Lecture Notes in Electrical Engineering 1146

Jitendra Agrawal
Rajesh K. Shukla
Sanjeev Sharma
Chin-Shiuh Shieh *Editors*

Data Engineering and Applications

Proceedings of the International Conference, IDEA 2K22, Volume 1

Springer

Lecture Notes in Electrical Engineering

Volume 1146

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering—quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

Jitendra Agrawal · Rajesh K. Shukla ·
Sanjeev Sharma · Chin-Shiuh Shieh
Editors

Data Engineering and Applications

Proceedings of the International Conference, IDEA 2K22, Volume 1

Springer

Editors
Jitendra Agrawal
Rajiv Gandhi Technical University
Bhopal, Madhya Pradesh, India

Sanjeev Sharma
Rajiv Gandhi Technical University
Bhopal, Madhya Pradesh, India

Rajesh K. Shukla
Oriental Institute of Science
and Technology
Bhopal, Madhya Pradesh, India

Chin-Shiuh Shieh
National Kaohsiung University of Science
and Technology
Kaohsiung, Taiwan

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-97-0036-3 ISBN 978-981-97-0037-0 (eBook)
https://doi.org/10.1007/978-981-97-0037-0

Contents

About the Editors

Dr. Jitendra Agrawal is a Director of the School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, Madhya Pradesh, India. He is also a program chair of IDEA2k22. His research interests include data mining and computational intelligence. He has authored two books and published more than 70 papers in international journals and conferences. He is a senior member of IEEE, a life member of CSI, ISTE, and a member of ACM and IAENG. He has served as part of the program committees for several international conferences organized in countries such as the USA, India, New Zealand, Korea, Indonesia, and Thailand.

Dr. Rajesh K. Shukla is a Director of the Oriental Institute of Science and Technology in Bhopal, India. He is also a program chair of IDEA2k22. With more than 23 years of teaching and research experience, he has authored and edited ten books, published/presented more than 45 papers in international journals/conferences, and six patents are published/granted. He received an ISTE UP Government National Award in 2015 and various prestigious awards from the Computer Society of India. His research interests include recommendation systems and machine learning. He is a fellow of IETE, a senior member of IEEE and ACM, a life member of ISTE, and ISCA, and a member of IE(I).

Dr. Sanjeev Sharma is a Professor in the School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, Madhya Pradesh, India. He is also a general chair of IDEA2k22. He has over 33 years of teaching and research experience and received the World Education Congress Best Teacher Award in Information Technology. His research interests include mobile computing, ad hoc networks, image processing, and information security. He has edited proceedings of several national and international conferences and published more than 160 research papers in reputed journals. He is a member of IEEE, CSI, ISTE, and IAENG.

Dr. Chin-Shiuh Shieh is an Associate Professor at the Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Taiwan. He is also a general chair of IDEA2k22. His research interests include computer networking, wireless communication, information security, computational intelligence, embedded systems, and medical informatics. He has edited proceedings of several international conferences and published more than 320 research publications including papers in reputed journals and conferences and book chapters. He has been the director of the Academic Division of the Signal Processing Society, IEEE Tainan Chapter, and the vice chair of the Signal Processing Society, IEEE Tainan Chapter. He has been the guest editor, a reviewer board member, and a reviewer for various international journals and conferences. He is a member of IEEE.

Machine Learning, Neural Networks, Evolutionary and Swarm Algorithms and Applications

Review of Methods for Handling Class Imbalance in Classification Problems

Satyendra Singh Rawat and Amit Kumar Mishra

Abstract Learning classifiers using skewed or imbalanced datasets can occasionally lead to classification issues; this is a serious issue. In some cases, one class contains the majority of examples while the other, which is frequently the more important class, is nevertheless represented by a smaller proportion of examples. Using this kind of data could make many carefully designed machine learning systems ineffective. High training fidelity was a term used to describe biases versus all other instances of the class. The best approach to all possible remedies to this issue is typically to gain from the minority class. The article examines the most widely used methods for addressing the problem of learning with a class imbalance, including data-level, algorithm-level, hybrid, cost-sensitive learning, and deep learning, including their advantages and limitations. The efficiency and performance of the classifier are assessed using a myriad of evaluation metrics.

Keywords Machine learning · Imbalance learning · Resampling techniques

1 Introduction

In the realms of machine learning and data mining, class imbalance learning is a significant problem. In recent years, increasing attention has been paid to the categorization of class-imbalanced data from a variety of fields of study. A balanced sample distribution across classes is generally achieved by traditional classification techniques. However, such a belief led to the majority class performing unfavorably. Any classifier learned from an imbalanced dataset would exhibit more classification errors in comparison to examples of minority classes since classifiers normally try to reduce the overall classification error [1].

