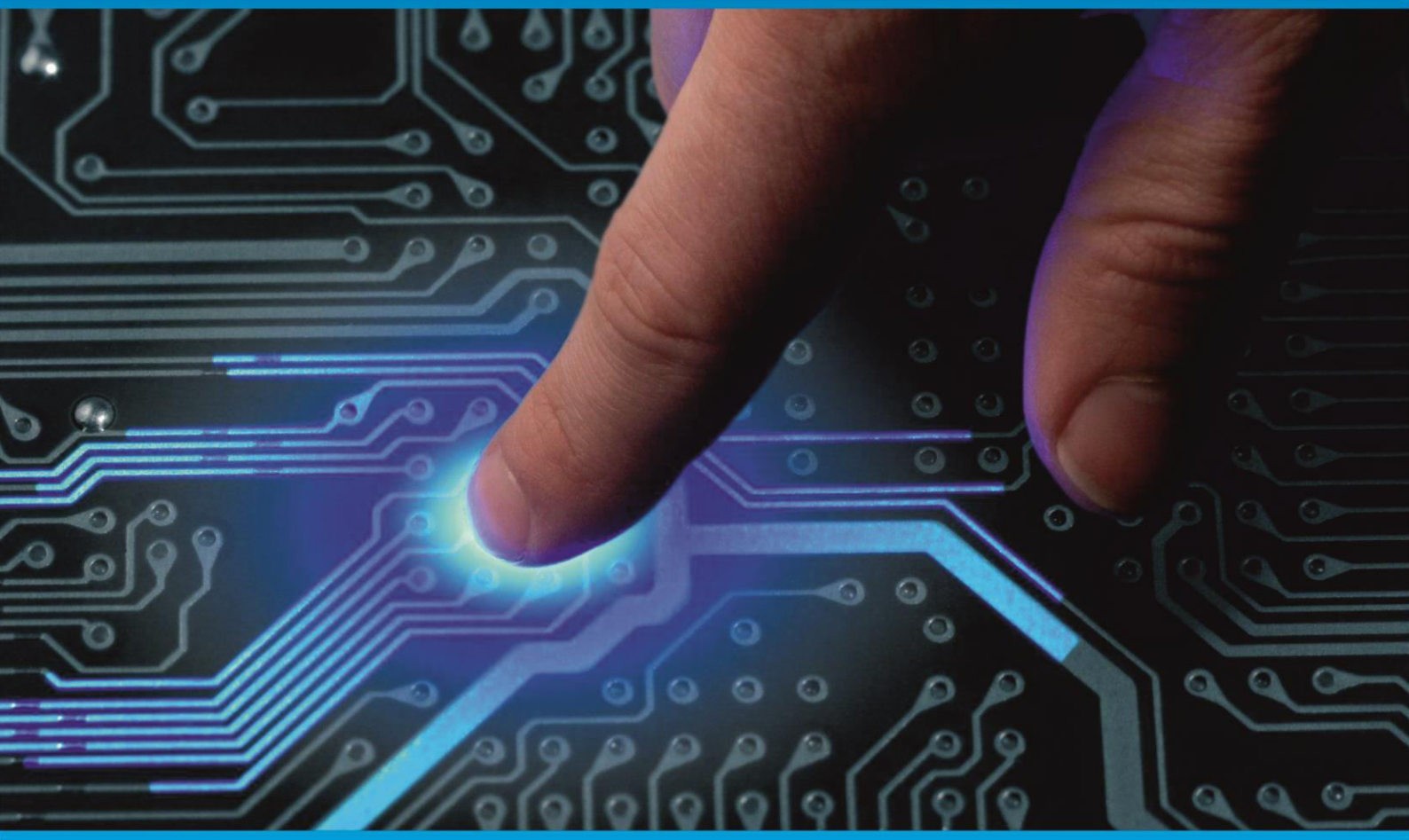
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 3, March 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021. 0903146|

Entropy Based Solution for the Imbalanced Data of the Product Using Sampling Methods

Divya.V[1] , Karthika.M[2] , Kilda Jenifer.A.P[3] , Veena.T[4]

