International Journal of Future Innovative Science and Engineering Research (IJFISER)
Volume – 3, ISSUE – I **ISSN (Online):2454- 1966**

Research Manuscript Title

Detecting Fraud Applications Using Ensemble Learning

Mr.A.Syed Mustafa, M.Aishwarya, D.Anusha Ganapathy, T.Gowri

Department of Information Technology,
K.S.Rangasamy College of Technology,
Tiruchengode, Tamilnadu, India.

Corresponding author E-Mail-ID: aishwaryamohan1196@gmail.com

March – 2017

www.istpublications.com

Detecting Fraud Applications Using Ensemble Learning

Mr.A.Syed Mustafa, M.Aishwarya, D.Anusha Ganapathy, T.Gowri

Department of Information Technology,
K.S.Rangasamy College of Technology,
Tiruchengode, Tamilnadu, India.

Corresponding author E-Mail-ID: aishwaryamohan1196@gmail.com

ABSTRACT

***ABSTRACT*- Ranking fraud in the mobile App refers to fraud activity. Indeed, it becomes more and more frequent for App developers to use shady means, such as inflating their Apps' sales or posting phony App ratings, to commit ranking fraud. While the importance of preventing ranking fraud has been recognized, there is limited understanding and research in this area. To this end, in this project, we provide a holistic view of ranking fraud and propose a ranking fraud detection system for mobile Apps. Specifically, it first proposes to accurately locate the ranking fraud by mining namely leading sessions, of mobile Apps. Such leading sessions can be leveraged for detecting the local anomaly instead of global anomaly of App rankings. We proposed a approach Ensemble learning using (NaïveBayes,IBK,SVM)classification algorithm suggest that ensemble based machine learning methods provide better performance in classification. Artificial neural networks (ANNs) are rarely being investigated in the literature of sentiment classification Furthermore, we investigate three types of evidences, i.e., ranking based evidences, rating based evidences and review based evidences, by modeling Apps' ranking, rating and review behaviors through statistical hypotheses tests. Finally, it evaluates the proposed system with real-world App data collected from the iOS App Store for a long time period. In the experiments, we validate the effectiveness of the proposed system, and show the scalability of the detection algorithm as well as some regularity of ranking fraud activities.**

Keywords: Ensemble learning, Reviews, Artificial neural Networks, Modeling Apps, iOS App.

1. INTRODUCTION

Ranking Fraud Detection:

The number of mobile Apps has grown at a breathtaking rate over the past few years. For example, as of the end of April 2013, there are more than 1.6 million Apps at Apple's App store and Google Play. To stimulate the development of mobile Apps, many App stores launches App leader boards, which demonstrates the rankings of popular Apps. Indeed, the App leader board is one of the most important ways for promoting mobile Apps. A higher rank on the leader board usually leads to a huge number of downloads and dollars in revenue. Therefore, App developers tend to explore various ways such as advertising campaigns to promote their Apps in order to have their Apps ranked as high as possible in such App leader boards. However, instead of depending on traditional marketing solution, App developers find out to fraud dataset means to boost their System [1] .This is usually implemented by using so-called "bot farms" or "human water armies" to inflate the App downloads, ratings and reviews in a very short time. For example, an article from Venture Beat reported that, when an

App was promoted with the help of ranking manipulation, it could be propelled from number 1,800 to the top 25 in Apple's top free leader board and more than 50,000-100,000 new users could be acquired within a couple of days. In fact, such despite fraud raises great concerns to the mobile App industry [2]. For example, Apple has warned of cracking on App developers who commit ranking fraud in the App store. Indeed, our correct fully observation reveals that mobile Apps are not always ranked high in the leaderboard, but only in some leading events, which form different leading sessions. Note that, we will introduce both leading [3].

Ensemble Modeling

Ensemble modeling is a powerful way to improve the performance of your model. It usually various form to apply ensemble learning over and above various models you might be building. People have used ensemble models in competitions like Kaggle and benefited from it. Ensemble learning is a huge form of dataset and is only confined by your own imagination. For the purpose of this article, it will cover the basic concepts and ideas of ensemble modeling. This should be enough for you to start creating ensembles at your own end. As usual, we have tried to keep things as simple as possible [4]. Ensemble is the art of combining huge set of learners (individual models) together to improvise on the stability and predictive power of the model. In the above example, the way it combine all the predictions together will be termed as Ensemble Learning [5].

