Predicting stars: application of three machine learning algorithms

Marco Antonio Andrade Barrera[*]

November 14, 2015

Introduction

In recent years, supervised machine learning has become a boon in the social sciences, supplementing assistants with a computer that can classify documents with comparable accuracy (Jurka 2012, 56). This types of works are classical applications of text mining, whose methods have one thing in common: text as input information (Feinerer, Hornik & Meyer 2008, 1).

Meanwhile, 'online recommendation communities, like Yelp, are valuable information sources for people' (Bakhshi, Kanuparthy & Shamma 2014, 1). However, 'user-generated reviews are usually inconsistent in terms of length, content, writing style and usefulness because they are written by unprofessional writers' (Fan & Khademi 2014, 1). Then, if we want to analyze this type of information, depth computational methods should be used in order to put the data in an usable format.

Many works have tried to address questions using the Yelp Dataset Challenge. To name a few, Ganu, Elhadad and Marian (2009, 1) proposed a way to improve rating predictions using an ad-hoc and regression based method, whose results 'show that using textual information results in better general or personalized review score predictions than those derived from the numerical star ratings given by the users'. In other work, Fan and Khademi (2014) use a combination of three feature generation methods as well as four machine learning models to find the best prediction result. Other researchers claim that combining topic modeling and sentiment analysis is possible to obtain better predictions.

The goal of this work is explore a sampling approach to predict users' rates using its free text alone. Due to, in general, machine learning algorithms for classification require a lot of time and computer resources, this way of analysis big datasets can be more flexible. In fact, we should make a balance between the gain obtained using full datasets and the loss of quality in predictions when using just samples.

Methods and Data

The dataset used in this papers is part of the Yelp Dataset Challenge, and corresponds to Round 6 of their challenge. 'Founded in 2004, Yelp is a large online recommendation

[*]mandradebs@gmail.com

community that is also a user-maintained business and service directory to help people find local business' (Bakhshi, Kanuparthy & Shamma 2014, 2). Although the available dataset contains a lot of variables with information about business, users, reviews, and tips, in this work only two variables are used: the review free text and the review's star rating, with 990,627 observations that consist of reviews about restaurants.

To reach the goal of this work, three machine learning algorithm are used for classification, we selected those considering that are low-memory algorithms (Jurka et al. 2013). Actually, by discounting the exploratory analysis, the next steps for this work are based on the start-to-finish product described by Jurka et al. (2013). Using the R package RTextTools, a document term matrix is generated, removing numbers, stem words, sparse terms and stop words. Then, the three algorithms are used to train data. The size of training and testing datasets were fixed to 80% and 20%, following the ideas of others works related to review text mining (Chada & Naik 2015, 2). Finally, we present precision and recall measures. 'Precision refers to how often a case the algorithm predicts as belonging to a class actually belongs to that class' (Jurka et al. 2013, 9), whereas that recall refers to the proportion of cases in a class the algorithm correctly assigns to that class.

The three trained algorithms are support vector machines (svm), glmnet and maximum entropy. The svm algorithm is a powerful technique for general (nonlinear) classification (Meyer 2015, 1). An intuitive explanation of support vector machines method can be found in Bennett & Campbell (2000). GLMNET 'fits a generalized linear model via penalized maximum likelihood'. The algorithm 'use cyclical coordinate descent, which successively optimizes the objective function over each parameter with others fixed, and cycles repeatedly until convergence' (Hastie & Qian 2014). Finally, maximum entropy is an algorithm that performance a multinomial logistic regression using an efficient C++ library that reduce memory consumption (Jurka 2012, 56).

Even with those methods the training process take a long time with low resources PCs. For that reason, we adopt an approach that uses only random samples. Although it is known that 'resulting predictions tend to improve (nonlinearly) with the size of the reference dataset' (Jurka et al. 2013), in following sections we will see that taking random samples is not a bad approximation to results using the full dataset. Three samples were taken, two of 5,000 reviews (0.5%) and one of 50,000 reviews (5.05%).