Present Address:
S. S. Rawat (✉) · A. K. Mishra
Amity University, Gwalior, India
e-mail: satyandra.rawat@s.amity.edu

A. K. Mishra
e-mail: akmishra1@gwa.amity.edu

J. Agrawal et al. (eds.), *Data Engineering and Applications*, Lecture Notes in Electrical Engineering 1146, https://doi.org/10.1007/978-981-97-0037-0_1

With the arrival of big data technology, through machine learning and data mining, we have a better understanding of the nature of imbalanced learning, but we are also facing new challenges [2]. Finding abnormal events can be considered a prediction task, as in the machine learning and data mining fields. As an impact of the rarity of these events in our real-life applications, the prediction task is affected by a lack of balanced data [3]. Big data makes it more difficult to lower class disparity due to the diverse and complex structure of the significantly larger datasets. These unbalanced datasets are generally common in real-world data, such as fraud detection, spam detection, and software defect prediction [4].

Detecting electronic fraud in transactions also poses an extremely challenging problem in class imbalance with overlap. In order to avoid scrutiny, fraudsters have spent a lot of effort in closely cloning a legitimate transaction. It is difficult to distinguish between legitimate and illegal transactions due to the huge amount of data that overlaps. For machine learning-based fraud transaction detection methods, overlapping problems have, however, received less attention than problems with class imbalance [5].

The rationale for the imbalanced data is biased in favor of the majority of class instances owing to high training accuracy. The generation of data from the minority class is consistently regarded as the solution to the issue that has the best chance of success [6].

1.1 Class Imbalance Problem

Classification problems commonly face serious issues of learning classifiers from skewed or unbalanced datasets. The majority of instances in these instances belong to one class, while the other class, which is more important contains a small number of instances. Traditional classifiers put all of the data into the majority class, which is typically the class with the lowest importance, leaving them obviously unsuited to handle unbalanced learning tasks [7].

A population with rare diseases, for example, can have medical data with few disease categories. Statistical and machine learning techniques are prone to encounter issues when some classes are glaringly underrepresented. Despite being learned, cases from the rare classes are lost amid the others. The resulting classifiers misclassified unknown rare cases, and descriptive models could have misrepresented the data. If a small class is difficult to identify due to its other characteristics, the learning task becomes significantly more difficult. A small class, for instance, may significantly overlap the other classes. The following depicts a small, difficult class as an interesting class numerous domains exhibit class imbalance, including fraud detection, spam filtering, disease prediction, software defect prediction, ransomware, detection, etc. [8].

This paper discusses the various techniques that are used to handle the class-imbalanced datasets in binary classification problems and also provides a comparative study of the most popular methods with their benefits and limitations. The rest of the

sections of the paper are organized as: a review of the literature is given in Sect. 2; existing methods are described in Sect. 3; in Sect. 4 important evaluation metrics are discussed, and finally, the conclusion is given in Sect. 5.

2 Review of Literature

The author has explored different facets of learning from imbalanced data, such as mining data streams, clustering, classification, regression, and big data analytics, and has given a thorough overview of new challenges in such fields. Such challenges relate to learning from imbalanced data and have their own roots in contemporary real-world applications [2]. An open-source Python toolbox called imbalanced-learn aims to offer a variety of solutions for the imbalanced dataset issue that frequently arises in pattern recognition and machine learning [9].

Imbalanced data have been given to sampling techniques like SMOTE in order to artificially balance the dataset for classifier training. In order to overcome SMOTE's limitation for nonlinear problems, a weighted kernel-based SMOTE (WK-SMOTE) that oversamples the feature space of the SVM classifier is implemented in this study [10]. Unfortunately, defective modules typically have a lower presence in software defect datasets than non-defective modules. For imbalanced software defect datasets, the MAHAKIL synthetic oversampling method is introduced, which is based on the chromosomal theory of inheritance [11].

An RK-SVM algorithm based on sample selection was proposed to address the class imbalance issue in the identification of breast cancer [12]. Noise and borderline cases are two important problems brought on by SMOTE's blind oversampling. According to the distance between the artificially generated new minority class examples and the original minority class examples, [13] proposed the advanced SMOTE, also known as A-SMOTE.

The fraud detection problem was unable to be successfully tackled by random undersampling using conventional binary classifiers due to a high-class imbalance. The imbalanced data problem was investigated using a variety of methods, and a novel method based on entropy-based undersampling laced with a dynamic stacked ensemble was developed [14]. Minority class data are transformed into a realistic data distribution when the minority class data are insufficient for GAN to process them effectively on its own [15]. The controlled sampling method QDPSKNN used in this study was developed to account for the uneven class distribution of user click data in the classification of fraudsters [16].

3 Existing Solutions

Four broad categories can be used to group the solutions to the class imbalance problem.

3.1 Data-Level (i.e., Resampling) Methods

Changes to the training set's distribution are made using data-level techniques, which keep the algorithm's overall structure, including the loss function and optimizer, undisturbed. In order to make popular learning algorithms, data-level methods try to alter the dataset [7].

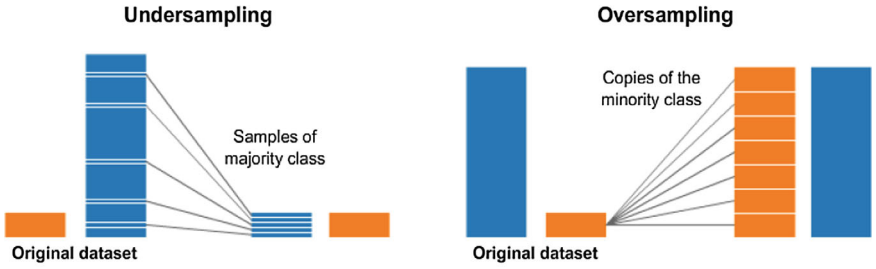
Resampling is a method that balances the number of majority and minority instances in training data. Undersampling techniques and oversampling techniques are the two kinds of resampling methods. Figure 1 depicts the concept of resampling given.

