

Predicting Small Business Failure

Murdock Tracy

University California, Irvine

Irvine, CA

mtracy@uci.edu

ABSTRACT

We were presented the problem of determining when a small business is going to go out of business with the information about the business from Experian. In this paper we will go over the two methods we choose to solve this problem with the data presented with us. A decision was made to take two different approaches which ended up being applying Naïve Bayes Classifier and k-means clustering, both a supervised and unsupervised learning algorithms. This paper will mainly cover the details of the results from applying k-means clustering to this problem.

Due to the nature of the data we had to address a couple different problems. One problem we needed to overcome is running time of k-means clustering for such a large data set. We run some analysis on the running time of two different algorithms. Second problem we have how to get the results we want from an unsupervised learning algorithm.

Predicting small business failure was not an easy task and unknown with the data provided to us. Experian gave us certain criteria for what it means for a business to go “bad” but we had no way to check if indeed a business failed.

1 INTRODUCTION

Small business failure currently is a problem because it cannot be defined easily. The current solution is to telephone a business to absolutely make sure the business status. Experian presented this problem to us and since we cannot use the data to show which business are out of business we will try to predict when a business starts to fail. There are different criteria that Experian has decided on what it means for a business to go “bad” or start to fail. For the rest of the paper I will describe business

in two different categories, either a business is in good standing or is in bad standing. For short it will be stated as a good business or a bad business.

The data given to us is in a time series per quarter. We have one year of data to read and then need to predict if the business will go bad in the next year. We are given two years of data so we can check the productiveness of our program. First year data will be read into the program and then report a result if they business is predicted to go bad or stay good. Business that already went bad is eliminated from the data set. Experian current system uses logistic regression that outputs a score which correlates to the change of a business going bad.

Our group took two different approaches to this problem. The approaches being a supervised and an unsupervised learning algorithm to predicting if a business will go bad. Naïve Bayes classifier was our choice for the supervised learning algorithm. Our group was split into two subgroups and Edward Wong and Brian Solloway worked on implementing that algorithm for this problem. Later in this paper I will post results that they have found. K-means clustering is the choice we made for the unsupervised learning algorithm. Willson Luu and I worked on implementing the k-means algorithm on this problem. Most of this paper will be focused around the result found from the k-means clustering approach.

One of the major issues with the data we were given is that it is completely foreign. In order to apply any algorithm to the data we first needed to understand the general layout of the data we are dealing with.

(more introduction)

2 ALGORITHM DESIGN

2.1 K-MEANS

(Explain what k-means is, make a point to show points and problem with using euclidean distance)

2.2 K-MEANS++

Benefit over k-means and show a graph on time analysis differences

3 DATA PREDICTION

One of the problems with using a unsupervised learning algorithm is one cannot specify what values they want to predict. With k-means, the input is points in space and the output is clusters of how those points are grouped. We first applied k-means to the data without changing the data. Initial results as expected showed no particular cluster or clusters of bad businesses. In order to get the algorithm to cluster bad business together we have to adjust the data to exploit fields that are closer related to predicting a bad business and then normalize or remove fields that have no discrimination between good or bad business.

Show example why need to normalize and weight data for this problem. Some type of graph that conveys the idea in elementary level.

3.2 PREDICTING ATTRIBUTES

First we selected which fields that has predictability and then removed all fields that didn't seem to have enough predictability for clustering the way needed. After selecting which fields we are keeping we then ran some tests. The tests showed as before, no clear clustering of bad businesses. Next we ordered which fields have the most predictability and

apply weights appropriately. The idea is separate the bad businesses from the good ones with a larger euclidean distance for the k-means algorithm. Applying weights did not have noticeable results. Still thinking about what we can do to manipulate the data to help the bad businesses cluster together, we decide to move on to the other variables that are important to getting results out of k-means.

4 CHOOSING K CLUSTERS

K-means' performance is greatly affected by choosing the correct number of k clusters. (Explain why choose clusters are important, then explain what we tried)

5 CONCLUSION

K-means was not a good choice for predicting the field we wanted no. No conclusive clustering. More on why I think that is and what might be a better algorithm choice or next algorithm would want to try.

ACKNOWLEDGEMENTS

Groups members were great help and what they did...

REFERENCES

Cite the papers and the sources I used