Results

As part of exploratory analysis, Figure 1 shows a histogram of the number of stars for all restaurants ($N = 990,627$). Defining X as the random variable that represents the number of rating stars for a new review, based of frequencies, the distribution of X is $P(X = 1) = 0.088$, $P(X = 2) = 0.0985$, $P(X = 3) = 0.1551$, $P(X = 4) = 0.3226$ y $P(X = 5) = 0.3358$. As it can be seen, if we try to predict the rating using only the more frequent category (5), then we expect to guess the 33.58% of the ratings. Of course, this case it is not useful because we could not 'guess' the other categories and it is presented only for comparative purposes.

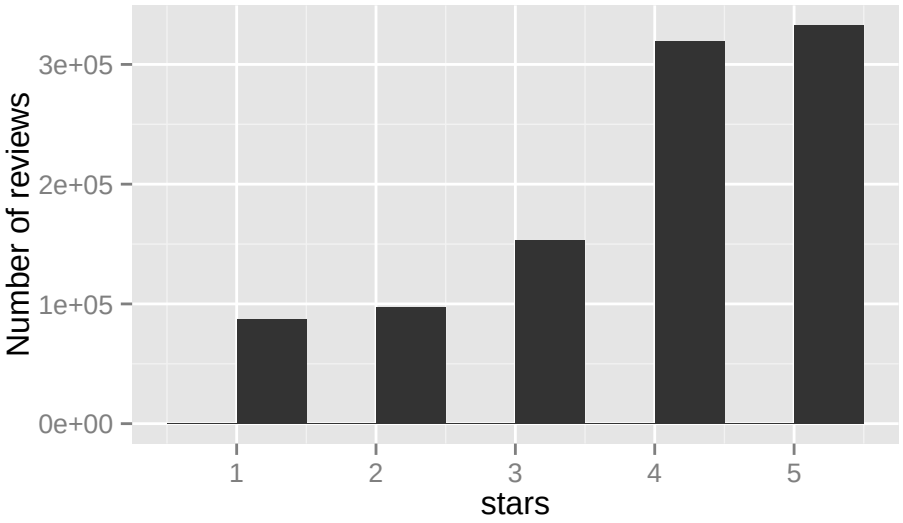

Figure 1: Histogram of stars for restaurants

Table 1 and Table 2 show the precision of each method, based on two random samples of 5,000 reviews each one (0.5% of total observations).

	1	2	3	4	5	General
SVM	0.51	0.46	0.45	0.45	0.56	0.49
GLMNET	0.52	0.39	0.46	0.45	0.52	0.47
MAXENTROPY	0.48	0.30	0.34	0.42	0.53	0.41

Table 1: First sample. Precision of each method by number of stars ($N = 5,000$)

	1	2	3	4	5	General
SVM	0.57	0.34	0.40	0.40	0.62	0.47
GLMNET	0.64	0.28	0.39	0.41	0.63	0.47
MAXENTROPY	0.56	0.29	0.39	0.40	0.58	0.44

Table 2: Second sample. Precision of each method by number of stars ($N = 5,000$)

In both cases, SVM and GLMNET get the best although poor precision. Particularly, the algorithms can predict better reviews ranked with one star (up to 62%). We get low precision, but nevertheless the results are comparable with others methods and they are not very different or better. For example, excluding the sentiment feature added by Chada and Naik (2015), the precision of some methods used in their work were 0.57 for logistic regression, 0.51 for multinomial naive bayes and 0.48 for nearest neighbors. But in the case of Cahda and Naik (2015) there is an additional difference, they trained the models using a data set of about 700,000 entries.

Table 3 and Table 4 show the recall measure of each method, for the same two random samples of 5,000 reviews.

Again, SVM is positioned in the best place, but in this case the higher recall measured is found in the five star ranking. This mean that SVM correctly assigned 66% and 70% of

	1	2	3	4	5	General
SVM	0.48	0.29	0.22	0.51	0.69	0.44
GLMNET	0.30	0.24	0.20	0.52	0.69	0.39
MAXENTROPY	0.43	0.34	0.34	0.41	0.53	0.41

Table 3: First sample. Recall of each method by number of stars ($N = 5,000$)

	1	2	3	4	5	General
SVM	0.49	0.31	0.22	0.53	0.62	0.43
GLMNET	0.36	0.12	0.17	0.60	0.69	0.39
MAXENTROPY	0.46	0.40	0.40	0.40	0.55	0.44

Table 4: Second sample. Recall of each method by number of stars ($N = 5,000$)

reviews, for first and second sample respectively, with five star ranking to that ranking.

