Improving Sentiment Analysis with Active Learning and the Crowd

Stephen Goodwin sgoodwin@cs.utexas.edu Dept. of Computer Science, UT-Austin

Introduction

Sentiment analysis is useful for identifying positive or negative feeling in text. This project explores building a sentiment classifier for user-generated product reviews. Instead of just finding positive/negative sentiment, it identifies the star-rating that the review would have received. The classifier takes advantage of labeled and unlabeled reviews using a label propagation technique. A number of unlabeled reviews are then selected to be annotated by the crowd to improve the sentiment classifier.

This project aims to answer two questions:

(1) Can the crowd accurately determine sentiment from online reviews?

(2) Are these sentiment classifications useful in this label propagation technique?

Dataset

The dataset is of product reviews from Amazon.[1] The reviews and ratings serve as nodes and labels in the graph, and the similarity between two reviews is the edge weight.

★★★★☆ **Big Performance Leap,** February 27, 2011

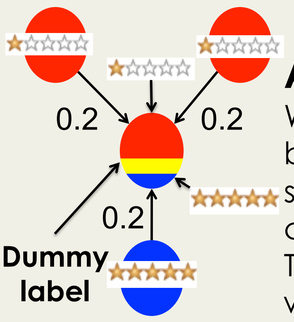
By **djsteve** - See all my reviews

Amazon Verified Purchase (What's this?)

This review is from: **Apple MacBook Pro MC725LL/A 17-Inch Laptop (Personal Computers)**

I upgraded to this model (early 2011 Quad Core i7 17" MacBook Pro) from the first in Apple's series of "unibody" laptops (late 2008 Core 2 duo 15" MacBook Pro). Both have been excellent laptops and have remarkably similar features given the years between them. However, what makes the difference here is pure raw speed.

The i7 processor has four cores each running at 2.2Ghz and the ability to "hyperthread" to operate in a virtual eight core mode. Applications must be written in a specific way to take full advantage of the multiple cores, however the processor has another trick up its sleeve in such cases. If some of the cores are underutilized then the processing speed of the those in use are bumped up. This translates into blazing speed across the board.



Active Learning

We selected nodes to crowdsource based on a calculated *uncertainty* score. The uncertainty is indicated by a confidence score of a dummy label. The dummy label is propagated along with actual labels.

Method

Our sentiment analysis approach uses a graph-based, semi-supervised learning, label propagation algorithm called Modified Adsorption.[2] The label propagation algorithm aims to propagate labels based on how similar a node is with the labeled nodes around it.

Figure 1: Example starting graph with 2 nodes (red & blue) seeded with labels

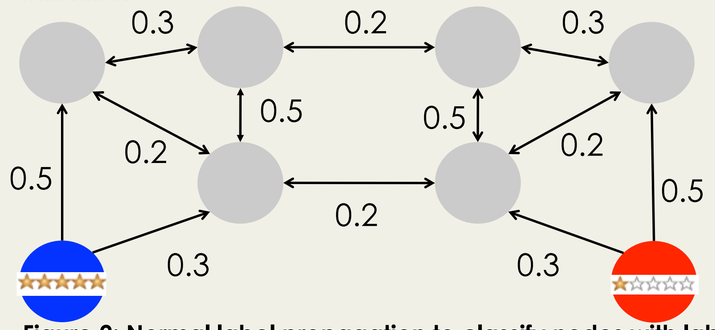


Figure 2: Normal label propagation to classify nodes with labels.

