OPEN ACCESS

1st International Electronic Conference on Entropy and Its Applications
3 - 21 November 2014

www.sciforum.net/conference/ecea-1

Conference Proceedings Paper – Entropy

Variations of Neighbor Diversity for Fraudster Detection in Online Auction

Laksamee Khomnotai [1,3] and Jun-Lin Lin [1,2,*]

[1] Department of Information Management, Yuan Ze University, 135 Yuan-Tung Road, Chungli, Taoyuan 32003, Taiwan
[2] Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taiwan
[3] Faculty of Management Science, Nakhon Ratchasima Rajabhat University, Nakhon Ratchasima, 30000, Thailand; E-Mails: k.laksamee@hotmail.com (L.K.)

***** Author to whom correspondence should be addressed; E-Mail: jun@saturn.yzu.edu.tw; Tel.: +886-3-4638800 (Ext.2611); Fax: +886-3-4352077.

Received: 12 September 2014 / Accepted: 21 October 2014 / Published: 3 November 2014

Abstract: Inflated reputation fraud is a serious problem in online auction. Recently, the neighbor diversity based on Shannon entropy has been proposed as an effective feature to discern fraudsters from normal users. In the literature, there exist many different methods to quantify diversity. This raises the problem of finding the most suitable method to calculate neighbor diversity for fraudster detection. In this study, we collect four different methods of quantifying diversity, and apply them to calculate neighbor diversity. We then use these various neighbor diversities for fraudster detection. Our experimental results against a dataset collected from a real world auction website show that, although these diversities are calculated differently, their performances on fraudster detection are similar.

Keywords: online auction; fraudster detection; neighbor diversity; entropy

---

1. Introduction

Online shopping/auction websites have gained increasing popularity for the past few years. This lucrative business opportunity has drawn not only the legitimate sellers to conduct their business online but also the fraudsters to commit fraudulent transactions. As a result, online shopping/auction websites often provide a reputation system to help their users to distinguish legitimate sellers from fraudsters. The reputation system requests the buyer and the seller of a transaction to give each other a

rating. Then, the reputation system calculates a reputation score of a user based on all the ratings the user received in his/her previous transactions. Intuitively, users with higher reputation scores are more trustworthy, and consequently are more likely to attract sales.

Because the reputation score of a user is based on all the ratings the user received in the past, a legitimate user requires time and effort to accumulate good ratings from other users. In contrast, a fraudster often commits the so-called "inflated reputation fraud" [1] to accumulate good ratings quickly, and cheats the reputation system into giving him/her a high reputation score . The inflated reputation fraud is accomplished by a group of collusive users who conduct many fake transactions for low-price merchandises and give each other good ratings. Because the reputation score is crucial for evaluating the trustworthiness of a user, detecting the inflated reputation fraud has become a key task for online shopping/auction websites.

In the literature, many methods had been proposed to detect fraudsters with inflated reputation in online auctions. Some of them adopted the concept of network graph to detect fraudsters who rely on their collaborators to boost up their reputations [1-5]. With this concept, social network analysis (SNA) has been found as an effective tool to detect fraudsters and their cohesive groups [1, 4, 5]. In our recent work [6], we proposed the concept of neighbor diversity to detect inflated reputation fraud. The neighbor diversity of a user quantifies the diversity of all traders that have transactions with the user. We showed that the neighbor diversity on the number of received ratings outperformed previous works that use *k*-core and/or center weight [1, 4, 5].

In [6], Shannon entropy [7] was adopted to quantify the neighbor diversity. However, different ways to define and calculate diversity exist in the literature. This motivates the idea of using various diversity definitions to calculate neighbor diversity for fraudster detection. Specifically, we adopt the four different definitions of diversity from Lin [8] to calculate the neighbor diversity. Our experimental results show, although these diversities are calculated differently, their performances on fraudster detection are similar.

The remaining of this paper is organized as follows. Section 2 reviews previous works on fraudster detection. Section 3 applies various definitions of diversity to calculate neighbor diversity. Section 4 describes the experimental settings, and Section 5 presents the experimental results. Finally, Section 6 concludes this paper.

