Springer Proceedings in Mathematics & Statistics

Frank M. Lin
Ashokkumar Patel
Nishtha Kesswani
Bosubabu Sambana *Editors*

Accelerating Discoveries in Data Science and Artificial Intelligence I

ICDSAI 2023, LIET Vizianagaram, India, April 24–25

Springer

Springer Proceedings in Mathematics & Statistics

Volume 421

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including data science, operations research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Frank M. Lin • Ashokkumar Patel •
Nishtha Kesswani • Bosubabu Sambana
Editors

Accelerating Discoveries in Data Science and Artificial Intelligence I

ICDSAI 2023, LIET Vizianagaram, India, April 24–25

Springer

Editors
Frank M. Lin
Information & Decision Sciences,
Jack H. Brown College of Business and
Public Administration
California State University, San Bernardino
San Bernardino, CA, USA

Nishtha Kesswani
Department of Computer Science
Central University of Rajasthan
Ajmer, Rajasthan, India

Ashokkumar Patel
College of Engineering
University of Massachusetts Dartmouth
Dartmouth, MA, USA

Bosubabu Sambana
Lendi Institute of Engineering and
Technology
Vizianagaram, Andhra Pradesh, India

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-031-51166-0 ISBN 978-3-031-51167-7 (eBook)
https://doi.org/10.1007/978-3-031-51167-7

Mathematics Subject Classification: 68T01, 68T09, 68T27, 68T35

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Preface: Volume 1

The 8th International Conference on Computers, Management & Mathematical Sciences (ICCM) 2022 brings together researchers from around the world whose primary goal is to build computational systems that reflect the management and mathematical processes of the mind. Since its establishment, the ICCM has been an international forum for researchers in areas related to computational methods, numerical modelling and simulation, and management techniques.

This volume 1 of ICCM 2022 conference aims at providing presentations on a wide range of topics to facilitate the inter-disciplinary exchange of ideas in computers, management and related mathematical disciplines, and foster various types of academic collaborations. The papers provide good examples of current research on relevant topics, covering intelligent information systems and management, technology and management, enterprise operation and management, data science and computational analytics, international economy and trading, commodity economics and services, and mathematics to AI. The success of the ICCM 2022 is due to the collective effort of all. Special thanks go to our invited speakers, and other professors, researchers and practitioners. In particular, we would like to express and record our gratitude and appreciation to the authors for their contributions.

San Bernardino, CA, USA	Frank M. Lin
Dartmouth, MA, USA	Ashokkumar Patel
Ajmer, Rajasthan, India	Nishtha Kesswani
Vizianagaram, Andhra Pradesh, India	Bosubabu Sambana

Organization

Program Committee Chairs	
Kesswani, Nishtha	
Program Committee Members	
Akula, Chandrasekhar	Lendi Institute of Engineering and Technology, Computer Science and Engineering, Vizianagaram, India
Adimalla, Dr. Rama Rao AnnanNaidu, P	Aditya Institute of Technology and Management, CSE, Srikakulam, India
Gajjala, Nirmala Joycee	St Ann's College for Women, Computer Science, Hyderabad, India
Kavuri, Karunakar	Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, CSE Department, Avadi-Vel Tech Road, Vel Nagar, Chennai, India
Kesswani, Nishtha Laxmikanth, Pydipala	Vignan Institute of Technology and Science, Department of Computer Science and Engineering, Yadadri Bhuvanagiri District, India
M, N P Patnaik	GITAM Deemed to be University, Visakhapatnam, Department of Computer Science Engineering, Visakhapatnam, India
Mohan, Dr. T Murali	Swarnandhra Institute of Engineering and Technology, Seethrampuram, Computer Science and Science, East Godavari, India
Mummana, Dr. Satyanarayana	Raghu Engineering College, CSE, Visakhapatnam, India
Muttipati, Appala Srinuvasu OWK, Dr. Mrudula Pyla, J V G Prakasa Rao	Lendi Institute of Engineering and Technology, CSE, Vizianagaram, India
R, Ramyadevi	SRM Institute of Science and Technology, Department of Computer Science and Applications, Chennai, India
Sriramulu, Mahanti	Dhanekula Institute of Engineering and Technology, Computer Science and Engineering, Vijayawada, India

(continued)

