Pattern Recognition Letters 30 (2009) 1489–1495

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Mutual information-based SVM-RFE for diagnostic classification of digitized mammograms

Sejong Yoon, Saejoon Kim *

Department of Computer Science and Engineering, Sogang University, 1 Shinsu-dong, Mapo-gu, Seoul 121-742, Republic of Korea

ARTICLE INFO

Article history:
Received 11 September 2008
Received in revised form 15 June 2009
Available online 7 July 2009

Communicated by Y. Ma

Keywords:
Digital mammography
CADx
Feature selection
SVM-RFE
Mutual information
Correlation

ABSTRACT

Computer aided diagnosis (CADx) systems for digitized mammograms solve the problem of classification between benign and malignant tissues while studies have shown that using only a subset of features generated from the mammograms can yield higher classification accuracy. To this end, we propose a mutual information-based Support Vector Machine Recursive Feature Elimination (SVM-RFE) as the classification method with feature selection in this paper. We have conducted extensive experiments on publicly available mammographic data and the obtained results indicate that the proposed method outperforms other SVM and SVM-RFE-based methods.

1. Introduction

According to the American Cancer Society (2008), breast cancer is the second largest cause of cancer deaths and the most frequently diagnosed cancer in women. The report estimates 26% of all new cancer cases and 15% of all cancer-related deaths are caused by breast cancer in 2008. While the exact cause of breast cancer is yet unknown, when detected early, breast cancer is a curable disease. For these reasons, early detection of breast cancer is becoming increasingly important for its timely treatment and eventual cure. The currently known most popular method for early detection of breast cancer is the digital mammography which has merits such as wide availability, relatively low cost, and non-invasiveness compared to other technologies such as ultrasound and magnetic resonance imaging (Elmore et al., 2005). With digital mammography, doctors can recognize non-palpable breast lesions at low costs and thus it aides doctors' ability to detect breast cancer.

On the other hand, abnormal lesions found are not always easily distinguished into benign or malignant ones, especially in the early stages of breast cancer. One solution that has been shown to be effective for this problem is the *computer-aided diagnosis* (CADx). CADx functions as a second opinion to help doctors make more accurate final decisions by generating computational assessments of mammograms and classifying benign and malignant findings in the films. Common classification methods used in CADx include many classic machine learning algorithms such as linear discriminant analysis (Lo et al., 2003), Bayesian networks (Fischer and Lo, 2004), artificial neural networks (Lo et al., 2002; Wei et al., 2005; Panchal and Verma, 2006) and support vector machines (SVMs) (El-Naqa et al., 2002; Land et al., 2003; Wei et al., 2005). It has been reported that 36.9% of unnecessary biopsies can be avoided through the use of CADx while maintaining the same level of sensitivity in the detection of malignancy findings (Isaac and Lederman, 2006). In addition, by improving CADx systems, radiologists using the systems are expected to reduce their variability in interpretations of mammograms (Jiang et al., 2001).

While CADx shows promising results for classifying an abnormality into benign or malignant, the diagnostic accuracy is not very high as it stands. This is partly because, like other classification problems, it faces the challenge of feature selection since the number of features that can be extracted from mammograms is theoretically infinite while using a feature selection scheme tailored for the specific classification algorithm is preferable (Lo et al., 2006). To this end, we propose a new feature selection-based support vector machines (SVM) for this problem in this paper. Our scheme is inspired by a mutual information-based feature selection method which minimizes redundancy among features and maximizes relevance to classes (mRMR) (Ding and Peng, 2005). We chose SVM-Recursive Feature Elimination (SVM-RFE) (Guyon et al., 2002) as the base algorithm and have tested our scheme on the dataset of mass and calcification lesions found in the Digital Database of Screening Mammography (DDSM) (Heath and Bowyer,

* Corresponding author. Tel.: +82 2 705 8931; fax: +82 2 704 8273.
E-mail addresses: sjyoon@sogang.ac.kr (S. Yoon), saejoon@sogang.ac.kr (S. Kim).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.06.012

2001). The result showed that our scheme outperforms, or at least is competitive to other classification methods.

The rest of this paper is organized as follows. In Section 2, we review SVM and SVM-RFE-based feature selection methods. In Section 3, we describe our method and in Section 4, we present the experiment framework and the results of CADx using data obtained from digitized mammograms of DDSM. Finally, we present conclusions and future works in Section 5.

2. Previous works

Consider a dataset of N examples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ each having P features represented as $\{f_1, f_2, \ldots, f_P\}$. The feature value of kth feature, $1 \leqslant k \leqslant P$, from the ith example, $1 \leqslant i \leqslant N$, is denoted by $x_{i,k}$. Let y_i be the class label of the ith example where $y_i \in \{+1, -1\}$. In this paper, we consider binary classification since we are interested in classifying benign and malignant examples.