In undersampling methods, by deleting a portion of the majority examples from the training data, an undersampling enables the balance of the majority and minority occurrences. During undersampling, the majority of class samples are removed one at a time until the size of the two classes is nearly equal. As seen in Fig. 1a. In Table 1, the few significant undersampling methods are compared along with their advantages and limitations.

The advancements in the classification of unbalanced data are closely examined in this review paper. This paper discusses the nature of the problem after showing numerous examples of application domains that the class imbalance problem disturbs. The well-known classifier learning algorithms, such as decision trees, back-propagation neural networks, Bayesian networks, nearest neighbors, support vector machines, and associative classification, are analyzed in order to gain an understanding of how difficult it is to use these algorithms to learn from unbalanced data [20].

As you can see once more in Fig. 1b, an oversampling process makes a similar proportion of synthetic minority samples to original minority samples until the sizes of both classes are almost equal. A few significant oversampling techniques are shown in Table 2, along with their advantages and limitations.

Existing solutions, like undersampling and oversampling, alleviate the problem of class imbalance, but they still have major limitations. For instance, undersampling results in the loss of samples containing valuable data about the majority class, while



(a) Undersampling technique. (b) Oversampling technique.

Fig. 1 Resampling methods [17]

Table 1 Undersampling methods

Undersampling methods	Dataset	Performance metrics	Compare algorithm	Advantages	Limitation
RIUS [18]	Glass, Haberman, iris0, vehicle, yeast	Sensitivity, specificity, G-mean, AUC	RUS1, UB4, SBAG4	It chooses the majority class's most pertinent examples	It is solely appropriate for binary class tasks
Downsampling [19]	Pima, Haberman, vehicle, yeast, synthetic, abalone, poker, letter	F-score, G-mean, AUC, AUC-PR	No sampling, TL, NCL, SMOTE, random downsampling, random oversampling	To minimize the impact of imbalanced class labels, it selected the samples that were most informative	It focuses on classification with a binary imbalance
EUStack [14]	Credit card dataset made by European cardholders in Sep. 2013	Precision, recall, F1-score, MCC	AdaBoost, gradient boost, XGBoost, LDA, Naïve Bayes, stacked ensemble	Picks the subset of samples from the dominant class that is most informative	It can serve as a fraud detection method

oversampling necessitates a large amount of computational time. The combination of these problems makes it really difficult to use the fraud detection model [14]. The benefits and disadvantages of under and oversampling-based algorithms are unique to them. It is suggested to use a hybrid resampling algorithm that combines oversampling and undersampling if you want results in data processing that are truly accurate. In reducing the proportion of majority samples while raising the number of minority samples, sample imbalance is largely minimized [24].

The few significant hybrid methods are listed in Table 3, along with their advantages and limitations.

3.2 Algorithmic-Level Methods

This study discussed a new method to unbalanced classification it utilizes a single-class classifier technique to accurately capture the properties of the minority class [7]. The RUSBoost algorithm is described by [26], as a novel hybrid sampling/boosting method for learning from skewed training data and this technique is used in place of SMOTEBoost [27], another technique that mixes boosting and data sampling. In this work, a new technique for the classification of noisy label-imbalanced data is proposed, based on the bagging of XGBoost classifiers [28]. The proposed technique, Weighted Ensemble with One-Class Classification with Oversampling and Instance

Table 2 Oversampling methods

Oversampling	Dataset	Performance metrics	Comparative algorithm(s)	Advantages	Limitation
WK-SMOTE [10]	Pima, Segment0, iris0, yeast, glass, *E. coli*	G-mean	SVM, SMOTE, borderline, ADASYN, PI-SMOTE, SVMDC	It balances the class distribution in an SVM classifier	It is mainly introduced for real-world industrial fault detection problems
MAHAKIL [11]	Ant, arc, camel, ivy, jedit, log4j, pbeans2, redactor	pf measure	SMOTE, borderline-SMOTE, ADASYN, random oversampling	A synthetic oversampling approach for software defect dataset	It does not work in local patches for multi-cluster datasets
GSMOTE-NFM [12]	Pima, Haberman Wisconsin glass, new thyroid, vehicle, *E. coli*	G-mean, F-measure	ROS, SMOTE, B-SMOTE1, B-SMOTE2, SL-SMOTE, GG-SMOTE, RNG-SMOTE	GSMOTE-NFM algorithm generally has better adaptivity and robustness	Its time complexity is generally higher than some other oversampling algorithms
SMOTE-FUN [21]	Pima, phoneme, Australian Bank, heart, oil-spill, abalone90	ROC, AUPRC, Wilcoxon signed—rank test	SWIM using Naïve Bayes, SVM, classifiers, SMOTE, ADASYN	It has no parameters tuning (such as k in SMOTE). So, it is used in real-life applications	It is affected by the fact that one minority class is isolated from other classes and treated as an outlier
SMOTE-tBPSO-SVM [22]	Dataset contains 10,153 Android applications, where 500 of them are ransomware	Sensitivity, specificity, and G-mean	SMOTE, borderline-SMOTE1, borderline-SMOTE2, ADASYN, SVM-SMOTE	It is used for ransomware detection	To handle the big data, it doesn't use more data or advanced models
Approx-SMOTE [23]	SUSY IR4, SUSY IR6, HIGGS	AUC, F1-score	No sampling, SMOTE-BD	Imbalanced learning in big data scenarios is handled	It is designed as an algorithm for the Apache Spark framework

