Hybrid SVM/CART Classification of Pathogenic Species of Bacterial Meningitis with Surface-Enhanced Raman Scattering

Chung-Yueh Huang*, Tsung-Heng Tsai†, Bing-Cheng Wen‡, Chia-Wen Chung‡,
Yung-Jui Li‡, Ya-Ching Chuang‡, Wen-Jie Lin‡, Li-Li Li*,
Juen-Kai Wang*, Yuh-Lin Wang*, Chi-Hung Lin‡, Da-Wei Wang‡

*Institute of Atomic and Molecular Sciences, Academic Sinica, Taipei, Taiwan, Email: bugmoon0509@iis.sinica.edu.tw

†Institute of Information Science, Academic Sinica, Taipei, Taiwan, Email: wdw@iis.sinica.edu.tw

‡Institute of Microbiology and Immunology, National Yang-Ming University, Taipei, Taiwan

Abstract—Bacterial meningitis is still a life-threatening disease, and early diagnosis of pathogen can be crucial to improving survival rate. Using the surface-enhanced Raman scattering (SERS) platform developed by our group, the pathogens can be differentiated on the basis of their SERS spectra which are believed to related to their surface chemical components. We collected the SERS spectra of ten pathogens: *Streptococcus pneumoniae* (Spn), *Streptococcus agalactiae* (group B streptococcus, GBS), *Staphylococcus aureus* (Sa), *Pseudomonas aeruginosa* (Psa), *Acinetobacter baumannii* (Ab), *Klebsiella pneumoniae* (Kp), *Neisseria meningitidis* (Nm), *Listeria monocytogenes* (Lm), *Haemophilus influenzae* (Hi), and *Escherichia coli* (E.coli). These samples were obtained from patients in National Taiwan University Hospital, and were believed to represent the real diversity of clinical pathogens. Using the support vector machine (SVM) method, the classification accuracy can achieve around 88%. However, we noted that SVM cannot distinguish between [E.coli, Kp] and [Sa, Hi] due to the fact that the global features of these two groups of pathogens are very similar. We therefore incorporated a classification tree method that can focus on local differences in classification rules. This improved the accuracy to 90%. To get a better understanding of the SERS signals, we also compared several other classification methods. In addition, rule extraction method which attempts to explain why classifier fail or succeed is also discussed. Our preliminary results are interesting, encouraging, and await more thorough investigation.

Keywords—SVM; Hybrid SVM; SERS; CART;

I. INTRODUCTION

Surface-enhanced Raman scattering (SERS) overcomes the limitation of low scattering efficiency of Raman spectroscopy and facilitates a fast and convenient identification of vibration signal of molecules from chemical or biological entities [1], [2]. In this paper, we focus on bacterial meningitis application, since bacterial meningitis is still a lethal disease, and early identification of the species of the pathogen plays a vital role in devising an effective treatment plan. An improved SERS platform [3], [4] were utilized to acquire SERS spectrum samples. Since the samples were taken from patients, this naturally represents a more realistic yet challenging scenario where ten common bacterial meningitis pathogens from National Taiwan University Hospital including Spn, GBS, Sa, Psa, Ab, Kp, Nm, Lm, Hi, and E.coli were collected for this study.

Support vector machine (SVM) have been applied to Raman spectrum classification [5], [6]. The previous studies focused on classification with Raman intensity feature. However, from our experience, we noticed that when global intensity profiles are similar, SVM suffers ambiguity. To overcome the limitation, our spectra were subjected to peak identification process [7], [8], to produce a peak position sequence - a sequence of integers where peaks appeared. With the intensity sequence and peak sequence, we address the classification problem by a hybrid SVM/CART model where two-level procedure is considered [9]. For better interpretation of these models, it is desired to generate or extract rules from models derived. We describe some preliminary observations and explanations of the behavior of trained classification models.

The remainder of the paper is organized as follows. In Section II, we review SVM and CART. The concept and implementation procedure of hybrid SVM/CART are described in Section III. The results of our approach are examined and compared with other methods in Section IV. Lastly, we discuss relevant issues and future work in Section V.

II. BACKGROUND

SVM is a supervised learning method proposed by Cortes and Vapnik in 1995 [10]. The idea of SVM is to map input vectors into a high-dimensional feature space, and then to find a hyperplane that separates the mapped features with maximum soft margin [11]. Being a binary classifier, SVM can be applied to problems of multi-classes with one-versus-one (OVO) or one-versus-all (OVA) strategy. The idea of the strategy is to decompose a multi-class problem into many pairwise problems, and combine these binary classification results for a final prediction. LIBSVM [12] was used as our SVM implementation tool and OVO strategy was applied. Classification and Regression Tree (CART) was proposed by Breiman in 1984 [13]. The purpose of CART is to construct prediction models from data. The models are trained by partitioning the data space recursively and search a simple “if then” rule within each partition. In this discussion, we want to search rules like “if there is a peak at certain

Table I
MENINGITIS SERS DATA SET USED IN THIS STUDY.