2.1. SVM

SVM is a classifier based on structural risk minimization principle. It searches for the hyperplane that maximizes the distance from the hyperplane to the nearest examples in each class. An attractive feature of SVM is that it can map linearly inseparable data into higher dimensional space where they can be linearly separated. SVM tries to find the decision hyperplane which can be written as $\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b = 0$ where $\mathbf{w}$ and b are classification model parameters and Φ is a mapping to a certain higher dimensional space in which $\mathbf{x}_i$ can be linearly separated. Then we can formalize the training task for the model as an optimization task $\min_{\mathbf{w}}(\|\mathbf{w}\|^2/2)$ subject to $y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) \geqslant 1$. Since the task is a convex optimization problem, we can rewrite the optimization formula to a Lagrangian function $L(\mathbf{w}, b, \boldsymbol{\lambda})$ and derive its dual form $\widetilde{L}(\boldsymbol{\lambda})$ as

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = (\|\mathbf{w}\|^2/2) - \sum_{i=1}^{N} \lambda_i (y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) - 1)$$

$$\widetilde{L}(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \lambda_i - (1/2) \sum_{i,j}^{N} \lambda_i \lambda_j y_i y_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

subject to the Karush–Kuhn–Tucker conditions

$$\lambda_i \geqslant 0, \quad \lambda_i \{y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) - 1\} = 0$$

where λ_i are Lagrangian multipliers. The multipliers can be calculated by exploiting quadratic programming techniques or faster heuristic algorithms. After they are calculated, we can determine model parameters $\mathbf{w}$ and b by using the fact that $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) = \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$ where $\mathbf{K}(\cdot, \cdot)$ is a kernel function. With all the multipliers and model parameters determined, we can classify a newly input test example $\mathbf{x}_{new}$ by investigating which side of the hyperplane it resides. In summary, we can write this non-linear SVM classifier's overall decision function h as

$$h(\mathbf{x}_{new}) = sign\left(\sum_{i=1}^{N} \lambda_i y_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}_{new}) + b\right)$$

for a predefined kernel function $\mathbf{K}$ where $sign(\cdot)$ is the sign function. In this paper, we will consider linear and Gaussian radial basis function (RBF) kernels.

2.2. SVM-RFE

SVM-RFE is a wrapper feature selection method which generates the ranking of features using backward feature elimination. Based on the Optimal Brain Damage (OBD) algorithm (LeCun et al., 1990), the ranking criterion of SVM-RFE is defined using the difference of the objective function for each removal of a feature, while the function to be minimized can be expressed as

$$J = (1/2) \sum_{i,j}^{N} \lambda_i \lambda_j \mathbf{H}(i,j) - \sum_{i=1}^{N} \lambda_i$$

where $\mathbf{H}(i,j) = y_i y_j \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$. Taking the difference of the function for each removal of the kth feature while leaving Lagrangian multipliers unchanged, we can compare the contribution provided by each feature to minimization of the objective function. Ranking score for the kth feature can be computed as

$$DJ(k) = (1/2)\left(\sum_{i,j}^{N} \lambda_i \lambda_j \mathbf{H}(i,j) - \sum_{i,j}^{N} \lambda_i \lambda_j \mathbf{H}(i,j,-k)\right) \quad (1)$$

where $\mathbf{H}(i,j,-k)$ is $\mathbf{H}(i,j)$ without the kth feature. The overall recursive procedure of SVM-RFE is described in Fig. 1.

Note that computation time needed by the algorithm can be reduced if one removes multiple features in each iteration. However, this may cause some information loss because eliminating a feature may change the importance of other features. Therefore, we removed one feature at a time in this paper.

While SVM-RFE usually finds a feature subset that yields good classification performance, the feature subset found may not be the best possible combination since SVM-RFE is a greedy method that can only hope to find the best possible combination for classification. In particular, SVM-RFE does not necessarily make an effort to minimize redundancy and maximize relevance of features for every feature subset in the feature list R.

2.3. Minimum redundancy maximum relevance (mRMR)

Ding and Peng (2005) introduced a criteria to measure relevance and redundancy of features by using mutual information called minimum redundancy maximum relevance (mRMR). To compute mutual information for discrete data, we first consider the number of possible values N_{f_k} and N_{f_l} each feature f_k and f_l can take, respectively, where $k, l \in \{1, \ldots, P\}$. Let $f_{k,q}$ and $f_{l,r}$ be the qth and rth possible value of feature f_k and f_l, respectively, where $q \in \{1, \ldots, N_{f_k}\}$ and $r \in \{1, \ldots, N_{f_l}\}$. Then we can define the mutual information between the two features f_k and f_l, $I(f_k, f_l)$, by using the joint probability distribution $p(f_{k,q}, f_{l,r})$ and individual marginal distributions $p(f_{k,q})$ and $p(f_{l,r})$ as

$$I(f_k, f_l) = \sum_{q=1}^{N_{f_k}} \sum_{r=1}^{N_{f_l}} p(f_{k,q}, f_{l,r}) \log\left(\frac{p(f_{k,q}, f_{l,r})}{p(f_{k,q}) p(f_{l,r})}\right).$$

