A Subspace Learning Algorithm For Microwave Scattering Signal Classification With Application To Wood Quality Assessment

2012 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 23–26, 2012, SATANDER, SPAIN

A SUBSPACE LEARNING ALGORITHM FOR MICROWAVE SCATTERING SIGNAL CLASSIFICATION WITH APPLICATION TO WOOD QUALITY ASSESSMENT

Yinan Yu, Student Member, IEEE, Tomas McKelvey, Senior Member, IEEE

Chalmers University of Technology
Department of Signals and Systems
412 96 Gothenburg, Sweden

ABSTRACT

A classification algorithm based on a linear subspace model has been developed and is presented in this paper. To further improve the classification results, the full linear subspace of each class is split into subspaces with lower dimensions and characterized by local coordinates constructed from automatically selected training data. The training data selection is implemented by optimizations with least squares constraints or L1 regularization. The working application is to determine the quality in wooden logs using microwave signals [1]. The experimental results are shown and compared with classical methods.

Index Terms— classification, linear subspace, sparse representation, training data selection

1. INTRODUCTION

Microwave signals are widely used for applications in a vast range of different domains [2][3][4][5][6]. In this paper, frequency domain measurements are used for wood quality classification. A classifier based on linear subspace settings has been developed, with the assumption that samples from one class lie on one of the linear subspaces and the sub-basis can be therefore derived from the corresponding data points. Namely, each class contains more than one such linear subspace, and the representation of each sample can be defined automatically using the least squares criterion or sparse regularization. The decision of the classifier is hence based on some criteria which involve the distance from the data to the estimated subspace. Moreover, these experimental signals have typically extremely high dimensionality, whereas the training sample size is usually very small. The global topological properties and statistical assumptions of the data points thus become extremely difficult to verify and the training of the classifier becomes very challenging. In this paper, two model assumptions are introduced, and a classification algorithm based on one of them is presented.

2. SIGNAL MODEL AND CLASSIFICATION HYPOTHESIS

Given N_c the total number of classes, let $\{\boldsymbol{x}_c^i\}$ be the training set of class $c \in \{1...N_c\}$, where $\boldsymbol{x}_c^i$ is D dimensional complex valued and $i \in \{1, \cdots, N\}$ is the sample index. First, without loss of generalization, let us consider the case where the total number of classes N_c is 2, namely $c \in \{1, 2\}$.

2.1. Model assumption 1

Each data point $\boldsymbol{x}_c^i$ drawn from class c is generated according to a linear model defined as:

$$\begin{aligned} \boldsymbol{x}_1^i &= \boldsymbol{U}_1\boldsymbol{\alpha}_1^i + \boldsymbol{e} \\ \boldsymbol{x}_2^i &= \boldsymbol{U}_2\boldsymbol{\alpha}_2^i + \boldsymbol{e} \end{aligned} \tag{1}$$

where the columns of $\boldsymbol{U}_c$, denoted as $\{\boldsymbol{u}_{c,l}\}$ represent the basis of the corresponding linear subspace with $l \in \{1, \cdots, D_c\}$; $\boldsymbol{\alpha}_c^i$ is the weighting vector; and $\boldsymbol{e}$ is random noise.

If $\boldsymbol{U}_c$ is given, we can compute the distance $d_c(\boldsymbol{x}^i)$ from $\boldsymbol{x}^i$ to the linear subspace spanned by its orthonormal columns $\{\boldsymbol{u}_{c,l}\}$,

$$d_c(\boldsymbol{x}^i) = \left\|\boldsymbol{x}^{\boldsymbol{i}} - \boldsymbol{P}_c\boldsymbol{x}^i\right\|_2 = \left\|\boldsymbol{x}^{\boldsymbol{i}} - \boldsymbol{U}_c\boldsymbol{U}_c^H\boldsymbol{x}^i\right\|_2 \tag{2}$$

where $\boldsymbol{P}_c$ denotes the projection matrix and $\boldsymbol{U}_c^H$ is the Hermitian of the matrix $\boldsymbol{U}$.