UG Scholar, Department of IT, S.A Engineering College, Chennai, Tamil Nadu, India[1,2,3]

Assistant Professor, Department of IT, S.A Engineering College, Chennai, Tamil Nadu, India[4]

ABSTRACT: Here in existing framework now days in shops they are keeping up old items and lapsed items if any one utilized those items in a few circumstances will be harmed. What's more, a portion of the shop people are changing that all dates or more the cover and making it like a unique item in the wake of terminating time they are changing that all spreads everything. Fundamentally these issues are occurring in healing facility drug additionally their specialists are giving distinctive sorts of medication for various sickness. At whatever point they will understand that therapeutic shop they will give for specific sickness diverse prescription. Here to overcome each one of those issue first client needs to keep up every one of the items with id. presently after login the businessperson account they need to transfer every one of the insights regarding items and they need to keep up make item and terminate date all they need to keep up in the wake of transferring all that these all data will goes to government group now government group will deal with that all data and they can investigate and they will give all the data about the item expire date if the item will lapse they will send a notice to retailer before 15days of item will terminate. At that point businessperson will make offer for that specific id items then just it won't be squander capable that items. It will demonstrate the fabricate date and terminate date in the event that it was phony it won't demonstrate any outcome. if like that any client discover like that, they can send a mail. To administrator they can make a move on that specific shop.

KEYWORDS: Product expire date, EID (Entropy imbalanced data), Alert message.

I.INTRODUCTION

Imbalanced learning has pulled in a lot of premiums in the examination network. The vast majority of the outstanding information mining and AI procedures are proposed to take care of grouping issues concerning sensibly adjusted class circulations. In any case, this supposition isn't in every case valid for a slanted class circulation issue existing in some true informational collections, in which a few classes (the greater parts) are over-spoken to by an enormous number of examples however some others (the minorities) are underrepresented by just a couple. The answers for the class imbalance issue utilizing customary learning methods predisposition the prevailing classes bringing about poor characterization execution. For amazingly multi-class imbalanced information set, imbalanced order execution might be given by conventional classifiers with an almost 100 percent precision for the larger parts and with near 0 percent precision for the minorities. Henceforth, the class-irregularity issue is considered as a noteworthy obstruction to the achievement of exact classifiers. So as to defeat this disadvantage, we present another metric, named entropy-based lopsidedness degree. It has been realized that data entropy can mirror the positive data substance of a given informational collection. Therefore, we measure the data substance of each class and acquire the distinctions among them, i.e., EID. So as to limit EID to adjust the informational index in data content, an entropy-based half and half examining methodology is proposed, joining both entropy-based oversampling and entropy-based under-sampling techniques.

II. RELATED WORKS

A. Hellinger distance based oversampling method to solve multi-class imbalance problem

Classification is a popular technique used to predict group membership for data samples in datasets. A multi-class or multinomial classification is the problem of classifying instances into more than two classes. With the emerging technology, the complexity of multi-class data has also increased thereby leading to class imbalance problem. With an imbalanced dataset, a machine learning algorithm cannot make an accurate prediction. Therefore, in this paper Hellinger distance based oversampling method has been proposed. It is useful in balancing the datasets so that minority class can be identified with high accuracy without affecting accuracy of majority class. New synthetic data is generated using this method to achieve balance ratio. Testing has been done on five benchmark datasets using two standard

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021. 0903146|

classifiers KNN and C4.5. The evaluation matrix on precision, recall and measure are drawn for two standard classification algorithms. It is observed that Hellinger distance reduces risk of overlapping and skewness of data. Obtained results show increase of 20% in classification accuracy compared to classification of imbalance multi-class dataset.

B. HIERARCHICAL FEATURE SELECTION FOR RANDOM PROJECTION

Random projection is a popular machine learning algorithm, which can be implemented by neural networks and trained in a very efficient manner. However, the number of features should be large enough when applied to a rather large-scale data set, which results in slow speed in testing procedure and more storage space under some circumstances. Furthermore, some of the features are redundant and even noisy since they are randomly generated, so the performance may be affected by these features. To remedy these problems, an effective feature selection method is introduced to select useful features hierarchically. Specifically, a novel criterion is proposed to select useful neurons for neural networks, which establishes a new way for network architecture design. The testing time and accuracy of the proposed method are improved compared with traditional methods and some variations on both classification and regression tasks. Extensive experiments confirm the effectiveness of the proposed method.