Since we cannot easily calculate the probabilities for each feature value if the data is in the continuous domain, mRMR uses F-statistics for relevance and Pearson correlation coefficient for redundancy. The F-test of a feature f_k across all c_m classes, $m = \pm 1$, where $c_1 = +1$ and $c_2 = -1$ can be formalized as

Require: Feature lists $R = [\,]$ and $S = [f_1, ..., f_P]$
1: **while** $S \neq [\,]$ **do**
2: Train a SVM with features in S
3: **for all** k-th feature f_k in S **do**
4: Compute $DJ(k)$ using Eq. (1)
5: **end for**
6: $e = \arg\min_k(DJ(k))$
7: $R = [f_e, R]$
8: $S = S$ - $[f_e]$
9: **end while**
10: **return** R

Fig. 1. Algorithm SVM-RFE.

$$F(f_k) = \left(\sum_{m=1}^{2} N_m(\mu_m - \mu)^2\right) \Big/ \sigma^2 \quad (2)$$

where μ is the mean of feature f_k over all examples, μ_m is the mean of feature f_k within the class c_m and $\sigma^2 = \left(\sum_{m=1}^{2}(N_m - 1)\sigma_m^2\right)/(N-2)$ where N_m is the number of examples of the class c_m and σ_m is variance of the class c_m. If we assume the optimal subset of features of $S = \{f_1, f_2, \ldots, f_P\}$ is R, the maximum relevance of the whole dataset can be written as $\max_R(Relevance_R)$ where

$$Relevance_R = (1/|R|)\sum_{f_k \in R} F(f_k).$$

For the Pearson correlation coefficient for redundancy, we just compute the correlation coefficient of all feature pairs. If we define $Corr(f_k, f_l)$ as the correlation coefficient value between two features f_k and f_l, the minimum redundancy of the whole dataset can be written as $\min_R(Redundancy_R)$ where

$$Redundancy_R = (1/|R|^2)\sum_{f_k f_l \in R} |Corr(f_k, f_l)|.$$

To get a single objective function to optimize both criteria, Ding and Peng (2005) suggested two kinds of combinations. One of them is difference based criterion (DBC) which can be computed by subtracting $Redundancy_R$ from $Relevance_R$, and the other is quotient based criterion (QBC) which is defined as the ratio of $Relevance_R$ to $Redundancy_R$, i.e.,

$$DBC = Relevance_R - Redundancy_R$$
$$QBC = Relevance_R / Redundancy_R$$

Then we only have to find a feature subset R that maximizes DBC or QBC. Since QBC showed better performance than DBC in the result of mRMR (Ding and Peng, 2005), we used QBC criterion in this paper.

2.4. SVM-RFE with mRMR

Recently, Mundra and Rajapakse (2007) suggested a modified SVM-RFE using mRMR criteria embedded into the algorithm. They slightly changed the ranking criterion so that SVM-RFE can find a subset of features with minimum redundancy and maximal relevance to classes. From here on, we will refer this SVM-RFE with mRMR criteria as "SVM-RFE (mRMR)." In addition to the ranking criterion defined by SVM-RFE, SVM-RFE (mRMR) calculates $mRMR(k)$, the QBC version of mRMR criterion for a single kth feature defined as

$$mRMR(k) = F(f_k)/(1/|S|)\sum_{f_l \in S} |Corr(f_k, f_l)| \quad (3)$$

where S is the set of features that are remained in each iteration. Then it normalizes both SVM-RFE criterion and mRMR criterion of each feature using maximum value of each criterion defined as

$$DJ^* = \max_k DJ(k) \quad (4)$$

$$mRMR^* = \max_k mRMR(k) \quad (5)$$

The overall algorithm of SVM-RFE (mRMR) is described in Fig. 2.

While SVM-RFE (mRMR) empirically showed slight improvement compared to SVM-RFE and SVM using mRMR as the feature selection method, it does not consider the characteristics of SVM-RFE as a wrapper method which utilizes the objective function of its base classification algorithm, namely, SVM. SVM-RFE (mRMR) assumes that the relevance of each feature to classes and the redundancy between features are of the same importance regardless of the base classifiers. However, if the base classifier only concentrates on one aspect more than the other or just ignores it, simple wrapper method is obviously not a preferred application. We present two intuitive explanation why SVM falls under this type of classifier.

Require: Feature lists $R = [\,]$ and $S = [f_1, \ldots, f_P]$
while $S \neq [\,]$ **do**
  Train a SVM with features in S
  for all k-th feature f_k in S **do**
    Compute $DJ(k)$ using Eq. (1)
    Compute $mRMR(k)$ using Eq. (3)
  end for
  Find DJ^* and $mRMR^*$ using Eq. (4) and Eq. (5) respectively
  for all k-th feature f_k in S **do**
    Compute $Score(k) = DJ(k)/DJ^* + mRMR(k)/mRMR^*$
  end for
  $e = \arg\min_k (Score(k))$
  $R = [f_e, R]$
  $S = S$ - $[f_e]$
end while
return R

Fig. 2. Algorithm SVM-RFE (mRMR).