Table 3 Hybrid methods

Hybrid resampling	Dataset	Performance metrics	Compare algorithm	Advantages	Limitation
RFMSE [24]	Spambase, abalone, contraceptive, diabetes, balance, Haberman	Sensitivity, specificity, F-value, MCC	SMOTE CCR, GSM, KSM, SMOTE-ENN	It is used to handle data imbalance in medical diagnosis	It still has a very large gap in the medical diagnosis thinking process of doctors
RK-SVM [10]	Pima, transfusion, iris	Accuracy, sensitivity, specificity, G-mean, AUC, MCC	RK-boosted C5.0, R-SVM, R-boosted C5.0	It improves performance significantly without increasing algorithm complexity	In the reality, the data label is very expensive to obtain
SA-CGAN [6]	Contraceptive, wine, dermatology, yeast	Recall, precision, accuracy, F1-score	GAN, SMOTE, ADASYN, SVM, K-NN, LR, DT	Handle overfitting problems, noisy synthetic, and unclear samples	Certain local data attributes weren't explored, including some local information
SMOTified-GAN [15]	Connect4, credit card, fraud, shuttle, spambase, abalone	Precision, recall, F1-score	Non-oversampled, SMOTE, GAN	Its time complexity is also reasonable for a sequential algorithm	It is an offline preprocessing technique
Hybrid bag-boost model with K-means SMOTE–ENN [25]	Glass, *E. coli*, yeast	AUC, Friedman test, Holm's test	SMOTE, SMOTE-ENN, K-means-SMOTE, K-means	Hybrid bag-boost model for handling noisy class imbalance datasets	It is only working for binary class noisy imbalanced datasets

Selection [29], combines a weighted ensemble classification with a method to tackle the class imbalance issue.

3.3 Cost-Sensitive Learning

Cost-sensitive learning (CSL), which takes into account the different misclassification costs for false negatives and false positives, seems to be another helpful method [30]. In [31] proposed a cost-sensitive (CoSen) deep neural network which can automatically learn acceptable feature representations for both the majority and minority classes. The results of experiments indicate that the function fitting strategy is more efficient than grid searching in obtaining the optimal cost weights for datasets showing imbalanced gene expression [32].

3.4 Deep Learning

Imbalance data classification is still a major challenge in data mining and machine learning, especially for multimedia data, despite research efforts. An extended deep learning approach was offered in [33] as a solution to this problem in order to find skewed multimedia datasets of promising outcomes. More information on the deep learning analysis of a software problem with class imbalance is revealed by this survey [34]. These studies show how to handle the uneven human activity from smart homes and improve the adaptability of the learning algorithms to the minority class using a data-level perspective and a temporal window technique [35]. Making sophisticated models with the DNN is an excellent method of gathering vital information for studies on drug discovery [36].

4 Evaluation Metrics

A confusion matrix, like the one in Table 4, can be used to quantify the performance of a binary classification problem. The minority class is marked by a positive label, whereas the dominant class is denoted by a negative label [34].

The base metrics for evaluation were false positives (FP), false negatives (FN), precision (P), recall (R), and F1-score.

Another useful metric of the accuracy of the prediction in unbalanced classes is the area under the precision-recall curve (AUC-PR), which is a single statistic that covers the precision-recall (PR) curve. As an alternative to the used receiver operating

Table 4 A confusion matrix for binary classification problems

		Truthful value	
		Positive (T)	Negative (N)
Estimated values	Positive (T)	True positive (TP)	False positive (FP)
	Negative (N)	False negative (FN)	True negative (TN)

Table 5 Evaluation metrics

Metric	Description
Precision	It determines how good the classifier is at detecting fraudulent cases
Recall	It evaluates the quality of a qualifier
Accuracy	It measures the efficiency of the algorithm
F-measure	It qualifies the quality of a classifier for the occasional classes
G-mean (geometric mean)	To balance the minority and majority classes, it evaluates a classifier's performance
ROC curve	It is used for evaluating the trade-offs between true positive and false positive error rates in the case of classification algorithms
AUC	It represents the area under a ROC curve
ROC convex hull	It is used as a method for identifying potentially excellent classifiers

characteristic (ROC) curve, which could present an overly optimistic picture of the performance for an unbalanced dataset, PR curves are suggested for dealing with heavily skewed data. With a score of 1.0 denoting a model with perfect ability, this score can then be used to compare various models using a binary classification problem. For comparison, the area under the curve (AUC), which is mostly used in many other articles, is also measured [37].

