

Cost-Sensitive Classification on Pathogen Species of Bacterial Meningitis by Surface Enhanced Raman Scattering

Te-Kang Jan^{*}, Hsuan-Tien Lin[§], Hsin-Pai Chen[¶], Tsung-Chen Chern^{*}, Chung-Yueh Huang^{*},
Bing-Cheng Wen[‡], Chia-Wen Chung[‡], Yung-Jui Li[‡], Ya-Ching Chuang[‡], Li-Li Li[†],
Yu-Jiun Chan[†], Juen-Kai Wang[†], Yuh-Lin Wang[†], Chi-Hung Lin[‡], Da-Wei Wang^{*},

^{*}Institute of Information Science, Academic Sinica, Taipei, Taiwan, Email: tekang@iis.sinica.edu.tw

[†]Institute of Atomic and Molecular Sciences, Academic Sinica, Taipei, Taiwan

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

[§]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

[¶]Department of Medicine, National Yang-Ming University Hospital, Yilan, Taiwan

Abstract—We propose a pathogen-classification system using the Surface-Enhanced Raman Scattering (SERS) platform. The system differentiates the pathogens based on their SERS spectra, which are believed to be related to the surface chemical components. The specialty of the system is to not only consider the usual classification accuracy, but also pay attention to the different types of costs during misclassification. For instance, due to the effectiveness of treatments, the cost of classifying a Gram-positive bacterium as another Gram-positive one should be lower than the cost of classifying a Gram-positive bacterium as a Gram-negative one. We express the task as the cost-sensitive classification problem, and take state-of-the-art cost-sensitive classification algorithms from the machine learning community to conquer the task. Our experimental study validates the usefulness of those algorithms on building the system.

Index Terms—SVM; Cost-sensitive Classification; SERS;

I. INTRODUCTION

Bacterial meningitis is a serious and often life-threatening form of meningitis infection. Delay in treatment increases patients morbidity and mortality rate. The proper treatment for bacterial meningitis relies on rapid diagnosis, early identification, and effective antibiotics therapy [1]. Surface Enhanced Raman Scattering (SERS) platform can perform a fast and accurate detection of molecules vibration signal from a single bacterium [2] and is potentially useful for prompt and reliable identification of bacterial pathogens [3].

Machine learning algorithms have been applied on SERS spectra to learn a good model to perform automatic classification for future SERS spectra. Previous studies [3]–[6] have shown that over 90% accuracy can be achieved using neural networks to classify with the intensity and peak features.

Nevertheless, the promising results are viewed solely by the accuracy, which does not always match the realistic needs of the clinical practice, where miscellaneous misdiagnosis will be charged with pre-determined cost according to the type of actual pathogen species. For example, misidentifying the Gram-positive *Staphylococcus aureus* as a Gram-negative bacterium (such as *Pseudomonas aeruginosa*) should be associated with

a high cost because the antimicrobial agents for *Pseudomonas aeruginosa* are totally ineffective for *Staphylococcus*. On the other hand, if the *Staphylococcus* is misidentified as another Gram-positive bacterium, such as *Streptococcus pneumoniae*, the cost is much lower because the antimicrobial agents may still be appropriate. Such a classification problem is called cost-sensitive classification. There are ongoing works in machine learning for developing algorithms to handle this type of classification with promising results [7]–[12]. However, to the best of our knowledge so far, no one has yet applied the cost-sensitive classification tools on the SERS data set.

In this work, we study the task of building a reliable identification system with SERS and cost-sensitive classification. First, we collect and analyze SERS spectra of ten species of meningitis-causing bacteria from National Taiwan University Hospital (NTUH). These pathogens are *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). Second, we assign a cost to each misclassification and feed the cost to state-of-the-art cost-sensitive classification algorithms on the SERS data set. These cost-sensitive algorithms include cost-sensitive one-versus-one (CSOVO) [13], cost-sensitive one-sided regression (CSOSR) [14] and cost-sensitive filter tree (CSFT) [15]. We couple these algorithms with support vector machine (SVM) [16] framework because it has been frequently used for meningitis infection studies [3] with promising results. Third, we carefully study how one can obtain a suitable cost-sensitive model by tuning the parameters in the cost-sensitive algorithms accordingly. Our experiment shows that CSOSR algorithms with a particular kernel in SVM can achieve the lowest cost on the SERS.