First, since the base classifier SVM is a maximal margin classifier that maximizes margin between classes and the hyperplane, the optimization process of SVM itself can be interpreted as a process of maximizing examples' relevance to classes. This means that the elements of model parameter **w** are optimized to features' relevance to classes, not redundancy between features.

Second, SVM does not always penalize redundant features as can be seen in the following example. Let's reconsider the formalized optimization task of SVM, specifically, $\min_{\mathbf{w}}(\|\mathbf{w}\|^2/2)$. For non-separable cases, SVM can employ slack variables ξ_i to penalize non-separable examples on the margin to trade off between training error and generalization error as $\min_{\mathbf{w},\Psi}(\|\mathbf{w}\|^2/2 + \Psi\sum_{i=1}^{n}\xi_i)$. However, since SVM inherently uses only a subset of examples, i.e., support vectors, these penalization factors only affect the examples on the margin. This means when determining each feature's importance by calculating the parameter **w**, SVM may ignore feature redundancy information that might have hidden in non-support vectors.

Li and Yang (2005) have partially explained this fact by using a modified logistic regression approximating SVM (Zhang et al., 2003) and their conclusion on the weakness of SVM in redundant feature elimination is identical to the intuitive explanation stated above: SVM penalizes redundant features of examples those are exactly on the margin only and ignores the others.

Thus we can infer that what SVM lacks of has more to do with feature redundancy and not with relevance to the classes. Therefore, the relevance and redundancy criteria of mRMR should be treated separately if we want to embed them into SVM-RFE. As a consequence, to obtain the subset of features that yield the highest classification accuracy with SVM-RFE and mRMR, we propose an algorithm that searches for features in two directions: a backward search that recursively eliminates least relevant features, and a forward search that finds a combination of features with minimum redundancy.

3. SVM-RFE with correlation

To achieve the goal just mentioned in the previous section, we first modified the algorithm of SVM-RFE (mRMR) and designed an iterative algorithm that separates the relevance and redundancy criteria. A sketch of the modified algorithm consists of the next five steps. First, compute the relevance measure (F-test) and the original SVM-RFE criterion ($DJ(k)$) for every kth feature. Second, normalize each by using the maximum values $DJ^*, F(S)^*$ defined as

$$DJ^* = \max_k DJ(k) \tag{6}$$

$$F(S)^* = \max_{f_k} F(f_k), \quad f_k \in S \tag{7}$$

Third, sort every feature in the current working feature set S by using the sum of the normalized values. Fourth, from the top-ranked feature, iteratively re-sort the other lower ranked features in S in ascending order, according to the absolute Pearson correlation coefficient between the top-ranked feature and each lower ranked one. Fifth, eliminate the bottom-ranked feature from S and put it into the head of ranked feature list R, and then continue iteration like in SVM-RFE. With this algorithm, we can obtain most non-redundant subset of features with respect to the top feature with maximum relevance to classes.

Though the above algorithm is quite straightforward, it is still possible that in some cases, chosen features may not yield the best accuracy in SVM. For example, consider four features, say, $f_1, f_2, f_3, f_4 \in S$ that are ranked in this order as a consequence of the third step of the above algorithm. Assume that f_2 is more correlated to f_3 than f_4 is. The above algorithm will first pick feature f_1 and sort the other as $[f_1, f_4, f_3, f_2]$. In the next step, the algorithm will pick feature f_4 and compare $Corr(f_4, f_2)$ and $Corr(f_4, f_3)$. Since $Corr(f_4, f_2) < Corr(f_4, f_3)$, the rank will be sorted as $[f_1, f_4, f_2, f_3]$. However considering f_1 and f_4 have already been chosen in the subset, feature f_3 is more preferable than f_2 in the sense of minimizing redundancy.

To resolve this issue, we introduce a notion of *average* to the second step of the algorithm. Specifically, instead of considering only the top feature in the working set, we will average correlation coefficients of features in the already chosen feature subset so that the newly added feature is not redundant to any of the features in the subset. From here on, we will refer to this SVM-RFE with averaged correlation as "SVM-RFE (Corr)". The overall modified algorithm is described in Fig. 3 where f'_l and f'_m are the lth and the mth feature, respectively, in the sorted feature list S and $ACorr(f'_m)$ is the average correlation coefficient value for feature f'_m. Note that we used notation $F(k)$ instead of $F(f_k)$ for better readability of the algorithm.

The asymptotic time complexity analysis explains practical merit of the method. Lines 1 through 8 can be computed in exactly the same time as that of SVM-RFE or SVM-RFE (mRMR) which depends on the sizes of both P and N. Additional computation, i.e., lines 9 through 16, may take additional P^2 times but note that the calculation of the F-test and the Pearson correlation coefficient, which only requires simple comparisons, can be cached in advance.