2. Related Work

Detecting fraudsters with inflated reputation is a critical issue for online shopping/auction websites. Many approaches have been proposed in the literature. Some earlier approaches used the properties derived from the transaction history [2, 9], e.g. sum, average, and standard deviation of buying or selling price of merchandises in a period of time. Most of the recent approaches used SNA to detect group of fraudsters [1-5, 10-15].

Fraudsters who want to increase their reputation scores quickly often have many transactions with the members in their collusive group. Consequently, many approaches applied SNA to detect fraudsters by searching for the cohesive groups in the transaction network. In the SNA literature, characteristics such as *k*-plex, clique, betweenness, and *k*-core are often used to detect cohesive groups. Among them, *k*-core has been found to be the most effective for detecting fraudsters [1, 5].

To calculate *k*-core, a transaction network is first created from the transaction history. In the network, each node represents a user account, and each edge connecting two nodes represents a transaction between two users. Then, SNA is applied to discover *k*-core components. Although fraudsters frequently usually appear in *k*-core with $k \geq 2$ [1], using *k*-core alone results in low precision [4]. Alternatively, applying both center weight (CW) and *k*-core improves the precision, but the recall is reduced [4].

The concept of neighbor diversity was proposed to improve both precision and recall [6]. As mentioned before, fraudsters mostly do businesses with their collaborators to boost up their reputation. Consequently, their collaborators may share some similar characteristics, and the neighbor diversity of a fraudster's neighbors on those characteristics is likely to be small. Based on this notion, Lin and Khomnotai [6] showed that the neighbor diversity on the number of received ratings provides an effective way to discern fraudsters from normal users.

3. Variants of Neighbor Diversity

In this study, we use the number of received ratings as the target attribute to quantify the neighbor diversity because this attribute achieves the best performance in our previous work [6]. Specifically, the number of received ratings is first calculated for each user. Let *x* denote a user. The neighbors of *x* are the users who gave at least one rating to *x*. The neighbors of *x* are partitioned into several classes based on the number of received ratings. Let *r* denote the number of received ratings of a user. If $0 \leq r < 50$, then the user is placed into class 1. If $50 \times 2^{i-2} \leq r < 50 \times 2^{i-1}$, then the user is placed into class *i*, where $i > 1$. Let $p_i(x)$ denote the proportion of the *x*'s neighbors in the *i*-th class, and *n* denote the total number of classes. Then, the following constraints must hold.

$$0 \leq p_i(x) \leq 1, \text{for } i = 1 \text{ to } n \tag{1}$$

$$\sum_{i=1}^{n} p_i(x) = 1 \tag{2}$$

Next, we can apply various definitions of diversity to calculate neighbor diversity, as described in the following subsections.

3.1. Shannon Entropy Diversity

In [6], Shannon entropy [7] was adopted to calculate the neighbor diversity. The neighbor diversity of *x* based on Shannon entropy is denoted as $D_s(x)$, and calculated as follows:

$$D_s(x) = -\sum_{i=1}^{n} p_i(x) \log_2 p_i(x) \tag{3}$$

3.2. Canonical Form of Diversity

The notion of diversity is also widely used in many different areas. For example, in portfolio management, diversity is used to avoid overly concentrated portfolios. Various diversity constraints were proposed, such as weight upper/lower bound constraint [16], L^p-norm constraint [17] and entropy constraint [18]. Lin [8] proposed a canonical form of these diversity constraints such that the value of

diversity is restricted to the same range for all these different definitions of diversity. In this paper, we adopt these canonical forms for calculating neighbor diversity. For problems related to various diversities, please refer to [16-19].

