Asian Journal of Information Technology 9 (4): 196-204, 2010
ISSN: 1682-3915

Identification of Urinary Tract Infection Bacteria by Fourier Transform Infrared (FT-IR) Spectroscopy

Bailing Zhang
Deparment of Computer Science and Software Engineering,
Xian Jiaotong Lierpool Unversity, 111 Ren Ali Road, Suzhou Industrial Park,
Suzhou Jiangsu Province, 215123, R.P. China

Abstract: Urinary Tract Infection (UTI) is a serious health problem affecting millions of people each year and it is significant to identify the causal agent prior to treatment. The bacteria typically associated with UTI include shape *Eschericha coli*, shape Klebsiella, shape *Proteus mirabilis*, shape *Citrobacter freundii* and shape *Enterococcus* sp. In recent years, a number of spectroscopic methods such as Fourier transform infrared (FT-IR) spectroscopy have been used to analyse the bacteria associated with UTI which are generally described as rapid whole organism fingerprinting. FT-IR typically takes only 10 sec per sample and generates holistic biochemical profiles from biological materials. In the past, multivariate analysis and artificial neural networks have been used to analyse and interpret the information rich data. In this study, The Support Vector Machine (SVM) applied to the FT-IR data for the automatic identification of UTI bacteria. Cross-validation test results indicate that the generalization performance of the SVM was over 98% to identify the UTI bacteria, compared to neural network's accuracy of 81%. Among the various multi-class SVM schemes tested, the Directed Acyclic Graph (DAG) method gives the best classification results. A Principal Component Analysis (PCA) based dimension-reduction could accelerate the training/testing time to a great extent, without deteriorating the identification performance.

Key words: Classification, support vector machine, machine learning, dimension reduction, urinary tract infection, Fourier Transform Infrared (FT-IR) spectroscopy

INTRODUCTION

Urinary Tract Infections (UTIs) are one of the major clinical problems which account for about 8.3 million doctor visits each year. Women are especially prone to UTIs for reasons that are not yet well understood. One woman in five develops a UTI during her lifetime. Nearly all UTIs are caused by bacteria that enter the urethral opening and move upward to the urinary bladder and sometimes the kidneys. In clinics, the bacteria typically associated with UTI are Eschericbia coli (causative organism of 50% of the cases), Klebsiella species (14%), other coliforms (4%), staphylococci (6%), Enterococcus faecalis (10%), Pseudomonas aeruginosa (3%) (Goodacre *et al*., 1998; Jarvis and Goodacre, 2004).

The common method of diagnosing urinary tract infections is based on the laboratory investigation of a mid-stream specimen of urine, often referred to as an MSU which usually comprises of microscopical examination of the urine sample followed by bacterial culture. A quantitative result is often used to confirm the clinical diagnosis and finding of 10^5 cfu mL^{-1} of urine is defined as significant bacteriuria (Morgan and McKenzie, 1993). Using conventional methods by laboratory examination of urine, however is expensive, time-consuming and labour-intensive; approximately 24 h incubation is required to obtain an accurate colony count. An additional 12-24 h is needed for organism identification and susceptibility testing, which may further delay administration of the most appropriate marrow-spectrum antibiotic.

With the developments in analytical instrumentation, the requirements for microbial characterization of UTI bacteria have been implemented by physico-chemical spectroscopic methods (Magee, 1993) including Pyrolysis mass Spectrometry (PyMS), Fourier-Transform Infrared spectroscopy (FT-IR) and UV resonance Raman spectroscopy. Often referred to as whole-organism fingerprinting, these methods measure primarily the bond strengths of molecules and the vibrations of bonds within Functional groups (FT-IR and Raman), thus offering quantitative information about the total biochemical composition of a sample. Among the three methods, the FT-IR spectra of micro-organisms has been regarded as robust and advantageous because it is reproducible and distinct for different bacteria and fungi (Maquelin *et al*., 2002; Winder *et al*., 2004). Therefore, there are numerous studies for applying the technique to classify or identify various micro-organisms.

For the differences between FT-IR spectra to be disclosed, proper pattern recognition systems should be employed, since the FT-IR spectral data are complex, nonlinear and overlapped. The interpretation of the high dimensional FT-IR data has typically been performed by multivariate analysis methods such as clustering, Principal Components Analysis (PCA) or Discriminant Function Analysis (DFA) (Goodacre *et al*., 1998; Jarvis and Goodacre, 2004). These methods are featured by their unsupervised learning characteristics by which the investigator can group objects based on their perceived closeness. While simple and convenient, this process also has the limitation of being subjective because it mainly relies upon the interpretation of complicated scatter plots and dendrograms. More recently, some flexible supervised learning methods have been applied to the analysis of these hyperspectral data for example, the Multiple Layer Perceptron (MLP) neural network model (Goodacre *et al*., 1998; Mouwen *et al*., 2006; Wenning *et al*., 2002). Neural networks have been largely used in the past as pattern classifiers in many applications. Their learning and generalization capabilities also make them as favourite options in the biomedical applications.