We present a brief description of our SERS platform in the next section. Thus, in section III, we briefly discuss the

four algorithms. In Section IV, we summarize the experiment results. Finally, we discuss relevant issues and future work in Section V.

II. MATERIALS

In this section we describe the method we used to acquire Meningitis SERS spectra. Our dataset contains 79 clinical samples of ten meningitis-causing bacteria species collected in NTUH. In addition, 17 standard bacteria samples from American Type Culture Collection (ATCC) are used for establishing the baseline. Raw spectra are collected with Raman spectromicroscope, (HR800, Jobin-Yvon) equipped with a HeNe laser at 632.8nm and NA 0.95 100x water-immersion objective lens. The Laser power intensity used was about $105W/cm^2$. A task is a batched experiment of scanning and collecting Raman spectra from a single specimen, and we often take 30 to 50 spectra in a single task. Raman signals in the bandwidth between 400 and 1600 cm^{-1} , the information-rich portion, was collected. The integration time was set from 1 to 3 seconds. Median filtering with noise estimation was used to reduce the cosmic ray signals, wavelet de-noising was used to filter out the thermal noise, and peak-clipping algorithm was used to remove background fluorescence. Finally, the spectra intensity was normalized to $[0, 1]$ to address multiplicative factors in the spectra. The details of the process can be found in our previous reports [17]. The number of collected tasks and spectra for each species are listed in Table 1.

TABLE 1
SAMPLES USED IN THIS STUDY

	species	Ab	Ecol	HI	Kp	Lm	Nm	Psa	Spm	Sa	GBS
#Task	ATCC	1	1	3	1	1	5	1	6	1	1
	NTUH	11	11	2	10	8	0	8	11	8	10
#Spectra	ATCC	50	50	91	27	34	0	60	141	17	50
	NTUH	326	400	100	349	283	135	298	313	350	439

In addition to the sample spectra obtained from SERS platform, we develop a cost matrix according to average weights provided by two physicians specializing in infectious diseases. A misidentification was assigned a cost of 10 when it is possible that the misidentification will lead to ineffective antimicrobial therapy, treatment failure or mortality of the patient, while a cost of 1 was given when the therapy are not supposed to be very different in antimicrobial spectrum despite misidentification of the causative microbes. We organize the resulting costs as a matrix, as shown in Table 2.

TABLE 2
COST MATRIX ON SERS

real class \ classify to	Ab	Ecoli	HI	KP	LM	Nm	Psa	Spn	Sa	GBS
Ab	0	1	10	7	9	9	5	8	9	1
Ecoli	3	0	10	8	10	10	5	10	10	2
HI	10	10	0	3	2	2	10	1	2	10
KP	7	7	3	0	4	4	6	3	3	8
LM	8	8	2	4	0	5	8	2	1	8
Nm	3	10	9	8	6	0	8	3	6	7
Psa	7	8	10	9	9	7	0	8	9	5
Spn	6	10	7	7	4	4	9	0	4	7
Sa	7	10	6	5	1	3	9	2	0	7
Gbs	2	5	10	9	8	6	5	6	8	0

III. METHODS

In this section, we introduce the core method and the algorithms used in this study, and then describe the validation and partition methods.

A. Core method

SVM is a widely used binary classifier which aims at producing a hyperplane that separates the two classes of examples with the maximal margin in a space introduced by a kernel function $K(\mathbf{x}, \mathbf{x}')$ that measures the similarity between input vectors \mathbf{x} and \mathbf{x}' . Given a training set $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$ with $y_n \in \{+1, -1\}$, the classifier is obtained by solving the following optimization problem

$$\begin{aligned} \min_{\alpha_1, \alpha_2, \dots, \alpha_N} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i y_i \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^N \alpha_i \\ \text{subject to} \quad & \sum_{i=1}^N y_i \alpha_i = 0; \\ & 0 \leq \alpha_i \leq C, \quad \text{for } i = 1, \dots, N. \end{aligned}$$

Here C is the parameter that controls the power of SVM. A proper use of SVM includes choosing K and C appropriately [18]. In this study, we adopt the linear kernel $K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$, which is one of the simplest choices, and the RBF kernel $K(\mathbf{x}, \mathbf{x}') = e^{-\gamma \|\mathbf{x} - \mathbf{x}'\|^2}$, the most popular one.