Sambana, Bosubabu	Lendi Institute of Engineering and Technology (A), Vizianagaram, Computer Science and Engineering, Vizianagaram, India
Shaik, Mahaboob Hussain	Vishnu Institute of Technology, Computer Science and Engineering, Bhimavaram, India
Simhadri, Kumaraswamy	Aditya Institute of Technology and Management, EEE, Tekkali, India
Sridhar, B	Lendi Institute of Engineering and Technology, ECE, Vizianagaram, India
T, Raghunadha Reddy	Matrusri Engineering College, CSE, Hyderabad, India
Tuduku, Vinod Kumar	Lendi Institute of Engineering and Technology, Management Studies, Vizianagaram, India
Thirupatirao, Tirlingi Usha Bala, Dr. Varanasi	Anil Neerukonda Institute of Technology and Sciences, Computer Science and Engineering, Visakhapatnam, India
Kantharaju, Kanaparti	VITAP University, Computer Science and Engineering, Andhra Pradesh, India

Contents

Analysis of Fraud Detection Approaches in Online Payment Systems

Kothapalli Mandakini and Kattula Shyamala

1 Introduction

Online selling has many benefits for merchants. E-commerce has some drawbacks, including various types of fraud online. Fraud detection in financial trades is one of the significant problems in financial companies. A reliable system for detecting online fraud is required, one that can identify these crimes and alert banks in time to stop them.

In 2021, e-commerce sufferers due to online theft are expected to total $20 billion in the United States alone. From $17.5 billion US dollars a year earlier, that's a rise of almost 14% in sales revenue this year. When compared to the year before the sanitary crisis, three-quarters of online businesses stated an overall surge in cyberattacks in 2021. However, each region in the world has been impacted differently by this problem, with Asia-Pacific and Latin America suffering more losses to fraud in terms of e-commerce revenue that year. An adaptive and predictive fraud risk score can be generated using machine learning methods, and fake incidences can be observed in real time as they occur. It also aids in the automation of more advanced preventive actions. Detection of fraud is typically centered on data analysis.

Fraud Detection Methods

1. *Statistical data analysis (SDA):* SDA suggests the different statistical operations ranging from fraud detection, collection, and validation. Some of the SDA methods are as follows: regression analysis, data matching, calculating statistical parameters, probability distributions and models, and so on are all types of statistical analysis techniques.

K. Mandakini (✉) · K. Shyamala
CSE Department, UCE, Osmania University, Hyderabad, India

F. M. Lin et al. (eds.), *Accelerating Discoveries in Data Science and Artificial Intelligence I*, Springer Proceedings in Mathematics & Statistics 421,
https://doi.org/10.1007/978-3-031-51167-7_1

2. *Artificial intelligence (AI):* AI is a front-runner in the race of efficient scam detection techniques. Some are as follows: data Mining, neural networks (NN), and machine learning**.**

Artificial intelligence has emerged as a crucial instrument for the prevention of fraud and the improvement of the effectiveness of fraud prevention and detection. Detecting fraud can also make use of other methods, including Bayesian networks, link analysis, sequence matching, and decision theory [1]. To identify online banking fraud, this research study evaluates the performance metrics gleaned from a variety of methodologies employed in previous studies and identifies the most effective techniques now in use.

The majority of contemporary fraud detection techniques involve a domain expert who has two tasks to complete: One is to collect historical transaction data, and the other is to provide features for basic or sophisticated machine learning models. By educating the public, frauds involving credit cards can be avoided [2]. Methods for detecting fraud include routine inspections at all branches, account reconciliation, submitting statement of accounts on a regular basis, etc. [3]. Frauds can be avoided by having financial knowledge, receiving proper compensation, and increasing KYC standards [4]. Competition among bank employees, employment uncertainty, insufficient internal control systems, a bad work atmosphere, and inadequate recruitment were obstacles to fraud detection [5].

There are several different types of e-commerce fraud, including application fraud, behavioral fraud, theft fraud, credit card fraud, computer intrusion fraud, and fraud involving bankruptcy [6]. A procedure for identifying fake transactions has been proposed in [2]. In the beginning, they employed the random forest method. The findings of the experiment conducted on Real data obtained from a bank revealed that the most recent Cardholder transactions have a countless impact, influential whether a operation is fraudulent in order to non-fraudulent. Two deep neural network-based approaches have been reported by many researchers. First, e-commerce aspects linked to fraud prediction and distribution are combined with in-person transactions in order to improve fraud forecasts. The domain tag and the fraud tag predictor's fundamental feature layer (fraudulent or genuine). Both methods are compared on a million real credit card transactions from a large issuer over five months. The second proposed method works faster than all the others. The suggested model employed SVM to identify features. Ensemble and Random Forest models' accuracy was found to be 97.41%, which was the greatest ever [7].