Given one unlabeled signal $\boldsymbol{x}^i$, the task is to estimate the class label $\hat{c}^i$ according to the following criterion:

$$\hat{c}^i = \arg\min_c d_c(\boldsymbol{x}^i) \tag{3}$$

2.2. Model assumption 2

Instead of a linear subspace spanned by $\boldsymbol{U}_c$, each $\boldsymbol{x}_c^i$ is considered to be generated from a linear subspace spanned by a

978-1-4673-1026-0/12/$31.00

'smaller' basis $\boldsymbol{U}_c^k$, where $k \in \{1, \cdots, K_c\}$, and K_c is the total number of such subspaces. By 'smaller' basis, one can imagine that the subspace spanned by the basis appeared in (1) is now a set of K_c linear subspaces spanned by some low dimensional bases.

Let $\mathscr{U}_c = \{\boldsymbol{x}^i : \boldsymbol{x}^i \in class\ c\}$. From the assumption 2, we have:

$$\mathscr{U}_c = \bigcup_{k \in \{1 \ldots K_c\}} \mathscr{U}_c^k \tag{4}$$

where $\mathscr{U}_c^k$ is a subset of $\mathscr{U}_c$ which is a linear subspace defined as:

$$\mathscr{U}_c^k = \left\{ \boldsymbol{x}^i : \ \boldsymbol{x}^i = \sum_{l=1}^{D_c^k} \beta_l \boldsymbol{u}_{c,l}^k \right\} \tag{5}$$

where D_c^k is the dimension of the subspace spanned by $\boldsymbol{U}_c^k = \left\{ \boldsymbol{u}_{c,l}^k \right\}$ with $l \in \{1, \cdots, D_c^k\}$, and β_l is the corresponding coefficient.

Therefore, the matrix $\boldsymbol{U}_c$ from (1) is no more defined as the basis of a linear subspace, but rather as a set of K_c bases:

$$\boldsymbol{U}_c = \left\{ \boldsymbol{U}_c^k \right\}, k \in \{1 \ldots K_c\} \tag{6}$$

Accordingly, the signal model becomes:

$$\begin{aligned} \boldsymbol{x}_1^i &= \boldsymbol{U}_1^{k_1(i)} \boldsymbol{\beta}_1^i + \boldsymbol{e} \\ \boldsymbol{x}_2^i &= \boldsymbol{U}_2^{k_2(i)} \boldsymbol{\beta}_2^i + \boldsymbol{e} \end{aligned} \tag{7}$$

where, the weighting vector $\boldsymbol{\beta}_c^i$ depends on the sample number i, the class label c, as well as the basis number $k_c(i)$. Note that the purpose of writing $k_c(i)$ is to show that k_c is a function of the sample index i, which means that the basis $\boldsymbol{U}_c^{k_c(i)}$ needs to be established for each $\boldsymbol{x}^i$ adaptively. Without ambiguity, we write k instead of $k_c(i)$ for convenience.

Therefore, $d_c^k(\boldsymbol{x}^i)$ is indicating the distance from $\boldsymbol{x}^i$ to the subspace $\boldsymbol{U}_c^k$:

$$d_c^k(\boldsymbol{x}^i) = \left\| \boldsymbol{x}^{\boldsymbol{i}} - \boldsymbol{U}_c^k (\boldsymbol{U}_c^k)^H \boldsymbol{x}^i \right\|_2 \tag{8}$$

And $\hat{c}$ can be estimated in the same way as in (3) with a slight modification.

$$\hat{c}^i = \arg \min_c d_c^k(\boldsymbol{x}^i) \tag{9}$$

A low dimensional example can be visualized in Fig. 1. Data point $\boldsymbol{x}^i$ is a high dimensional vector and we can imagine $D = 3$ for convenience. The solid and dash lines represent one dimensional subspaces contained in class 1 and 2 respectively. This $d_c^k(\boldsymbol{x}^i)$ indicates the distance from $\boldsymbol{x}^i$ to the corresponding subspace $\mathscr{U}_c^k$. In this example, the meaning of the model assumption 1 and 2 can be clearly visualized: according to assumption 1, the subspace of class c is constructed from both lines with the same type, which is a two dimensional hyperplane; whereas if we consider model assumption 2, each subset $\mathscr{U}_c$ is the union of two one dimensional lines:

$$\mathscr{U}_c = \mathscr{U}_c^{k_c=1} \cup \mathscr{U}_c^{k_c=2} \tag{10}$$

Note that we use k_c to indicate the subspace in the example from Fig. 1. However, the subscript for k is usually omitted for convenience.

