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

Te-Kang Jan

Institute of Information Science, Academic Sinica, Taipei

October 31, 2011

joint work with 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
Motivation and objective

Motivation

- Meningitis is a lethal disease and early identification of species play an vital role in devising an effective treatment.
- Antibiotic abuse and the cost of antibiotic is high.

Objective

- To set up a platform which is close to clinical setting
- To research which model can achieve the lower cost
SERS platform
Spectrum preprocessing

1. Raw spectra
2. Cosmic ray removal
3. Smoothing
4. Normalizing
5. Baseline removal
Specimen

- National Taiwan University Hospital: 21 tasks
- American Type Culture Collection (ATCC): 79 tasks
- We often take 30 to 50 spectra from one task. Here a task means an experiment.

**Table:** Samples used in this study

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

Our database contains the real world data.
What is the previous result before?
(Hung et al, BIBM, 2010; Bocklitz T et al, JRS, 2009)

Key idea

Using support vector machine (SVM), classification tree or other methods to get over the 90% accuracy

- *Good* but is viewed solely by the accuracy
- Not close to the real world’s need

Are all misclassification costs equal?
Identification system on SERS

- **Category:** Sa (gram-positive bacteria), Entrococcus faecails (another gram-positive bacteria), E.coli (gram-negative bacteria)
- The antibiotic agent of gram-positive is not effective for gram-negative bacteria.
- Sa mis-classified as Sa bacteria: Correct
- Sa mis-classified as Entrococcus faecails bacteria: Relatively OK
- Sa mis-classified as E.coli bacteria: High cost
Cost-sensitive classification

Cost-sensitive
Each misclassification error with different cost

Problem
How to choose a better classifier that pays low cost?
Assorted algorithms

Regular algorithms
- One-versus-one SVM

Cost-sensitive classification algorithms
- Cost-sensitive filter tree
- Cost-sensitive one-versus-one
- Cost-sensitive one sided regression
One-versus-one SVM (OVOSVM) and cost-sensitive one-versus-one SVM (csOVOSVM)

OVOSVM

1. It is very popular in machine learning.
2. It has been frequently used for classification in surface enhanced Raman spectroscopy.

csOVOSVM

1. It is the cost-sensitive extension of the OVOSVM.
2. It has the great results among cost-sensitive algorithms.

Comparison speed (in our system, $N = 10$)

<table>
<thead>
<tr>
<th></th>
<th>Training speed</th>
<th>Testing speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVOSVM</td>
<td>$\binom{N}{2}$ classifiers</td>
<td>$\binom{N}{2}$ classifiers</td>
</tr>
<tr>
<td>csOVOSVM</td>
<td>$\binom{N}{2}$ classifiers</td>
<td>$\binom{N}{2}$ classifiers</td>
</tr>
</tbody>
</table>
These two algorithms are great, but the speed is somehow not satisfied.

We then choose two fast and promising cost-sensitive algorithms.
Cost-sensitive one sided regression SVM (csOSRSVM) and cost-sensitive filter tree SVM (csFTSVM)

**csOSRSVM**
- It is faster than OVOSVM and csOVOSVM.

**csFTSVM**
- It is simple and even faster than csOSRSVM.

**Comparison speed (in our system, \(N = 10\))**

<table>
<thead>
<tr>
<th></th>
<th>Training speed</th>
<th>Testing speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>csOSRSVM</td>
<td>(N) classifiers</td>
<td>(N) classifiers</td>
</tr>
<tr>
<td>csFTSVM</td>
<td>(N - 1) classifiers</td>
<td>(\log N) classifiers</td>
</tr>
</tbody>
</table>
Experiments

Set up

- To validate each algorithm and kernel combination, we adapt 20 random runs and present their average as the result.
- For SVM kernels, we choose linear and RBF kernel.
Is linear kernel or RBF kernel good?

RBF kernel performance is better than linear kernel
Cost-sensitive algorithms perform better than regular algorithm.
Result extraction case study: OVOSVM and csOSRSVM

- Blue line: mean spectrum of each spectra
- Red line: overlapping misclassified spectra by both OVOSVM and csOSRSVM
- These spectra are very hard to be classified correctly.
csOSRSVM can still predict the lower cost.
Result extraction case study: OVOSVM and csOVOSVM

Cost-sensitive algorithms trade more expensive costs with cheaper ones.
Conclusion

- The results show that it is feasible to use SERS spectra to identify pathogen species.
- The cost-sensitive algorithms can reach the lower cost.

Thank you. Questions?