Discovering Fraud in Online Classified Ads

Alan McCormick and William Eberle

Department of Computer Science, Tennessee Technological University, Cookeville, TN USA
ammccormic21@students.tntech.edu and WEberle@tntech.edu

Abstract

Classified ad sites routinely process hundreds of thousands to millions of posted ads, and only a small percentage of those may be fraudulent. Online scammers often go through a great amount of effort to make their listings look legitimate. Examples include copying existing advertisements from other services, tunneling through local proxies, and even paying for extra services using stolen account information. This paper focuses on applying knowledge discovery concepts towards the detection of online, classified fraud. Traditional data mining is used to extract relevant attributes from an online classified advertisements database and machine learning algorithms are applied to discover patterns and relationships of fraudulent activity. With our proposed approach, we will demonstrate the effectiveness of applying data mining techniques towards the detection of fraud in online classified advertisements.

Introduction

Online classified advertisements are a popular way to sell goods or services. The popularity of online classified ad websites such as Craigslist (www.craigslist.org), Backpage (www.backpage.com), Oodle (www.oodle.com), and eBay Classifieds (www.ebayclassifieds.com) is continuing to increase. The World Wide Web provides a convenient and easily accessible medium for users to list and browse advertisements when compared to more traditional media such as newspapers and printed booklets. The wide spread accessibility of the web has an unwanted effect of attracting online scammers who pose as genuine sellers by posting fake advertisements in an effort to defraud would be buyers. Scammers have the ability to steal millions of dollars from unsuspecting users and threaten the reputation and utility of online ad services.

There is no standard reporting of market or fraud statistics for online classified ads. Classified ad companies usually do not make public disclosures regarding revenue or fraud numbers. Victims may also not report occurrences of fraud because of embarrassment or uncertainty of where to make the report [National Consumers League 2012]. To estimate the amount of fraud that occurs in online classifieds we may consider the amount of revenue and popularity of such sites.

Revenue from online classifieds needs to be differentiated from the amount of money that changes hands in online classified transactions. For example, the Internet Advertising Revenue Report (IAB) conducted by PriceWaterhouseCoopers lists online classified ads revenue at $2.6 billion for the year 2011[Price Waterhouse Coopers 2012]. It defines ad revenue as the fees advertisers pay to internet companies to list specific products or services. AIM Group's Classified Intelligence Reports projects the popular site, Craigslist, to have revenue of $126 million in the year 2012, an increase of 9.7 percent from the previous year [Zollman 2012]. However, a large majority of ads placed on classified sites are free. Considering that only a very small percentage of ads are paid and that the person listing a paid ad expects a return or profit, it is reasonable to assume that the total amount of money exchanged through classified ad transactions is much greater than the site's revenue.

Craigslist is the most popular classified ads website. According to the web information service, Alexa, it ranks 9th in the U.S. and 42nd worldwide among all websites in overall popularity [Alexa 2012]. Craigslist's factsheet states that the site receives more than 50 billion page views and well over 100 million classified ad postings each month [CraigsList 2012]. Other large classified ad sites that are not far behind include eBay, Naspers (www.naspers.com), and Schibsted (www.schibsted.com). In some areas, smaller local classified sites are more popular. With billions of advertisements placed each year involving billions of dollars' worth of transactions, even if only a small percentage of those ads are fraud, it has the potential to cheat users out of many millions of dollars.

With transactions on the order of millions, it is imperative that these web-sites monitor for and attempt to detect potentially fraudulent activity. Not only is their bottom-line at stake, but even for those companies that do not charge for posting advertisements, their reputation can be compromised. This research focuses on applying data mining techniques to discover patterns and relationships in classified ad data. Fraudulent listings can then be detected

Figure 1: eBay Classifieds Homepage [eBay Classifieds]

by comparing newly placed ads with a known model of fraudulent activity.

Background

Existing methods for fraud or spam detection include community moderation, proprietary automated systems, GeoIP services and commercial fraud protection services.

In community moderating, other users of the site flag suspect listings. If an advertisement receives enough votes, it is either removed or sent to the operators of the site for review. Community moderation has the advantage of being easy to implement. It crowd sources detection and removal of fraud by relying on the judgment and intelligence of other human users. The community is likely able to adapt and recognize new forms of fraud as they are discovered. Relying on community moderation has several drawbacks. Users must be exposed to fraudulent ads before they can be flagged. Not all information may be available to other users. For example, IP addresses and other machine-centric data, or personal account details are unlikely to be public. While human intuition and judgment can be beneficial, users are not experts at recognizing fraud and therefore may not be the best at detecting it. Users may also abuse the system by flagging content for any reason, including competition or retribution against other users.