The prediction only receives a high score if it accurately predicts in each of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), based on the size of the dataset's positive and negative elements, respectively. Matthew's correlation coefficient is a statistical statistic (MCC) [38]. The model is perfect whenever the coefficient is + 1; when it is 0 or equal to a random hypothesis; when it is − 1, the model is totally failed. Contrary to the F1-score, the MCC metric is more reliable [19].

Table 5 gives a short description of important evaluation metrics used for the classifier's performance analysis.

5 Conclusion

In this work, we assessed a few cutting-edge methods for handling class imbalance in classification problems. Every method has advantages and limitations. On imbalanced datasets, a variety of methods are used, such as deep learning, context-sensitive learning, algorithm-level methods, and data-level methods. On the training set, data-level methods such as oversampling, undersampling, and hybrids are used. Undersampling algorithms incur information loss, while oversampling algorithms suffer overfitting issues. Despite hybrid algorithms being more effective than resampling methods, they are indeed computationally more expensive and difficult to use.

Practical use of algorithmic techniques, such as one-class learning and ensemble learning, are applied at the classifier level (i.e., bagging and boosting algorithms). To tackle class imbalance issues in complex datasets, techniques such as deep learning and cost-sensitive learning are also used. To assess the classification accuracy and performance of the classifiers, various evaluation metrics are used.

References

1. Barua S, Murase K (n.d.) A novel synthetic minority oversampling technique for imbalanced data set learning
2. Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 5(4):221–232. https://doi.org/10.1007/s13748-016-0094-0
3. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 73:220–239. https://doi.org/10.1016/j.eswa.2016.12.035
4. Huda S, Liu K, Abdelrazek M, Ibrahim A, Alyahya S, Al-Dossari H, Ahmad S (2018) An ensemble oversampling model for class imbalance problem in software defect prediction. IEEE Access 6:24184–24195. https://doi.org/10.1109/ACCESS.2018.2817572
5. Li Z, Huang M, Liu G, Jiang C (2021) A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection. Expert Syst Appl 175. https://doi.org/10.1016/j.eswa.2021.114750
6. Dong Y, Xiao H, Dong Y (2022) SA-CGAN: an oversampling method based on single attribute guided conditional GAN for multi-class imbalanced learning. Neurocomputing 472:326–337. https://doi.org/10.1016/J.NEUCOM.2021.04.135
7. Kotsiantis S, Kanellopoulos D, Pintelas P (n.d.) Handling imbalanced datasets: a review
8. Laurikkala J (2001) Improving identification of difficult small classes by balancing class distribution
9. Lemaître G, Nogueira F, Aridas CK (2017) Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 18. http://jmlr.org/papers/v18/16-365.html
10. Mathew J, Pang CK, Luo M, Leong WH (2018) Classification of imbalanced data by oversampling in kernel space of support vector machines. IEEE Trans Neural Netw Learn Syst 29(9):4065–4076. https://doi.org/10.1109/TNNLS.2017.2751612
11. Bennin KE, Keung J, Phannachitta P, Monden A, Mensah S (2018) MAHAKIL: diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction. IEEE Trans Softw Eng 44(6):534–550. https://doi.org/10.1109/TSE.2017.2731766
12. Cheng K, Zhang C, Yu H, Yang X, Zou H, Gao S (2019) Grouped SMOTE with noise filtering mechanism for classifying imbalanced data. IEEE Access 7:170668–170681. https://doi.org/10.1109/ACCESS.2019.2955086
13. Hussein AS, Li T, Yohannese CW, Bashir K (2019) A-SMOTE: a new preprocessing approach for highly imbalanced datasets by improving SMOTE. Int J Comput Intell Syst 12(2):1412–1422. https://doi.org/10.2991/ijcis.d.191114.002
14. Laveti RN, Mane AA, Pal SN (2021) Dynamic stacked ensemble with entropy based undersampling for the detection of fraudulent transactions. In: 2021 6th international conference for convergence in technology, I2CT 2021, 2 Apr 2021. https://doi.org/10.1109/I2CT51068.2021.9417896
15. Sharma A, Singh PK, Chandra R (2022) SMOTified-GAN for class imbalanced pattern classification problems. IEEE Access 10:30655–30665. https://doi.org/10.1109/ACCESS.2022.3158977