B. Algorithms

OVOSVM: One-versus-one is a method for extending SVM for multi-class problems. One-versus-one SVM (OVOSVM) considers different pairs of classes and each pair is handled by one binary SVM classifier. Each classifier is assigned to learn which of the two classes is more likely. Let M represents the number of classes ($M = 10$ in our system), OVOSVM involves $\binom{M}{2}$ SVM classifiers. The prediction of OVOSVM is based on letting each SVM classifier vote, and the final decision is the class that gets the most votes. OVOSVM is designed to achieve decent accuracy and does not consider costs in its learning process. Next, we introduce several approaches that do consider costs during learning.

CSOVOSVM: CSOVOSVM [14] extends OVOSVM to cost-sensitive classification. CSOVOSVM also involves $\binom{M}{2}$ binary classifiers, each of which also working on a pair of classes. During training, the cost is embedded as the weights of training examples that can be learned by weighted SVM [19], a simple extension of binary SVM. The prediction procedure is the same as OVOSVM.

CSFTSVM: CSFTSVM [15] is another cost-sensitive classification algorithm based on SVM. Unlike CSOVOSVM, which takes $\binom{M}{2}$ comparisons in predicting the best class, CSFTSVM uses a single-elimination tournament, which can be represented as a binary tree of M leaves, for its prediction. The tree structure allows a prediction time of $O(\log M)$, faster than the $O(M^2)$ of CSOVOSVM. During training, the

cost is also embedded as the weights of training examples and each internal node of the tree is trained with a weighted SVM. There are $M - 1$ internal nodes in the binary tree and thus training CSFTSVM takes $O(M)$ SVM classifiers.

CSOSRSVM: CSOSRSVM [14] is a state-of-the-art cost-sensitive learning algorithm. Unlike CSOVOSVM and CSFTSVM, CSOSRSVM embeds the cost in the real-valued labels instead of the weights. Learning the real-valued labels in CSOSRSVM is usually referred to as an one-sided-regression problem. CSOSRSVM trains and predicts with M one-sided-regression SVM models, both taking $O(M)$ in times.

C. Validation Method

To validate each algorithm and kernel combination, we adopt 20 random runs and present their average as the result. The results, mean cost and accuracy, will be shown in the following tables. The number after \pm is the standard error of 20 runs. In all the experiments, for the RBF kernel, we select the parameter C within $\{2^{-5}, 2^{-3}, \dots, 2^{13}\}$ and the γ parameter within $\{2^{-8}, 2^{-7}, \dots, 2^0\}$ with a 5-fold cross-validation on the training set. For the linear kernel, we select C within $\{2^{-5}, 2^{-3}, \dots, 2^{13}\}$ with a 5-fold cross-validation on the training set.

D. Partition Strategy

We experiment with two partition strategies and conclude that the partition strategy affects the performance notably on the SERS data set. With the traditional partition strategy, we mix samples from all tasks, and then randomly selects 80% of all samples for training and the rest for testing. With the task-based partition strategy, we use 80% of the tasks for training and the other 20% for testing. The accuracy of task-based partition is 75.4%, which is lower than the traditional partition strategy 89.75%. In our data set, traditional partition strategy causes machine learning algorithms prone to identify rules from coincident thermal noise or background pollutants of the same task. Thus, the observed performance results are overly optimistic and misleading. Therefore, we choose to adopt the task-based partition strategy

E. Re-balance Data

Our SERS data set is inherently imbalanced. If we train classifiers with these original data set, classifiers will be biased toward classes with plentiful samples. The training strategy of SVM tends to ignore classes with fewer samples, which causes SVM to misclassify the minor classes often.

Furthermore, the common pathogens of bacterial meningitis vary over time and space [20]. Considering biased pathogens species from a single region, NTUH for example, is not an ideal strategy to prepare a model with appropriate generalization. In order to build a classifier which can generalize over unseen samples of any class with equal prior probability, we applied random undersampling to equalize species probability [21].

This process balances class distribution by randomly removing majority class samples until the number of majority class

samples equal to the number of minority class samples. In this study, we have selected 5 tasks for each species, and 10 spectra for each task.

IV. RESULTS

The mean cost and accuracy over 20 runs are summarized in Table 3 and Table 4. The p-value is calculated from a single tailed t-test over 20 runs.

First, we take a look at the performance of the linear kernel in Table 3, the cost of CSOSRSVM is 1.177, which is obviously better than OVOSVM, 1.251. The results indicate that it is promising to use cost-sensitive algorithms. As for CSOVOSVM and CSFTSVM, they can not lower the test costs because their accuracy rates are too low.