This study [8] compares subjective and objective rating fraud. Then it analyzes block chain's success in objective fraud and limitations in subjective fraud, specifically rating fraud. Finally, it evaluates blockchain-based reputation systems for each sort of rating fraud. Experimentations presented in [9] have shown that those binary classifiers that were trained using the cohesive features perform significantly better than those that were trained with the original characteristics. In this research [10], the authors suggest changing the anomaly detection problem into a problem using a pseudo-recommender system and then using an embedding-based strategy to address the problem. As a result, the concept of collaborative filtering is automatically applied to make use of data from comparable users, and the learnt preference matrices and attribute embedding offer a clear method for further usage.

This study [11] proposes a two-level fake credit card transaction detection approach from extremely imbalanced datasets using artificial bee colony and semantic fusion to improve classification accuracy and accelerate detection convergence. The results of the experiments reveal that the methodology that was suggested can lead to an improvement in risk classification for suspicious transactions compared to the methods that were used previously.

An online transaction fraud detection model based on deep forest architecture is presented in this paper [12] in order to address sample imbalance and strong concealing of online transaction data. The Bagging Balance technique is used. The findings of the experiments indicate that the proposed model outperforms the random forest model by more than ten percent and the original deep forest model by more than five percent when compared using the metric of precision and recall rate. These findings are based on an evaluation of the proposed model in comparison to the random forest model.

Both online fraud prevention and detection are highly specialized disciplines. In [13], found that in typical fraud detection tasks, there is a significant imbalance between the proportions of legitimate and fraudulent classes. Consequently, learning from unbalanced datasets is a difficult research subject that must be analyzed. Observations indicate that an uneven dataset decreases the performance of typical learning algorithms such as random forest and SVM [14]. The unbalanced nature of data leads to poor performance in fraud detection methods. Data unbalanced problems can be solved using different techniques, such as resampling, SMOTE, Random Over Sampling (ROS), Random Under Sampling(RUS), etc.

2 Fraud Detection and Sector-Specific Fraud

The fraud recognition methods can be put into two main categories: techniques based on statistical data analysis and techniques based on artificial intelligence. Regression analysis, probability distribution modeling, data matching, and statistical parameter computation are the four categories that fall under the umbrella of statistical data analysis techniques. Methods such as data mining, pattern recognition, neural networks, and machine learning are examples of AI-based techniques.

The several stages that make up the fraud detection cycle are depicted in Fig. 1.

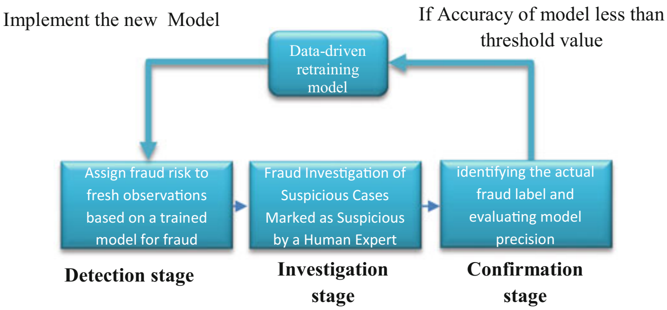

Fig. 1 Fraud detection cycle

Industry-Specific Fraud Detection

Financial and banking services: With the growth of online and ATM transactions, which has coincided with the global adoption of the digital trend, scams have been rising. The following are the most typical forms of banking fraud:

Application programming interface (API) fraud: The Payment Services Directive 2 requires some European financial institutions to make their services available through APIs. It also expands the attack surface.

Credit card theft/fake card fraud: Scammers steal a user's card info and use this info to create a fake card. Using a small, difficult-to-notice device, the fraudster attaches it to the payment terminal in order to steal credit card information. The most common technique used by fraudsters to scam consumers of their money is website cloning. ATM fraud is a scam in which convicts use the ATM card of another person to rapidly withdraw money from that person's account.

E-business and Retail

Users are being targeted through e-commerce channels more often than ever because to the COVID-19 pandemic's development in the e-commerce industry. Among the most popular techniques are: When a specific client, seller, or partner organization takes benefit of a deal, they are engaging in promo abuse, also known as coupon fraud. Redeeming coupons numerous times or just using them to get money and extra-valued goods and services might help scammers. Payment fraud: One of the most popular fraudulent actions is electronics commerce payment fraud typically; the victim is an online user who loses money, attention, profound information, or private property as a result of becoming a target of a cybercrime. Identity theft and friendly fraud are two categories of delivery fraud. Identity thieves utilize viruses, phone websites, emails, and short messages to try and get hold of sensitive user information.

Markets and Online Advertisements

Fraud of this nature is generally perpetrated through abusive referral and promotion practices as well as through the use of fake reviews. The purpose of spreading fake reviews is to bring down the reputation of companies, undermine consumers' trust in those companies, and provide examples of negative experiences that did not actually take place.