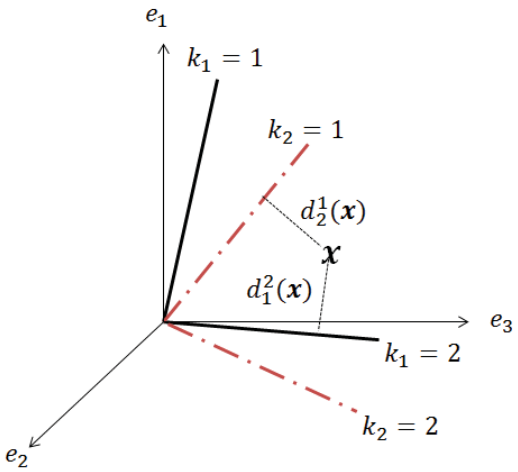


Fig. 1. A 3 dimensional example is shown. The solid and dash lines indicate linear subspaces domained by the data points from class 1 and 2 respectively.

3. PROPOSED METHOD

The proposed approach is based on the Model assumption 2, where the topological space of signal $\boldsymbol{x}$ is assumed to be a collection of some linear subspaces according to (6). To estimate $\hat{c}$ with respect to (9), the method is discussed in this section and a proposed algorithm is presented.

3.1. Adaptive training data selection for $\boldsymbol{x}^i$

According to (7), each sample from class c is assumed to be lying in one of the K_c subspaces. Therefore, before computing the distance by (8), we need to select the 'correct' training set for $\boldsymbol{x}^i$ in both classes. The 'correct' training set, denoted as $\mathscr{O}_c^i$, is defined as the data set dominating the subspace which has a smaller distance to $\boldsymbol{x}^i$ over all K_c subspaces. This can be illustrated in Fig. 1. As we can see, given an unlabeled data $\boldsymbol{x}^i$, although the lines with the same type represent the same class, we still need to select one of them to compute the

distance for $\boldsymbol{x}^i$. In this case, the selected subspaces are line $k_2 = 1$ and $k_1 = 2$.

We define the training set $\mathscr{O}_c^i$ of $\boldsymbol{x}_c^i$ as an open set containing the points from class c which are lying on the closest linear subspace spanned by a basis $\boldsymbol{U}_c^k$ with respect to $\boldsymbol{x}_c^i$. The metric is the usual distance computed by (2). Note that $\mathscr{O}_c^i \subset \mathscr{U}_c^k$.

In another word, $\mathscr{O}_c^i$ is selected with a linear subspace setting. Namely, it is the set of data points which are linearly dependent of $\boldsymbol{x}^i$ in some directions. Practically, this selection can be done by choosing D_c data from the training set using different criteria, where D_c is pre-defined by cross validation.

First, consider a measurement matrix constructed by placing all training data from one class as its columns:

$$\boldsymbol{X}_c = \left[\boldsymbol{x}_c^1,\ \boldsymbol{x}_c^2,\ \cdots,\ \boldsymbol{x}_c^{N_c}\right] \tag{11}$$

The task is to select D_c relevant columns from $\boldsymbol{X}_c$, such that $\boldsymbol{x}^i$ can be written as a linear combination of the basis which spans the subspace dominated by these data. The selection is carried out by computing a weighting vector $\boldsymbol{w}_c^i$ whose j^{th} element represents the importance of the respective column $\boldsymbol{x}^j$ with respect to reconstructing the $\boldsymbol{x}^i$. The more significant data points are then chosen to be the correct training data of $\boldsymbol{x}^i$. To simplify the expression, we call $\mathscr{O}_c^i$ the training set of $\boldsymbol{x}^i$ from now on.

- **Formulation using least square criterion**

$$\boldsymbol{w}_c = \arg\min_{\boldsymbol{w}_c} \left\|\boldsymbol{x}^i - \boldsymbol{X}_c\boldsymbol{w}_c\right\|_2 \tag{12}$$

  However, without constraints on the number of non-zero elements, any linear combinations of data points with insignificant directions are allowed. This results in relatively arbitrary selections. To resolve this problem, a sparse representation [9] is needed to maintain the significance of any selected columns.