3.3.1. Max Weight Diversity and Min Weight Diversity

The max weight diversity, denoted as $D_{max}(x)$, is the maximum of all $p_i(x)$ for i=1 to n, as shown below.

$$D_{max}(x) = \max_{i=1\ to\ n} p_i(x) \tag{4}$$

The min weight diversity, denoted as $D_{min}(x)$, is calculated using the minimum of all $p_i(x)$ for i=1 to n, as shown below.

$$D_{min}(x) = 1 + (1-n) \min_{i=1\ to\ n} p_i(x) \tag{5}$$

3.3.2. Canonical L^p-norm Diversity

The Canonical L^p-norm diversity, denoted as $D_{pow}(x)$, is similar to the L^p-norm except the outer exponent is $\frac{1}{pow-1}$ instead of $\frac{1}{pow}$, as shown below.

$$D_{pow}(x) = \left(\sum_{i=1}^{n} |p_i(x)|^{pow}\right)^{\frac{1}{pow-1}} \tag{6}$$

For the value of *pow*, the cases of *pow* = 2 and 3 are commonly used [8]. Hence, we consider only $D_2(x)$ and $D_3(x)$ in this study.

3.3.3. Canonical Shannon Entropy Diversity

The canonical Shannon entropy diversity, denoted as $D_{cs}(x)$, is the reciprocal of the natural exponential function of Shannon entropy, as shown below.

$$D_{cs}(x) = e^{-D_s(x)} = e^{\sum_{i=1}^{n} p_i(x) \log_2 p_i(x)} \tag{7}$$

Notably, the canonical entropy defined in [8] uses the natural logarithm to ensure that the range of the canonical entropy is $[\frac{1}{n},1]$. In Eq.(7), we use $\log_2$ instead of the natural logarithm such that $D_{cs}(x)$ can be easily calculated from $D_s(x)$.

4. Experimental Settings

To compare the performance of various neighbor diversities, we collected a dataset from Ruten (www.ruten.com.tw), which is one of the largest online auction websites in Taiwan [14]. Similar to the previous works [4-6], the dataset grows from a list of suspended users, and then conducts a level-wise expansion to include more users. The dataset consists of 4,407 users, where 1,080 are fraudsters and 3,327 are non-fraudsters (i.e. normal accounts). Notably, This dataset was also used in our previous study [6].

After collecting the dataset, we calculated $D_s(x)$, $D_{max}(x)$, $D_{min}(x)$, $D_2(x)$, $D_3(x)$, and $D_{cs}(x)$ for each user x in the dataset, as described in Section 3. Then, we used each of these neighbor diversities

to build a classifier to compare their performance on detecting fraudsters. Three classification algorithms (J48 decision tree, Neural Networks (NN), and Support Vector Machine (SVM)) from Weka [20] were used to perform 10-fold cross-validation.

5. Experimental Results

The experimental results include two parts. Part one uses only one of the neighbor diversities to build classifiers, and the results are shown in Tables 1, 2 and 3 for J48, NN, and SVM, respectively. The best results of each classification algorithm are shown in bold. In Tables 1, 2 and 3, the min weight diversity D_{min} performs the worst. The performances of the remaining five diversities (i.e., D_s, D_{max}, D_2, D_3, and D_{cs}) are similar. In spite of its simplicity, the max weight diversity D_{max} achieves competitive performance.

Table 1. J48 Performance (Part one)

Diversity	Accuracy(%)	Recall	Precision	F_1-measure
D_s	84.1843	0.8019	0.6420	0.7131
D_{max}	84.1616	0.8009	0.6417	0.7125
D_{min}	82.0059	0.6639	0.6251	0.6439
D_2	84.1162	0.7944	0.6422	0.7103
D_3	84.1162	**0.8028**	0.6403	0.7124
D_{cs}	**84.2523**	**0.8028**	**0.6432**	**0.7142**

Table 2. Neural Network performance (Part one)

Diversity	Accuracy(%)	Recall	Precision	F_1-measure
D_s	83.1405	0.7620	0.6287	0.6890
D_{max}	**83.8212**	**0.8120**	0.6323	**0.7110**
D_{min}	82.0286	0.6648	0.6254	0.6445
D_2	83.7077	0.7870	**0.6353**	0.7031
D_3	83.7985	0.7991	0.6346	0.7074
D_{cs}	83.5943	0.7713	0.6364	0.6974

Table 3. Support Vector Machine performance (Part one)