IT and Telecommunications

Sophisticated phone scams are used to perpetrate these scams. Phone fraud, often known as communications fraud, involves the fraudulent use of telecommunications equipment and services in order to defraud a telecommunications firm or even its consumers of money they are lawfully owed. Call forwarding fraud is a frequent VoIP telecom scam. In this, crooks hack a companie's voice mail or private branch exchange. They are able to send calls to an exclusive long-distance target while still earning money due to a revenue-sharing agreement. Fraud committed by letting the phone ring only once and then cutting the line is known as "one-ring-and-cut" (Wangiri). The Wangiri phone fraud technique is a method of making quick money with the use of a single ring.

3 Scenarios of Credit Card Fraud

Fraud on credit cards statistics shows that credit card fraud is most likely to happen to people in their 30s. By 2023, card-not-present transactions will cost stores about $130 billion each year. Last year, there was 135% more credit card information on the dark web than there was the year before. Now, card-not-present fraud is 81% more communal than fraud at the point of sale [15]. In the past ten years, a significant amount of research has been carried out as a direct result of the fact that ML strategies are able to effectively meet the issues that are presented by credit card fraud identification. According to the information presented in Fig. 2, between the years 2018 and 2022, thousands of articles connected to this subject were published, with the year 2020 alone accounting for the publication of over 1300 papers.

By outlining the major research problems and the essential machine learning ideas that can be applied to solve them, this section seeks to give a summary of this corpus of recent research. To have an understanding of the state of the ML research for credit card fraud detection at this time. The last 5 years' work of reviews and research on this issue have been retrieved from Google Scholar.

The fact that so many surveys came out in such a short amount of time shows how quickly the field of machine learning (ML) for finding credit card fraud is changing

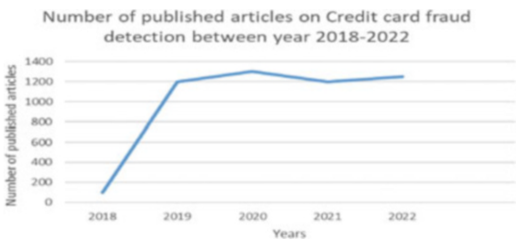

Fig. 2 Number of publication in 5 years between 2018 and 2022

and how much teams of independent researchers felt the need to sum up the state of research in this field. It is a difficult task to discover credit card fraud because it needs to evaluate a significant amount of transaction data. Large quantities of data and developing fraudulent techniques make it difficult for humans to investigate. Machine learning, which looks for and finds patterns in massive volumes of data, has increasingly been used in credit card theft in the previous decade. It has been demonstrated that machine learning algorithms may considerably increase the efficiency of fraud discovery systems and provide assistance to fraud investigators in the process of discovering fake transaction.

Challenges
Inaccessibility of datasets: The absence of publicly accessible datasets is unique to the utmost significant difficulties related to the matter [16, 17]. Credit card organizations keep the databases of their transactions, but for reasons relating to privacy and security, they are unable to make this information available to the general public. There are some outcomes that have been obtained by working with data that was generated artificially [11, 14]. However, earlier results reveal the sorts of data that were engaged in the classifier models. Dynamic Deceitful behaviour Fakes frequently alter their behaviour over the course of time so as to circumvent the existing detection systems by shifting the pattern of their fraudulent activities. As a result of this, the pattern of normal transactions and fake transactions shifts all the time. It is not out of the question that there were fake transactions in the history, but they may now conform to the pattern of typical (legal) transactions [14]. The datasets used to study credit card fraud are significantly skewed, with the vast popular of samples being real transactions and the small minority representing fraudulent ones.

In the majority of instances, greater than 99 percent transactions are legitimate, and hence, fewer than 1 percent are fraudulent [18]. When evaluating the performance of a classifier, one of the most typical metrics that is considered is its degree of accuracy. Even with a model with a high degree of accuracy, the majority of fraudulent transactions could be misclassified as legitimate because of the skewed nature of the datasets. It is essential to examine the recall and precision of such models. Recall refers to the ability of the model to classify fraudulent transactions (classifying legitimate transactions).

4 Analysing Unbalanced Dataset

Classification issues where one class has significantly fewer instances than the others are dealt with via unbalanced learning. It is difficult to learn from datasets that are unbalanced, because most training and learning algorithms stand put up to handle considerable variances in the number of examples belonging to certain classes. It is difficult to learn from datasets that are unbalanced, because most learning algorithms aren't built to handle substantial differences in the number of examples belonging to certain classes [19]. The fraction of credit card fraud in real-world statistics might be as low as 0.01 percent [20], making it an example of an unbalanced problem.