4. Results

4.1. Datasets

In this study, we used data from the Digital Database of Screening Mammography (DDSM) which is the largest publicly available database of mammographic data. DDSM contains more than 2500 cases of mammograms in total that are obtained between 1988 and 1999 from medical institutions in the US. We collected both mass and calcification data from Massachusetts General Hospital (MGH), Washington University in St. Louis (WU) and Wake Forest University School of Medicine (WFUSM) to evaluate the classification accuracies. Each case of mammogram contains one or more abnormality findings each of which is either malignant or benign. Mammography images in the database are digitized images from film-screening mammograms using three different types of digitizers: Howtek 960 of MGH, Howtek MultiRad850 of WU and Lumisys 200 Laser of WFUSM. Table 1 summarizes the number of abnormality findings from each institution.

Table 1
Dataset information.

Institution	Mass		Calcification	
	Benign	Malignant	Benign	Malignant
MGH	482	365	381	323
WU	154	115	41	98
WFUSM	163	255	188	159
Total	799	735	610	580

Require: Feature lists $R = [\,]$ and $S = [f_1, ..., f_P]$
1: **while** $S \neq [\,]$ **do**
2: Train a SVM with features in S
3: **for all** k-th feature f_k in S **do**
4: Compute $DJ(k)$ using Eq. (1)
5: Compute $F(k)$ using Eq. (2)
6: **end for**
7: Find DJ^* and $F(S)^*$ using Eq. (6) and Eq. (7) respectively
8: Sort S by $DJ(k)/DJ^* + F(k)/F(S)^*$ in descending order
9: **for** $l = 1$ to $(|S| - 1)$ **do**
10: Set S' as a list of features in S that were already considered in this loop: $S' = [f'_1, f'_2, ..., f'_l]$, $f'_k \in S$, $k \in \{1, ..., l\}$
11: **for** $m = l + 1$ to $|S|$ **do**
12: $ACorr(f'_m) = (1/|S'|) \sum_{k=1}^{l} Corr(f'_k, f'_m)$
13: **end for**
14: Sort $f_m \in S$ by $ACorr(f'_m)$ in ascending order where $l+1 \leq m \leq |S|$
15: $l = l + 1$
16: **end for**
17: $S = S - [f_e]$, f_e is the last element in S
18: $R = [f_e, R]$
19: **end while**
20: **return** R

Fig. 3. Algorithm SVM-RFE (Corr).

Table 2
BI-RADS mammographic features.

Feature type	Description or numeric value
Mass shape	no mass(0), round(1), oval(2), lobulated(3), irregular(4)
Mass margin	no mass(0), well circumscribed(1), microlobulated(2), obscured(3), ill-defined(4), spiculated(5)
Calcification Type	no calc.(0), milk of calcium-like(1), eggshell(2), skin(3), vascular(4), spherical(5), suture(6), coarse(7), large rod-like (8), round (9), dystrophic (10), punctate(11), indistinct(12), pleomorphic(13), fine branching(14)
Calcification distribution	no calc.(0), diffuse(1), regional(2), segmental(3), linear(4), clustered(5)
Density	1, 2, 3, 4
Assessment	1, 2, 3, 4, 5

Breast Imaging Reporting and Data System (BI-RADS) of the American College of Radiology is a widely used standard for reporting mammograms by trained radiologists (Balleyguier et al., 2007). It is reported to be effective in predicting malignancy (Orel et al., 1999) and for this reason, BI-RADS descriptors shown in Table 2 along with subtlety value and patient age were used as the set of features describing each abnormality finding in this study. This set of eight features all can be extracted from the annotation files of the mammographic images in DDSM. Four of the six BI-RADS descriptors, namely, mass shape, mass margin, calcification type and calcification distribution, were encoded into numerical values, shown inside the parentheses in the table, using a rank ordering system (Lo et al., 2003) in order to be used in a CADx system. In addition to these eight features, we used 14 statistical features utilized in a study which have shown their effectiveness for CADx systems (Panchal and Verma, 2006). These statistical features use gray level values of abnormal lesion findings in mammographic images and their exact formulae are summarized in the reference (Verma and Zhang, 2007). After extracting and calculating feature values, we normalized these features because raw values of some statistical features are enormous compared to those of the BI-RADS-based features and thus inhibit SVM from learning the whole dataset effectively. For all datasets, we performed five-fold cross-validation and computed the averaged area under ROC curves, A_z, generated by the classification algorithms described in Sections 2 and 3.

Table 3
Comparison of kernels in terms of maximum Az value.

Institution	Lesion type	Linear	RBF
MGH	Mass	**0.90055**	0.88805
	Calcification	0.72712	**0.77497**
WU	Mass	0.91115	**0.93642**
	Calcification	**0.95215**	0.91710
WUFSM	Mass	0.88476	**0.92474**
	Calcification	0.79821	**0.89738**

4.2. Empirical results

First, we compared several kernels' performance to choose which kernel is most suitable for the DDSM dataset. We applied simple SVM classifier to all datasets we have prepared and found optimal parameters for the kernels. The optimization was done by exhaustive search using Leave-One-Out cross-validation error on each dataset. The result is summarized in Table 3. While we actually compared three kinds of kernels, namely, linear, polynomial and RBF, we only present linear and RBF results here because polynomial kernels were extremely slow to train, showed poor performance, and there are few studies reasoning polynomial kernel to be effective for mammogram classifications. Note that in four out of six datasets, RBF kernels outperformed linear kernels significantly while average difference is almost twice if we compare the difference between datasets that RBF kernel is better and datasets that linear kernel is better. Moreover, the training time took much shorter with RBF kernel than the linear counterpart. For this reason, we consider RBF kernels only hereafter.