Diversity	Accuracy(%)	Recall	Precision	F_1-measure
D_s	83.1405	0.7306	0.6358	0.6799
D_{max}	**83.5716**	**0.7556**	**0.6395**	**0.6927**
D_{min}	82.0059	0.6639	0.6251	0.6439
D_2	83.0270	0.7222	0.6352	0.6759
D_3	83.2539	0.7361	0.6370	0.6830
D_{cs}	82.6639	0.6944	0.6334	0.6625

Because previous works suggest using *k*-core and CW for fraudster detection [4], part two of the experiment uses both *k*-core and CW and one of the neighbor diversities to build classifiers, and the

results are shown in Tables 4, 5 and 6 for J48, NN, and SVM, respectively. Compared to Part one, the addition of *k*-core and CW slightly improves the classification performance. The improvement on accuracy is most significant with J48 (between 1.5657% and 2.1103%), and less significant with NN and SVM (between -0.5673% and 1.4522%).

Table 4. J48 Performance (Part two)

Diversity	Accuracy(%)	Recall	Precision	F_1-measure
k-core+CW+D_s	85.8180	0.8731	0.6590	0.7511
k-core+CW+D_{max}	85.8861	0.8731	0.6604	0.7520
k-core+CW+D_{min}	84.1162	0.8278	0.6349	0.7186
k-core+CW+D_2	86.1130	0.8685	0.6662	0.7540
k-core+CW+D_3	**86.2038**	0.8704	**0.6676**	**0.7556**
k-core+CW+D_{cs}	85.8180	**0.8741**	0.6588	0.7513

Table 5. Neural Network performance (Part two)

Diversity	Accuracy(%)	Recall	Precision	F_1-measure
k-core+CW+D_s	83.7758	0.7787	0.6386	0.7017
k-core+CW+D_{max}	**84.1616**	**0.8083**	**0.6400**	**0.7144**
k-core+CW+D_{min}	82.3916	0.6620	0.6350	0.6482
k-core+CW+D_2	83.9120	0.7981	0.6371	0.7086
k-core+CW+D_3	83.9573	0.8028	0.6370	0.7104
k-core+CW+D_{cs}	83.8212	0.7843	0.6383	0.7038

Table 6. Support Vector Machine performance (Part two)

Diversity	Accuracy(%)	Recall	Precision	F_1-measure
k-core+CW+D_s	**84.4112**	**0.7685**	**0.6551**	**0.7073**
k-core+CW+D_{max}	83.0043	0.6835	0.6428	0.6625
k-core+CW+D_{min}	83.4581	0.7630	0.6353	0.6933
k-core+CW+D_2	83.2539	0.7370	0.6368	0.6833
k-core+CW+D_3	83.2993	0.7398	0.6372	0.6847
k-core+CW+D_{cs}	83.0951	0.7426	0.6320	0.6828

6. Conclusions

The concept of diversity has been widely used in many domains, e.g., ecology [21-24] and portfolio management [8, 18, 25]. Various ways to quantify diversity exist in the literature [8, 22]. In this work, we apply the diversity of the neighbors of each trader for fraudster detection in online auction. Specifically, we use various methods to calculate diversity, and study whether these methods cause significant difference on the classification performance of fraudster detection. Our experimental results show that the diversity D_{min} performs the worst. Also, the remaining five diversities (i.e., D_s, D_{max}, D_2, D_3 and D_{cs}) achieve similar performance.

The addition of *k*-core and CW only slightly improves the classification performance of the neighbor diversity (2.1103% in accuracy, at most). Therefore, finding new features to work better with the neighbor diversity for fraudster detection is planned for future work.

Acknowledgments

This research is supported by the National Science Council under Grant 102-2221-E-155-034-MY3.