- **Formulation using sparse representation**

$$\boldsymbol{w}_c = \arg\min_{\boldsymbol{w}_c} \left\|\boldsymbol{x}^i - \boldsymbol{X}_c\boldsymbol{w}_c\right\|_{l_2} + \lambda\left\|\boldsymbol{w}_c\right\|_{l_1} \tag{13}$$

  Theoretically, the training set of $\boldsymbol{x}^i$ can be selected by the columns in $\boldsymbol{X}_c$ corresponding to the non-zeros elements in the vector $\boldsymbol{w}_c^i$. Namely, the data points with non-zero correlation. In practice, this is implemented by sorting $\boldsymbol{w}_c^i$ and selecting the columns corresponding to the first D_c elements with higher values.

Let $\tilde{\boldsymbol{w}}_c^i$ be the sorted version of $\boldsymbol{w}_c^i$. The data points from $\boldsymbol{X}_c$ corresponding to the first D_c elements of $\tilde{\boldsymbol{w}}_c^i$ are selected to construct $\boldsymbol{O}_c^i$. Let J denote the set of the indices of $\boldsymbol{X}_c$ associated to $\tilde{\boldsymbol{w}}_c^i(1:D_c)$, we then have:

$$\boldsymbol{O}_c^i = \left[\boldsymbol{x}^j\right]_{j\in J} \tag{14}$$

One parameter in this setting is the dimension of the subspace D_c, which reflects the variation of the signal to some extent. Namely, the signal with higher variance is assumed to dominate a higher dimensional subspace. For instance, in the wood qualification application, we assume that the dimension of the 'rotten subspace' D_2 is larger than the 'normal subspace' D_1 and the assumption is verified by cross-validation.

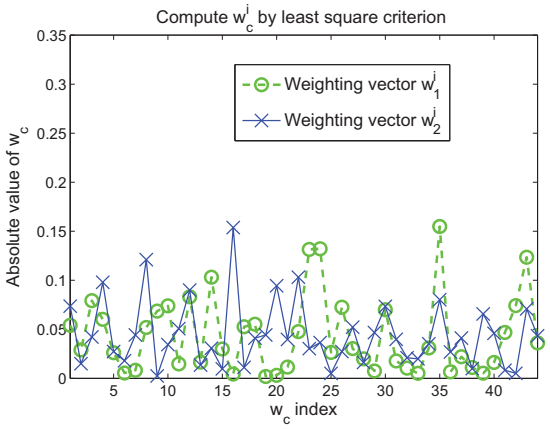


Fig. 2. The weighting vector $\boldsymbol{w}_c^i$ estimated with respect to least squares criterion.

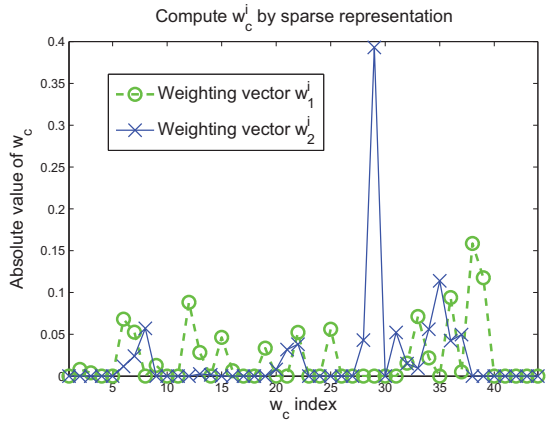


Fig. 3. The weighting vector $\boldsymbol{w}_c^i$ estimated using sparse representations.

3.2. Estimate basis $\boldsymbol{U}_c^k$

Once $\mathscr{O}_c^k$ is identified, the basis that spans the corresponding subspace can be estimated from the matrix $\boldsymbol{O}_c^i$.

Let $\boldsymbol{O}_c^i$ be the matrix constructed by (14), the basis of the subspace $\mathscr{U}_c^k$ can then be estimated by the left singular vectors of $\boldsymbol{O}_c^i$ computed from the singular value decomposition (SVD) and the subspace distance is obtained by (8).

