to the sample (feature vector) $\boldsymbol{x}_i \in R^n$. The proposed method is performed in a nonlinear feature space, $\boldsymbol{\phi} : \boldsymbol{x} \in R^n \mapsto \boldsymbol{\phi}(\boldsymbol{x}) \in \kappa \subseteq R^n$, defined by a kernel, which is a function $K(\cdot,\cdot)$ that satisfies $K(\boldsymbol{x}_a, \boldsymbol{x}_b) = \langle \boldsymbol{\phi}(\boldsymbol{x}_a), \boldsymbol{\phi}(\boldsymbol{x}_b) \rangle$ for all samples $\boldsymbol{x}_a, \boldsymbol{x}_b \in R^n$, where $\langle \cdot,\cdot \rangle$ denotes the inner product function. In the kernel-defined feature space κ, the distance between any two feature vectors $\boldsymbol{\phi}(\boldsymbol{x}_a), \boldsymbol{\phi}(\boldsymbol{x}_b) \in \kappa$ is computed as

$$d(\boldsymbol{x}_a, \boldsymbol{x}_b) = K(\boldsymbol{x}_a, \boldsymbol{x}_a) - 2K(\boldsymbol{x}_a, \boldsymbol{x}_b) + K(\boldsymbol{x}_b, \boldsymbol{x}_b). \quad (1)$$

2.1 Supervised Compact Hyperspheres

By employing the information from both the feature vectors and the labels of the training samples, we propose the concept of "supervised compact hypersphere" for binary classification: Positive SCH is the smallest hypersphere that contains the maximum possible number of positive training patterns and the minimum number of negative training patterns. Negative SCH is the smallest hypersphere that contains the maximum possible number of negative training patterns and the minimum number of positive training patterns. The following computations are derived to determine the two SCHs.

2.1.1 Positive SCH

The positive SCH can be obtained by solving the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{c}_+, r_+} \quad & r_+^2 + \frac{1}{2}C_1^+ \sum_{i=1}^{l_+} \xi_{1i}^2 + \frac{1}{2}C_2^+ \sum_{i=1}^{l_-} \xi_{2i}^2, \qquad (2) \\ s.t. \quad & \|\boldsymbol{\phi}(\boldsymbol{x}_i^+) - \boldsymbol{c}_+\|^2 \le r_+^2 + \xi_{1i},\ i = 1, 2, \ldots, l_+, \\ & \|\boldsymbol{\phi}(\boldsymbol{x}_i^-) - \boldsymbol{c}_+\|^2 \ge r_+^2 - \xi_{2i},\ i = 1, 2, \ldots, l_-, \end{aligned}$$

where $\boldsymbol{x}_i^+$ and $\boldsymbol{x}_i^-$ denote the positive and negative training samples, respectively; l_+ and l_- denote the numbers of the positive and negative training samples, respectively; r_+ and $\boldsymbol{c}_+$ denote the radius and center of the positive SCH in the kernel-defined feature space κ, respectively; ξ_{1i} and ξ_{2i} are slack variables to reduce the solution's sensitivity to outliers in the training samples; and C_1^+ and C_2^+ are the regularization parameters controlling the trade-off between minimizing the radius and controlling the slack variables. The following Lagrangian is maximized instead:

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}) \quad = \quad & -\sum_{i=1}^{l_+}\sum_{j=1}^{l_+} \alpha_i \alpha_j K(\boldsymbol{x}_i^+, \boldsymbol{x}_j^+) \qquad (3) \\ & -\sum_{i=1}^{l_-}\sum_{j=1}^{l_-} \beta_i \beta_j K(\boldsymbol{x}_i^-, \boldsymbol{x}_j^-) - \frac{1}{2C_1^+}\sum_{i=1}^{l_+} \alpha_i^2 \\ & +2\sum_{i=1}^{l_+}\sum_{j=1}^{l_-} \alpha_i \beta_j K(\boldsymbol{x}_i^+, \boldsymbol{x}_j^-) - \frac{1}{2C_2^+}\sum_{i=1}^{l_-} \beta_i^2 \\ & +\sum_{i=1}^{l_+} \alpha_i K(\boldsymbol{x}_i^+, \boldsymbol{x}_i^+) - \sum_{i=1}^{l_-} \beta_i K(\boldsymbol{x}_i^-, \boldsymbol{x}_i^-), \end{aligned}$$