In our project, we will discuss about a few ensemble techniques widely used in the industry. Before we get into techniques, let's first understand how do it actually get different set of learners. Data types can be different from each other for a variety of reasons, starting from the population they are built upon to the modeling used for building the model. Here are the top 4 reasons for a model to be different. They can be different because of a mix of these factors as well [6]:

1) *1. Difference in population*
2) *2. Difference in hypothesis*
3) *3. Difference in modeling technique*
4) *4. Difference in initial seed*

Sentiment analysis is an interdisciplinary area which comprises of natural language processing, text analysis and computational linguistics to identify the text sentiment. Web has been a rapidly growing platform for online users to express their sentiment and emotion in the form of text messages. As the opinionated texts are often too many for people to wade through to make a decision, an automatic sentiment classification method is necessary the effectiveness of the neural networks based methods in sentiment classification as the interest of this study for three reasons [7]. First, neural network based models has been very successfully applied to text classification and many other supervised learning tasks. The deep architectures of neural networks with layers (hidden) represents intelligent behavior more efficiently than ''shallow architectures" like support vector machines (SVMs) [8].

The major features of neural networks such as adaptive learning, parallelism, fault tolerance, and generalization provide superior performance. In spite of the above mentioned features of neural network methods, the research work on sentiment classification addressed the importance of integrating classification results provided by multiple classifiers. In addition, not such investigate has been carried out in sentiment classification to evaluate the benefits of combining neural network algorithms in order to increase the accuracy[9]. Moreover, most existing studies in sentiment classification used the traditional measures for performance evaluation. A recent study, however, showed that various quality measures can be proposed to evaluate the accuracy of classification models in another domain like software fault prediction. This further motivates this study to evaluate the various performance evaluation metrics [10].

2. EXISITING SYSTEM

In case of the existing system the fraud is detected after the complaint of the Mobile apps holder. And also now a days lot of online purchase are made so we don't know the person how is using the Mobile apps online, we just capture the IP address for verification purpose. So there need help from the cyber crime to investigate the fraud [11]. To avoid the entire above disadvantage it proposes the system to detect the fraud in a best and easy way.a holistic view of ranking fraud and propose a ranking fraud detection system for mobile Apps. Specifically, it first proposed to accurately locate the ranking fraud by mining the active periods of mobile Apps. Such leading sessions can be leveraged for detecting the local anomaly instead of global anomaly of App rankings [12]. Further, we investigate three types of evidences, they are ranking based evidences, rating based evidences and review based evidences. In addition, we propose an optimization based aggregation method to integrate evidences for fraud detection [13].

Finally, it evaluates the proposed system with real-world App data collected from the App Store for a long time period. In the experiments, we validate the effectiveness of the proposed system, and show the scalability of the detection algorithm as well as some regularity of ranking fraud activities [14]. First, ranking fraud does not always happen in the whole life cycle of an App, so we need to detect the time when fraud happens. Such challenges can be regarded as finding the local anomaly instead of global anomaly of mobile Apps [15]. Second, due to the huge number of mobile Apps, it is difficult to manually label ranking fraud for each App, so it is important to have a scalable way to detect ranking fraud. Finally, due to the dynamic nature of rankings, it is not easy to identify. Indeed, our careful observation reveals that mobile Apps are not always ranked high in the leader board, but only some leading events, which form different leading sessions. Note that it will introduce both leading events and leading sessions in detail later. In other words, ranking fraud usually happens in these leading sessions. Therefore, detecting ranking fraud of mobile Apps is actually to detect ranking fraud of mobile Apps. Then, it was found that the fraud Apps often have different ranking patterns in each leading session compared with Apps. Nonetheless, the ranking based evidences can be affected by App developers' reputation and some legitimate marketing campaigns, such as "limited-time discount" [16]. As a result, it is not sufficient to use rank based evidences. Therefore, it proposes two types of fraud evidences based on Apps rating and review history, which reflect some anomaly patterns from Apps' historical rating and review records. In addition, it develops three types of evidences for evaluating the credibility of leading sessions from mobile Apps.

DRAWBACK

- The detection of the fraud use of the Mobile apps is did not easy to found.
- In this system even the original Mobile apps holder is also checked for fraud detection.

We didn't find the most accurate detection on fraud apps using this technique.