The issue of an unbalanced dataset has received numerous solutions, both for ensemble methods and traditional learning algorithms. Cost-sensitive and resampling strategies can be used to widely classify imbalanced learning systems. The algorithms are modified in order to improve their detection of the underrepresented group. In most cases, this necessitates an adjustment to the optimization function that is used during the training phase of the learning algorithm. A pre-processing phase is added to the dataset prior to using a training algorithm in resampling approaches, which allows for an even distribution of data prior to the training algorithm. Resampling can be done by deleting common class examples (undersampling), oversampling, or using both [19, 21]. Simple classification test with synthetic data in cost-sensitive learning is presented, as in Fig. 3.

The dataset has 6000 0 and 1 samples. About 95% of samples are from class 0, 5% from class 1. Using the same manner as in cost-sensitive learning, the k-fold CV with classifier function calculates a baseline classifier's performance without resampling presented in Table 1.

The decision threshold for the initial classifier in the cross-validation is shown in Fig. 4. When two classes are overlapping, the classifier returns identical probability for both.

To rebalance the dataset, oversampling techniques create new samples for the underrepresented groups in the population. SMOTE and random oversampling are the two most used approaches. This paper goes into greater detail on how these strategies can be put into practice and how they can be used to shift power in favour of the underrepresented minority group. For random oversampling construct a sampler for randomized oversampling with a dataset that has an imbalance ratio of 1. Make use of the (train df function) DataFrame in the first cross-validation fold by utilizing this sampler. Class 0 samples number 4537, while class 1 samples

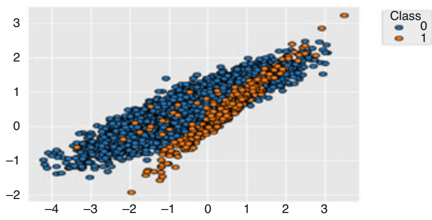

Fig. 3 Class distribution with two classes

Table 1 Baseline classifier performance with k-fold cv classifier function

Index	Fit time in second	Score time in second	Balanced accuracy	Average precision	AUC ROC
Decision tree – Baseline	0.009 ± 0.002	0.004 ± 0.0	0.64 ± 0.051	0.421 ± 0.082	0.872 ± 0.038

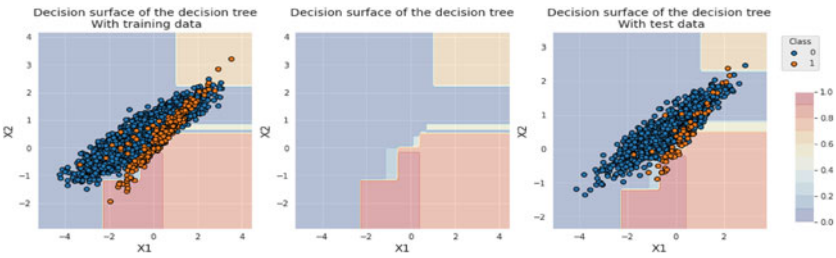

Fig. 4 Decision boundaries for two class

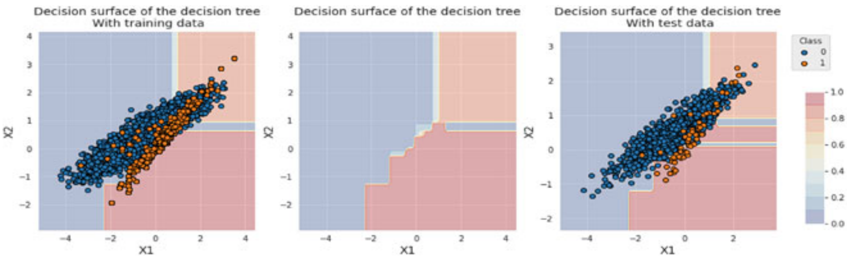

Fig. 5 Decision boundaries with resampling

number 263. The dataset is resampled by invoking the sampler object's fit method. Data from classes 1 and 0 (4537 samples in each class) have been added to the resampled DataFrame. Using pipelines, samplers, and sklearn estimators may be integrated and the baseline classifier's performance evaluated in conjunction with random oversampling. It is possible to notice in Fig. 5 how the decision boundary was shifted in favour of the minority group due to resampling.

Classification results demonstrate an improvement in balanced accuracy. The dataset -2 contains 5000 samples with two classes, labelled 0 and 1. About 95% of the samples are associated with class 0 and 5% of the samples with class 1. The DataFrame contains 3784 samples of class 0 and 216 samples of class 1. The resampling allowed shifting the decision boundary towards the minority class. The dataset used was provided by ResearchGate page – Joint collaboration: MLG ULB and Worldline (Table 3).