subject to

$$\sum_{i=1}^{l_+} \alpha_i - \sum_{i=1}^{l_-} \beta_i = 1.$$

where $\alpha_i \ge 0$ and $\beta_i \ge 0$ are Lagrange multipliers. This quadratic programming (QP) problem can be solved by modifying the sequential minimal optimization (SMO) algorithm [15]. Letting $\left[\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*\right]^T$ denote the optimal solution of Eq.(3), the optimal center of the positive SCH is derived by

$$\boldsymbol{c}_+^* = \sum_{i=1}^{l_+} \alpha_i^* \boldsymbol{\phi}(\boldsymbol{x}_i^+) - \sum_{i=1}^{l_-} \beta_i^* \boldsymbol{\phi}(\boldsymbol{x}_i^-), \qquad (4)$$

2.1.2 Negative SCH

The negative SCH can be obtained by solving the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{c}_-, r_-} \quad & r_-^2 + \frac{1}{2}C_1^- \sum_{i=1}^{l_+} \xi_{1i}^2 + \frac{1}{2}C_2^- \sum_{i=1}^{l_-} \xi_{2i}^2, \qquad (5) \\ s.t. \quad & \|\boldsymbol{\phi}(\boldsymbol{x}_i^+) - \boldsymbol{c}_-\|^2 \ge r_-^2 - \xi_{1i},\ i = 1, 2, \ldots, l_+, \\ & \|\boldsymbol{\phi}(\boldsymbol{x}_i^-) - \boldsymbol{c}_-\|^2 \le r_-^2 + \xi_{2i},\ i = 1, 2, \ldots, l_-, \end{aligned}$$

where r_- denotes the radius of the negative SCH, $\boldsymbol{c}_-$ denotes the center of the negative SCH in the transformed feature space κ, and C_1^- and C_2^- are the regularization parameters controlling the trade-off between minimizing the radius and controlling the slack variables. The optimal center of the negative SCH, denoted as $\boldsymbol{c}_-^*$, can be derived by employing the same computation as used to obtain $\boldsymbol{c}_+^*$ in Section 2.1.1.

2.2 Classification

The distance between an input sample $\boldsymbol{x}$ and each optimal center $\boldsymbol{c}_+^*$ and $\boldsymbol{c}_-^*$ in the kernel-defined feature space κ, denoted as $d_+(\boldsymbol{x})$ and $d_-(\boldsymbol{x})$, respectively, can be simply calculated based on Eq. (1). The label of the input sample can be predicted by seeking a separating function $f(\boldsymbol{x})$ using LDA [16] in the distance space with $d_+(\boldsymbol{x})$, $d_-(\boldsymbol{x})$, and $\frac{d_+(\boldsymbol{x})}{d_-(\boldsymbol{x})}$ as the new input features.

3. CONFIDENCE OF MALIGNANCY

A classifier predicts the label for an input sample by applying the step function with a threshold on its corresponding output value. Different input samples may have different levels of the output value, even though they have the same predicted label. We propose an approach to compute the confidence of malignancy for each mass:

1. Sort the benign masses in the descending order based on their corresponding output values; and divide the benign masses with negative output values into N partitions $\{B_i\}_{i=1}^N$ along the sorted order, each with the same number of benign masses N_{B}.
2. Sort the malignant tumors in the ascending order based on their corresponding output values; and divide the malignant tumors with positive output values into N partitions $\{M_i\}_{i=1}^N$ along the sorted order, each with the same number of malignant tumors N_{M}.
3. Calculate the numbers of the misclassified malignant tumors with the corresponding output values falling into each benign partition B_i, denoted as $E_{\mathrm{B}i}$.
4. Calculate the numbers of the misclassified benign masses with the corresponding output values falling into each malignant partition M_i, denoted as $E_{\mathrm{M}i}$.
5. In each benign partition B_i, the probability of being benign is established by $P_{\mathrm{B}i}^{(B)} = \frac{N_{\mathrm{B}}}{N_{\mathrm{B}}+E_{\mathrm{B}i}}$, and the probability of being malignant by $P_{\mathrm{B}i}^{(M)} = \frac{E_{\mathrm{B}i}}{N_{\mathrm{B}}+E_{\mathrm{B}i}}$.
6. In each malignant partition M_i, the probability of being benign is established by $P_{\mathrm{M}i}^{(B)} = \frac{E_{\mathrm{M}i}}{N_{\mathrm{M}}+E_{\mathrm{M}i}}$, and the probability of being malignant by $P_{\mathrm{M}i}^{(M)} = \frac{N_{\mathrm{M}}}{N_{\mathrm{M}}+E_{\mathrm{M}i}}$.
7. For a given test mass $\boldsymbol{x}$, one needs to decide which partition $\boldsymbol{x}$ belongs to based on its corresponding output value,

and assign the probabilities of being benign and malignant in that partition to $\boldsymbol{x}$, denoted as $P_B(\boldsymbol{x})$ and $P_M(\boldsymbol{x})$, respectively.

