Data Analysis Projects Summary

Robert Yates

In the following passage, I will summarize my data analysis course projects covering topics from identifying fraudulent users and automating music classification to forecasting wind power.

While taking Social and Information Network Analysis (CS224W) at Stanford, two classmates and I analyzed user trust and helpfulness in Amazon.com product reviews. The relationship is shown in in Figure 1. Many reviews are dishonest spam entries written to make a given product more favorable or a competing product less favorable. We calculated user helpfulness by taking sum of the positive helpfulness votes over all of a user's reviews minus the sum of the unhelpful votes over all of that user's reviews. Using the algorithm described in Wang et al. [1], we determine the following four properties: a. User Trust, the tendency of a user to supply honest reviews; b. Review Honesty, the tendency of a review to have a high review agreement value for highly reliable or highly unreliable products; c. Product Reliability, the tendency of a product to get high scores from trusted users; d. Review Agreement, the tendency of a review to agree with most reviews for the same item by trusted reviewers. We used MATLAB to plot the distribution of users' helpfulness votes with their computed trustworthiness scores (property a.) as shown in Figure 2. The data showed a strong correlation between users with low helpfulness and low trust, and high helpfulness and high trust. 87% of users with helpfulness less than -100 have trust ratings less than -0.9, which could prove useful as indicator of a spam user.

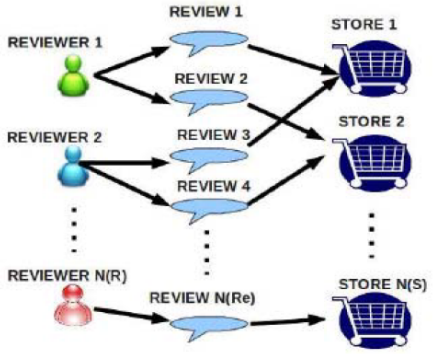

Figure 1: Review Graph (Wang et al. [1])

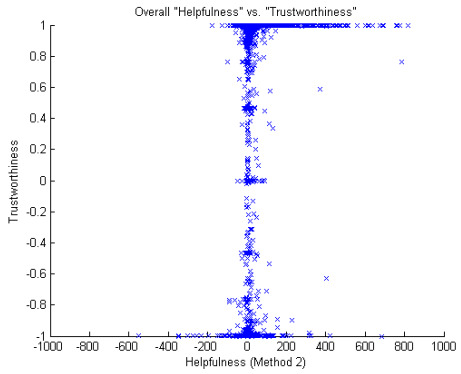

Figure 2: Helpfulness vs. Trustworthiness

My class partner and I wanted to improve musical organization for the final project of Artificial Intelligence (CS221). For instance, the artist "The Tallest Man on Earth" is labeled a 'folk' singer by Groveshark, iTunes considers his music 'alternative', and at one point Spotify had featured him in a 'blues rock' playlist. For our project, we predicted song genres by analyzing their lyrics. Using data from the Million Song Dataset (Bertin-Mahieux et al. [2]), we are given a dictionary of the word frequency in the song lyrics. From the

EchoNest.com API, we are provided with multiple genres for a given song, so we use the first and most prominent genre in our list. Our dataset consisted of over 2000 songs distributed over 16 genres: 'metal', 'pop', 'r&b/soul', 'country', 'folk', etc. Analyzing the lyrics, we found Spanish words were a good indicator a song was Latin (correctly predicted 92% of the time). Our baseline decision tree algorithm written in Python achieved 29% accuracy in correctly classifying a song as one of 16 genres, while the best algorithm, a version of logistic regression, achieved a classification success rate of 47%. Given randomly guessing song genres would give $\sim 6\%$ accuracy, we concluded that lyrical analysis is helpful for classifing songs into genres.

For the Machine Learning (CS229) final project, we analyzed the dataset from the analytics competition website Kaggle: A wind power forecasting problem: predicting hourly power generation up to 48 hours ahead at 7 wind farms. [3] Predicting future energy generation precisely for a specific wind power generation field is useful to properly design the power network and achieve full compliance of customers energy demands. For our training data, we were provided with the 48-hour ahead forecast for wind speed, wind direction, and the zonal and meridional wind components at each of the seven farms. For our prediction task, we were given generated power measurements (normalized between 0 and 1) for the first 36 hours periods and we needed to predict the power for the next 48 hours. Using ARIMAX time series prediction model with wind speed features correlated with the generated wind power implemented in the statistical language R, we beat the benchmark method by almost a factor of two.

Implementing modern data analysis techniques for predicting fake reviewers, song genres, and wind energy generated demonstrates my versatility and experience in data analysis projects.

References

[1] Wang, Guan, Sihong Xie, Bing Liu, and Philip S. Yu. "Review graph based online store review spammer detection," in *Data Mining (ICDM), Proceedings of the 11th International Conference on*, 2011.

[2] Bertin-Mahieux, Thierry, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. "The Million Song Dataset," in *Music Information Retrieval Conference (ISMIR), Proceedings of the 12th International Society for*, 2011.

[3] Hong, Tao, Shu Fan, and Pierre Pinson. "Wind Forecating" in *Global Energy Forecasting Competition (GEFCom)*, 2012.