3. PROPOSED SYSTEM

In proposed system, we present A Novel Ensemble Learning using PCA .which does not require fraud signatures and yet is able to detect frauds by considering a Mobile apps holder's spending habit. Mobile apps transaction processing sequence by the stochastic process of an EL(Ensemble Learning). The details of items purchased in transactions are usually not known to any Fraud Detection System (FDS) running at the bank that issues detecting Mobile apps to the Mobile apps holders. Hence, we feel that EL(Ensemble Learning) is an ideal choice for addressing this problem [17]. Another important advantage of the EL (Ensemble Learning)-based approach is a reduction in the number of False Positives transactions identified as malicious by an FDS. An FDS runs at a detecting Mobile apps issuing bank. FDS gets the Mobile apps details and the value of purchase to verify, whether the transaction is genuine or not. The types of goods that are bought in the transaction are not known to the FDS. If the FDS confirms the transaction to be of fraud, it raises an alarm, and the issuing bank declines the transaction [18].

i. Perform data pre-processing and segregate the features (product attributes).

ii. Develop word vector for model I using unigram features, model II using unigram and bigram features and model III using unigram, bigram and trigram features.

iii. Perform PCA on the model I, II and III to produce reduced feature set for all the models.

iv. Develop the classification methods using the respective training data set with the dimension reduced feature set.

a. Develop support vector machine model.

b. Develop linear discriminant analysis model.

c. Develop the BPN based neural network model.

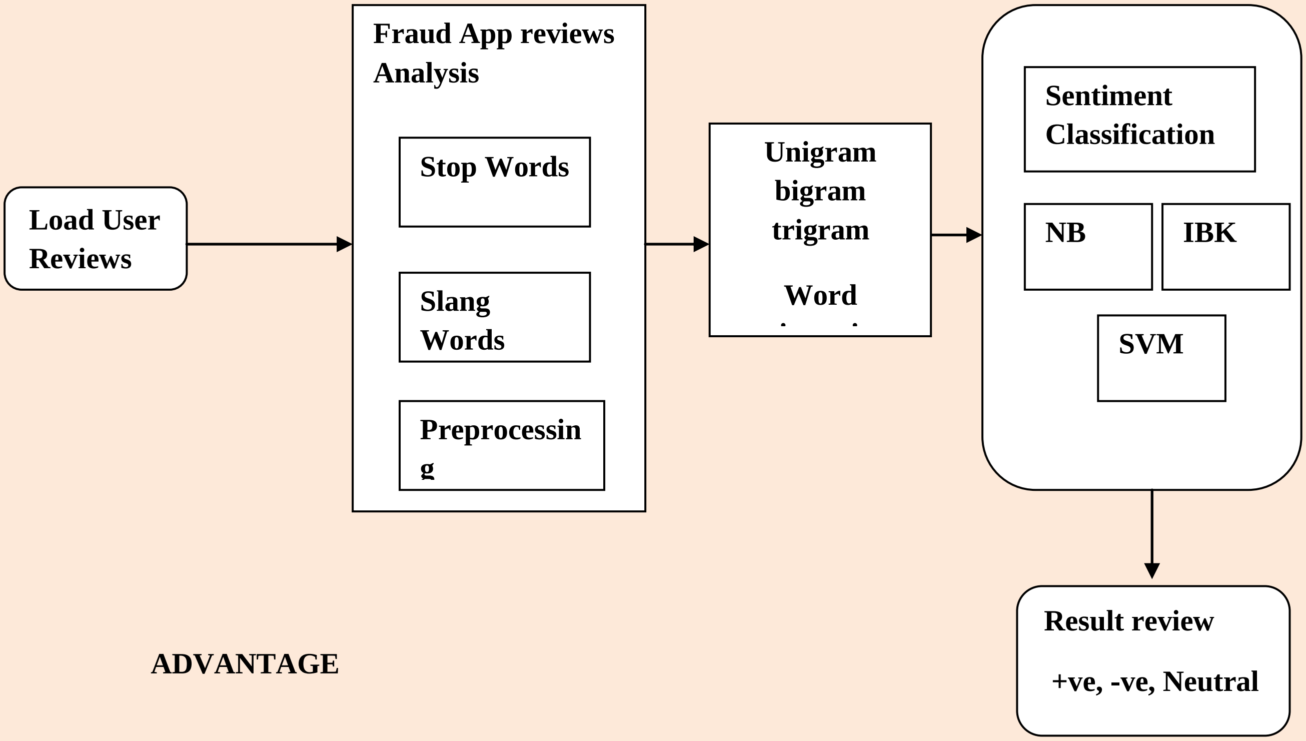
d. Develop the PNN based neural network model.

e. Develop the homogeneous ensemble model based on PNN.
v. Classify the class (positive or negative) of each review in the test data set.
vi. Compare the classification results with actual results.
vii. Compute the quality parameters such as the overall error rate (misclassification), completeness, correctness, efficiency and effectiveness and compare the classification accuracy of the methods and compute training time of learning models.