Next, we move on to the RBF kernel. Table 5 shows that RBF kernel results in lower cost than linear kernel generally, which indicates that the SERS data set needs more sophisticated classifiers than SVM with the linear kernel. This also shows the relationship between our label and the features is non-linear. Again, CSOSRSVM has lowest mean cost 1.071 among the four algorithms in our SERS dataset. The t-test shows cost of CSOSRSVM with RBF kernel is significantly lower than our baseline, OVOSVM.

It's worth mentioning that CSOSRSVM achieves similar accuracy with a lower cost compared to OVOSVM. The reason may be that some of our spectra are very difficult to be properly classified. We examine the misclassified spectra from both algorithms and find that 73% of the misclassified spectra of OVOSVM overlap with the misclassified spectra of CSOSRSVM. In Figure 1, for each species, we plot a blue line to indicate the average spectra, and red lines for each spectra misclassified by both OVOSVM and CSOSRSVM. Although these samples are mostly noises. CSOSRSVM only predicts cost 4.66, which is lower than OVOSVM, 4.98.

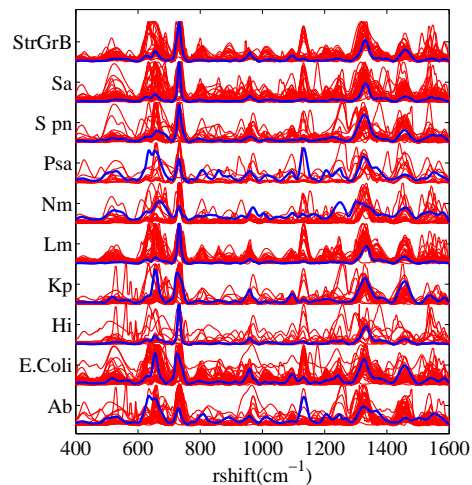


Fig. 1. SERS spectra of meningitis bacterial pathogen. For each species, blue line is the mean spectrum and the red thin lines are the overlapping misclassified spectra by both OVOSVM and CSOSRSVM algorithms.

In summary, the results suggest that CSOSRSVM with RBF kernel is better than the other algorithm and kernel combinations and is suitable for the SERS data set.

TABLE 3
EXPERIMENT RESULTS IN LINEAR KERNEL

	Accuracy	Cost	p-value
OVOSVM	75.35 ± 1.49	1.251 ± 0.087	N/A
CSOSRSVM	73.3 ± 1.87	1.177 ± 0.093	0.128
CSOVOSVM**	57.3 ± 2.03	1.477 ± 0.111	0.9779
CSFTSVM**	54.65 ± 2.32	1.831 ± 0.099	1

** OVOSVM significantly better
(single-tailed pairwise t-test on cost with $\alpha = 0.05$)

TABLE 4
EXPERIMENT RESULTS IN RBF KERNEL

	Accuracy	Cost	p-value
OVOSVM	75.45 ± 1.58	1.232 ± 0.087	N/A
CSOSRSVM*	75.6 ± 1.63	1.071 ± 0.081	0.004
CSOVOSVM	68.10 ± 2.03	1.209 ± 0.089	0.3538
CSFTSVM**	64.55 ± 1.83	1.489 ± 0.095	0.9987

*OVOSVM significantly worse ** OVOSVM significantly better
(single-tailed pairwise t-test on cost with $\alpha = 0.05$)

TABLE 5
THE DIFFERENCE BETWEEN TWO KERNELS

	Cost(linear kernel)	Cost(RBF kernel)	p-value
OVOSVM	1.251 ± 0.087	1.232 ± 0.087	0.441
CSOSRSVM	1.177 ± 0.093	1.071 ± 0.081	0.204
CSOVOSVM*	1.477 ± 0.111	1.209 ± 0.089	0.038
CSFTSVM*	1.831 ± 0.099	1.489 ± 0.095	0.010

* RBF kernel significantly better
(single-tailed pairwise t-test on cost with $\alpha = 0.05$)

V. CONCLUSION AND FUTURE WORK

We use empirical data and domain knowledge to design the cost matrix and the platform for comparing cost-sensitive classification algorithms. We demonstrate traditional algorithms is insufficient for clinical bacterial meningitis pathogen identification practice when considering the cost of misclassification, since they only uses the accuracy to validate the model and do not deal with the difference of assorted error types. We compared three cost-sensitive algorithms in order to find a proper cost-sensitive algorithm that reflects the unequal misdiagnosis cost in clinical practice. The result shows CSOSRSVM with RBF kernel achieve the lowest cost among OVOSVM, CSOVOSVM and CSFTSVM.