Species	# ATCC Tasks	# Clinical Tasks	# Tasks	# Spectra
Ab	1	11	12	376
E.coli	1	11	12	456
Hi	3	2	5	191
Kp	1	10	11	376
Lm	1	8	9	317
Nm	5	0	5	135
Psa	1	8	9	358
S pn	6	11	17	454
Sa	1	8	9	367
GBS	1	10	11	489

wavelengths, then this spectrum belongs to some species.” for each pairwise classification. CART was pruned according to Occam’s razor principle, and the number of leaf nodes is equal to the number of sample classes after CART pruning in this study. Only two leaf nodes were preserved for a binary-class CART. This CART model used only one rule, or one peak, to separate two-species spectrums. This two leaf CART model has lower complexity than SVM with intensity feature. We wish this CART model can focus on local yet reproducible features and compensate for the overfitting risk of SVM especially when global spectral patterns are similar and the number of training data is limited.

III. MATERIAL AND METHODS

Meningitis SERS spectra were collected, classified and the performance and rules of final classifiers were compared.

A. Samples

National Taiwan University Hospital collected 79 clinical samples comprising ten species of bacteria, which are common pathogens causing meningitis. In addition to these clinical samples, we also analyzed 17 standard bacteria samples from American Type Culture Collection (ATCC). For each sample, a set of SERS spectra were acquired in one experiment. We call such set of spectra a *task*.

B. SERS Platform

Bacterial suspension (10 μ l) were placed on the Ag-nanoparticles imbedded in anodic aluminum oxide substrate with nanochannels developed by our laboratory [3]. The spectrum was collected by Raman spectromicroscopy, (HR800, Jobin-Yvon) equipped with a HeNe laser at 632.8 nm and NA 0.95 100x water-immersion objective lens. The Laser power intensity is about 10^5 W/cm². Raman signals were collected from the information-rich part of the spectrum between 400 and 1600 cm⁻¹ for integration time of 1 to 3 seconds. A series of preprocessing steps were performed and details can be found in our previous reports [7], [8]. The number of tasks and number of spectra for each species are listed in Table I. As implied by the number of tasks, these results are preliminary.

C. Features and Classification Methods

Spectral intensity features are used for the hybrid SVM/CART model in the first level classification. In the second level classification, spectrum is further processed, going through the peak identification proposed by our previous works [7], [8], to produce a peak sequence. Thus, each raw spectrum is associated to two representative feature vectors, namely the intensity sequence and the peak sequence. We want to search rules like “*If there is a peak at a particular wavelength, then the spectrum belongs to a specific species*” in the second level. Unfortunately, as is the case with most biological spectrum, peak positions often shift. Thus, before finding rules associated to the above description, we have to first perform peak alignment. In our approach, we simply assume that peak shift is within a certain range, i.e. δ nm. When peak location difference between peaks of any two spectra is smaller than δ nm, we consider these peaks as identical. The same idea is used in [14]. We hope δ is small enough to have irrelevant peaks excluded. On the other hand, we hope δ is large enough to reasonably capture the peak shift. Here we have simply chosen δ as ten times the spectrum resolution units, i.e., about 11.36 nm, according to our empirical observations of peak distribution. Only the peak position feature was considered; other peak information, such as intensity and width, were not used here.

As the number of attributes, 1060, is larger than the number of binary training samples, less than 800, we simply chose linear kernel SVM in the first level since it is not necessary to map our sample vectors into higher dimensional feature spaces. The pretest result of radial basis function kernel SVM also supported this argument. Another advantage of using linear kernel SVM is that we can extract rules simply from the normal vector of hyperplane [5]. In the second level classification, a set of less confident binary SVM models are identified and the worse binary SVM models will be replaced with CART. In our discussion, less confident models represent those yielded G score on the validation are lower than $\mu - \sigma$, where G is the geometric mean of sensitivity and specificity [15], μ is the average G score of all binary SVM, and σ is the G score standard deviation of all binary SVM. We only checked the less confident binary SVM to reduce training time. Four-fold cross-validation was applied to model selection in this stage. After finishing this comparison, we retrained the preferred binary classifiers, SVM or CART, with all available training data. For the testing process, each binary classifier would label all samples in each task and the majority of predicted results in that task will be the final task label.