The findings from results presented in Table 2 represent the classification that show an increase in balance as well as average precision.

The goal of ensemble methods is to make accurate predictions by combining the results of numerous prediction models that have been trained on the same data set [19].

There are two key categories of ensemble methods are parallel-based and iterative-based [20]. Each baseline learner in parallel-based ensembles is trained using a fraction of the training data, training features, or both. The summarized and

Table 2 Comparison of balance accuracy of decision tree Baseline and ROS

Index	Score time in second	AUC ROC	Fit time (S)	Balanced accuracy	Average precision
Decision tree – Baseline	0.004 ± 0.0	0.872 ± 0.038	0.009 ± 0.002	0.64 ± 0.051	0.421 ± 0.082
Decision tree – ROS	0.004 ± 0.0	0.872 ± 0.022	0.014 ± 0.002	0.836 ± 0.025	0.434 ± 0.039

Table 3 Comparison of balance accuracy of decision tree Baseline and SMOTE, RUS, Edited Nearest Neighbour (ENN)

Index	Score time in second	AUC ROC	Fit time (S)	Balanced accuracy	Average Precision
Decision tree – Baseline	0.008 ± 0.0	0.906 ± 0.025	0.008 ± 0.002	0.786 ± 0.046	0.528 ± 0.072
Decision tree – ROS	0.008 ± 0.0	0.88 ± 0.038	0.019 ± 0.002	0.888 ± 0.03	0.456 ± 0.062
Decision tree – SMOTE	0.008 ± 0.0	0.913 ± 0.032	0.022 ± 0.002	0.91 ± 0.019	0.499 ± 0.056
Decision tree – RUS	0.006 ± 0.0	0.913 ± 0.02	0.005 ± 0.001	0.896 ± 0.023	0.408 ± 0.058
Decision tree – ENN	0.009 ± 0.0	0.879 ± 0.038	0.023 ± 0.004	0.857 ± 0.041	0.474 ± 0.08

Table 4 Row-wise performance metrics, column-wise ensemble approaches

	Balanced bagging	Baseline RF	Baseline XGBoost	Balanced RF	Weighted XGBoost	Baseline bagging
AUC ROC	0.887 ± 0.01	0.870 ± 0.02	0.871 ± 0.01	0.870 ± 0.02	0.871 ± 0.01	0.848 ± 0.02
Average precision	0.69 ± 0.01	0.667 ± 0.01	0.689 ± 0.01	0.687 ± 0.01	0.689 ± 0.01	0.654 ± 0.01
Card Precision@100	0.315 ± 0.01	0.298 ± 0.01	0.302 ± 0.01	0.312 ± 0.01	0.3 ± 0.01	0.289 ± 0.01

simulated results from ensemble learning on the transaction dataset are presented in the below Table 4.

AUC ROC and CP@100 bagging performance were both improved by using balanced bagging, which resulted in the best results. For weighted XGBoost and XGBoost, little or no improvement was found. Overfitting and unbalanced data do not affect the performance of random forest and XGBoost.

Conclusion

In order to detect the fraud in online payment systems, a lot of research has been done. But it is a dynamic test bed for researchers to develop an accurate and efficient model to detect and predict the fraud in online payment systems.

This study discussed the use of unbalanced learning in different fraud detection approaches in online payment systems. Methods such as cost-sensitive resampling and ensemble analysis were also studied. A comparison was done with baseline as well as ensemble methods. Ensembles show that the experimental results are not significantly better than those of the baselines. The combination of ensemble methods and imbalanced learning techniques like resampling methods are significantly more accurate than others.