The domain chosen for the study is digital pics art reviews.

A Java web crawler was developed to download 970 positive reviews and 710 negative reviews randomly. In the crawled reviews, it is found that, there are borderline and neutral reviews in between along with the clear positive and negative reviews [19]. We discard a review if it is not clearly aligned toward positive or negative sentiment. Outliers analysis is performed (Briand and Wust, 2002) [20]. Twenty-five sentences are identified as outliers and are not considered for further processing. As a result, there are 950 positive and 705 negative reviews. For our binary classification problem, to avoid the imbalanced class distribution, we selected 600 positive and 600 negative reviews randomly to establish the data set.

4. BLOCK DIAGRAM



ADVANTAGE

- The detection of the fraud use of the Mobile apps is found much faster that the existing system because we focus by consumer opinons.
- In case of the existing system even the original Mobile apps holder is also checked for fraud detection.

5. METHOD AND SYSTEM

DATA PRE-PROCESSING

Previous studies revealed that pre-processing of text messages can improve the performance of text classification. The steps involved in data pre-processing are tokenization and transformation to reduce ambiguity. Then stop words are filtered to remove common English words such as ‘a’ and ‘the’ etc [21]. Porter stemmer is then used for stemming. After pre-processing, the reviews are represented as unordered collections of words (bag of words).

FEATURE IDENTIFICATION

Each product has its own set of features. As product reviews are about product features (also defined as product attributes) the product features are good indicators in classifying the sentiment of product reviews for product review based sentiment classification. Hence, the right features can be selected based on product features [22].

To construct a feature space for product feature based sentiment classification, product features can be included and treated as features in the feature space. For each of the positive and negative review sentences represented as bag of words, the product features in the review sentences are collected. From the machine learning perspective, it is useful for the features to include only relevant information and also to be independent of each other .The unique characteristic of a product feature is that they are mostly nouns and noun phrases by part of speech tagging. In order to identify the nouns and noun phrases, part of speech (Stanford POS) tagging is applied and then association mining is done on the review sentences of nouns and noun phrases to identify frequent features. Compactness pruning and redundancy pruning are applied on the frequent features to obtain more accurate features [23]. The product features extracted from review sentences are unigram, bigram and trigrams.

FEATURE VECTOR

Converting a piece of text into a feature vector model is an important part of machine learning methods for sentiment classification. A word vector representation of review sentences is created using the features identified. The feature vector model is constructed by using term presence method. Another focus of the work is to compare the influence of using different n-gram schemes. For this reason, the product features identified are grouped based on the word granularity as unigram, bigram and trigram (Table 2). In order to find the effect of the word size in the classification, three different models are developed with varying levels of word granularity.

FEATURE REDUCTION

Principal component analysis is a linear technique for dimensionality reduction which performs a linear mapping of the data to a lower dimensional space. The mapping is done in such a way that the variance of the data in the low dimensional representation is maximized thus resulting in new principal component variables (PC’s). A leading session is composed of several leading events. Therefore, we should first analyze the basic characteristics of leading events for extracting fraud evidences [24].

GUARANTEED RATING BASED EVIDENCES:

The ranking based evidences are useful for ranking fraud detection. However, sometimes, it is not sufficient to only use ranking based evidences. Specifically, after an App has been published, it can be rated by any user who downloaded it. Indeed, user rating is one of the most important features of App advertisement. An App which has higher rating may attract more users to download and can also be ranked higher in the leader board. Thus, rating manipulation is also an important perspective of ranking fraud. Intuitively, if an App has ranking fraud in a leading session s, the ratings during the time period of s may have anomaly patterns compared with its historical ratings, which can be used for constructing rating based evidences [25].

REVIEW BASED EVIDENCES:

Besides ratings, most of the App stores also allow users to write some textual comments as App reviews. Such reviews can reflect the personal perceptions and usage experiences of existing users for particular mobile App

6. CONCLUSION

Performances of neural network based approaches are compared with two statistical approaches. The homogeneous ensemble method performs better than other classification methods used. Among the individual neural network approaches used, PNN was highly robust. The performance was analyzed through the five quality parameters along with traditional techniques. The proposed approach of combining the neural network with PCA shows its superiority not only in quality measures, but also in training time. This indicates that feature reduction is an essential issue for learning methods in sentiment classification. Our experimental analysis shows that a hybrid combination of PNN and PCA could be a better solution for reducing the training time and increasing the classification performance. Our analysis also shows that the compound combination of unigram, bigram and trigram performs better for almost all the prediction models. The possible reason for the better performance of PNNs is because of the combined effect of the computational capability and flexibility, by retaining its simplicity. The prediction accuracy of the ensemble method can still be increased by increasing the number of classifier combinations. To test the limitations of the proposed method, future works could use different data domains and classification approaches probably with a data set of much a larger number of reviews.