Though efficient for some applications including the identification of UTI bacteria, MLP classifier has been shown the limitations due to the problems like local minima in the optimization. Over the last few years, another particular machine learning algorithm, Support Vector Machines (SVMs) has shown promise in a variety of classification tasks. SVM is based on a variation of regularization techniques for regression (Cristianini and Shawe-Taylor, 2000). Because SVM seeks a globally optimized solution and avoids over-fitting, it has the ability to construct predictive models with larger generalization power, thus obtaining extensive applications including medical diagnosis (Comak *et al*., 2007) and bioinformatics (Furey *et al*., 2000). As compared with probabilistic models and classical neural networks, SVM provide a well-understood regularization mechanism which makes learning from few examples in high-dimensional feature spaces possible. In that way, SVM and related methods can effectively cope with the curse of dimensionality which has been difficult for the more traditional tools in machine learning.

In this study, it is proposed that to apply SVM technique for the automated identification of UTI bacteria using the FT-IR spectra. Cross-validation test results indicate that the generalization performance of the SVM was on average over 90% to identify the UTI bacteria, compared to neural network's accuracy of 80%. A number of different multi-class SVM schemes have been studied including using a direct multiclass SVM, Directed Acyclic Graph (DAG) or combining multiple binary SVM classifiers via the general ECOC scheme classification. Among these methods, the DAG scheme offers the best performance. A Principal Component Analysis (PCA) based dimension-reduction could accelerate the training/testing time to a great extent without deteriorating the identification performance. The study provides the foundation for successful applications of SVMs to many real world microorganisms classification tasks.

FOURIER-TRANSFORM INFRARED (FT-IR) SPECTROSCOPY FOR URINARY TRACT INFECTION (UTI) BACTERIA

The FT-IR data for a group of 59 bacteria isolated from the urine of patients with Urinary Tract Infection (UTI) was provided by the researchers of (Goodacre *et al*., 1998) which were collected from Bronglais General Hosiptal, Alberystwyth. By conventional biochemical tests, all isolates were typed to belong to *E. Coli*, *Pr. mirabilis*, *Klebsiella* sp. *Ps. aeruginosa* and *Enterococcus* sp. The cultivation details for the strains have been described by Goodacre *et al*. (1998).

For the completeness, briefly introduce FT-IR spectra data collection process given by (Goodacre *et al*., 1998). About 10 μL of each bacteria sample was evenly applied onto a sand-blasted aluminum plate. Prior to the spectra analysis, the samples were oven-dried at 50°C for 30 min. Samples were then run in triplicate. The instrument used was a Bruker IFS28 FT-IR spectrometer equipped with an MCT detector cooled with liquid N_2. The aluminum plate was then loaded onto the motorized stage of a reflectance TLC accessory. Spectra were collected over the wave number range 4000-600 cm^{-1}. The spectra were acquired at a rate of 20 sec^{-1} and the spectral resolution used was 4 cm^{-1}. Each sample was displayed in terms of absorbance as calculated from the reflectance-absorbance spectra. The typical FT-IR spectra are shown in Fig. 1.

To minimize the problems arising from baseline shifts, the following procedure was implemented following the steps in (Goodacre *et al*., 1998; Jarvis and Goodacre, 2004) the spectra were normalized so that the smallest absorbance was set to 0 and the highest to +1 for each spectrum; the normalized spectra were detrended by subtracting a linearly increasing baseline from 4000-600 cm^{-1}; finally the smoothed first derivatives of these normalized and detrended spectra were calculated using the Savitzky-Golay algorithm with 5 point

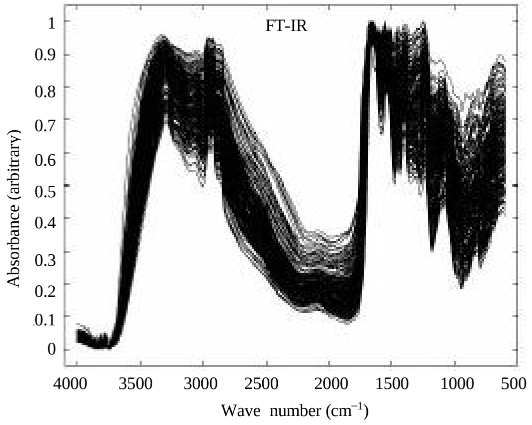

Fig. 1: FT-IR spectra for UTI bacteria

smoothing. The multivariate analysis Matlab routines provided by the researcher of (Goodacre *et al.*, 1998) are exploited.