D. Validation

To validate our models, we separated spectra into training/testing sets according to their task labels. For examples, if a spectral data set which included 200 spectra of two species bacteria was collected from ten tasks, task based

Table II
MEAN ACCURACY FOR MULTI-CLASS PROBLEMS.

10 Classes	Spectral Accuracy	Task Accuracy
MCSVM(s)	0.816 ± 0.072	0.880 ± 0.068
MCSVM(p)	0.705 ± 0.041	0.880 ± 0.068
CART(p)	0.615 ± 0.073	0.820 ± 0.067
INN(s)	0.676 ± 0.063	0.800 ± 0.084
Linear Naive Bayes(s)	0.632 ± 0.049	0.770 ± 0.087
SVM(s)+CART(p)	0.827 ± 0.069	0.900 ± 0.063

partition strategy for five-fold cross-validation would assign samples of 8 tasks, 4 tasks for each species, for training, and assign samples of 2 tasks, 1 task for each species, for testing each turn. This strategy can guarantee that classifiers would not "see" training samples that came from the same task of testing data and it is also much closer to the case of clinical application than task-regardless strategy.

We compared the five-fold cross-validation accuracy of several models: KNN, multi-class c-SVM with spectrum and peak feature, linear naive Bayesian classifier (LNBC) with spectrum feature, and CART with peak feature. In the KNN model, K ranges from 1 to 10, and we only listed the best result. In the first level classification of hybrid SVM/CART, the linear kernel and $C=10$ (searched through $10^{-5}, 10^{-4}, \dots, 10^5$ with first training data set) were used. In the second level CART classification, task based partition method was also implemented. We used the Matlab 2008.b Statistic Toolbox to perform KNN, along with the linear naive Bayesian classifier, and CART [16].

IV. RESULTS AND DISCUSSION

The five-fold cross-validation accuracy for various machine learning algorithms are summarized in Table II. We used the notation *method(feature)* to denote the classification method and feature type respectively. We note that the average task accuracy of the two-level classifier, SVM(s)+CART(p), is $90.00 \pm 0.063\%$ (the number after " $\pm$ " is the standard deviation of five-fold accuracy. "s" represent intensity feature, and "p" represent peak feature.), which is the best accuracy rate among the other one level classifiers listed in Table II. This result is consistent with previous discussion [9] that showed advantage of a hybrid model. Next, we want to investigate why hybrid model is better. We examined all 45 pairwise comparisons among ten species and noted that SVM usually performed well, only 2 pairs out 45 with accuracy lower than 85%. They are [E.coli, Kp] and [Hi, Sa]. We therefore focused on these 2 pairs and the results are shown in Table III. The binary CART classifier with the peak feature is preferred when in consideration of classification between E.coli and Kp. The accuracy of CART(p) was 0.857 ± 0.035 , which was better than the accuracy of the binary SVM, 0.802 ± 0.100 . And "the peak at 1244 nm" was usually a criterion of the single-node CART in each cross-validation run. We plotted the peak distribution of each spectrum of E.coli and Kp in Fig. 1, where each row

Table III
MEAN ACCURACY FOR TWO-CLASS PROBLEMS.

Species	E.coli vs. Kp	Hi vs. Sa
SVM(s)	0.802 ± 0.100	0.832 ± 0.194
SVM(p)	0.775 ± 0.035	0.783 ± 0.132
CART(p)	0.857 ± 0.035	0.604 ± 0.208
LNBC(s)	0.631 ± 0.164	0.777 ± 0.218
KNN(s)	0.645 ± 0.126	0.867 ± 0.189

corresponds to a spectrum and a dot in a row represents a peak appeared in that position. It is noted that E.coli has a consistent peak at around 1244 nm while Kp does not (as indicated by the arrow in Fig. 1). We tried to extract human comprehensible rules from the SVM decision function. Since we are using linear kernel, the decision value for a testing sample is simply the inner product of the feature vector of the sample and the normal vector minus a constant [11]. Therefore, if normal vector have a high absolute value at certain wavelength, the difference at that wavelength is more important for classification. If we check the normal vector

Figure 1. Peak distribution of *Escherichia coli* and *Klebsiella pneumoniae* SERS spectrum. We can see the main peak difference that the decision tree identified between the two species is at 1240 nm.

of the decision function for [E.coli, Kp], line 3 in Fig. 2, the weight around wavelength 1244 nm is also significant. However, in the mean spectrum of E.coli (second dash line in Fig. 2), only a minor peak showed out around 1244 nm, the intensity difference is not significant enough to dominate the decision. In other words, although SVM can pick out regions around 1244 nm by assigning large weight, but due to the high complexity of these SVM models and relatively small sample size, other random effects might make SVM to learn from innocent coincidence pattern (marked by arrows in Fig. 2). [Hi, Sa] is another tough case for SVM. As shown in bottom of Fig. 2, the mean spectra of Hi and Sa are very similar, and the important features that the normal vector points out (marked by circle on bottom blue line in Fig. 2) are not apparent to the naked eye. Although the accuracy, 83.24%, is fairly good, it is doubtful whether the characteristic features determined by SVM are reasonable and further examinations are needed.