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

1. Wang, J. C.; Chiu, C. Q., Detecting Online Auction Inflated-Reputation Behaviors Using Social Network Analysis. In *Annual Conference of the North American Association for Computational Social and O rganizational Science (Online Publication, http://www.casos.cs.cmu.edu/events/conferences/2005/2005_proceedings/Wang.pdf).* Notre Dame, Indiana, USA, 2005.
2. Chau, D. H.; Pandit, S.; Faloutsos, C., Detecting Fraudulent Personalities in Networks of Online auctioneers. In *Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases: PKDD 2006*, Springer-Verlag: Berlin, Germany, 2006; pp 103-114.
3. Pandit, S.; Chau, D. H.; Wang, S.; Faloutsos, C., Netprobe: A Fast and Scalable System for Fraud Detection in Online Auction Networks. In *Proceedings of the 16th international conference on World Wide Web*, ACM: Banff, Alberta, Canada, 2007; pp 201-210.
4. Wang, J.-C.; Chiu, C.-C., Recommending Trusted Online Auction Sellers Using Social Network Analysis. *Expert Systems with Applications* **2008,** *34*, pp.1666-1679.
5. Chiu, C. C.; Ku, Y. C.; Lie, T.; Chen, Y. C., Internet Auction Fraud Detection Using Social Network Analysis and Classification Tree Approaches. *International Journal of Electronic Commerce* **2011,** *15*, 123-147.
6. Lin, J.-L.; Khomnotai, L., Using Neighbor Diversity to Detect Fraudsters in Online Auctions. *Entropy* **2014,** *16*, 2629-2641.
7. Shannon, C. E., A mathematical theory of communication. *Bell System Technical Journal, The* **1948,** *27*, 379-423.
8. Lin, J.-L., On the Diversity Constraints for Portfolio Optimization. *Entropy* **2013,** *15*, 4607-4621.
9. Chau, D. H.; Faloutsos, C., Fraud Detection in Electronic Auction. In *European Web Mining Forum, held as part of ECML/PKDD*, Porto, Portugal, 2005.
10. Morzy, M., New Algorithms for Mining the Reputation of Participants of Online Auctions. In *Internet and N etwork Economics*, Deng, X.; Ye, Y., Eds. Springer Berlin Heidelberg: 2005; Vol. 3828, pp 112-121.
11. Morzy, M., Cluster-Based Analysis and Recommendation of Sellers in Online Auction. *Computer Systems Science and Engineering* **2007,** *22*, 279-287.
12. Bin, Z.; Yi, Z.; Faloutsos, C., Toward a Comprehensive Model in Internet Auction Fraud Detection. In *Proceedings of the 41st Annual Hawaii International Conference on Sy stem Sciences,* , 2008; pp 79-79.
13. Yu, C. H.; Lin, S.-J., Web Crawling and Filtering for On-line Auctions from a Social Network Perspective. *Information Systems and e-Business Management* **2012,** *10*, 201-218.

14. Yu, C. H.; Lin, S. J., Fuzzy Rule Optimization for Online Auction Frauds Detection Based on Genetic Algorithm. *Electronic Commerce Research* **2013,** *13*, 169-182.
15. Lin, S.-J.; Jheng, Y.-Y.; Yu, C.-H., Combining Ranking Concept and Social Network Analysis to Detect Collusive Groups in Online Auctions. *Expert Systems with Applications* **2012,** *39*, 9079-9086.
16. Frost, P. A.; Savarino, J. E., For Better Performance: The Journal of Portfolio Management. *The Journal of Portfolio Management* **1988,** *15*, 29-34.
17. DeMiguel, V.; Garlappi, L.; Nogales, F. J.; Uppal, R., A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms. *Management Science* **2009,** *55*, 798-812.
18. Huang, X., An Entropy Method for Diversified Fuzzy Portfolio Selection. *International Journal of Fuzzy Systems* **2012,** *14*, 160-165.
19. Jagannathan, R.; Ma, T., Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. *The Journal of Finance* **2003,** *58*, 1651-1684.
20. Witten, I. H.; Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc.: 2011.
21. Hill, M. O., Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **1973,** *54*, 427-432.
22. Tuomisto, H., A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia* **2010,** *164*, 853-860.
23. Jost, L., Entropy and diversity. *Oikos* **2006,** *113*, 363-375.
24. Tuomisto, H., A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography* **2010,** *33*, 2-22.
25. Usta, I.; Kantar, Y. M., Mean-variance-skewness-entropy measures: A multi-objective approach for portfolio selection. *Entropy* **2011,** *13*, 117-133.