C: LEARNING A DISTANCE METRIC BY BALANCING KL-DIVERGENCE FOR IMBALANCED DATASETS

In many real-world domains, datasets with imbalanced class distributions occur frequently, which may confuse various machine learning tasks. Among all these tasks, learning classifiers from imbalanced datasets is an important topic. To perform this task well, it is crucial to train a distance metric which can accurately measure similarities between samples from imbalanced datasets. Unfortunately, existing distance metric methods, such as large margin nearest neighbor, information-theoretic metric learning, etc., care more about distances between samples and fail to take imbalanced class distributions into consideration. Traditional distance metrics have natural tendencies to favor the majority classes, which can more easily satisfy their objective function. Those important minority classes are always neglected during the construction process of distance metrics, which severely affects the decision system of most classifiers. Therefore, how to learn an appropriate distance metric which can deal with imbalanced datasets is of vital importance, but challenging. In order to solve this problem, this paper proposes a novel distance metric learning method named distance metric by balancing KL-divergence (DMBK). DMBK defines normalized divergences using KL-divergence to describe distinctions between different classes. Then it combines geometric mean with normalized divergences and separates samples from different classes simultaneously. This procedure separates all classes in a balanced way and avoids inaccurate similarities incurred by imbalanced class distributions. Various experiments on imbalanced datasets have verified the excellent performance of our novel method.

D. SDE: A NOVEL CLUSTERING FRAMEWORK BASED ON SPARSITY-DENSITY ENTROPY

Clustering of data with high dimension and variable densities poses a remarkable challenge to the traditional density-based clustering methods. Recently, entropy, a numerical measure of the uncertainty of information, can be used to measure the border degree of samples in data space and also select significant features in feature set. It was used in our new framework based on the sparsity-density entropy (SDE) to cluster the data with high dimension and variable densities. First, SDE conducts high-quality sampling for multidimensional data and selects the representative features using sparsity score entropy (SSE). Second, the clustering results and noises are obtained adopting a new density-variable clustering method called density entropy (DE). DE automatically determines the border set based on the global minimum of border degrees and then adaptively performs cluster analysis for each local cluster based on the local minimum of border degrees. The effectiveness and efficiency of the proposed SDE framework are validated on synthetic and real data sets in comparison with several clustering algorithms. The results showed that the proposed SDE framework concurrently detected the noises and processed the data with high dimension and various densities.

E. RUSBOOST: IMPROVING CLASSIFICATION PERFORMANCE WHEN TRAINING DATA IS SKEWED.

Constructing classification models using skewed training data can be a challenging task. We present RUS Boost, a new algorithm for alleviating the problem of class imbalance. RUS Boost combines data sampling and boosting, providing a simple and efficient method for improving classification performance when training data is imbalanced. In addition to performing favorably when compared to SMOTE Boost (another hybrid sampling/boosting algorithm), RUS Boost is computationally less expensive than SMOTE Boost and results in significantly shorter model training times. This combination of simplicity, speed and performance makes RUS Boost an excellent technique for learning from imbalanced data.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021. 0903146|

F. HELLINGER DISTANCE BASED OVERSAMPLING METHOD TO SOLVE MULTI-CLASS IMBALANCE PROBLEM.

Classification is a popular technique used to predict group membership for data samples in datasets. A multi-class or multinomial classification is the problem of classifying instances into more than two classes. With the emerging technology, the complexity of multi-class data has also increased thereby leading to class imbalance problem. With an imbalanced dataset, a machine learning algorithm cannot make an accurate prediction. Therefore, in this paper Hellinger distance based oversampling method has been proposed. It is useful in balancing the datasets so that minority class can be identified with high accuracy without affecting accuracy of majority class. New synthetic data is generated using this method to achieve balance ratio. Testing has been done on five benchmark datasets using two standard classifiers KNN and C4.5. The evaluation matrix on precision, recall and measure are drawn for two standard classification algorithms. It is observed that Hollinger distance reduces risk of overlapping and skewness of data. Obtained results show increase of 20% in classification accuracy compared to classification of imbalance multi-class dataset.