The above random samples are very small with only 0.5% of the total reviews. If we increase ten times the size of sample, as we can see below, the results do not change dramatically. Table 5 and Table 6 show the precision and recall measures for the three algorithms, respectively, using a random sample of 50,000 reviews, 5.05% of the total entries.

	1	2	3	4	5	General
SVM	0.62	0.45	0.48	0.52	0.64	0.54
GLMNET	0.64	0.47	0.48	0.43	0.57	0.52
MAXENTROPY	0.57	0.37	0.44	0.53	0.64	0.51

Table 5: Third sample. Precision of each method by number of stars ($N = 50,000$)

The most favored method when we increase the size sample is MAXENTROPY, since its precision goes from about 0.40 in small samples to about 0.50 in a bigger sample. Almost the same occurs in the recall measure. The other two methods had smaller improvements.

Discussion

In this work three machine learning algorithms were tested with three samples of the Yeld Reviews Dataset. It was only an exercise and it is clear that here we are not discovering anything. Notwithstanding, some key point can be highlighted.

The first problem that we can easily see when analysis this kind of data is the size, the amount of information. Common computers can take days training just one machine learning algorithm, and for that reason, some authors recommend clod computing services for larger datasets (Jurka et al. 2013).

As a first approximation, this paper shows that using samples we can reach similar results of those obtained using the full datasets. However, in order to improve this paper, future work may explore this approach in deep, determining size of samples and the exact effect of taken just samples.

As a final remark, the precision and recall measures presented in paper are not very promising. In some cases, probability of guess the correct ranking is even lower that tossing a coin. Then,

	1	2	3	4	5	General
SVM	0.59	0.33	0.34	0.57	0.73	0.51
GLMNET	0.39	0.13	0.18	0.55	0.75	0.40
MAXENTROPY	0.59	0.38	0.40	0.49	0.70	0.51

Table 6: Third sample. Recall of each method by number of stars ($N = 50,000$)

other tools should be used trying to improve the results. In this regard, others tools like sentimental analysis (Chada & Naik 2015), ensemble agreement (Jurka et al. 2013) or simply considering additional variables (Bakhshi, Kanuparthy & Shamma 2014) have shown better results.

References

- Bakhshi, Saeideh, Partha Kanuparthy and David A. Shamma. 2014. If it is funny, it is mean: Understanding social perceptions of Yelp online reviews. https://s.yimg.com/ge/labs/v2/uploads/main3.pdf.
- Bennett, Kristin P. and Colin Campbell. 2000. Support Vector Machines: Hype or hallelujah? *SIGKDD Explorations* 2, n. 2, 1:13.
- Chada, Rakesh and Chetan Naik. 2015. Data mining Yelp Data - Predicting rating stars from review text. http://www3.cs.stonybrook.edu/~cnaik/files/data_mining_report.pdf.
- Fan, Mingming and Maryam Khademi. 2014. Predicting a business' star in Yelp from its reviews' text alone. ArXiv e-prints: 1401.0864.
- Ganu, Gayatree, No´emie Elhadad y Amélie Marian. 2009. Beyond the Stars: Improving Rating Predictions using Review Text Content. Twelfth International Workshop on the Web and Databases (WebDB 2009). USA.
- Hastie, Trevor and Junyang Qian. 2014. Glmnet Vignette. Stanford. https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html.
- Jurka, Timothy P. 2012. maxent: An R Package for low-memory multinomial logistic regression with support for semi-automated text classification. *The R Journal* 4, n. 1, 56:59.
- Jurka, Timothy P., Loren Collingwood, Amber E. Boydstun, Emiliano Grossman, and Wouter van Atteveldt. 2013. RTextTools: a supervised learning package for text classification. *The R Journal* 5, no. 1, 6-12.
- Meyer, David. 2015. Support vector machines. The interface to libsvm in package e1071. Austria: FH Technikum Wien.