3.3. Algorithm

Given the training data sets $\left\{\boldsymbol{x}_1^j\right\} \in \mathscr{U}_1$ and $\left\{\boldsymbol{x}_2^j\right\} \in \mathscr{U}_2$, the measurement matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can be constructed by (11). The dimensions of the subspaces $\mathscr{U}_1^k$ and $\mathscr{U}_2^k$ are estimated as D_1 and D_2 by cross validation. A classification algorithm for multi-class case based on Model assumption 2 is presented in Algorithm 1.

Algorithm 1 Classification algorithm (Model assumption 2)

- Produce the feature vector $\boldsymbol{x}^i$ by pre-processing the data;
- $\forall c \in \{1, 2, \cdots, N_c\}$, compute the weight vector $\boldsymbol{w}_c^i$:

$$\boldsymbol{w}_c^i = \arg\min_{\boldsymbol{w}_c^i} \left\|\boldsymbol{x}^i - \boldsymbol{X}_c \boldsymbol{w}_c^i\right\|_2 + \lambda \left\|\boldsymbol{w}_c^i\right\|_{l_1}$$

- Pick up the first D_c larger elements of $\boldsymbol{w}_c^i$ and identify the set of corresponding indices J_c of $\boldsymbol{X}_c$, with D_c being defined by cross-validation;
- Construct the matrix:

$$\boldsymbol{O}_c^i = \left[\boldsymbol{x}_c^j\right], \quad j \in J_c$$

- Estimate the basis $\boldsymbol{U}_c^k$ spanning $\boldsymbol{O}_c^i$:

$$\left[\boldsymbol{U}_c^k, \ \sim, \ \sim\right] = \text{SVD}(\boldsymbol{O}_c^i)$$

- Compute the distance:

$$d_c^k(\boldsymbol{x}^i) = \left\|\boldsymbol{x}^{\boldsymbol{i}} - \boldsymbol{U}_c^k((\boldsymbol{U}_c^k)^H \boldsymbol{x}^i)\right\|_2$$

- Estimate the label of $\boldsymbol{x}^i$:

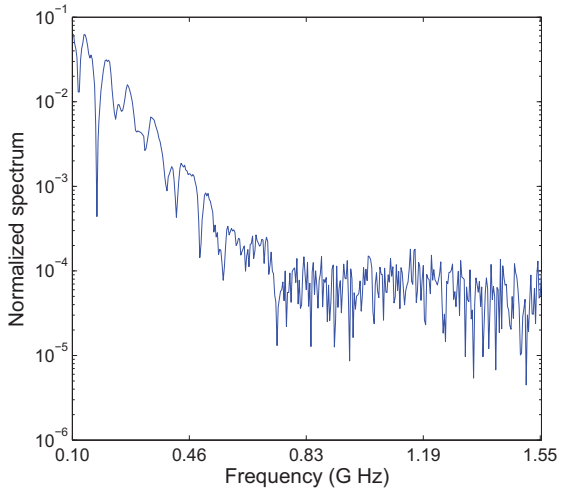
$$\hat{c}^i = \arg\min_c d_c^k(\boldsymbol{x}^i)$$

Fig. 4. An illustration of the experimental setup. Each cross indicates the position of one antenna, playing a role both as transmitter and receiver. The signal is then measured as the S parameters in frequency domain. The antennas are labeled as $1, 2, \cdots, N_a$ in a counterclockwise order.

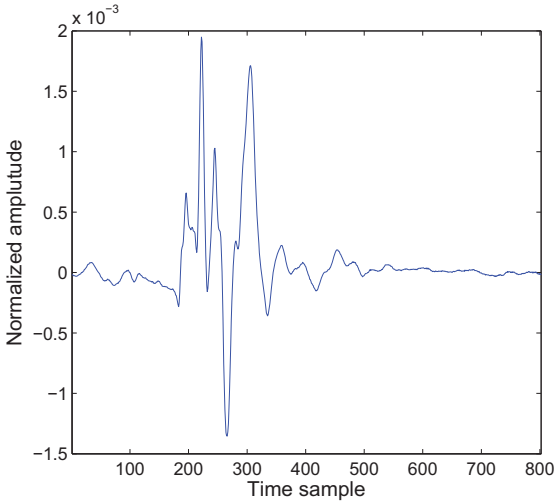
4. APPLICATIONS AND RESULTS

4.1. Signal description

An illustration of the set up is shown in Fig. 4. Each green cross indicates the position of one antenna playing both the role of transmitter and receiver. The raw signals are scattering (S) parameters (the ratio between the received and transmitted energy of one antenna measured in frequency domain).