III. EXISITNG SYSTEM

In the existing structure, the inspecting strategies have shown their in-adequacy, for instance, causing the issues of over-age and over-lapping by oversampling techniques or the absurd loss of enormous information by under-looking at frameworks, which implies if the administrator is sending the alarm to the retailer implies, the businessperson may supplant or may not supplant the item. This has been an immense issue for the client to discover the items and get them. Tragically, existing inspecting strategies have indicated their inadequacies, for example, causing the issues of over-age and over-lapping by oversampling procedures, or the exorbitant loss of critical data by under examining methods.

ADVANTAGE:

This system is used to identifying the expire dates of the products. . the expire dates are handled by the admin and the admin will give the alert message to the shopkeeper once the particular product has been expired. . by using this system, the shopkeeper will replace the expired products by the new products.

DISADVANTAGE:

the disadvantage is that the execution time is more as wasted in checking the expire date every time. • the disadvantage of the apriori algorithm is the algorithm , this process takes a lot of time and memory , especially if the pattern it too many and long. • apriori algorithm suffers from some weakness , inspite of being clear and simple. the main limitation is costly and wasting of time . • and if the admin is sending the alert to the shopkeeper means, the shopkeeper may replace or may not replace the product. this has been the huge problem for the customer to find the products and buy it.

IV. PROPOSED SYSTEM

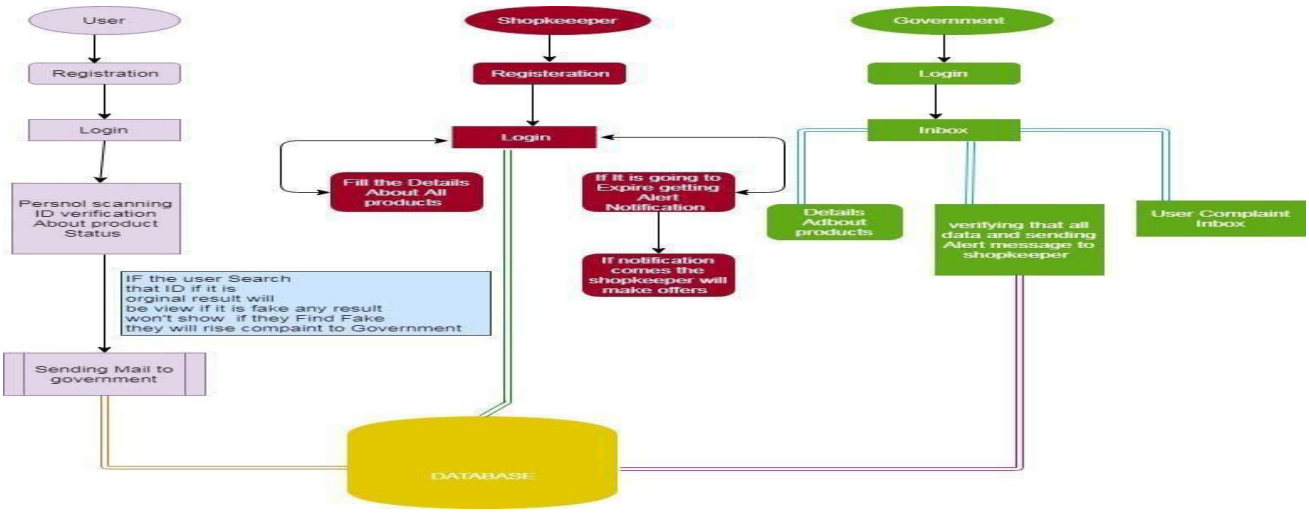
This paper presents three examining based methodologies, each essentially improving the general mining cost by diminishing the number of copies produced. By utilizing these examining methods, the public authority can without much of a stretch send the item termination date to the businessperson effectively and causes them to effortlessly locate the lapsed items. To conquer this disadvantage, we present another measurement, named entropy-based unevenness degree. It has been realized that data entropy can mirror the positive data substance of a given informational collection. In this way, we measure the data substance of each class and get the distinctions among them, i.e., EID.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021. 0903146|