SUPPORT VECTOR MACHINE: A SHORT INTRODUCTION

Brief review: For binary pattern classification, the essence of SVM is to find the optimal separating hyperplane that separates the positive and negative examples with maximal margin (Scholkopf and Smola, 2000). Usually, the classification decision function for a linearly separable problem can be represented by:

$$f = \text{sign}(w \cdot x + b) \tag{1}$$

SVM is based on the structural risk minimization principle (Cristianini and Shawe-Taylor, 2000) which determines the classification decision function by minimizing the empirical risk:

$$R = \frac{1}{l}\sum_{i=1}^{l}\left|f(x_i) - y_i\right| \tag{2}$$

where, l represent the size of examples. The optimal separating hyperplane is determined by giving the largest margin of separation between different classes. This optimal hyperplane bisects the shortest line between the convex hulls of the two classes. The optimal hyperplane is required to satisfy the following constrained minimization as:

$$\begin{aligned} &\text{Minimize}: \frac{1}{2}w^T w \\ &\text{subject to} \ : y_i(w \cdot x_i + b) \geq 1, i = 1,\ldots,l \end{aligned} \tag{3}$$

The SVM classifier is then obtained by the inner product $x_i^T x$, where $i \in S$, the set of support vectors. However, it is not necessary to use the explicit input data to form the classifier. Instead, all that is needed is to use this inner products between the support vectors and vectors of the feature space. That is by defining the kernel:

$$K(x_i, x) = x_i^T x \tag{4}$$

And a nonlinear classifier can be obtained as:

$$f(x) = \text{sign}\{\alpha_0 y_i K(x_i,) + b_0\} \tag{5}$$

There are three most popular kernels, the polynomial, Gaussian and the tanh kernel. The polynomial kernel of degree d is given by:

$$k\ (x,z) = (x \cdot z + c)^d \tag{6}$$

where, c is a constant, d can be user defined. When d is 1, the kernel becomes linear. The Gaussian kernel is:

$$k\ (x,z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) \tag{7}$$

Where the parameter σ controls the support region of the kernel. The tanh kernel is given by:

$$k\ (x,z) = \tanh\ (x \cdot z + b)$$

Multi-class support vector machine: The aforementioned support vector machines were primarily designed for binary pattern classification problems. A variety of schemes have been proposed in the literature for solving multi-class problem (Hsu and Lin, 2002; Franc and Hlavac, 2002). Commonly used techniques include: one-against-all (OAA), one-against-one (OAO); the Directed Acyclic Graph (DAG); Error Correcting Output Coding (ECOC) and Multiclass objective function by adding Bias to the objective function (BSVM). Their brief description is given in the following.

One-against-all: In this approach, a SVM is constructed for each class by discriminating that class against the remaining (l-1) classes. The number of SVMs used in this approach is l. A test pattern x is classified by using the winner-takes-all decision strategy, i.e., the class with the maximum value of the discriminant function f (x) is assigned to it.

One-against-one: This strategy consists in constructing one SVM for each pair of classes. Thus, for a problem with l classes, l (l-1)/2 SVMs are trained to distinguish the samples of one class from the samples of another

Table 1: Code matrix for 5 class with 15 bits

Code matrix	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}
C_1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C_2	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
C_3	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1
C_4	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1
C_5	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

class. Usually, classification of an unknown pattern is done according to the maximum voting where each SVM votes for one class.

Directed acyclic graph: The Directed Acyclic Graph (DAG) SVM was proposed in (Platt *et al*., 2000). In the training phase, it works as the one-against-one method solving l (l-1)/2 binary SVMs. However, in the testing phase, it uses a rooted binary DAG which has l (l-1) = 2 internal nodes and l leaves. Given a test sample x, starting at the root node, a pair-wise SVM decision is made and either class is rejected. Then it moves to either left or right depending on the result and continues until reaching to one of leaves which indicates the predicted class. So the DAG SVM requires l-1 comparisons and hence is more efficient than the one-against-one method.

Error correcting output coding: Error Correcting Output Coding (ECOC) was proposed in (Dietterich and Bakiri, 1995) by decomposing the l-class problem into a set of binary subproblems, training the resulting base classifiers and then combining their outputs to predict the class label. Each of the involved classes is assigned a binary codeword following some specified coding scheme. Several methods have been proposed to generate such error correcting code. Among them the exhaustive coding (Dietterich and Bakiri, 1995) was often utilized. For l classes, all the possible different classifier arrangements are exhaustively used in the code matrix. For the FT-IR spectra classification task, there are 5 classes, resulting in 2^{5-1}-1 = 15 different arrangements in total.

Table 1 shows the code matrix for a task with 5 classes (C_i) using 15 base classifiers (S_i). In the code matrix above, each class C_i is associated with a codeword (i.e., the column vector). Each classifier S_i is then trained to perform a binary classification task that is to distinguish the two subsets of the classes labeled with 1 and 0, respectively. During testing, a vector of scores is generated by the 5 binary classifiers for each test sample. This vector is then compared to each codeword and the one with the minimum distance is chosen as the hypothesis.