The overall classification result of digitized mammograms from the DDSM is summarized in Table 4. It is observed that SVM-RFE (Corr) outperforms all other SVM-RFE-based methods in most cases. SVM-RFE (Corr) also showed better performance than SVM (mRMR) and the SVM classifier using mRMR criterion as the filter method of Ding and Peng (2005). Figs. 4–6 illustrate the overall performance of the algorithms for mass lesions with respect to different number of features used. They also show the convergence speed of the classification algorithms where the term convergence here refers to the number of features needed to reach within 5% of the maximum value of A_z. For example, we interpret the convergence speed of one method to be faster than the other when the number of features needed to be within 5% of the maximum value of A_z is smaller than the other. In terms of the convergence speed, SVM-RFE (mRMR) dominated others in many cases.

Figs. 7–9 illustrate the similar performance curves for calcification lesions in which SVM-RFE (Corr) clearly outperforms SVM-RFE

Table 4
Comparison of methods by Az value and number of features used.

Institution	MGH		WU		WUFSM	
Lesion Type	Mass	Calcification	Mass	Calcification	Mass	Calcification
SVM	0.88805 22	0.77497 22	0.93642 22	0.91710 22	0.92474 22	0.89738 22
SVM (mRMR)	0.91369 11	0.78024 19	0.93642 22	**0.97432** 17	0.92474 22	0.89738 22
SVM-RFE	0.88849 11	0.77497 22	0.94173 20	0.93436 19	0.93037 17	0.89859 21
SVM-RFE (mRMR)	0.91639 7	0.77497 22	0.94356 17	0.95023 11	**0.93764** 16	0.90364 15
SVM-RFE (**Coir**)	**0.91958** 14	**0.80051** 15	**0.94389** 14	0.97114 13	0.92958 16	**0.90402** 18

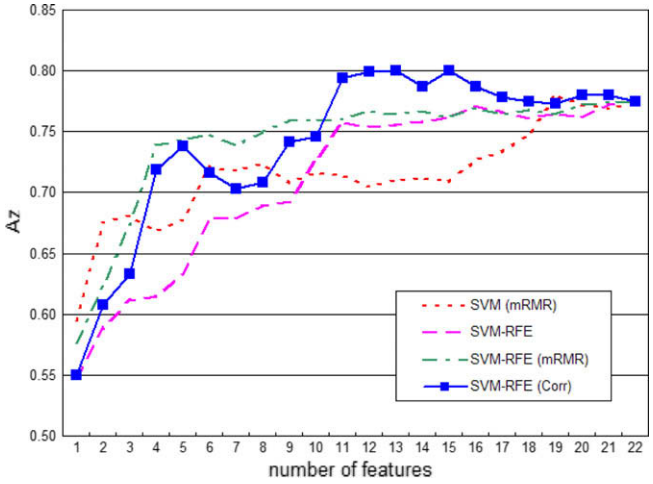

Fig. 4. Az with different number of features for mass of MGH.

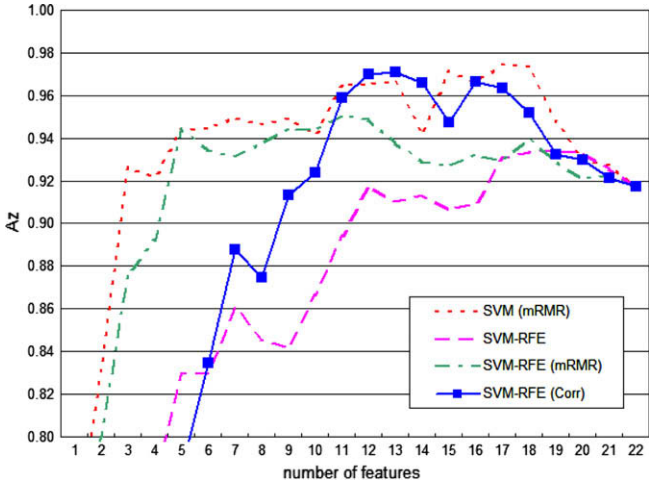

Fig. 5. Az with different number of features for mass of WU.

