Business Anomaly Detection Method of Power Dispatching Automation System Based on Clustering Under-Sampling in the Boundary Region

Junliang Li¹², Jianhang Xu¹², Xu Huang³*, Bing Ren³, Tianqi Dai¹², Zemin Zhang¹² and Rui Su¹²
¹ State Grid Electric Power Research Institute Company Limited, Beijing, China
² Beijing Kedong Power Control System Company Limited, Beijing, China
³ School of Automation, Beijing University of Posts and Telecommunications, Beijing, China
Email: 913498518@qq.com

Abstract. Timely detecting business anomaly in the power dispatching automation system is significant for the steady operation of the power grid. Though the imbalanced binary classification method in machine learning is an effective way to achieve the business anomaly detection of the system, the overlap of boundary samples is an urgent issue affecting the classification effect. An under-sampling method by removing the clustering noises of the majority samples in the boundary region is proposed. Firstly, KNN is used to search adjacent points of the majority class, and the boundary region and the safety region are divided according to the proportion of the majority samples in adjacent points. Secondly, DBSCAN is used to cluster the majority samples in the boundary region, and noise points are removed. Finally, it’s combined with the method based on model dynamic selection driven by data partition hybrid sampling (DPHS-MDS). The purpose of reducing the overlap degree of boundary samples, balancing the dataset and improving the classification effect is achieved. Experimental results show that the proposed method is superior to the relevant mainstream methods under F-measure and G-mean.

Keywords. Machine learning; anomaly detection; under-sampling.

1. Introduction
The power dispatching automation system is of great significance for the steady operation of the power grid, and it is very important to detect the system anomaly in time. Due to the stable operation of the system, the number of normal samples and abnormal samples varies greatly, and the imbalanced binary classification method in machine learning is one of the effective means to solve this problem [1]. In recent years, one of the research hotspots is the combination of data preprocessing and algorithm improvement. The relevant methods include RUSboost [2], SMOTEboost [3], EasyEnsemble [4], BalancedBagging [5], BRAF [6], DTE-SBD [7]. The ensemble imbalanced classification method based on model dynamic selection driven by data partition hybrid sampling (DPHS-MDS) proposed by Xin Gao is typical [8]. However, only minority classes in the boundary region are sampled by BMW-smote, which does not reduce the overlap of samples in the boundary region. Therefore, this paper proposes an under-sampling method by removing the majority clustering noise in the boundary region. Combined with DPHS-MDS [8], the overlap degree of samples in the boundary region is reduced and the dataset is balanced, so as to improve the classification effect.
Through comparison with mainstream methods, sufficient experiments show that the proposed method is more effective.

2. Methodology

2.1. Analysis on the Improvement of the Method

In 2020, Xin Gao proposed an ensemble imbalanced classification method (DPHS-MDS) [8]. This method only conducts BMW-smote sampling for the minority samples in the boundary region, and does not reduce the overlap degree of samples in the boundary region [8].

In order to solve the above problem, this paper proposes an under-sampling method by removing the majority clustering noise in the boundary region. On the basis of BMW-smote sampling for the minority class in the boundary region, we hope to reduce the sample overlap by under-sampling the majority class [8]. But random under-sampling will lose important samples. Therefore, density based clustering algorithm (DBSCAN) is used for clustering [9]. The samples that cannot be clustered are regarded as clustering noise points. We argue that in the neighborhood of these points, the existence of the majority sample is not of universal significance, and the degree of relative importance is low, so we remove such points and effectively retain the relatively important majority class points, at the same time, the dataset is balanced. The method is mainly divided into two steps: KNN algorithm is used to divide majority class samples in the boundary region [10]. Then, clustering noise points are removed after clustering majority class points in the boundary region by DBSCAN algorithm [9].

2.2. Dividing Majority Class Samples into the Boundary Region Based on KNN

Suppose that the distribution of some samples in the two-dimensional space of the dataset is shown in figure 1a, taking k = 5 as an example, find k nearest points of majority class points and count the number of majority classes of them. If adjacent points are all of the majority classes, the sample point is divided into the safety region, otherwise the sample point is divided into the boundary region.

The specific algorithm steps are as follows: according to the imbalanced dataset $D$ and the label of majority class $L_{maj}$, the majority class dataset $D_{maj}$ is divided. By traversing the majority class dataset $D_{maj}$, KNN algorithm is used to search k adjacent points of each majority class point $d_i$, and the number of majority samples $N_i$ in the adjacent points is counted [10]. By judging $N_i = k$, the majority sample point $d_i$ is divided into the majority class safety region $D_{safe}$, otherwise, the point is divided into the majority class boundary region $D_{border}$.

2.3. Under-Sampling of the Majority Clustering Noise in The Boundary Region Based On DBSCAN

After obtaining majority class samples in the boundary region, DBSCAN algorithm is used for clustering [9]. The basic principle of clustering is as follows: search whether there are other sample points in the neighborhood of each sample point, if not, mark the sample point as the clustering noise point. As shown in figure 1b, if two points exist in each other's neighborhood, then they are defined to be connected by density. A series of points connected by density are clustered into one class, and clustering noise points that cannot be clustered are removed to achieve the effect of under-sampling.