To calculate the probabilities of being benign and malignant based on a set of classifiers, the average is employed. Assuming N classifiers are employed, and letting $P_B^{(i)}(\boldsymbol{x})$ and $P_M^{(i)}(\boldsymbol{x})$ denote the probabilities obtained by the ith classifier, the combined probabilities are calculated by

$$P_B(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} P_B^{(i)}(\boldsymbol{x}), \tag{6}$$

$$P_M(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} P_M^{(i)}(\boldsymbol{x}), \tag{7}$$

The confidence of malignancy for an input mass is defined as its corresponding probability of being malignant.

4. FEATURE PREPARATION

The WDBC dataset [17] consists of 569 breast masses with 357 benign and 212 malignant cases. In order to evaluate the size, shape, and texture of each cell nuclei, ten characteristics were derived and described as follows [2, 3]:

- **Radius** is computed by averaging the length of radial line segments from the center of mass of the boundary to each of the boundary points.
- **Perimeter** is measured as the sum of the distances between consecutive boundary points.
- **Area** is measured by counting the number of pixels on the interior of the boundary and adding one-half of the pixels on the perimeter, to correct for the error caused by digitization.
- **Compactness** combines the perimeter and area to give a measure of the compactness of the cell, calculated as $\frac{\text{perimeter}^2}{\text{area}}$.
- **Smoothness** is quantified by measuring the difference between the length of each radial line and the mean length of the two radial lines surrounding it, calculated by
$$\frac{\sum_{\text{points}} |r_i - (r_i + r_{i+1})/2|}{\text{perimeter}},$$
where r_i is the length of the line from the center of mass of the boundary to each boundary point.
- **Concavity** is captured by measuring the size of any indentations in the boundary of the cell nucleus.
- **Concave points** is similar to concavity, but counts only the number of boundary points lying on the concave regions of the boundary, rather than the magnitude of such concavities.
- **Symmetry** is measured by finding the relative difference in length between pairs of line segments perpendicular to the major axis of the contour of the cell nucleus, calculated by
$$\text{symmetry} = \frac{\sum_i |\text{left}_i - \text{right}_i|}{\sum_i (\text{left}_i + \text{right}_i)},$$
where left_i and right_i denote the lengths of perpendicular segments on the left and right of the major axis, respectively.
- **Fractal dimension** is approximated using the "coastline approximation" described by Mandelbrot [18]. The perimeter of the nucleus is measured using increasingly larger "rulers". As the ruler size increases, the precision of the measurement decreases, and the observed perimeter decreases. Plotting these values on a log-log scale and measuring the downward slope gives the negative of an approximation to the fractal dimension.
- **Texture** is measured by finding the variance of the gray-scale intensities in the component pixels.

The mean value, standard error, and the extreme (largest or "worst") value of each characteristic were computed for each image, which resulted in 30 features of 569 images, yielding a database of 569×30 samples. A total of 250 samples were randomly selected for training (150 benign masses and 100 malignant tumors), and the remaining 319 samples for test.

5. EXPERIMENTS, RESULTS, AND COMPARATIVE ANALYSIS

5.1 Feature Selection

Feature selection was performed using GA, with the objective function set as the alignment between the inner product matrix of features in the original feature space and the target label matrix, computed with the training samples, given as [14]

$$A_{\mathbf{E}} = \frac{\langle \mathbf{E}, \boldsymbol{y}\boldsymbol{y}^T \rangle_F}{\sqrt{\langle \mathbf{E}, \mathbf{E} \rangle_F \langle \boldsymbol{y}\boldsymbol{y}^T, \boldsymbol{y}\boldsymbol{y}^T \rangle_F}}, \tag{8}$$

where $\mathbf{E}$ denotes the original inner product matrix, and $\boldsymbol{y}$ denotes the column vector of the labels of the training samples. The Frobenius product $\langle \cdot, \cdot \rangle_F$ between two Gram matrices $\mathbf{M}$ and $\mathbf{N}$ is defined as

$$\langle \mathbf{M}, \mathbf{N} \rangle_F = \text{tr}(\mathbf{MN}), \tag{9}$$

where "tr" denotes the trace of a matrix. A MATLAB toolbox [19] was employed to implement GA by encoding the possible feature combination into a 30-bit binary string. The fitness function was calculated by linear ranking. The population size was set as 50. The generation gap was set as 0.6. Two-point crossover was employed, with the corresponding crossover rate set as 0.8. Bit flip was used for mutation, with the corresponding mutation rate set as 0.05. The maximum number of generations was set as 300. A subset of nine features was selected as the optimal combination. The dimensionality of features was reduced from 30 to 9.