Multi-class BSVM algorithm: Instead of creating many binary classifiers to determine the class labels, this method attempts to directly solve a multiclass problem (Hsu and Lin, 2002; Weston and Watkins, 1998) by adding bias to the binary class objective function. The modified objective function allows simultaneous computation of multiclass classification and is given by:

$$(w, b, \xi) = \operatorname{argmin}_{w,b} = \frac{1}{2}\sum_{y}(\|w_y\|^2 + b_y^2) + C\sum_{i\in I}\sum_{y\in Y\setminus\{y\}} \varsigma^r$$

Subject to the constraints,

$$w_{y_i} \cdot x_i + b_{y_i} \geq w_r \cdot x_i + b_r + 2 - \varsigma_i^T$$

where, $\xi(\xi \geq 0)$ is a vector of slack variables and I = {1(...)1} is set of indices.

SOME CONVENTIONAL MULTI-CLASS PATTERN CLASSIFICATION METHODS COMPARED

In machine learning, there are some conventional methods that have been extensively applied for multi-class classification problems (Duda *et al*., 2001; Theodoridis and Koutroumbas, 2000) for example, the k-Nearest Neighbor classifier (kNN), the Nearest Mean Classifier (NMC) and artificial neural networks (e.g., Perceptron, multilayer Perceptrons). In the following we briefly summarize a few of the most commonly used methods for comparison purpose.

K-nearest neighbor classifier: kNN classifier is a prototype-based classifier among the oldest types of classification methods. It is based on a distance function, for example, the Euclidean distance for pairs of data samples.

The kNN classifies a test sample on the basis of the training set by first finding the k closest samples in the training set and then predicting the class by the majority voting. In other words, the class that is most common among those k neighbors is chosen as the predicted label. Obviously the kNN classifier needs to access all learning data at the time when a new test case is to be classified.

Nearest mean classifier: The Nearest Mean Classifier is another traditional prototype-based classifier (Duda *et al*., 2001). Being different with kNN classifier which uses all training data to label a test sample, the NMC abstract training data first by only storing the mean of each class, i.e., one prototype per class. It then classifies a test sample with the label of the nearest class prototype.

Perceptron: The classical perceptron algorithm (Duda *et al*., 2001) is a very simple classification algorithm which maps an input x to an output value f (x):

$$f(x) = \begin{cases} 1 & \text{if } \ w \cdot x + b > 0 \\ 0 & \text{else} \end{cases}$$

Where:

w = Avector of real-valued weights
b = The bias, a constant term that does not depend on any input value

The value of f (x) (0 or 1) is used to classify x as either a positive or a negative instance in the case of a binary classification problem.

There are several adaptations of the Perceptron algorithm to multiclass settings (Haykin, 1998) for example by the Kesler's construction which convert a multi-class error-correction procedures to two-class procedures.

Multi-layer perceptron: The Multilayer Perceptrons (MLP) is a common type of neural network classifier which is often trained by the error back-propagation algorithm (Haykin, 1998). It consists of a layer of input nodes, each linked by weighted connections to every one of a layer of hidden nodes, each of which is linked in turn to a set of output nodes. It has been shown that MLPs can virtually approximate any function with any desired accuracy provided that enough hidden units and enough data are given (Haykin, 1998). Therefore, it can also implement a discrimination function that separates input data into classes. Such an ability of an MLP to learn from data makes it a practical classifier for many classification tasks.

There are a number of articles where MLPs have been employed in the identification of microorganisms using a variety of data (Goodacre *et al*., 1998; Jarvis and Goodacre, 2004; Maquelin *et al*., 2002; Mouwen *et al*., 2006; Oberreuter *et al*., 2002; Rebuffo *et al*., 2006; Perkins *et al*., 2005; Timmins *et al*., 1998; Wenning *et al*., 2002; Rosch *et al*., 2005; Lasch *et al*., 2006; Ellis and Goodacre, 2006; Ellis *et al*., 2002). Though successful for many applications, MLP classifier has several limitations and training an MLP network involves a considerable degree of empiricism. And the performance often depends on the nature and quality of the data on which it is trained for example, the classification accuracies may be sensitive for different class frequencies in the training set.

EXPERIMENTS

The FT-IR spectra data Eq. 1 belong to five isolates, i.e., *E.coli*, *Pr. mirabilis*, *Klebsiella* sp. *Ps.aeruginosa and Enterococcus* sp. There are total 236 samples available, each with 882 wave numbers. The distribution of different isolates are shown in the Table 2.

Table 2: Isolate sample distributions

Species	Values
E. coli	68
Pr.mirabilies	40
Klebsiella sp.	40
Ps. æruginosa	40
Enterococcus sp.	48
Total	236

All of the data were pre-processed following the steps by (Goodacre *et al*., 1998), consisting of normalization and smoothing. The experiment settings for all the classifiers are summarized as follows. For MLP, we experimented with a three-layer network with the same structure and algorithm as used in (Goodacre *et al*., 1998). Specifically, the number of inputs is the same as the number of features (i.e., 882 for raw FT-IR spectra or the first several PCA projections), one hidden layer with 20 units and a single linear unit representing the class label. It is the experience that varying the number of hidden units in such an MLP usually does not significantly change the performance. All of the support vector machine classifier were optimized by quadratic programming (Dietterich and Bakiri, 1995). For kNN classifier, chosen k = 3 after testing a range of different values of k with the similar results.

