Data Mining Project Proposal

Group Member: Naixin Zhu

Data set: Data Scientist Job Market in the U.S.[1]

The data set contains 6,964 data science jobs, which contain 5 columns: position, company, job description, reviews, and location.

Goal: Predict the type and location of the data scientist job

There are 2,214 unique company values, thus company is not a good choice for label. However, from brief observations, companies from all kinds of industries are hiring data scientists. Thus, we can engineer a feature called company industry. From the job descriptions, we can also use text analysis to generate a few features, which captures the nature of the job. Using the company industry and nature of the job, we can engineer a feature that we want to predict, the type of job. For example, a type could be consumer trend prediction, which is common in e-commerce companies and uses online ordering data in their daily job.

Literature Review

Very few researchers have done a similar analysis before, but below are some more general papers that one should look at before proceeding:

- Pavel Berkhin[2]
- Raymond J. Mooney and Razvan Bunescu[3]
- Vishal Gupta and Gurpreet S. Lehal[4]

Data Preprocessing

[1] https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us/version/4

[2] http://www.miv.t.u-tokyo.ac.jp/ishizuka/pr-class/clustering_survey(Berkhin2002).pdf

[3] http://www.cs.utexas.edu/~ml/papers/text-kddexplore-05.pdf

[4] http://www.jetwi.us/uploadfile/2014/1230/20141230112729939.pdf

I am going to clean, impute the missing values, engineer additional features, standardize and arrange the data into relevant file types.

Algorithm

I plan to conduct both supervised learning and unsupervised learning. Using decision tree, random forest, support vector machine, K nearest neighbors and K means to predict the location and type of the job.