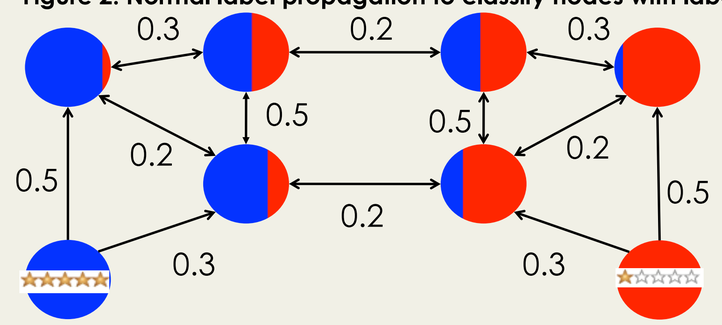
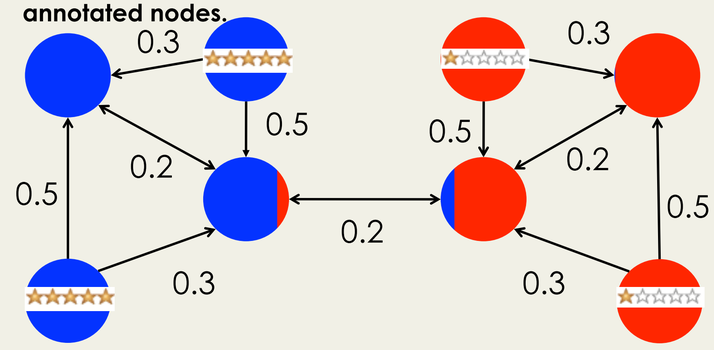


Figure 3: Label Propagation seeded with 2 labeled nodes & 2 crowd annotated nodes.

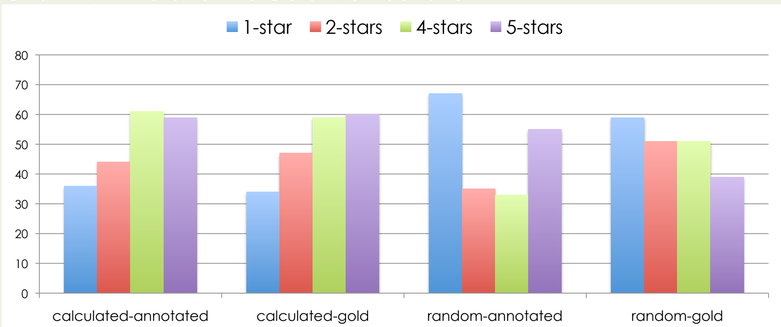


Experiments: We ran experiments on 200 labeled instances and 800 unlabeled instances. Each experiment had (10, 25, 50, 100, 200) of the unlabeled instances annotated by the Crowd.[3]

Results

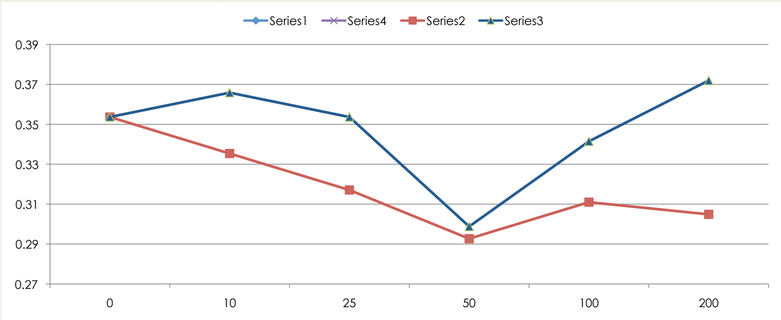
The chart below compares the calculated uncertainty-calculated vs. the random (a baseline) distribution annotations by the Crowd. The uncertainty-calculated yielded a much more similar distribution.

Chart 1: Annotation vs Gold Distributions



The uncertainty-calculated annotations produced a slightly better sentiment classifier than the random annotations.

Chart 2: Uncertainty-Annotated vs. Randomly-Annotated



Conclusion

The Crowd can be effective in annotating, when the uncertainty calculation is used. However, the results using the uncertainty-calculated annotations were not better than random annotations. Without further work, the current active learning method is ineffective.

References

[1] John Blitzer, Mark Dredze, Fernando Pereira. *Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification.* Association of Computational Linguistics (ACL), 2007.

[2] Partha Talukdar and Koby Crammer. *New Regularized Algorithms for Transductive Learning.* In Machine Learning and Knowledge Discovery in Databases, 2008.

[3] Using Crowdflower.com