V. SYSTEM ARCHITECTURE



VI. MODULES

- User Interface design
- Shopkeeper uploading details about products.
- Government inbox
- Government view and maintain the product
- Customer complaint inbox
- Shopkeeper product status inbox
- Customer verification
- Sending compliant to Government

VII. CONCLUSION

In this paper, we present three new entropy-based learning approaches, for multi-class unevenness learning issues. For a given imbalanced informational index, the proposed techniques utilize new entropy-based unevenness degrees to gauge the class irregularity as opposed to utilizing conventional unevenness proportion. EOS depends on the data substance of the biggest dominant part class. EOS oversamples different classes until their data substance accomplish the biggest one. EHS depends on the normal data substance of the considerable number of classes, and oversamples the minority classes just as under samples the greater part classes as indicated by EID. The viability of our proposed three techniques is exhibited by the unrivaled learning execution both on manufactured and real-world informational collections. Moreover, since entropy-based half and half examining can all the more likely safeguard information structure than entropy-based oversampling and entropy-based under-sampling by creating less new minority tests just as expelling less greater part tests to adjust informational indexes, it has more predominance than entropy-based oversampling and entropy-based under-sampling.

VIII. FUTURE ENHANCEMENT

In the future, we might want to investigate the hypothetical properties of our proposed lopsidedness measure and broaden it just as our three imbalanced learning techniques for other grouping issues, for example, picture arrangement what's more, move realizing

REFERENCES

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021. 0903146|

[1] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on knowledge and data engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
[2] Z. Wan, H. He, and B. Tang, "A generative model for sparse hyperparameter determination," IEEE Transactions on Big Data, vol. 4, no. 1, pp. 2–10, March 2018.
[3] C.-T. Lin, T.-Y. Hsieh, Y.-T. Liu, Y.-Y. Lin, C.-N. Fang, Y.-K. Wang, G. Yen, N. R. Pal, and C.-H. Chuang, "Minority oversampling in kernel adaptive subspaces for class imbalanced datasets," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 5, pp. 950– 962, 2018.
[4] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, "Confusion-matrix-based kernel logistic regression for imbalanced data classification," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 9, pp. 1806–1819, 2017.
[5] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time ev charging scheduling based on deep reinforcement learning," IEEE Transactions on Smart Grid, pp. 1–1, 2018.
[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority oversampling technique," Journal of artificial intelligence research, vol. 16, pp. 321–357, 2002.
[7] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," Pattern Recognition, vol. 72, pp. 327–340, 2017.
[8] K. E. Bennin, J. Keung, P. Phannachitta, A. Monden, and S. Mensah, "MAHAKIL: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction," IEEE Transactions on Software Engineering, 2017.
[9] Z. Wan and H. He, "Answernet: Learning to answer questions," IEEE Transactions on Big Data, pp. 1–1, 2018.
[10] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, 2009, pp. 475–482.
[11] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in IEEE International Joint Conference on Neural Networks, 2008, pp. 1322–1328.
[12] S. Chen, H. He, and E. A. Garcia, "RAMO Boost: ranked minority oversampling in boosting," IEEE Transactions on Neural Networks, vol. 21, no. 10, pp. 1624–1642, 2010.
[13] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 2, pp. 405–425, 2014.
[14] H. Han, W.-Y. Wang, and B.-H. Mao, "BorderlineSMOTE: a new over-sampling method in imbalanced data sets learning," in International Conference on Intelligent Computing, 2005, pp. 878–887.
[15] X. Yang, Q. Kuang, W. Zhang, and G. Zhang, "Amdo: an over-sampling technique for multi-class imbalanced problems," IEEE Transactions on Knowledge and Data Engineering, vol. PP, no. 99, pp. 1–1, 2017.
[16] L. Li, H. He, J. Li, and W. Li, "Edos: Entropy difference based oversampling approach for imbalanced learning," in 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018, pp. 1–8.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details