16. Sisodia D, Sisodia DS (2022) Quad division prototype selection-based k-nearest neighbor classifier for click fraud detection from highly skewed user click dataset. Eng Sci Technol Int J 28. https://doi.org/10.1016/j.jestch.2021.05.015
17. Agarwal R (2020) Sampling [online image]. Kdnuggets.com. https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanceddatasets?scriptVersionId=1756536&cellId=12
18. Hoyos-Osorio J, Alvarez-Meza A, Daza-Santacoloma G, Orozco-Gutierrez A, Castellanos-Dominguez G (2021) Relevant information undersampling to support imbalanced data classification. Neurocomputing 436:136–146. https://doi.org/10.1016/j.neucom.2021.01.033
19. Lee W, Seo K (2022) Downsampling for binary classification with a highly imbalanced dataset using active learning. Big Data Res 28. https://doi.org/10.1016/j.bdr.2022.100314
20. Sun Y, Wong AKC, Kamel MS (2009) Classification of imbalanced data: a review. Int J Pattern Recogn Artif Intell 23(4). www.worldscientific.com
21. Tarawneh AS, Hassanat ABA, Almohammadi K, Chetverikov D, Bellinger C (2020) SMOTE-FUNA: synthetic minority over-sampling technique based on furthest neighbour algorithm. IEEE Access 8:59069–59082. https://doi.org/10.1109/ACCESS.2020.2983003
22. Almomani I, Qaddoura R, Habib M, Alsoghyer S, Al Khayer A, Aljarah I, Faris H (2021) Android ransomware detection based on a hybrid evolutionary approach in the context of highly imbalanced data. IEEE Access 9:57674–57691. https://doi.org/10.1109/ACCESS.2021.3071450
23. Juez-Gil M, Arnaiz-González Á, Rodríguez JJ, López-Nozal C, García-Osorio C (2021) Approx-SMOTE: fast SMOTE for big data on Apache Spark. Neurocomputing 464:432–437. https://doi.org/10.1016/j.neucom.2021.08.086
24. Xu Z, Shen D, Nie T, Kou Y (2020) A hybrid sampling algorithm combining M-SMOTE and ENN based on random forest for medical imbalanced data. J Biomed Inform 107. https://doi.org/10.1016/j.jbi.2020.103465
25. Puri A, Gupta MK (2022) Improved hybrid bag-boost ensemble with K-means-SMOTE–ENN technique for handling noisy class imbalanced data. Comput J 65(1):124–138. https://doi.org/10.1093/comjnl/bxab039
26. Seiffert C, Khoshgoftaar TM, van Hulse J, Napolitano A (2010) RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans Syst Man Cybern Part A Syst Hum 40(1):185–197. https://doi.org/10.1109/TSMCA.2009.2029559
27. Chawla NV, Lazarevic A, Hall LO, Bowyer K (n.d.) SMOTEBoost: improving prediction of the minority class in boosting
28. Ruisen L, Songyi D, Chen W, Peng C, Zuodong T, Yanmei Y, Shixiong W (2018) Bagging of XGBoost classifiers with random under-sampling and tomek link for noisy label-imbalanced data. IOP Conf Ser Mater Sci Eng 428(1). https://doi.org/10.1088/1757-899X/428/1/012004
29. Czarnowski I (2022) Weighted ensemble with one-class classification and over-sampling and instance selection (WECOI): an approach for learning from imbalanced data streams. J Comput Sci 61. https://doi.org/10.1016/j.jocs.2022.101614
30. López V, Fernández A, Moreno-Torres JG, Herrera F (2012) Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. Expert Syst Appl 39(7):6585–6608. https://doi.org/10.1016/j.eswa.2011.12.043
31. Khan SH, Hayat M, Bennamoun M, Sohel F, Togneri R (2015) Cost sensitive learning of deep feature representations from imbalanced data. http://arxiv.org/abs/1508.03422
32. Lu H, Xu Y, Ye M, Yan K, Gao Z, Jin Q (2019) Learning misclassification costs for imbalanced classification on gene expression data. BMC Bioinform 20. https://doi.org/10.1186/s12859-019-3255-x
33. Yan Y, Chen M, Shyu ML, Chen SC (2016) Deep learning for imbalanced multimedia data classification. In: Proceedings—2015 IEEE international symposium on multimedia, ISM 2015, pp 483–488. https://doi.org/10.1109/ISM.2015.126
34. Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6(1). https://doi.org/10.1186/s40537-019-0192-5
35. Hamad RA, Kimura M, Lundström J (2020) Efficacy of imbalanced data handling methods on deep learning for smart homes environments. SN Comput Sci 1(4). https://doi.org/10.1007/s42979-020-00211-1

36. Korkmaz S (2020) Deep learning-based imbalanced data classification for drug discovery. J Chem Inf Model 60(9):4180–4190. https://doi.org/10.1021/acs.jcim.9b01162
37. Davis J, Goadrich M (n.d.) The relationship between precision-recall and ROC curves
38. Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom 21(1). https://doi.org/10.1186/s12864-019-6413

Course Material Recommendation System Using Student Learning Behavior and Course Material Complexity Score for Slow Learner Students

Kamal Bunkar, Chhaya Arya, and Sanjay Kumar Tanwani

Abstract Educational data mining (EDM) is not only a process of applying data mining algorithms on academic data. That is a process of exploring and providing solutions at various levels of educational system. That is useful for students, educators, management, and administration for decision-making and preparing futuristic strategies. The proposed EDM framework is motivated to enhance the student academic performance. The focus is on threefold: first to identify the student learning behavior to support the students, and teacher provide the remedial actions on the weak students. We propose clustering method for the study of students' learning behavior associated with positive and negative outcomes (in exams) by utilizing data mining techniques. Second, there is some classifying approach to characterize student based on performance measure that they earn in examination. Applying supervised classifier on the datasets, we have found significant improvement in results. Finally, the study turns toward developing a technique for recommending appropriate study material by calculating readiness of student and complexity of course material. To achieve the objectives, three models are proposed and combined into one for designing accurate and efficient course material recommendation model. In initial steps, popular data mining algorithms, i.e., K-Means, fuzzy c-means (FCM), and kernel-based FCM (K-FCM), are implemented to cluster students according to their learning behaviors. The comparative study demonstrates that K-FCM-based pattern identification is providing more accuracy with respect to other two algorithms. On the other hand to design a model for student performance prediction, two supervised classifiers, i.e., C4.5 and CART, are implemented. During experiments, we found that the C4.5 decision works well for student performance dataset and for predicting the performance. Using both the components, a course study material recommendation model