In many applications, the number of variables in a multivariate data set needs to be reduced due to the existence of irrelevant, noisy and redundant information in the data which has been proved to be the detrimental elements leading to the inaccuracies in classification performance. Moreover, as the number of features used for classification increases, the number of training samples required for statistical modelling and/or learning systems grows exponentially (Duda *et al*., 2001). Principal Component Analysis (PCA) is an efficient dimensionality reduction technique which carry out linear transformation of data and project it to a lower dimensional subspace in such a way that most of the information is retained while discarding the noisy component of data.

The use of PCA scores as inputs to MLP classifier has been previously exploited for the identification of bacteria from their FT-IR spectra (Goodacre *et al*., 1998; Jarvis and Goodacre, 2004). Following the discussion in (Goodacre *et al*., 1998) is chosen the first 20 PCs (96.88% of total variance retained) for the FT-IR spectra data for comparison purpose.

There are many standard procedures to test the performance of a pattern classification scheme. The commonly used ones are holdout and k-fold cross-validation methods. The k-fold cross-validation is an established technique for estimating the accuracy of a classifier. In general, all of the examples are partitioned into k subsamples of which the kth subsamples is retained for testing the model while the remaining k-1 subsamples are used as training data. The cross-validation is then repeated k times with all of the k subsamples used exactly once as the validation data. The cross-validation estimation of accuracy is the overall number of correct classifications divided by the number of instances in the data set. The holdout method is the simplest kind of cross-validation (2-fold) with the data set being separated into training set and testing set.

Classification performance from holdout experiment: In the first experiment, we compared a SVM classifier with several other methods including kNN, NMC, perceptron and MLP based on the random divisions of the dataset into a training set (80%) and a test set (20%). The SVM applied is based on the DAG scheme with values of the regularization parameter C = 10 and sigma parameter ($\sigma^2 = 1$) when using the radial basis function kernel. The values are from the so-called grid search (Cristianini and Shawe-Taylor, 2000).

About 100 random splitting were repeated and the classification results on the testing set were recorded and averaged. For each random testing, the same set of training/testing were applied to the five classifiers. The experiments were conducted independently with original FT-IR spectra data and the first 20 PCA scores, respectively. The results are displayed in the boxplots as shown in Fig. 2 which gives the statistical means and standard deviations of accuracy over the 100 repeated random sampling.

It can be observed that for both of the original UT-IR spectra data and the PCA scores, support vector machine gives the best classification performance with regard to the accuracy and the standard deviation. With both of the original data and the PC scores, the nearest mean classifier gives the worst result (65%).

The benefit of applying PCA projection is obvious that all the classifiers applied produce improved classification accuracies and gap between SVM and MLP become narrowed. Without PCA projection, the mean classification accuracies from MLP and SVM are 81 and 98%, respectively. With PCA preprocessing, the accuracies are 97 and 99.2%.

We also numerically studied the performance of the five different multi-class SVM methods discussed in the previous section namely, DAG, ECOC, OAA, OAO and BSVM. For all these methods, the RBF kernel is employed. Each classifier requires the selection of two hyperparameters: a regularization parameter C and a kernel parameter σ^2. the procedure made by Hsu and Lin is followed (Hsu and Lin, 2002) and take the C = 10 and $\sigma^2 = 1$ of all the binary classifiers within a multiclass method to be the same.

From Fig. 3 it is shown that DAG SVM gives the best classification results. For DAG approach, an accuracy of 97% is achieved. The OAA method is very close to the BSVM with accuracy of 90% which are all better than the OAO SVM. Further, results using exhaustive technique based ECOC approach are not significantly better in comparison to OAA or BSVM in terms of classification accuracy. The results from applying PCA projections are consistent with that of applying raw data as illustrated in Fig. 2.

