Final Course Project CSE 5334 Data Mining

Fake reviews detect on Yelp Dataset with deep learning

Son Nguyen & Kanishka Tyagi

Phd Student, Department of Electrical Engineering

Objective and overview

Yelp challenge dataset is a multifold dataset that has been analyzed predominantly from a database systems and data mining point of view. As in [2], by minting the data after storing it in a relational database, several interesting business behaviors are analyzed thru query search. [3] looks at it from a data mining point of view, where the entire dataset is classifier into relevant categories.

In the spirit of finding useful information in unlabelled dataset, we are specifically interested in detecting fake reviews. Reviews are becoming valuable information for data mining or search ranking. But unfortunately, there are fake reviews which will decrease accuracy of search's results. Detecting a review is fake or not become an interesting topic. [1] reports recent advances in this regards by using linear and non linear Kernel machines.

Projected Tasks

Owing to our research in designing powerful neural networks for both shallow and deep learning architecture, we plan to apply them to the Yelp dataset. We will focus mainly on reviews data set. There are 2 main task: Features extraction and classification. Regarding features extraction, we might use technique on [1] or design our own. For classification task, we plan to use our own designed non linear classifier that will be based on [4]. It has shown considerably good performance on conventional datasets and it'll be interesting to see it scalability.

The key task we would like to accomplish in the classification is the multiclass classification. There has been a growing interest in multi label multi instance classification or extreme classification. So much aso that NIPS in 2013 had a separate workshop on it extreme classification leads to many real life applications that can be dealt with classifiers. In this case each instance belong to many classes at the same time. Since Yelp Dataset provides a suitable scenario to try these classifiers, we would like to investigate more in this direction.

Output products (Deliverables)

We are planning to deliver a software that automatically classify a review is fake or not. If the result are good enough, we will try to publish the results in a peer reviews conference.

Challenges of project

In our opinion, there are three main challenges in the project:

1) the dataset is huge, so it's desirable to convert data formatted as JSON to text files that can be easily imported in tables.

2) The feature engineering will require a scalable algorithm owing to the large size of data. not only that it should also run fast.

3) The data annotation process is very time consuming and we have to either use deep learning based data abstraction techniques or self taught learning approach that we have been working in our research.

Plan to address the challenges

We have background on neural network so we might use some technique of neural network for big data, example Stochastic Gradient Descent (SGD), random coordinate descent and other gradient based algorithms. We also have expertise on second order non gradient algorithms that have performed considerably better than conventional gradient based algorithms. Refer [5]-[7] for some recent results.

Regarding data annotation, we can use labels from previous work like [1], this act will also make it easier to compare our results with others.

Evaluation Metrics

We use Precision and Recall to measure the performance of our classifier. We will compare our result with recent Yelp dataset challenge results in terms of time and space complexity.

Partition the tasks and coordinate among team members:

As a team we are still figuring out to divide the whole task but the probably plan is that one team-mate will do features extraction and the other will look into classification.

References

[1] Mukherjee, Arjun, et al. "What yelp fake review filter might be doing?." *ICWSM*. 2013.

[2] http://courses.cs.washington.edu/courses/cse544/13sp/final-projects/p09-tingtinz.pdf

[3] http://www.ics.uci.edu/~vpsaini/

[4] Cai, Xun, Kanishka Tyagi, and Michael T. Manry. "Training multilayer perceptron by using optimal input normalization." *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*. IEEE, 2011.

[5] K.Tyagi, X.Cai, M.Manry, "An Optimal Construction and Training Algorithm for Radial Basis Neural Network based on Second Order Algorithm", IEEE IJCNN'11, San Jose, California, USA.

[6] K.Tyagi, X.Cai, M.Manry, "Fuzzy C-Means clustering based Construction and Training Algorithm for Second Order RBF Network", Oral presentation, IEEE Intl. Conf. on Fuzzy Systems, June 27-31, 2011, Taipei, Taiwan.

[7] X.Cai, K.Tyagi, M.Manry, "An Optimal Input Gain with Bias Algorithm for MLP Backpropagation", Oral presentation, IEEE Intl. Conf. on Fuzzy Systems, June 27-31, 2011, Taipei, Taiwan.

[8] http://research.microsoft.com/en-us/um/people/manik/events/XC13/index.html