5.2 Classification and Diagnostic Decision

The SCH-based classifier was compared with linear discriminant analysis (LDA), support vector machines (SVMs), and kernel Fisher's discriminant analysis (KFDA), using the selected features. Radial basis function (RBF) kernel was employed for the SCH-based classifier, SVMs, and KFDA. Parameters of each classifier was decided by cross validation within the training samples. The corresponding classification performance in accuracy is recorded in Table 1, as well as the computing time in seconds. The proposed SCH-based classifier provided the highest classification accuracies (100% for training and 98.8% for test), but the slowest computing speed, as there are two QP problems to be solved. However, the computing time of the SCH-based classifier were still comparable to that of SVMs and KFDA. The distribution

Classifiers	LDA	KFDA	SVM	SCH
Training	95.6%	97.6%	99.2%	**100**%
Test	95.6%	96.2%	97.5%	**98.8**%
Time (s)	0.013	0.303	0.302	0.536

Table 1: Performance comparison in classification accuracy and computing time for different classifiers.

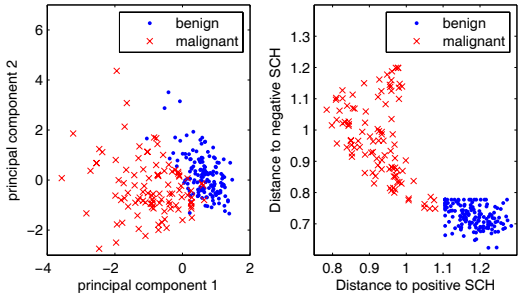

Figure 1: Distribution of the first two principal components of the training samples, and the distribution of the distances between the training samples and the centers of the positive and negative SCHs.

of the first two principal components of the original features was compared with the distribution of the distances between samples and the centers of the two SCHs, in Figure 1 and Figure 2. By learning the positive and negative SCHs, better dispersion was reached in the SCH-defined distance space, for both the training and test samples.

5.3 Confidence Assignment

We calculated the combined confidence of malignancy for the test masses with the selected features, using LDA, KFDA, SVM, and SCH-based classifiers. The training samples were divided into four partitions, with two related to benign (B_1 and B_2) and two to malignant (M_1 and M_2). The separate probabilities of being benign and malignant in each partition, computed with the training samples, are recorded in Table 2. The SCH-based classifier provided 100% confidence for each mass. By applying a threshold of 50% on values of the confidence of malignancy for the test masses, only one malignant tumor was misclassified, leading to an improved test accuracy of 99.5%, with sensitivity of 99.1% and specificity of 0.0%. The result is better than the accuracy of 95.6% obtained by the edited nearest-neighbor with pure filtering [20] and the best accuracy of 98.9% obtained by genetic programming [5], with the same dataset. The value of the confidence of malignancy was 30% for the misclassified malignant tumor.

6. DISCUSSION AND CONCLUSION

We have investigated and presented results of an automatic classification methodology to analyze breast masses from FNA. The SCH-based classifier determined a positive SCH and a negative SCH to cover each class region through solving two QP problems. Such a method outperformed LDA, KFDA, and SVM, with the training accuracy of 100% and the test accuracy of 98.8%, using the GA-selected features. For the first time, the confidence of malignancy was com-

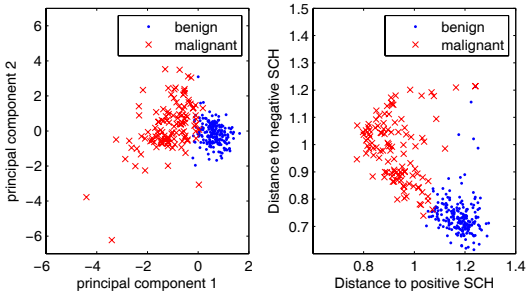

Figure 2: Distribution of the first two principal components of the test samples, and the distribution of the distances between the test samples and the centers of the positive and negative SCHs.