Websites also employ their own proprietary systems for fraud removal. These systems have the advantage of being automated and can incorporate domain specific knowledge. Proprietary systems may contain a set of hand crafted rules. Each rule considers an attribute of the ad and either increases or decreases its score. New listings are scored by each rule and then its value is compared to a defined threshold. If the ad exceeds the threshold then it is marked suspect. Hand crafted rules rely on a domain expert to have sufficient expertise to be able to create rules that effectively discriminate between fraudulent and normal listings. It may be difficult to craft those rules and best weights manually. If the rules are successful at preventing fraudulent ads, scammers will likely adapt and change their methods, resulting in the need to update the rule set.

GeoIP services, such as MaxMind [MaxMind 2012], are able to translate a user's IP address into a geographic location. This location can be compared to the location of the advertisement. If the IP address corresponds to a different location it may indicate the advertisement is fraudulent. While GeoIP services are beneficial, more information is available and may be needed to detect fraud. It is also not uncommon for users to post ads while out of town when their location may not match the listing.

Tran et. al. proposed a method to detect spam in online classified advertisements [Tran et al. 2011]. They collected advertisements placed on Craigslist over several months. Volunteers then created a training set from a small sample of those advertisements by labeling the ads as spam or not spam. A decision tree classifier was trained on features extracted from the advertisements to detect instances of spam. They showed significant improvement with their classification method over traditional methods of web spam detection. While the authors give an example of an advertisement that is clearly fraudulent, they do not give a clear description of the difference between spam and fraud.

It is important to make a distinction between fraud and spam. Fraud and spam may have different motives and intentions and therefore different signatures. Methods that work to detect spam may not be suitable for detecting fraud and vice versa. Spam is unsolicited advertisements that flood services, such as classifieds. Contrary to classifieds placed by individuals to sell items or services to other local individuals, spam is commercial and global in nature. Spam often includes dubious products such as knockoffs, cheap prescription medications, or get-rich-quick schemes. There may be no attempt to disguise the intent of spam advertising. The goal is to entice users to pay for the products being advertised and they may be selling actual products. In contrast, the goal of fraudulent advertisements is to deceive the user by appearing to be a normal listing. They attempt to blend into the hosting service by posting similar goods and services to others that are already available. Usually, no real product exists. The entire advertisement and backstory are designed to lure users into a scam. For example, the seller may pretend to be out of town and request funds be transferred electronically. After receiving the funds, the seller disappears and the item being sold is nowhere to be found.

In Tran et. al.'s experiment, the training data provided by the volunteers indicated 17% of the ads were spam. It seems unlikely that such a large percentage of all ads are fraudulent. Without a clear distinction, it might be assumed that the labeled spam included a mixture of spam and fraud. Also, similar to community moderation, the volunteers used to label the training data are probably not experts at detecting fraud. It is possible, that only the most conspicuous instances of fraud were labeled. Considering that spam may not be disguised and that only clear examples of fraud may have been detected, it is likely easier to train a classification model to detect those instances.

In our approach we will examine data provided by a classified ads website. The site operators have also provided a list of ads that were previously determined fraudulent. It is our belief that this should provide us the most accurate data available because the site operators are probably best experienced at labeling fraud. In addition, because the data is being provided from an internal source, features that are not made public can be extracted from the data. These attributes include such information as the IP address and date the user joined. We use this data to build a classification model. This model could be used to detect fraudulent ads the moment they are placed which prevents users from being exposed to the fraudulent ads, as in the case of community moderation. Because the user's information does not need to be shared and tracked by a third party, there are less concerns regarding user privacy.

Data

Collection

In order to validate our approach to detecting fraud in online classified advertisements, a database of advertisements and associated data was provided by a company that maintains an online classified ads website. The company currently marks fraudulent ads using a system of handmade rules and human review that is labor intensive. Each advertisement contains fields for the date the ad was placed, ad title, description of the item, category, price, location, paid placement options, images, and user identification. The ad database also includes user account data and logs of user activity.