Lastly let us examine a case, [Ab, Sa], which SVM(s) performed well; the accuracy is 0.990 ± 0.051 which was

Figure 2. Normal vector of SVM hyperplane (blue lines) and mean spectra of [*Escherichia coli*, *Klebsiella pneumoniae*], [*Acinetobacter baumannii*, *Staphylococcus aureus*], and [*Haemophilus influenzae*, *Staphylococcus aureus*]. Important features are marked by red circles. Arrows shows false important features learned by SVM.

better than that of previous two cases. From first blue line in Fig. 2, we can see wavelength regions with relative high weight (marked by circles on first blue line in Fig. 2) really capture the differential region of two species that are apparent to the naked eye (marked by circles on first dash and red lines in Fig. 2).

V. CONCLUSION AND FUTURE WORK

In this paper, we present preliminary results on classification of different pathogenic species of bacterial meningitis using SERS in a process that is close to the clinical setting. We collected samples from hospital, and used task based classification to reflect the fact that clinical pathogens cannot be acquired beforehand. The results show that it is feasible to use SERS spectra of the pathogens to identify their species. Rules of employed classifiers can be extracted and compared, so that we can have more confidence and a better understanding of these classifiers. The hybrid SVM/CART model yields better performance than standard multi-class c-SVM (with intensity feature). However, we must point out that our data set is small, and we are still in the process of collecting more samples. We are also trying to reduce the complexity of our target classifier set, that is to reduce its VC dimension, so that the theoretical generalization error bound can be applied and produce reasonable bounds.

ACKNOWLEDGMENT

This work was supported by the National Science Council (NSC 95-3114-P-001-007-MY3), Taiwan, R.O.C.

REFERENCES

- [1] B. Kennedy, S. Spaeth, M. Dickey, and K. Carron, "Determination of the distance dependence and experimental effects for modified SERS substrates based on self-assembled monolayers formed using alkanethiols," *J. Phys. Chem. B*, vol. 103, no. 18, pp. 3640–3646, 1999.
- [2] S. Efrima and B. Bronk, "Silver colloids impregnating or coating bacteria," *J. Phys. Chem. B*, vol. 102, no. 31, pp. 5947–5950, 1998.
- [3] H.-H. Wang *et al.*, "Highly Raman-enhancing substrates based on silver nanoparticle arrays with tunable sub-10 nm gaps," *Advanced Materials*, vol. 18, no. 4, pp. 491–495, 2006.
- [4] T.-T. Liu *et al.*, "A high speed detection platform based on surface-enhanced Raman scattering for monitoring antibiotic-induced chemical changes in bacteria cell wall," *PloS One*, vol. 4, no. 5, 2009.
- [5] P. Rösch *et al.*, "Chemotaxonomic identification of single bacteria by micro-Raman spectroscopy: application to clean-room-relevant biological contaminations," *Applied and environmental microbiology*, vol. 71, no. 3, p. 1626, 2005.
- [6] E. Widjaja, W. Zheng, and Z. Huang, "Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines," *International journal of oncology*, vol. 32, no. 3, pp. 653–662, 2008.
- [7] T.-H. Tsai, T.-T. Liu *et al.*, "A multiscale approach for surface-enhanced Raman spectroscopy (SERS) spectrum representation and its application to bacterial discrimination," in *BMEI 2008.*, 2008, pp. 328–333.
- [8] T.-H. Tsai *et al.*, "Multiscale peak identification for bacterial SERS spectra," in *Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference*, 2009, pp. 1–5.
- [9] C. Silva and B. Ribeiro, "Two-level hierarchical hybrid SVM-RVM classification model," 2006, pp. 89–94.
- [10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [11] C. Bishop, *Pattern recognition and machine learning*. Springer New York:, 2006.
- [12] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [13] L. Breiman, *Classification and regression trees*. Chapman & Hall/CRC, 1984.
- [14] V. Makinen, "Peak alignment using restricted edit distances," *Biomolecular Engineering*, vol. 24, no. 3, pp. 337–342, 2007.
- [15] S. Lessmann, "Solving imbalanced classification problems with support vector machines," in *Proc. International Conference on Artificial Intelligence*, 2004, pp. 214–220.
- [16] B. Jones, "Statistics toolbox for use with matlab," *The Math-Works, Inc.*, vol. 24.