		B_2	B_1	M_1	M_2
LDA	P_B (%)	100	89.2	4.2	0.0
	P_M (%)	0.0	10.8	95.8	100
SVM	P_B (%)	100	97.4	0.0	0.0
	P_M (%)	0.0	2.6	100	100
KFDA	P_B (%)	100	93.7	2.0	0.0
	P_M (%)	0.0	6.3	98.0	100
SCH	P_B (%)	100	100	0.0	0.0
	P_M (%)	0.0	0.0	100	100

Table 2: Values of probability of being benign (P_B) and probability of being malignant (P_M), calculated with the training samples, in the four partitions B_1, B_2, M_1, and M_2.

puted for each mass based on the corresponding output values obtained by multiple classifiers, e.g. LDA, KFDA, SVM, and SCH-based classifiers. By applying a threshold of 50% on these confidences, only one malignant tumor was misclassified, leading to an improved test accuracy of 99.5%. It is believed that the methods we have proposed for computing the confidence of malignancy for a given mass address the important need in the assignment of a degree of confidence in the CAD labels or marks placed on an image being analyzed, and thereby make an important contribution to CAD of breast cancer.

REFERENCES

[1] L. Duijm, J. H. Groenewoud, F. H. Jansen, J. Fracheboud, M. Beek, and H. J. de Koning. Mammography screening in the Netherlands: delay in the diagnosis of breast cancer after breast cancer screening. *British Journal of Cancer*, 91:1795–1799, 2004.

[2] W. H. Wolberg, W. N. Street, and O. L. Mangasarian. Breast cytology diagnosis via digital image analysis. *Analytical and Quantitative Cytology and Histology*, 15(6):396–404, 1993.

[3] W. H. Wolberg, W. N. Street, and O. L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. *Cancer Letter*, 77:163–171, 1994.

[4] W. H. Wolberg, W. N. Street, and O. L. Mangasarian. Image analysis and machine learning applied to breast

cancer diagnosis and prognosis. *Analytical and Quantitative Cytology and Histology*, 17(2):77–87, 1995.

[5] H. Guo and A. K. Nandi. Breast cancer diagnosis using genetic programming generated feature. *Pattern Recognition*, 39:980–987, 2006.

[6] I. Anagnostopoulos and I. Maglogiannis. Neural network-based diagnostic and prognostic estimations in breast cancer microscopic instances. *Medical and Biological Engineering and Computing*, 44(9):773–784, 2006.

[7] T. Mu and A. K. Nandi. Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM–RBF classifier. *Journal of the Franklin Institute*, 344(3-4):285–311, 2007.

[8] J. J. Sylvester. A question in the geometry of situation. *Quarterly Journal on Mathematics*, pages 1–79, 1857.

[9] N. Megiddo. Linear-time algorithms for linear programming in R^3 and related problems. *SIAM Journal on Computing*, 12:759–776, 1983.

[10] E. Welzl. Smallest enclosing disks (balls and ellipsoids). In H. Maurer, editor, *New Results and New Trends in Computer Science*, pages 359–370. Springer-Verlag, 1991.

[11] I. W. Tsang, J. T. Kwok, and P-M Cheung. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

[12] D. M. J. Tax and R. P. W. Duin. Support vector domain descriptionn. *Pattern Recognition Letters*, 20(14):1191–1199, 1999.

[13] D. M. J. Tax and R. P. W. Duin. Uniform object generation for optimizing one-class classifiers. *Journal for Machine Learning Research*, 2:155–173, 2001.

[14] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor. On optimizing kernel alignment. Technical Report NC-TR-01-087, Royal Holloway Univeristy of London, London, UK, 2001.

[15] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 185–208. MIT Press, 1998.

[16] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, New York, NY, 2nd edition, 2001.

[17] University of Wisconsin-Madison. *Machine Learning for Cancer Diagnosis and Prognosis*. http://pages.cs.wisc.edu/ olvi/uwmp/cancer.html.

[18] B. B. Mandelbrot. *The Fractal Geometry of Nature, Chapter 5*. W. H. Freeman and Company, New York, 1997.

[19] A. J. Chipperfield, P. J. Fleming, H. Pohlheim, and C. M. Fonseca. *Genetic Algorithm Toolbox for use with MATLAB (version 1.2)*. University of Sheffield, Sheffield, UK, 1994.

[20] W. Lam, C. Keung, and C. X. Ling. Learning good prototypes for classification using filtering and abstraction of instances. *Pattern Recognition*, 35(7):1491–1506, 2002.