In the future, we will incorporate species distribution into our experiments. We will also consider time-varying and region-varying solutions using transfer learning and on-line learning techniques to extend the capability of our cost-sensitive models.

ACKNOWLEDGMENT

This work was supported by the National Science Council (NSC 99-2120-M-001-003-CC1 and NSC 99-2628-E-002-017), Taiwan, R.O.C.

REFERENCES

- [1] A. Tunkel, B. Hartman, S. Kaplan, B. Kaufman *et al.*, "Practice guidelines for the management of bacterial meningitis," *Clinical Infectious Diseases*, vol. 39, no. 9, pp. 1267–1284, 2005.
- [2] T.-T. Liu, Y.-H. Lin, C.-S. Hung, T.-J. 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, p. e5470, 2009.
- [3] C.-Y. Huang, T.-H. Tsai, B.-C. Wen, C.-W. Chung *et al.*, "Hybrid svm/cart classification of pathogenic species of bacterial meningitis with surface-enhanced raman scattering," in *International Conference on Bioinformatics and Biomedicine*, 2010, pp. 406–409.
- [4] P. Rösch, M. Harz, M. Schmitt, K. Peschke, O. Ronneberger *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, pp. 1626–1637, 2005.
- [5] P. Rösch, M. Harz, K. Peschke, O. Ronneberger *et al.*, "Identification of single eukaryotic cells with micro-Raman spectroscopy," *Biopolymers*, vol. 82, no. 4, pp. 312–316, 2006.
- [6] T. Bocklitz, S. Putsche, M. C., J. K äs, A. Niendorf *et al.*, "A comprehensive study of classification methods for medical diagnosis," *Journal of Raman Spectroscopy*, vol. 40, no. 12, pp. 1759–1765, 2009.
- [7] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 155–164.
- [8] C. Elkan., "The foundations of cost-sensitive learning," in *International Joint Conference on Artificial Intelligence*, vol. 17, 2001, pp. 973–978.
- [9] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in *Proceedings of the 3rd IEEE International Conference on Data Mining*, 2003, pp. 435–442.
- [10] N. Abe, B. Zadrozny, and J. Langford, "An iterative method for multi-class cost-sensitive learning," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 3–11.
- [11] J. Langford and A. Beygelzimer, "Sensitive error correcting output codes," in *Learning Theory: 18th Annual Conference on Learning Theory*, 2005, pp. 158–172.
- [12] Z.-H. Zhou and X.-Y. Liu, "On multi-class cost-sensitive learning," in *In Proceeding of the 21st National Conference on Artificial Intelligence*, vol. 21, 2006, pp. 567–572.
- [13] H.-T. Lin, "A simple cost-sensitive multiclass classification algorithm using one-versus-one comparisons," 2010, downloaded from <http://www.csie.ntu.edu.tw/~htlin/paper/doc/csovo.pdf>.
- [14] H.-H. Tu and H.-T. Lin, "One-sided support vector regression for multiclass cost-sensitive classification," in *Machine Learning: Proceedings of the 27th International Conference*, 2010, pp. 1095–1102.
- [15] A. Beygelzimer, J. Langford, and P. Ravikumar, "Multiclass classification with filter trees," 2007, downloaded from <http://hunch.net/~jl>.
- [16] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
- [17] T.-H. Tsai, D.-W. Wang, T.-T. Liu, Y.-H. Lin *et al.*, "Multiscale peak identification for bacterial sers spectra," in *International Conference on Bioinformatics and Biomedical Engineering*, 2009, pp. 1–5.
- [18] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667–1689, 2003.
- [19] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 464–471, 2002.
- [20] K. Dawson, J. Emerson, and J. Burns, "Fifteen years of experience with bacterial meningitis," *The Pediatric infectious disease journal*, vol. 18, no. 9, pp. 816–822, 1999.
- [21] S. Kotsiantis and P. Pintelas, "Mixture of expert agents for handling imbalanced data sets," *Annals of Mathematics, Computing & Teleinformatics*, vol. 1, no. 1, pp. 46–55, 2003.