A

Appendix: SIAM Text Mining Competition 2007

Overview

The 2007 Text Mining Workshop held in conjunction with the Seventh SIAM International Conference on Data Mining was the first to feature a text mining competition. Members of the Intelligent Data Understanding group at NASA Ames Research Center in Moffett Field, California, organized and judged the competition. Being the first such competition held as a part of the workshop, we did not expect the large number of contestants that more established competitions such as KDDCUP[1] have, but we did receive five submissions, though one person later withdrew.

Matthew E. Otey, Ashok N. Srivastava
Santanu Das, and Pat Castle

A.1 Classification Task

The contest focused on developing text mining algorithms for document classification. The documents making up the corpus used in the competition are aviation safety reports documenting one or more problems that occurred on certain flights. These documents come from the Aviation Safety Reporting System (ASRS), and they are publicly available in their raw form at http://asrs.arc.nasa.gov. Since each of these documents can describe one or more anomalies that occurred during a given flight, they can be tagged with one or more labels describing a class of anomalies. The goal is to label the documents according to the classes to which they belong, while maximizing both precision and recall, as well as the classifier's confidence in its labeling. This second criterion concerning confidence is useful for presenting results to end-users, as an end-user may be more forgiving of a misclassification if he or she knew that the classifier had little confidence in its labeling.

[1] See http://www.kdnuggets.com/datasets/kddcup.html.

The competition entries were scored using the following cost function that accounts for both the accuracy and confidence of the classifier. When describing the cost function, we will use the following notation. Let D be the number of documents in the corpus, and let L be the number of labels. Let F_j be the fraction of documents having label j. Let $t_ij \in \{-1, +1\}$ be the target (true) value for label j of document i. Let $p_ij \in \{-1, +1\}$ be the predicted value for label j of document i, and let $q_ij \in [0, 1]$ be the classifier's confidence of the value p_ij. Finally, let A_j be the area under the ROC curve for the classifier's predictions for label j. We define the intermediate cost function Q_j for label j as

$$Q_j = (2A_j - 1) + \frac{1}{D}\sum_{i=1}^{D} q_{ij}t_{ij}p_{ij} \; . \tag{A.1}$$

This function has a maximum value of 2. The final cost function Q is the average Q_j for all labels:

$$Q = \frac{1}{C}\sum_{j=1}^{C} Q_j \; . \tag{A.2}$$

In case of ties, we also formulated a figure of merit (FOM), defined as

$$FOM = \frac{1}{C}\sum_{j=1}^{C} \frac{(F - F_j)}{F} Q_j \tag{A.3}$$

where

$$F = \sum_{j=1}^{C} F_j. \tag{A.4}$$

This figure of merit measures the quality of predictions across all anomaly categories, giving a lower weighting to those categories making up a larger fraction of the dataset. Hence, better FOM are achieved by obtaining higher accuracies and confidences on rarer classes.

We wrote a small Java program implementing these functions to use during our judging of the submissions. Before the contest deadline, we made this program and its source code available to the contestants so that they could validate its correctness, and so that they could use our implementation to tune their algorithms.

A.2 Judging the Submissions

A training dataset was provided over a month in advance of the deadline, giving the contestants time to develop their approaches. Two days before the deadline we released the test dataset. Both of these datasets had been run through PLADS, a system that performs basic text processing operations such as stemming and acronym expansion. The contestants submitted their labeling of the test dataset, their confidences in the labeling, and the source code implementing their approach. The scores of the

submissions were calculated using the score function described above. In addition to scoring the submissions, we ran each contestant's code to ensure that it worked and produced the same output that was submitted, and we inspected the source code to ensure that the contestants properly followed the rules of the contest.

A.3 Contest Results

The submissions of the contestants all successfully ran and passed our inspection, and we announced our three winners. First place was awarded to Cyril Goutte at the NRC Institute for Information Technology, Canada, with a score of 1.69. A team consisting of Edward G. Allan, Michael R. Horvath, Christopher V. Kopek, Brian T. Lamb, and Thomas S. Whaples of Wake Forest University, and their advisor, Michael W. Berry of the University of Tennessee, Knoxville, came in second with a score of 1.27. The third place team of Mostafa Keikha and Narjes Sharif Razavian of the University of Tehran in Iran, and their advisor, Farhad Oroumchian of the University of Wollongong, Dubai, United Arab Emirates, scored a 0.97. At NASA, we evaluated Schapire's and Singer's BoosTexter approach, and achieved a maximum score of 0.82 on the test data, showing that the contestants made some significant improvements over standard approaches.

Index