References

1. V. Kanade, *What is fraud detection? definition, types, applications, and best practices |Spiceworks*. Spiceworks (2021, June 11.); www.spiceworks.com. https://www.spiceworks.com/it-security/vulnerability-management/articles/what-is-fraud-detection/
2. D.A. Williams, Credit card fraud in Trinidad and Tobago. J. Financ. Crime **14**(3), 340–359 (2007). https://doi.org/10.1108/13590790710758521
3. S. Mahdi, A. Zhila, Fraud detection and audit expectation gap: Empirical from Iranian bankers. Int. J. Bus. Manag **3**(10), 65–67 (2008)
4. C. Singh, Frauds in the Indian Banking Industry. Working Paper, IIMB, WP N0. 505, March 2016
5. B.A. Badejo, B.A. Okuneye, M.R. Taiwo, Fraud detection in the banking system in Nigeria challenges and prospects. J. Econ. Bus. **2**(3), 255–282 (2017)
6. Y. Lucas, J. Jurgovsky, Credit card fraud detection using machine learning: a survey. arXiv preprint arXiv:2010.06479 (2020)
7. B. Alghamdi, F. Alharby, An intelligent model for online recruitment fraud detection. J. Inf. Secur. **10**(03) (2019). https://doi.org/10.4236/jis.2019.103009
8. Y. Cai, D. Zhu, Fraud detections for online businesses: A perspective from blockchain technology. Financ. Innov **2**(1) (2016). https://doi.org/10.1186/s40854-016-0039-4
9. J. Cui, C. Yan, C. Wang, Learning transaction cohesiveness for online payment fraud detection. ACM International Conference Proceeding Series, PartF168982 (2021). https://doi.org/10.1145/3448734.3450489
10. J. Cui, C. Yan, C. Wang, ReMEMBeR: Ranking metric embedding-based multicontextual behavior profiling for online banking fraud detection. IEEE Trans. Comput. Soc. Syst **8**(3) (2021). https://doi.org/10.1109/TCSS.2021.3052950
11. S.M. Darwish, An intelligent credit card fraud detection approach based on semantic fusion of two classifiers. Soft. Comput. **24**(2) (2020). https://doi.org/10.1007/s00500-019-03958-9
12. M. Huang, L. Wang, Z. Zhang, Improved deep forest mode for detection of fraudulent online transaction. Comput. Inform **39**(5) (2021). https://doi.org/10.31577/CAI_2020_5_1082
13. F. Mohammed Aamir Ali, M.A. Azad, M.P. Centeno, F. Hao, A. van Moorsel, Consumer-facing technology fraud: Economics, attack methods and potential solutions. Futur. Gener. Comput. Syst. **100**, 408 (2019)
14. S.K. Saddam Hussain, E.S.C. Reddy, K.G. Akshay, T. Akanksha, Fraud detection in credit card transactions using SVM and Random Forest Algorithms, in *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, (2021), pp. 1013–1017. https://doi.org/10.1109/I-SMAC52330.2021.9640631
15. N. Cveticanin, *Credit card fraud statistics: What are the odds? | DataProt*. Dataprot; dataprot.net (2022, March 8). https://dataprot.net/statistics/credit-card-fraud-statistics/
16. J. Chunhua, N. Wang, Research on credit card fraud detection model based on similar coefficient sum, in *First InternationalWorkshop on Database Technology and Applications, DBTA 2009, Wuhan, Hubei, China, April 25-26, 2009, Proceedings*, (2009), pp. 295–298

17. E.W.T. Ngai, H. Yong, Y.H. Wong, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decis. Support. Syst. **50**(3), 559–569 (2011)
18. Kanika, J. Singla, A survey of deep learning based online transactions fraud detection systems, in *Proceedings of International Conference on Intelligent Engineering and Management, ICIEM 2020*, (2020). https://doi.org/10.1109/ICIEM48762.2020.9160200
19. A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, F. Herrera, *Learning from Imbalanced Data Sets* (Springer, 2018)
20. G. Haixiang, L. Yijing, G. Jennifer Shang, H.Y. Mingyun, G. Bing, Learning from class-imbalanced data: Review of methods and applications. Expert Syst. Appl. **73**, 220–239 (2017)
21. S. Jha, M. Guillen, J.C. Westland, Employing transaction aggregation strategy to detect credit card fraud. Expert Syst. App **39**(16), 12650–12657 (2012)

Investigating Context-Aware Sentiment Classification Using Machine Learning Algorithms

P. Ashok Kumarr, B. Vishnu Vardhan, and P. Chirankjeevi

1 Introduction

Context-aware sentiment classification is a trending research area whose aim is to improve the accuracy and reliability of sentiment analysis by considering the context in which the sentiment expressions are used. Sentiment analysis involves identifying the sentiment polarity of a given tweet, which can be positive polarity, negative polarity, or neutral polarity. However, the sentiment polarity of a tweet can depend on the context in which it is used, such as the type of entity being discussed, the temporal and spatial aspects, and the social factors. Ignoring the context can lead to inaccurate sentiment analysis results, which can affect decision-making processes in different domains, which are marketing, restaurant, politics, and social media tweets.

Context-aware sentiment classification is a subfield of NLP, the context in which a sentiment expression is used to observe the sentiments in the text accurately. The polarity of a sentence in text can be negative, positive, or neutral, but it can also be influenced by the context in which it is used, such as the type of entity being discussed, the temporal and spatial aspects, and the social factors. Ignoring the context can lead to inaccurate sentiment analysis results, which can affect decision-making processes in different domains, which are marketing, restaurant, politics, and social media tweets.