(a) S parameter measured at channel $\{1, 5\}$



(b) Corresponding time domain signal

Fig. 5. The absolute value of the measured S parameters $\boldsymbol{S}_{1,5}$ using the antennas at the 1^{st} and the 5^{th} positions as input and output sensors respectively over all the frequency points.

The transmitter and the receiver number are indicated by q and p respectively, and the pair is referred as channel $\{p, q\}$.

One example of measurement is shown in Fig. 4.1. The S parameter measured at frequency point $\omega_n = 2\pi f_n$ at channel $\{p, q\}$ can be written as:

$$S_{pq}(\omega_n) = e^{\eta + j\gamma} \tag{15}$$

where the real part η represents the dumping and the imaginary part $j\gamma$ gives the phase information. Note that $S_{pq}(\omega_n) = S_{qp}(\omega_n)$

The signal for a given channel $\{p, q\}$ can be expressed as follows:

$$\boldsymbol{S}_{pq} = \begin{bmatrix} S_{pq}(\omega_1) \\ \vdots \\ S_{pq}(\omega_{N_\omega}) \end{bmatrix} \tag{16}$$

Therefore, the full measurement $\boldsymbol{x}$ for all the channels can be written as:

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{S}_{11} \\ \vdots \\ \boldsymbol{S}_{1N_q} \\ \boldsymbol{S}_{21} \\ \vdots \\ \boldsymbol{S}_{2N_q} \\ \vdots \\ \boldsymbol{S}_{N_pN_q} \end{bmatrix} \tag{17}$$

Furthermore, according to our setup, we have $N_p = N_q = N_a$, where N_a is the number of antennas. The vectorized signal $\boldsymbol{x}$ is therefore considered as a D dimensional vector, where the dimension D is determined by

$$D = \frac{1}{2} N_\omega (N_a^2 + N_a). \tag{18}$$

4.2. Pre-processing

Different types of signal pre-processing procedures can be applied before the signals are used as the input of the classifier. In this work, the main operations are 1) feature selection by frequency points; 2) logarithm transformation; 3) normalization of each channel.

- Frequency point selection:
  For each channel $\boldsymbol{S}_{pq}$ in (16), $N_\omega = 401$ frequency points are measured. However, only the first 180 points (corresponding to approximately $0.1 \sim 1.3$ GHz) are assumed to be containing the main information are selected as the input of the classifier for computational efficiency. This is determined empirically and verified by cross-validation. More sophisticated feature selection can be further developed to improve the performance. The signal $\boldsymbol{x}^i$ in (17) is then constructed by using only 180 frequency points from each channel.
- Logarithm transform
  We take $\log(\boldsymbol{x}^i)$ as the new signal vector instead of the $\boldsymbol{x}^i$ defined in (17) to retrieve the complex number $\eta + j\gamma$.
- Normalization
  The reflection S_{pq}, where $p = q$, is typically much stronger than the transmission where $p \neq q$. However, the later one might carry more information of the object. Therefore, to unify the contribution of different channels, a channel-wise normalization is implemented on the signal $\boldsymbol{S}_{pq}$ in (16) to ensure that they contain the same energy level.

$$\boldsymbol{S}_{pq} = \frac{\boldsymbol{S}_{pq}}{\|\boldsymbol{S}_{pq}\|_2} \tag{19}$$

4.3. Experimental Results

The experiment is based on the setup shown in Fig. 4. There are $N_a = 12$ antennas and the number of selected frequency points N_ω is 180. From (18), we have the dimension of the vectorized signal $\boldsymbol{x}^i$ is 14040. In this experiment, 54 and 108 samples for normal and rotten timbers are measured respectively.