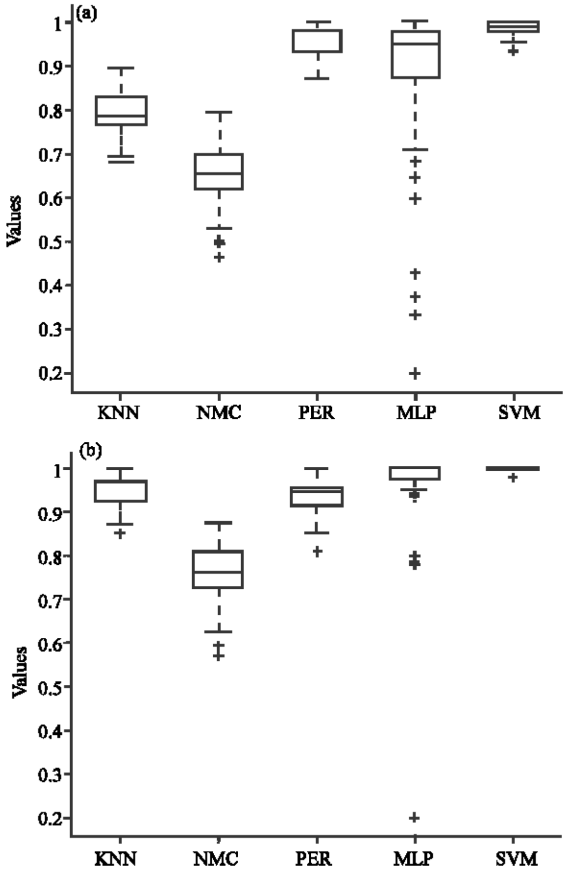

Fig. 2: Boxplots of the classification accuracies from the five different classifiers. About 80% of data were used for training while the remaining for testing. The results were from the average of 100 tests. The SVM used is based on DAG scheme. (a) Results from original UTI data and (b) Results from using first 20 PCA projections

The confusion matrices that summarize the details of comparisons of MLP and SVM from the above experiment are shown in Table 3 and 4. SVM model selection: Tuning the hyperparameters of a Support Vector Machine (SVM) classifier is a crucial step in order to establish an efficient classification system. Generally, at least two parameter values have to be chosen carefully in advance. They concern respectively the regularization parameter C which sets the trade-off cost between the training error and the complexity of the model and the kernel function parameter (s), reduced to the bandwidth in the classical case of a radial basis function kernel (σ). The problem of choosing

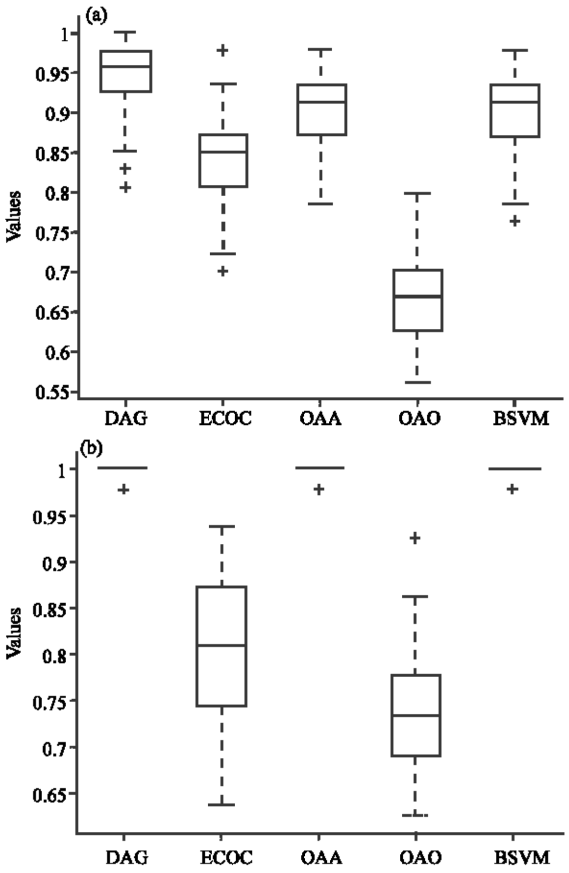

Fig. 3: Boxplots of classification accuracies from the different multiclass SVM schemes: DAG, ECOC, OAA, OAO and BSVM. About 80% of data were used for training while the remaining for testing. The results were from the average of 100 tests. (a). Results are from original UTI data; (b) Results from using first 20 PCA projections

these parameters values is called model selection in the literature and its results strongly impact the performance of the classifier. Cross-validation has often been applied to SVM model selection with the guideline that the average cross-validation error is minimum. In the study we followed the conventional grid search method which select the parameters empirically by trying a finite number of values and keeping those that provide the least test error. The next experiment applied the k-fold cross-validation for two multi-class SVM schemes, i.e., DAG SVM and BSVM with different kernels and regularization parameter C are compared.

As explained earlier, k-fold cross-validation describes the performance of a give SVM model with the chosen regularisation parameter with a separate test set that is not used during training. Figure 4a, b display the boxplots of classification results from 10-fold cross-validation and 5 fold cross-validation of the support vector machines with different kernel and regularization parameters. The abbreviations are: DAG1, BSVM1-RBF kernel with $\sigma = 0.5$, $C = 10$, DAG2, BSVM2-- RBF kernel with $\sigma = 0.5$, $C = 100$, DAG3, BSVM3-polynomial kernel with $d = 2$, $C = 10$, DAG4, BSVM4-polynomial kernel with $d = 2$, $C = 100$.