K. Bunkar
Institute of Computer Science, Vikram University, Ujjain, India

C. Arya
Pt. JNIBM, Vikram University, Ujjain, India

S. K. Tanwani (✉)
SCS&IT, DAVV, Indore, India
e-mail: sanjay_tanwani@hotmail.com

J. Agrawal et al. (eds.), *Data Engineering and Applications*, Lecture Notes in Electrical Engineering 1146, https://doi.org/10.1007/978-981-97-0037-0_2

is proposed. Thus an application is implemented to demonstrate the student's categorization according to their current performance. That also indicates the learning behavior of the student. On the other hand, the course syllabus is used to offer relevant study material. First the syllabus keywords are extracted. These keywords are used as query to find content from the internet. The identified material is offloaded, and then preprocessing of the contents is carried out. The aim is to recover two kinds of features, first by using the natural language processing (NLP)-based text parser to compute the material's complexity score. Second feature is calculated using TF-IDF as the content features. The calculated features are working as transaction set for Apriori algorithm. The Apriori algorithm is used for the frequent pattern mining. The obtained contents from this mining help to filter the data according to their content relevance. Secondly, the complexity score of the document is used for suggesting the suitable course study material according to student's readiness compatibility. The experimental evaluation of the proposed ML-based EDM framework demonstrates effectiveness, for supporting students to getting performance feedback, to enhance their learning, and obtaining relevant and personalized study material. The performance analysis in terms of precision, recall, and F-score demonstrates the accurate outcomes of the recommendation model.

Keywords EDM · Data mining · Clustering · Learning behavior · Classification and prediction · Student performance prediction · Study material recommendation · Readiness · Complexity of course material

1 Introduction

In a world of increasing complexities and dynamism, evolving new thoughts and applying them in the teaching and learning practices is important and seems inevitable. The ability to think creatively and drive innovations in the education sector particularly, higher education is one of the important prospects toward academician. Education improves the overall ability of a learner. It helps the learner, understands, and solves problems in the real world. The primary goal of higher education policy is to improve the quality of students' academic suitability and to increase their educational base, extensive outreach, and better use of information and resources. In this work, an attempt is made to improve the quality of education through the modern techniques of computer science, like data mining (DM) and machine learning (ML).

Data mining is a technique for exploring essential and fruitful patterns on raw data using mathematical algorithms [1, 2]. This technique is used to understand the relationship between two or more attributes, classifying data, categorizing them, and making predictions. In this context, different kinds of algorithms such as classifiers, clustering algorithms, association rule mining, soft computing, and other kinds of methods and algorithms are applied and that help in various applications and industries to make effective decisions, i.e., banking, credit companies, medicine, decision support system, and others. Similarly, data mining techniques can be applied

to educational data for making decisions and performing predictions for various aspects.

Various techniques specifically for mining educational data have given rise to a new field, called educational data mining (EDM) [3–7]. EDM is defined as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings and using these methods to better understand students and the settings in which they learn in [3]. EDM covers all the methods of data mining and machine learning that consider education data as input and give the output that helps us to understand student requirements.

1.1 Purpose of EDM Research

By enabling them to evaluate the students' performance and monitor their learning progress, instructors and tutors will be better able to comprehend their students' learning behavior [8]. "EDM research has also sought to shed light on the components of a course's structure that require revision in order to enhance learning. The main goal of this kind of research is to aid educators" [6, 7, 9–12]. Other EDM study has focused on offering advantages to teachers as well as their pupils [12–16]. Data mining can be applied for a variety of purposes:

- **Students**: to find learning objectives, resources, and activities based on behavior and comparable paths in order to promote learning,
- **Educators**: to get feedback from students for teaching, assess the course's structure and its impact on student learning, categories students, look for trends, learn more about how to better adapt and customize the course, etc.

The fact that educational data is hierarchical is a significant and distinctive characteristic. Data are nested at the level of the question–answer session, the student interaction, the classroom engagement, the teaching patterns, and the activities of colleges [12]. Advancements of the EDM techniques concentrates on the longitudinal data modeling and hierarchical data mining techniques. Researchers in educational data mining [13]; see the following objectives as important:

1. Making data models that contain specifics of students' knowledge, motivation, meta-cognition, and attitudes can be used to predict students' future learning behavior,
2. Identifying or improving data models to demonstrate the course outcome that is to be learn and enhance the instructional flow,
3. Investigating the results of various pedagogical supports offered by learning software, and
4. Developing computational models that include educational domain, data models of the learner, and automation of pedagogy to understand the scientific approach of learning and learner.