In order to be able to work with the website data, personal user information was anonymized in accordance with the website's privacy policy. Therefore, names, street addresses, e-mail address, and other personal account details were removed prior to receiving the data. Each advertisement record links to a user record containing a unique integer identifier, date the user registered, city, state, and zip code. The site also maintains a list of user login records. Each time a user logs in a record is created containing the date of login, session id, IP address, user agent, breadcrumb, and geoip data. The breadcrumb is a 128-bit randomly generated identifier stored in a cookie on the user's machine. The purpose of the identifier is to uniquely identify user activity across multiple sessions and even different user accounts. Geoip data is extracted and stored with each login record from a web service that provides additional information about an IP address. This geoip data includes the country code, region, city, postal code, latitude, longitude, and ISP associated with the IP address. The company also provided a table of advertisements that were previously marked as fraudulent using their current methods of detection.

Table I shows a breakdown of the number of ads provided. While fraud represents less than 1% of the advertisements provided, it is still presents a significant problem in day to day operations and presents a risk to users of the website. The challenge lies in finding the proverbial needle in the haystack by identifying those fraudulent ads.

Table I. Breakdown of number of advertisements and fraud

	Normal	Fraud	Total
Current Ads	18,936	64	19,000
Expired Ads	417,564	2,436	420,000

Preparation

The provided database had some missing information which can be attributed to deletion of old records or changes in the site structure over time. Advertisements that were missing important attribute data were deleted from the sample database. The advertisements category id indicates in which category the ad was placed (e.g. electronics, pets, or automobiles). This category could likely be of interest since scammers often target particular types of items. Approximately 283,000 ads did not contain the category that they were originally placed and were removed from the sample database. To match an advertisement record to a user login record, an entry in the login table must exist for the user who placed the ad prior to the time the ad was created. The user login history had been cleared in the past, leaving many advertisements without corresponding user login information. This information is likely to be useful because it contains the aforementioned geoip data and unique crumb. Approximately 93,000 ads that were missing their corresponding user login data were also purged from the database. After purging those records, there were 61,377 remaining ads in the dataset. Future work with this company will attempt to recover some of the missing information in order to retain those records.

Location information for each ad is directly input by the user. This can leave many advertisements with misspelled city names. Further, many cities in the database were blank for unknown reasons. To correct the blank cities, a table was created that counted the most frequent city value for each zip code, then each missing city was updated by default to the most common city in that area. At this time, it was decided not to correct misspelled city names, because this may be relevant in fraud detection, i.e. someone not local to the area may be more likely to misspell the name of the city.

Feature Extraction

In order to facilitate pattern recognition for fraud detection, several features were extracted from the existing attribute data:

time_since_join: For each advertisement, the corresponding user record was located and the difference between the creation date of the ad and the registration date of the user was calculated and stored in minutes.

has_url: A regular expression was used to match any URL that may exist in the description of an advertisement. A Boolean attribute, has_url, was created and set to true for any ad that matched otherwise false.

has_email: Similarly to has_url, a regular expression was created to match any e-mail address and executed on all advertisements descriptions. A Boolean was added to each ad and set to true for any ad who matched the regular expression, else it was set to false.

geo_city, geo_state, geo_zip, geo_country, geo_ISP: Each advertisement includes the user id which corresponds to the user who posted that ad. The user logs table was searched for the login that immediately preceded the posting of the advertisement. Then, the geoip information was split into its respective fields and added to the attributes for that ad.

match_city, match_state, match_zip, match_US: Each prior mentioned geoip location attribute was compared to each advertisements posted location and a flag was stored for each part that matched. For country, a flag was created and set to true iff the *geo_country* field matched the value "US".

same_crumb: For each advertisement, the corresponding user login record was located and the uniquely generated crumb identified. Using this crumb value, a query counted the number of distinct users that have logged in with the same crumb value. This number was stored for each advertisement.

The user crumb keeps a unique identifier on a user's machine. If an advertisement is marked as fraudulent, other fraudulent advertisements can be directly found by finding those which share the same crumb. While this can be a powerful tool in finding fraud, it does not generalize well to finding new instances of fraud that are unknown. Therefore, the *same_crumb* attribute was extracted as an indirect way to use this field without relying on prior knowledge of specific fraud.