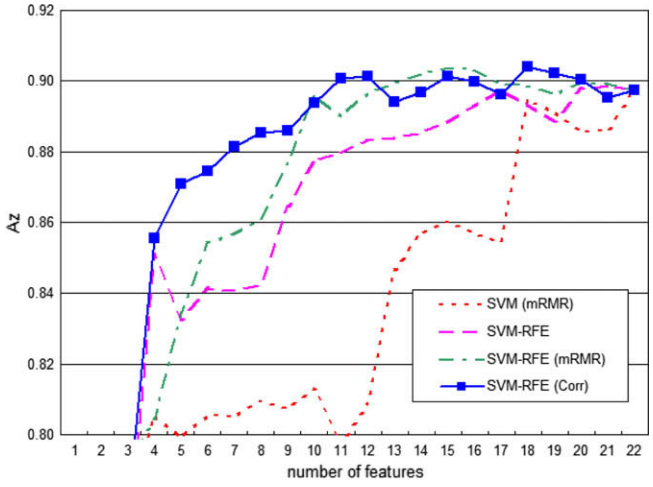

Fig. 6. Az with different number of features for mass of WFUSM.

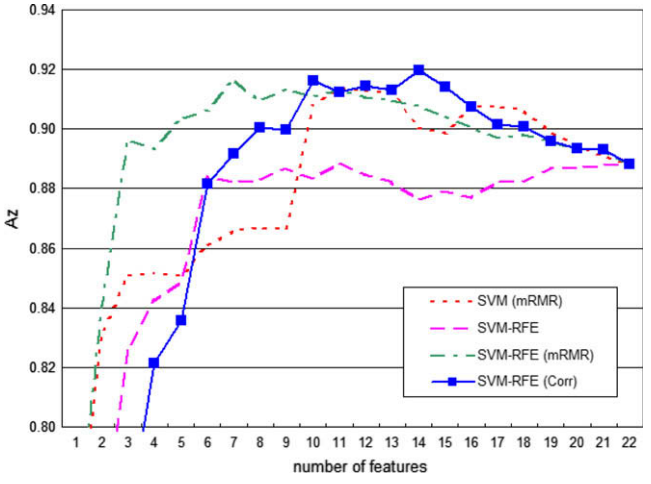

Fig. 7. Az with different number of features for calcification of MGH.

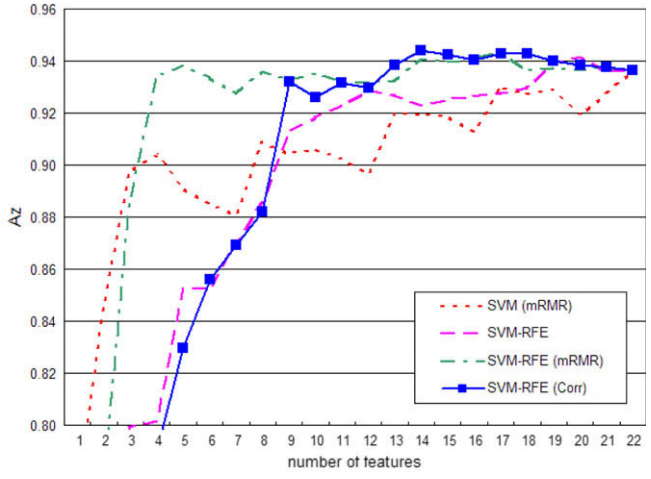

Fig. 8. Az with different number of features for calcification of WU.

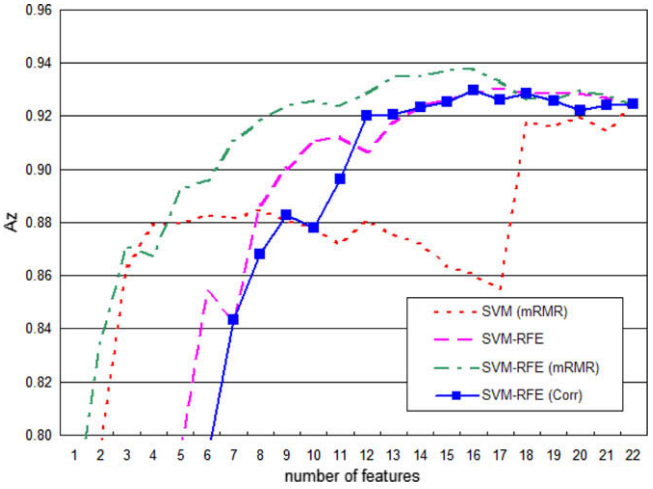

Fig. 9. Az with different number of features for calcification of WFUSM.

(mRMR). In Fig. 9, SVM-RFE (Corr) outperformed SVM-RFE (mRMR) even in terms of the convergence speed. Moreover, unlike SVM (mRMR) which yielded good performance only in the calcification dataset from WU while presenting extremely poor or unstable result in the others, SVM-RFE (Corr) showed relatively stable performance in all datasets we have used for the experiments.

5. Conclusion

In this work, a modified SVM-RFE-based feature selection method for the mammography classification problem was proposed. With two baseline algorithms, extensive experiments using real datasets were conducted to estimate the effectiveness of the proposed method. From the empirical results, we found the following: SVM-RFE (Corr) showed the best performance in most of the datasets we used. In particular, SVM-RFE (Corr) outperformed SVM-RFE (mRMR) in terms of classification accuracy supporting our intuition that the separation of redundancy and relevance might be helpful in feature selection process of SVM-RFE. As a side note, we observed that SVM-RFE (mRMR) converges faster than other methods while it does not yield the best performance.