In the implementation of a k-fold cross-validation of 236 samples, the indices containing approximately equal proportions of the integers 1 through k were first generated in each test, which define a partition of the 236 samples into k disjoint subsets. Thus for 10 fold cross- validation for each division of the samples, a model is developed with 213 samples and tested in the rest 23 samples. This process of randomly generated partitions is repeated 20 times with randomly chosen training and testing sets giving up 200 unbiased estimates of discriminant ability. As the test samples are

Table 3: Confusion matrix from the classification by MPL using the original data

	Predicated class					
Actual class	*E. coli*	*Pr. mirabilies*	*Klebsiella* sp.	*Ps. aeruginosa*	*Enterococcus* sp.	Accuracy (%)
E. coli	49	1	2	1	1	89.1
Pr. mirabilies	2	27	1	2	2	79.4
Klebsiella sp.	3	1	25	2	2	75.8
Ps. aeruginosa	4	2	2	24	1	72.7
Enterococcus sp.	1	0	2	0	36	92.3

Table 4: Confusion matrix from the classification by DAG SVM using the original UTI data

	Predicated class					
Actual class	*E. coli*	*Pr. mirabilies*	*Klebsiella* sp.	*Ps. aeruginosa*	*Enterococcus* sp.	Accuracy (%)
E. coli	54	0	1	0	0	98
Pr. mirabilies	1	31	0	1	0	94
Klebsiella sp	2	0	29	0	1	90
Ps. aeruginosa	1	2	1	29	0	85
Enterococcus sp.	1	0	0	0	37	97

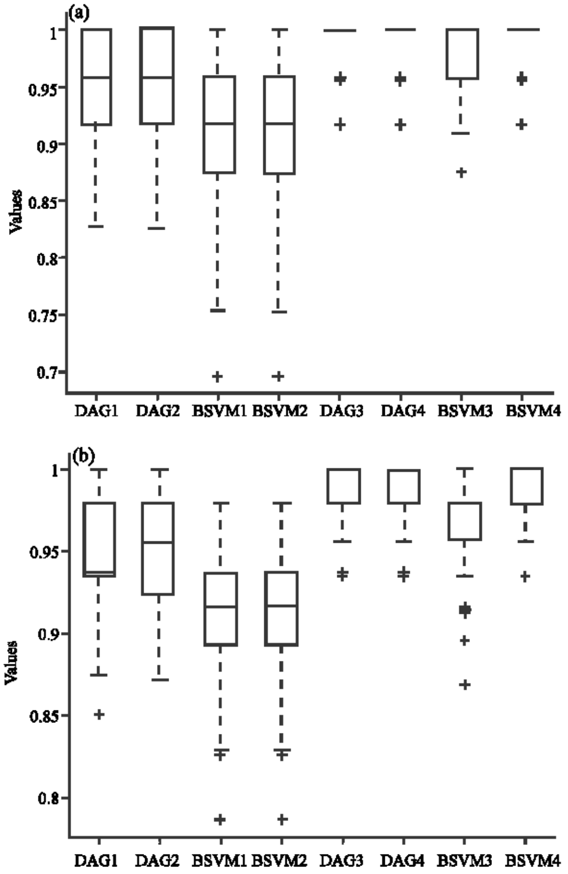

Fig. 4: Boxplots for comparing classification accuracies in terms of kernel functions and regularization parameter. (a) From 10 fold cross-validation; (b) From 5 fold cross-validation

independent of the training data, the results derived from this 10-fold cross-validation are reliable. The experiments demonstrated that with all of the original features, the polynomial kernels showed better performance compared to the RBF kernel and the regularization parameter C can be chosen in a wide range without obvious change of the performance.

CONCLUSION

The ability to identify pathogenic organisms rapidly provides significant benefits to clinicians for example it may help to give better prescription practices and tracking of recurrent infections. Conventional bioassays require a long period (35 days) before identification of an organism can be made, thus compromising the effectiveness with which patients can be treated for bacterial infections.

In this study we have investigated utilizing SVMs for multiclass UTI bacteria classification. Various multi-class SVM techniques have been studied, including one-against-one, one-against-all, direct multiclass SVM objective function (BSVM), the Directed Acyclic Graph and the combination of multiple binary SVM classifiers via the general ECOC scheme. Experiment results demonstrated that the DAG SVM is much better than the other SVM schemes. The study confirmed that SVM as an advanced machine learning system can deliver state-of-the-art performance in the rapid identification of FT-IR based UTI bacteria.

REFERENCES

Comak, E., K. Polat, S. Gunes and A. Arslan, 2007. A new medical decision making system: Least square support vector machine (LSSVM) with fuzzy weighting pre-processing. Expert Syst. Appl., 32: 409-414.

Cristianini, N. and J. Shawe-Taylor, 2000. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. 1st Edn., Cambridge University Press, Cambridge.

Dietterich, T. and G. Bakiri, 1995. Solving multi-class learning problem via error-correcting output codes. J. Artificial Intel. Res., 2: 263-286.

Duda, R.O., P.E. Hart and D.G. Stork, 2001. Pattern Classification. 2nd Edn., John Wiley and Sons, New York.

Ellis, D. and R. Goodacre, 2006. Metabolic fingerprinting in disease diagnosis: Biomedical applications of infrared and Raman spectroscopy. Analyst, 131: 875-885.

Ellis, D.I., D. Broadhurst, B.D. Kell, J.J. Rowland and R. Goodacre, 2002. Rapid and quantitative detection of the microbial spoilage of meat by fourier transform infrared spectroscopy and machine learning. Applied Environ. Microbiol., 68: 2822-2828.