To accomplish these goals, EDM research uses the five categories of technical methods [14]:

1. **Prediction** requires creating a model that can use a single data point from a combination of other data points called successively predicted variables and predictor variables. Prediction include identifying student actions when they are manipulating the system, behaving inattentively, or answering a question incorrectly despite possessing a talent. The prediction demonstrates anticipating and comprehending educational outcomes for students, such as achievement on post-tests following tutoring [15].
2. **Clustering** refers to the process of identifying data points that correlate and that can act as a similar group as per the similar characteristics among them [17].
3. **Relationship** mining is the process of identifying connections between variables in a dataset and encoding those connections as rules for subsequent usage. Relationship mining, for instance, can reveal connections between items bought online [18].
    - **Association rule mining** can be used to identify common errors made by students, link user types to content to create recommendations for the necessary content, or modify teaching methods.
    - **Sequential pattern mining** creates rules that account for the relationships between the occurrences of sequential events, such as identifying temporal sequences like student errors followed by support requests.
4. **Distillation** for human judgment is a technique that involves presenting data in a way that makes it possible for a person to swiftly recognize or categorize its qualities.
5. **Discovery** using a verified model of a phenomenon (created by prediction, clustering, or human knowledge engineering) as a part of the study is an approach known as using models [19].

1.2 Motivation

India is a growing country and one of the youngest nations having a large percentage of the population in the young generation. A large proportion of the population is studying in schools, colleges, and other higher educational institutions. Students study to gain knowledge and evolve a deeper understanding of the subject. Not all students have similar learning capabilities and potential to learn. Therefore, it is required to understand the strength and weaknesses of every student. Additionally, need to understand the gap between the learning strategy and available resources [20]. That is required for making improvements in learning patterns, teaching skills, and providing the appropriate study or learning materials. In this context, the proposed EDM system works in three aspects:

1. Understanding the level of student's learning ability by analyzing their previous performance and by predicting the performance of the students for the near future.

2. Identifying the key issues and weaknesses in the learning process and suggesting ways to improve them.
3. Providing the personalized learning material according to their needs and learnability by using a collaborative filtering approach that first understands the behavior and then suggests the material to be used.

These methods can help the teachers and organizations to understand the quality of learning, possible improvements on the offered services, and make future sustainable resources for next-generation teaching and learning systems.

1.3 Research Objectives

To explore and investigate data mining techniques for designing an adoptive EDM framework with the following key objectives to work:

- **To propose a framework for efficient and accurate knowledge extraction for higher education**: The different attributes of the student's performance are identified, and according to observations, a dataset is prepared/obtained.
- **To explore the productivity of teachers**: The student's performance is reflections of teacher's efforts and productivity, by categorizing the student's performance in three groups (elaborate group bright, medium, or weak) the teacher's productivity and interaction of a teacher in a classroom can be measurable.
- **To identify teaching patterns and gaps in learning with the help of efficient data mining techniques**: Using the group of subject-wise weak students, the teacher can make more effort on that student to scale their performance.
- **To give teachers feedback so they can decide how to improve learning of the learner and take preventative or corrective action**: The attempts to offer fresh perspectives on students' learning habits that both teachers and students can find helpful.
- **Personalizing student's behavior and learning ability to recommend study material and future adaptable solutions**: We wish to implement a recommendation system that analyzes behavior and learning ability for students and suggests more appropriate learning material for the students to boost their performance of learners.

1.4 Contributions

The EDM is a wide domain of research and involves a large user community from teachers, educationists, policymakers, and various stakeholders of the education system and has been applied for many outcomes in education research. In the current context, few limited, but essential aspects of EDM are proposed, which are explained as follows:

1. **Student learning patterns**: Learning performance of each student is different and needs a different level of explanation and material to understand. The proposed work is intended to recover the student learning patterns for improving their learning skills and suggest appropriate material for reading.
2. **Student performance and success rate in corporate**: Educational institutes wish to maximize their student's success. The performance predictions of students and their future projections help to understand their future possibilities of growth and learning ability.
3. **Identifying students learning behavior and recommending appropriate course material**: The recommendation systems are providing ease in various real-world applications and problems to offer appropriate services and products in e-commerce platforms. In this work, the aim is to understand the learning behavior of students and recommend relevant study materials according to their learning habits and their level of understanding.

2 Proposed Work

The proposed work is helping to improve the learning of students and teaching skills of educators using the EDM techniques. That may help to educators and educational institutes to prepare their strategies for future growth and productivity of students and teachers, respectively. This section provides a discussion about the operational and functional aspects of the proposed EDM system.

2.1 System Overview

Data exploration and pattern discovery using data mining techniques are beneficial for creating new applications and improving existing ones. Many different types of algorithms are available for data analysis using data mining techniques. Algorithms are chosen for use based on the types of data they will process (i.e., structured or unstructured). The educational data, also known as EDM, is utilized with data mining techniques in this work.

The work primarily focused on examining educational data in order to empower students, increase teacher productivity, identify student learning patterns, and suggest pertinent reading or course materials. The job is separated into three primary components as a result. The clustering technique is applied on the student's previous performance records in the first module to identify the group of ineffective, average, and productive pupils. These groupings assist us in identifying the pupils' learning preferences. In the second module, we provide a prediction technique using the performance data of the pupils to comprehend their future development and growth. Finally, a recommendation model is put into place to identify the student's learning behavior