Acknowledgement

The work of S. Kim was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0074611) and by the Special Research Grant of Sogang University 200811028.01.

References

American Cancer Society, 2008. Cancer Facts and Figures.

Balleyguier, C., Ayadi, S., Vannguyen, K., Vanel, D., Dromain, C., Sigal, R., 2007. Birads classification in mammography. Eur. J. Radiol. 61 (2), 192–194.

Ding, C., Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biological 3 (2), 185–205.

El-Naqa, I., Yang, Y., Wernick, M., Galatsanos, N., Nishikawa, R., 2002. A support vector machine approach for detection of microcalcifications. IEEE Trans. Med. Imag. 21, 1552–1563.

Elmore, J., Armstrong, K., Lehman, C., Fletcher, S., 2005. Screening for breast cancer. J. Amer. Med. Assoc. 293, 1245–1256.

Fischer, E., Lo, J., Markey, M., 2004. Bayesian networks of bi-rads descriptors for breast lesion classification. In: Proc. 26th IEEE EMBS, San Francisco, CA, USA, vol. 2, pp. 3031–3034.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learn. 46 (1-3), 389–422.

Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, W., 2001. The digital database for screening mammography. In: Yaffe, M. (Ed.), Proc. 5th IWDM. Medical Physics Publishing, pp. 212–218.

Isaac, L., Lederman, R., Buchbinder, S., Srour, Y., Bamberger, P., Sperber, F., 2006. Computerized classification can reduce unnecessary biopsies in bi-rads category 4a lesions. In: Astley, S.M., Brady, M., Rose, C., Zwiggelaar, R. (Eds.), Digital Mammography/IWDM, Lecture Notes in Computer Science, vol. 4046. Springer, pp. 76–83.

Jiang, Y., Nishikawa, R., Schmidt, R., Toledano, A., Doi, K., 2001. Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. Radiology 220, 787–794.

Land Jr., W.H., Mckee, D., Velazquez, R., Wong, L., Lo, J., Anderson, F., 2003. Application of support vector machines to breast cancer screening using mammogram and clinical history data. In: Sonka, M.F.J. (Ed.), Proc. SPIE, Medical Imaging 2003: Image Processing, vol. 5032, pp. 546–556.

LeCun, Y., Denker, J.S., Solla, S.A., 1990. Optimal brain damage. In: Advances in Neural Information Processing Systems. Morgan Kaufmann, pp. 598–605.

Li, F., Yang, Y., 2005. Analysis of recursive gene selection approaches from microarray data. Bioinformatics 21 (19), 3741–3747.

Lo, J., Bilska-Wolak, A., Baker, J., Tourassi, G., Floyd, C., Markey, M., 2006. Computer-aided diagnosis in breast imaging: Where do we go after detection? In: Suri, J., Rangayyan, R. (Eds.), Recent Advances in Breast Imaging, Mammography and Computer-aided Diagnosis of Breast Cancer. SPIE Press, pp. 871–900.

Lo, J., Gavrielides, M., Markey, M., Jesneck, J., 2003. Computer-aided classification of breast microcalcification clusters: Merging of features from image processing and radiologists. In: Sonka, M., Fitzpatrick, J. (Eds.), Medical Imaging 2003: Image Processing, vol. 5032. SPIE Press, pp. 882–889.

Lo, J., Markey, M., Baker, J., Floyd Jr., C., 2002. Cross-institutional evaluation of bi-rads predictive model for mammographic diagnosis of breast cancer. Amer. J. Roentgenol. 178, 457–463.

Mundra, P.A., Rajapakse, J.C., 2007. SVM-REF with relevancy and redundancy criteria for gene selection. In: Rajapakse, J.C., Schmidt, B., Volkert, L.G. (Eds.), PRIB, Lecture Notes in Computer Science, vol. 4774. Springer, pp. 242–252.

Orel, S., Kay, N., Reynolds, C., Sullivan, D., 1999. Bi-Rads categorization as a predictor of malignancy. Radiology 211, 845–850.

Panchal, R., Verma, B., 2006. Characterization of breast abnormality patterns in digital mammograms using auto-associator neural network. In: King, I., Wang, J., Chan, L., Wang, D.L. (Eds.), ICONIP (3), Lecture Notes in Computer Science, vol. 4234. Springer, pp. 127–136.

Verma, B., Zhang, P., 2007. A novel neural-genetic algorithm to find the most significant combination of features in digital mammograms. Appl. Soft Comput. 7 (2), 612–625.

Wei, L., Yang, Y., Nishikawa, R., Jiang, Y., 2005. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. IEEE Trans. Med. Imag. 24, 371–380.

Zhang, J., Jin, R., Yang, Y., Hauptmann, A.G., 2003. Modified logistic regression: An approximation to SVM and its applications in large-scale text categorization. In: Fawcett, T., Mishra, N. (Eds.), ICML. AAAI Press, pp. 888–895.