P. A. Kumarr (✉)
ACE Engineering College, Hyderabad, Telangana, India

B. V. Vardhan
JNTUH College of Engineering, Manthini, Telangana, India

P. Chirankjeevi
Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India

F. M. Lin et al. (eds.), *Accelerating Discoveries in Data Science and Artificial Intelligence I*, Springer Proceedings in Mathematics & Statistics 421,
https://doi.org/10.1007/978-3-031-51167-7_2

Context-aware sentiment analysis involves several sub-tasks, including word identification, word-based sentiment analysis, and context representation. Word identification involves identifying the relevant words or entities in each context that contribute to sentiment expression. Aspect-based sentiment classification involves performing sentiment classification on large amount of unstructured data, which requires considering the context and the sentiment expressions associated with each aspect. Context representation involves representing the context of a given text accurately, which can include temporal, spatial, and social aspects.

Context-aware sentiment analysis is a challenging research area that involves several issues and challenges, such as subjectivity and ambiguity, data quality and quantity, domain adaptation, multi-lingual sentiment analysis, handling sarcasm and irony, and model interpretability. Addressing these challenges requires interdisciplinary collaborations between natural language processing, machine learning, and domain experts, as well as the development of novel algorithms and evaluation metrics. The aim of context-aware sentiment classification is to enhance accuracy and reliability of sentiment analysis and to give a more fine-tuned way of the sentiment expressions in different contexts.

1.1 Context Overview

The word context has been used with a variety of meanings. Context is the situation in which something happens that helps you understand the context. The context in text documents depends on one or more words and can be obtained from the different words of history. Context is made up of the environment, causes, conditions, identities of the user, location, time, etc. Context is extracted from two or more words in a text.

Sentiment classification is the step of finding the emotional feeling of a piece of text, such as a review or social media post. The context of the text can be having an important impact on the sentiment analysis outcome. Some of the different types of contexts that can affect sentiment analysis are:

Textual context: Textual context is a significant factor in sentiment classification, as it consists of text data associated with the sentiment expressed. Here are some different types of textual contexts that can be used for sentiment analysis:

Topic or domain context: The topic or domain of the text can influence sentiment analysis results. For example, the sentiment of a review of a hotel might be different from the sentiment of a review of a movie.

Historical context: The historical context of a text, such as the time period in which it was written, can affect sentiment analysis results. For example, a positive review of a movie from the 1960s might not be seen as positive by today's standards.

Cultural context: Cultural differences can affect sentiment analysis results. For example, the way that people express positivity or negativity can vary across cultures.

User context: The user who wrote the text can affect sentiment analysis results. For example, a tweet from a celebrity might be interpreted differently than a tweet from an average person.

Linguistic context: The linguistic context of a text can affect sentiment analysis results. For example, the use of sarcasm or irony can make it poor to find the true sentiments in a tweet.

Polarity context: The polarity of the sentiment expressed in the text can also affect sentiment analysis results. For example, a text with mixed sentiments can be challenging to interpret using a simple positive/negative sentiment analysis model.

It is important to consider these different contexts shown in Table 1 when conducting sentiment analysis to ensure that the results are accurate and reliable.

2 Literature Review

The sentiment expressed in a given text, has become a critical research area in NLP due to its wide range of applications such as marketing, political analysis, public opinion analysis, and customer feedback analysis. Traditional sentiment classification approaches are mostly based on the bag-of-words (BOW) approach, which ignores contextual information and often leads to poor accuracy. In recent years, researchers have explored context-aware sentiment classification using ML algorithms to improve the accuracy of sentiment classification.

Context-aware sentiment classification involves considering the context of a word or phrase when determining its sentiment. For instance, the word "love" can have a positive sentiment in some contexts, such as "I love this movie," while it can have a negative sentiment in others, such as "I love you, but I can't be with you." Context-aware sentiment classification aims to capture such nuances in sentiment by analyzing the surrounding context.

Liu et al. [1] provided an overview of sentiment analysis in opinion extraction mining. The authors discussed different techniques for sentiment classification, including lexical-based, machine and deep learning-based, and hybrid processes. They also discussed the hidden challenges in sentiment classification, which are sarcasm and irony.

"Sentiment Analysis using Support Vector Machines with Different Feature Representations" by Pang et al. [2] investigated the use of support vector machines (SVMs) for sentiment analysis. The authors compared different feature representations, including bag-of-words, n-grams, and word embeddings. They found that word embeddings performed better than the other feature representations.