REFERENCES

1) L. Azzopardi, M. Girolami, and K. V. Risjbergen, “Investigating the relationship between language model perplexity and ir precision-recall measures,” in Proc. 26th Int. Conf. Res. Develop. Inform. Retrieval, 2003, pp. 369–370.
2) D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., pp. 993–1022, 2003.
3) Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou, “A taxi driving fraud detection system,” in Proc. IEEE 11th Int. Conf. Data Mining, 2011, pp. 181–190.
4) D. F. Gleich and L.-h. Lim, “Rank aggregation via nuclear norm minimization,” in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011, pp. 60–68.
5) T. L. Griffiths and M. Steyvers, “Finding scientific topics,” Proc. Nat. Acad. Sci. USA, vol. 101, pp. 5228–5235, 2004.
6) G. Heinrich, Parameter estimation for text analysis, “ Univ. Leipzig, Leipzig, Germany, Tech. Rep., http://faculty.cs.byu.edu/~ringger/ CS601R/papers/Heinrich-GibbsLDA.pdf, 2008.
7) N. Jindal and B. Liu, “Opinion spam and analysis,” in Proc. Int. Conf. Web Search Data Mining, 2008, pp. 219–230.
8) J. Kivinen and M. K. Warmuth, “Additive versus exponentiated gradient updates for linear prediction,” in Proc. 27th Annu. ACM Symp. Theory Comput., 1995, pp. 209–218.
9) A.Klementiev, D. Roth, and K. Small, “An unsupervised learning algorithm for rank aggregation,” in Proc. 18th Eur. Conf. Mach. Learn., 2007, pp. 616–623.
10) A.Klementiev, D. Roth, and K. Small, “Unsupervised rank aggregation with distance-based models,” in Proc. 25th Int. Conf. Mach. Learn., 2008, pp. 472–479.
11) E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, “Detecting product review spammers using rating behaviors,” in Proc. 19thACMInt. Conf. Inform. Knowl. Manage., 2010, pp. 939–948.
12) Y.-T. Liu, T.-Y. Liu, T. Qin, Z.-M. Ma, and H. Li, “Supervised rank aggregation,” in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 481–490.
13) A.Ntoulas, M. Najork, M. Manasse, and D. Fetterly, “Detecting spam web pages through content analysis,” in Proc. 15th Int. Conf. World Wide Web, 2006, pp. 83–92.
14) K. Shi and K. Ali, “Getjar mobile applicat ion recommendations with very sparse datasets,” in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 204–212.
15) N. Spirin and J. Han, “Survey on web spam detection: Principles and algorithms,” SIGKDD Explor. Newslett., vol. 13, no. 2, pp. 50– 64, May 2012.
16) Abbasi, A., Chen, H., Thoms, S., Fu, T., 2008. Affect analysis of web forums and blogs using correlation ensembles. IEEE Trans. Knowl. Data Eng. 20 (9), 1168–1180.

17) Ahmed, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. 26 (3).

18) Briand, L., Wust, J., 2002. Empirical studies of quality models in object-oriented systems. In: Zelkowitz, Marvin (Ed.), . In: Advances in Comput-ers, 56. Academic Press, pp. 1–44.

19) Cambria, E., Mazzocco, T., Hussain, A., 2013. Application of multidimensional scaling and artificial neural networks for biologically inspired opinion mining. Biol. Insp. Cogn. Archit. 4, 41–53.

20) Chaovalit, P., Zhou, L., 2005. Movie review mining: a comparison between supervised and unsupervised classification approaches. In: Proceedings of the 38th Annual HICSS.

21) Chen, Long-Sheng, Liu, Cheng-Hsiang, Chiu, Hui-Ju, 2011. A neural network based approach for sentiment classification in the blogosphere. J. Inf. 5, 313–322.

22) Dave, K., Lawrence, S., Pennock, D.M., 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: The 12th WWW.

23) Gamon, M., 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In: Proceedings of the 20th International Conference on Computational Linguistics, p. 841.

24) Ghiassi, M., Olschimke, M.,Moon, B., Arnaudo, P., 2012. Automated text classification using a dynamic artificial neural network model. Exp. Syst. Appl. 39 (12), 10967–10976.

25) Ghiassi, M., Skinner, J., Zimbra, D., 2013. Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network. Exp. Syst. Appl.