Classification results			
	Method	Corr. Rate	FAR
Classical methods	NN	64.2%	37.0%
	SVM	77.7%	29.0%
Subspace model	Assum.1	81.6%	29.0%
	Assum.2(LS)	84.2%	29.0%
	Assum.2(L1)	87.9%	29.0%

Table 1. The classification rate and false alarm rate of rotten log obtained by randomized N-fold testing.

A randomized N-fold testing procedure has been performed, where the samples are randomized, and among which 44 are used for training and the rest for testing. The randomization is repeated 30 times. The results are evaluated by both the classification rate for the rotten wood and the false alarm for the normal wood, then compared with classical methods such as support vector machines (SVM) [7] and nearest neighbors approach (NN) [8]. The results of classification rate with a fixed false alarm rate are shown in Table 1 and the ROC curve can be found in Fig. 6. The effect of different formulations of $\boldsymbol{w}_c^i$ discussed in Sec. 3.1 on the classification result (referred as LS and L1 in Table 1) is similarly compared.

5. CONCLUSION

A classification algorithm has been proposed in this paper to differentiate healthy timber from the rotten ones. We introduce the model where the signals from one class are drawn

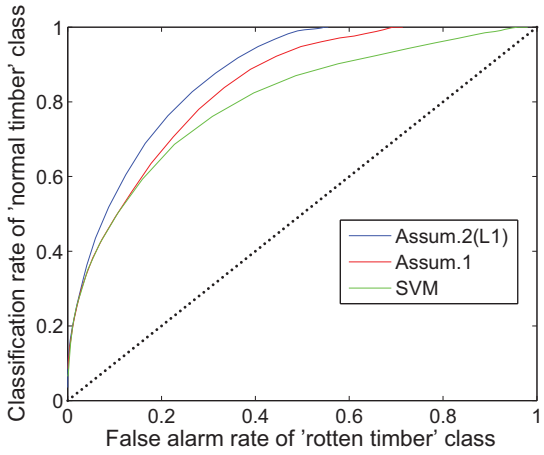


Fig. 6. The ROC curves of selected classifiers introduced in the paper.

from several different subspaces. It gives the most promising results when the correct training set is estimated by a sparse representation. One potential of the classifier is that we estimate the basis independently and the classification results only depend on the estimated basis and the unlabeled data point. Therefore, this approach can be extended to multi-class cases with no extra effort. That is, as long as the basis for each class is estimated, the distance can then be computed accordingly. Related experiments and analysis are under progress. Issues are remaining to be investigated as a subject of the future work, such as automatic determination in the dimension of the subspaces.

6. REFERENCES

[1] Persson M., McKelvey T. , Fhager A., Lui H., etc. *Advances in Neuro Diagnostic based on Microwave Technology, Transcranial Magnetic Stimulation and EEG source localization*, Asia Pacific Microwave Conference, 2011.

[2] Arye Rosen, Maria A. Stuchly, and Andr Vander Vorst, *Applications of RF/Microwaves in Medicine*, IEEE Trans. on Microwave Theory AND Techniques, vol. 50, No. 3, pp. 963 - 974, March 2002.

[3] Williams, T. C., Sill, J.M., Fear, E.C., *Breast surface estimation for radar-based breast imaging systems*, IEEE Transactions on Biomedical Engineering, Vol.55, No.6, pp. 1678-1686, May, 2008.

[4] Himanshu.K. Solanki, Vipul D.Prajapati, Girish K. Jani, *Microwave Technology - A Potential Tool in Pharmaceutical Science*, International Journal of PharmTech Research, Vol. 2, No.3, pp 1754-1761, July-Sept 2010

[5] Kappe, C. O., Stadler, *A. Microwaves in Organic and Medicinal Chemistry*, Wiley-VCH, Weinheim, 2005

[6] Fhager, A.; McKelvey, T.; Persson, M., *Stroke Detection Using a Broad Band Microwave Antenna System*, 4th European Conference on Antennas and Propagation, Barcelona, Spain, pp. C13P1-2, April, 2010.

[7] Cristianini N., and Shawe-Taylor J., *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, First Edition, Cambridge University Press, Cambridge, 2002.

[8] Belur V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, IEEE Computer Society, Dec., 1990.

[9] Figueiredo M.A.T., *Adaptive sparseness for supervised learning*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 25, no. 9, pp. 1150 1159, sep. 2003.