Franc, V. and V. Hlavac, 2002. Multi-class support vector machine. Proc. 16th Int. Conf. Patt. Recognit., 2: 236-239.

Furey, T., N. Cristianini, N. Duffy, D. Bednarski, M. Schummer and D. Haussler, 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16: 906-914.

Goodacre, R., E.M. Timmins, R. Burton, N. Kaderbhai, A.M. Woodward, D.B. Kell and P.J. Rooney, 1998. Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. Microbiology, 144: 1157-1170.

Haykin, S., 1998. An Introduction to Neural Networks: A Comprehensive Foundation. 2nd Edn., Prentice-Hall, Upper Saddle River, NJ.

Hsu, C.W. and C.J. Lin, 2002. A comparison on methods for multi-class support vector machines. IEEE Trans. Neural Netw., 13: 415-425.

Jarvis, R.M. and R. Goodacre, 2004. Ultra-violet resonance Raman spectroscopy for the rapid discrimination of urinary tract infection bacteria. FEMS Microbiol. Lett., 232: 127-132.

Lasch, P., M. Diem, W. Hansch and D. Naumann, 2006. Artificial neural networks as supervised techniques for FT-IR microspectroscopic imaging. J. Chemom., 20: 209-220.

Magee, J.T., 1993. Whole-Organism Fingerprinting. In: Handbook of New Bacterial Systematics, Goodfellow, M. and A.G. O'Donnell (Eds.). Academic Press, London, pp: 383-427.

Maquelin, K., C. Kirschner, L.P. Choo-Smith, N. van de Braak, H.P. Endtz, D. Naumann and G.J. Puppels, 2002. Identification of medically relevant microorganisms by vibrational spectroscopy. J. Microbiol. Meth., 51: 255-271.

Morgan, M.G. and H. McKenzie, 1993. Controversies in laboratory diagnosis of community acquired urinary tract infection. Eur. J. Clin. Microbial. Infect. Dis., 12: 491-504.

Mouwen, D.J., R. Capita, C. Alonso-Calleja, J. Prieto-Gomez and M. Prieto, 2006. Artificial neural network based identification of *Campylobacter* species by Fourier transform infrared spectroscopy. J. Microbiol. Meth., 67: 131-140.

Oberreuter, H., H. Seiler and S. Scherer, 2002. Identification of coryneform bacteria and related taxa by Fourier-transform infrared (FT-IR) spectroscopy. Int. J. Syst. Evol. Microbiol., 52: 91-100.

Perkins, D.L., C.R. Lovell, B.V. Bronk, B. Setlow, P. Setlow and M.L. Myrick, 2005. Classification of endospores of *Bacillus* and *Clostridium* species by FT-IR reflectance microspectroscopy and autoclaving. Proceedings of the IEEE International Workshop on Measurement Systems for Homeland Security, Contraband Detection and Personal Safety Workshop, March 29-30, Orlando, Florida, USA., pp: 81-87.

Platt, J., N. Cristianini and J. Shawe-Taylor, 2000. Large margin DAGs for multiclass classification. Proc. Neural Inform. Process. Syst., 12: 547-553.

Rebuffo, C.A., J. Schmitt, M. Wenning, F. von Stetten and S. Scherer, 2006. Reliable and rapid identification of *Listeria monocytogenes* and *Listeria* species by artificial neural network-based Fourier transform infrared spectroscopy. Applied Environ. Microbiol., 72: 994-1000.

Rosch, P., M. Harz, M. Schmitt, K.D. Peschke and O. Ronneberger *et al*., 2005. Chemotaxonomic identification of single bacteria by micro-raman spectroscopy: Application to clean-room-relevant biological contaminations. Applied Environ. Microbiol., 71: 1626-1637.

Scholkopf, B. and A.J. Smola, 2000. Learning with Kernels. MIT Press, Cambridge.

Theodoridis, S. and K. Koutroumbas, 2000. Pattern Recognition. Academic Press Inc., San Diego, Calif.

Timmins, E.M., S.A. Howell, B.K. Alsberg, W.C. Noble and R. Goodacre, 1998. Rapid differentiation of closely related *Candida* species and strains by pyrolysis-mass spectrometry and fourier transform-infrared spectroscopy. J. Clin. Microbiol., 36: 367-374.

Wenning, M., H. Seiler and S. Scherer, 2002. Fourier-transform infrared microspectroscopy, a novel and rapid tool for identification of yeasts. Applied Environ. Microbiol., 68: 4717-4721.

Weston, J. and C. Watkins, 1998. Multi-Class Support Vector Machines, CSD-TR-98-04, Royal Holloway. 1st Edn., Department of Computer Science, University of London, London.

Winder, C.L., E. Carr, R. Goodacre and R. Seviour, 2004. The rapid identification of *Acinetobacter* species using fourier transform infrared spectroscopy. J. Applied Microbiol., 96: 328-339.