Each advertisement includes fields which indicate paid placement options for an advertisement. It may be thought, that fraudulent users would be less likely to pay for services. However, in conversations with the company who provided the data, it was stated that fraudulent users have used stolen account information to purchase placement options. When the company discovers the fraud, they must refund the charges. Therefore, if the company believes an ad may be fraudulent, then they skip the paid placement options entirely. While this makes good business sense, it directly biases the usefulness of such attributes. Combined with this knowledge and considering only a very small percentage of legitimate users pay for placement options, these fields were not used.

Experiments

Weka [Hall et al. 2009], a machine learning software suite, was used to train several classifiers to detect fraud. A program was implemented to query the ad database, and export each advertisement's respective attributes and fraud classification to the Weka specified ARFF format. An experiment was created to train and test the following classifiers: Naïve Bayes, Multilayer Perceptrons (artificial neural network), J48 (C4.5) decision trees, and random forests.

Naïve Bayes is a simple classifier that assumes that each attribute is independent of the other attributes when determining the presence of the class (e.g., fraud) [John and Langley 1995]. For example, it would assume that *match_city* and *match_state* were unrelated and that each attribute independently contributes to the classification. Even if this assumption is incorrect, Naïve Bayes usually performs well. Naïve Bayes is often used as a simple and efficient classifier and serves as a baseline comparison for the other algorithms.

An artificial neural network was chosen for its ability to generalize and approximate any arbitrary function. Artificial Neural Networks (or ANNs) consist of a graph of interconnected neurons or nodes. Each neuron uses a function that coverts its weighted inputs into its output activation. The training phase of the network uses a method called back-propagation to assign weights to each input value and neuron in the network [Haykin 1998]. This allows the network to effectively learn the importance of each attribute when presented labeled training data. After the network is built, it can be evaluated for a new set of inputs. The resulting output is a score that represents the strength of the prediction that the instance is a member of the class.

However, with an ANN it is often not possible or very difficult to extract rules for classification from the model. Therefore, it was decided to include the J48 decision tree algorithm, which is an implementation of C4.5 [Quinlan 1993], which is similar to the approach used by Tran et. al discussed earlier [Tran et al. 2011].

Finally, random forests is a method of aggregating multiple decision tree classifiers into one classifier [Breiman 2001]. Multiple decision trees are trained and vote on the classification of each instance. This concept is called bagging [Breiman 1996] and it was included to examine if this would offer an improvement in performance over the other algorithms.

Results

Table II. Classifier Statistics

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Naïve Bayes	0.912	0.075	0.102	0.912	0.183	0.969
ANN	0.356	0.001	0.696	0.356	0.471	0.974
J48	0.612	0.002	0.703	0.612	0.655	0.911
Random Forest	0.623	0.002	0.724	0.623	0.67	0.964

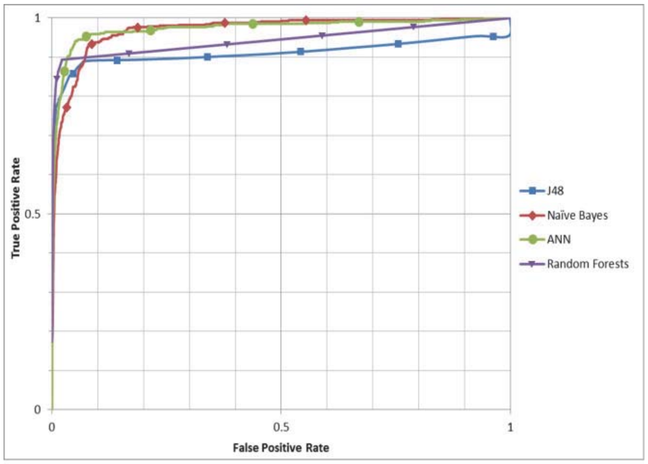

Figure 2: ROC graph showing performance of classifiers

In Figure 2, at first glance, all of the classifiers have very high accuracy. However, when looking closer at Table II, we can see that the recall rate and precision of the classifiers are far from perfect. Recall is the ability of the classifier to correctly identify instances of fraud, while precision is the ratio of true positive to false positives. The difficulty with recall lies in the makeup of the initial data. Fraud constitutes less than 1% of the total set of data. Therefore, even if each classifier were to miss a large percentage of fraudulent ad classifications it still may have an accuracy of over 99%. The ultimate goal is to maximize the recall rate for fraud, while preventing loss in precision. We can use the ROC graph (Figure 2) to determine how each algorithm could be adjusted with respect to this trade off. As we adjust the classification threshold we can move along the curve for each classifier. As the true positive rate increases, when moving up the y-axis, we can correctly identify a larger percentage of fraud, which increases recall rate. But, any shift in the curve to the right increases the false positive rate and greatly affects precision since a large majority of the advertisements are not fraud.

