4-1-2012

Predicting Disease Outbreaks using a Support Vector Machine model

Nicolae Dragu
Trinity College, nicolae.dragu@trincoll.edu

Follow this and additional works at: http://digitalrepository.trincoll.edu/theses

Recommended Citation
Dragu, Nicolae, "Predicting Disease Outbreaks using a Support Vector Machine model". Senior Theses, Trinity College, Hartford, CT 2012.
Trinity College Digital Repository, http://digitalrepository.trincoll.edu/theses/158
Predicting disease outbreaks using a Support Vector Machine model

Author: Nicolae Dragu  Advisers: Ralph Morelli, Takunari Miyazaki

Motivation
Can Artificial Intelligence provide a way for people to be able to predict and prevent disease outbreaks?

This project proposes a potential solution to this problem by providing users with a user interface to check news articles for potential outbreaks.

Overview
The project consists of a user interface which takes a health-related news article as input, parses it and a Support Vector Machine (SVM) makes a prediction based on the content. The output is then presented in the form “The SVM indicates that this article does/does not indicate a disease outbreak”.

Support Vector Machines: An SVM represents a machine learning algorithm which can classify data into two or more classes. In this case, our two classes are articles which indicate disease outbreaks and articles which do not.

Example illustrating an SVM classification of two classes (red & blue). In order to make the classification, the SVM attempts to find a linear decision boundary.

Details
After analyzing many health-related articles from the World Health Organization archive, a list of words (36 in total) has been created denoting the most common disease outbreak terminology found in these articles.

A java parser is then used to automatically match these words with the content of an article and whenever a match is found, a simple negation detection algorithm is used to determine whether the word is negated or not. The parser then returns a 36 dimensional vector. Vector values of 0 correspond to words which were not found or were negated and values of 1 correspond to words which were found.

Results
This process was repeated for 158 articles resulting in 158 vectors, out of which 98 were used to train the SVM and 60 were used to test the SVM.

Note: An SVM is trained in order for it to be able to classify future articles.

<table>
<thead>
<tr>
<th>Training Set (98 Vectors)</th>
<th>Testing Set (60 Vectors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 0 2 3 1 351 361</td>
<td>1 0 2 3 1 350 360</td>
</tr>
<tr>
<td>0 1 0 2 3 0 350 360</td>
<td>0 1 0 2 3 0 350 360</td>
</tr>
<tr>
<td>1 0 2 3 1 351 361</td>
<td>1 0 2 3 1 351 361</td>
</tr>
</tbody>
</table>

54/60 correctly classified 90% Accuracy

Web Interface
The goal of this interface is to allow users to test any news article for a potential disease outbreak threat. The content of the article is pasted into a text box and after submitting the article, the output is presented as seen in the adjacent image.

Acknowledgements
My advisers: Prof. Ralph Morelli, Prof. Takunari Miyazaki

References
Developers of LIBSVM - Chih-Chung Chang and Chih-Jen Lin
A Library for Support Vector Machines: http://www.csie.ntu.edu.tw/~cjlin/lbsvm/