The Naïve Bayes classifier does fairly well. However, it does not perform as well as the other classifiers in the region of the graph between 0 and 10% false positive rate. This region of the graph is of the most interest because it still maintains a reasonable level of precision. The ANN could arguably have the best performance. Near the knee of the ROC curve, the classifier has a recall rate of over 90% and a precision of about 33%. Random Forests out performs J48 as expected, since it contains a vote of many decision trees, and an interesting feature is that it seems to have higher precision than the ANN up until about 90% true positive rate in which the ANN surpasses its performance.

Overall the classifiers are very accurate when classifying ads as fraud or not fraud. However, most of the correct classifications are of advertisements that are not fraud which represent over 99% of the dataset. For example, the ANN has an accuracy exceeding 99%, but a low recall rate of 35%. Using the ROC curve information, we can adjust the threshold of the classifier to detect 90% of the fraudulent ads but at a cost of reducing the precision to 33%. With that threshold we would detect 9 out of 10 advertisements that were fraudulent, but we would also have on average about 2 false positives for every hit. This is likely a reasonable tradeoff. The cost of a false positive is human labor. Each ad the classifier suggests is fraud is an advertisement that a human must review. At a rate of 2:1, falsely classified ads to actual fraudulent ads, with a small percentage of ads being fraud, this is still manageable amount of work. However, any improvement to the classifiers will be beneficial in reducing the amount of labor. This is especially true if the website grows and the number of advertisements being considered becomes much larger.

Conclusions and Future Work

This work presents an approach for discovering fraud in classified ads. Using well known machine learning algorithms, our initial data mining approach was comparable in performance to the current methods of detection used by the company. In addition, we believe that our approach can be improved when we have access to more of the company's information, much of which was not provided or was intentionally discarded.

Comparing the overall performance of the tested classifiers, given our current data, random forests appears to be the best choice. While the ANNs recall rate eventually surpasses random forests it is not until a point in which the precision of the classifiers suffer a large decrease. Using random forests may also provide more insight into the usefulness of each attribute by examining the generated decision trees and aid in the selection and extraction of future attributes. It may be possible that this relationship will change as we find and extract more useful features.

In the future, we plan to modify our data mining approach to extract other relevant features from the advertisement data. With new attributes and perhaps some refinement to the training of classifiers, we believe we can work to further improve the recall rate for fraud while minimizing the loss of precision. After fine tuning our method of detection, we hope to create a tool to detect fraudulent ads the instant they are placed.

Acknowledgements

Special thanks to Rob Weeden and Alex Zlaten for their help in providing classified ad data and their ongoing support.

References

Alexa. 2012. http://www.alexa.com .

Breiman, L. 1996. Bagging predictors. *Mach. Learn.* 24, 2 (August 1996), 123-140.

Breiman, L. 2001. Random Forests. *Mach. Learn.* 45, 1 (October 2001), 5-32

CraigsList. 2012. Craigslist Factsheet. http://www.craigslist.org/about/factsheet

eBay Classifieds. 2012. http://www.ebayclassifieds.com

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*. Volume 11, Issue 1.

Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Second Edition.

John, G,, Langley, P. 1995. Estimating Continuous Distributions in Bayesian Classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo*, 338-345.

MaxMind. 2012. *Fraud Detection through IP Address Reputation and a Mutual Collaboration Network*. (http://www.maxmind.com/Maxmind_WhitePaper.pdf)

National Consumers League, 2012, http://www.nclnet.org/.

PriceWaterhouseCoopers. 2012. Internet Advertising Revenue Report. sponsored by The *Internet Advertising Bureau, 2011 Full Year Results, April 2012.*

Quinlan, J.R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Tran, H., Hornbeck, T., Ha-Thuc, V., Cremer, J., and Srinivasan, P. 2011. Spam detection in online classified advertisements. In *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality* (WebQuality '11). ACM, New York, NY, USA, 35-41.

Zollman, P. 2012. Craigslist 2012 revenues increase 9.7%. *AIM Group. http://aimgroup.com/2012/11/07/craigslist-2012-revenues-increase-9-7-big-four-battle-for-global-classified-lead/*