The specific algorithm steps are as follows: first, according to the dataset $D$ and sample labels $L_{maj}$ and $L_{min}$, the number of majority class samples $N_c$ and the number of minority class samples $N_r$ are calculated, and the imbalance rate $r$ of the dataset is calculated. The appropriate neighborhood threshold $\epsilon$ is calculated according to equation (1):

$$
\epsilon = (\log_{10} \frac{1}{r^{2/3}}) + 2
$$

Using the above parameters, DBSCAN is used to assign clustering labels $c\_label$ to majority class samples in the boundary region $D_{border}$, and delete the sample points whose clustering labels are noise $L_{noise}$. Finally, the processed dataset of majority class in the boundary region $D_{border\_us}$ is obtained.
2.4. **DPHS-cnus-MDS Overall Framework**

The overall framework of the proposed method (DPHS-cnus-MDS) is as follows: as shown in figure 1c, we remove samples in the minority class noise region from the original dataset and form the filtered dataset. After under-sampling the clustering noise of majority classes in the boundary region, over-sample the minority classes in the boundary region by BMW-SMOTE, and randomly under-sample the majority classes in the safety region after clustering them. The minority class samples in the safety region are retained, and the balanced dataset is formed by merging the processed samples from the above regions. The original random forest model and the biased ones are generated by the filtered dataset and the balanced dataset respectively. The base classifiers (decision trees) of two models are integrated to obtain the hybrid random forest model. Finally, the type of test points is judged according to the proportion of majority class training samples in the neighborhood of test points, and the suitable model is dynamically selected for different types of test points to classify [8].

3. **Experiment**

The experimental data are from UCI, KEEL and power dispatching business. In order to fully compare the proposed method with relevant methods, Friedman test was carried out [11]. This paper uses F-measure and G-mean as evaluation indexes to evaluate the classification effect of the method.

3.1. **Experiments on Public Datasets**

In order to verify that the proposed method (DPHS-cnus-MDS) can effectively improve the classification effect of imbalanced data, seven relevant mainstream methods are compared, including RUSboost [2], SMOTEboost [3], EasyEnsemble [4], BalancedBagging [5], BRAF [6], DTE-SBD [7], DPHS-MDS [8]. Tables 1 and 2 show the F-measure and G-mean of DPHS-cnus-MDS and above relevant methods on the selected 8 public datasets.

According to the analysis in tables 1 and 2, the proposed method is superior to other relevant methods under F-measure and G-mean, and achieves 7 best results on 8 datasets. The average rank of 1.13 and 1.25 rank first in Friedman test. This shows that the proposed method not only ensures the accuracy of the minority class classification, but also reduces the misjudgment of the majority class.
Table 1. Comparison of results between DPHS-cnus-MDS and related methods under F-measure.

| Datasets      | RUS boost | SMOTE boost | Easy Ensemble | Balanced Bagging | BRAF | DTE-SBD | DPHS-MDS | DPHS-bus-MDS |
|---------------|-----------|-------------|---------------|------------------|------|---------|----------|-------------|
| vehicle1      | 0.4678    | 0.5130      | 0.6193        | 0.6169           | 0.5218 | 0.5862  | 0.5849   | **0.6345**   |
| ecoli1        | 0.6544    | 0.5270      | 0.7378        | 0.8101           | 0.7823 | 0.7801  | 0.8080   | **0.8143**   |
| ecoli2        | 0.6474    | 0.5924      | 0.7597        | 0.7856           | 0.8102 | 0.7826  | 0.8352   | **0.8456**   |
| yeast3        | 0.5250    | 0.5569      | 0.6947        | 0.7467           | 0.7348 | 0.7500  | 0.7744   | **0.8001**   |
| ecoli3        | 0.3604    | 0.3482      | 0.5831        | 0.5887           | 0.6126 | 0.6080  | 0.5988   | **0.6770**   |
| vowel0        | 0.7819    | 0.7295      | 0.8691        | 0.8222           | 0.9623 | 0.9097  | 0.9569   | **0.9654**   |
| yeast-1vs7    | 0.2593    | 0.1391      | 0.3228        | 0.3634           | 0.3849 | 0.2857  | 0.3578   | **0.4025**   |
| ecoli-0-1-3-7vs2-6 | 0.1479 | 0.1035   | 0.1483    | **0.7369**     | 0.5000 | 0.4469  | 0.5000   | 0.7100      |
| Average       | 0.4805    | 0.4387      | 0.5919        | 0.6838           | 0.6636 | 0.6437  | 0.6770   | **0.7312**   |
| Friedman rank | 7.25      | 7.75        | 5.25          | 3.5              | 3.44  | 4.38    | 3.31     | **1.13**     |

Table 2. Comparison of results between DPHS-cnus-MDS and related methods under G-mean.

| Datasets      | RUS boost | SMOTE boost | Easy Ensemble | Balanced Bagging | BRAF | DTE-SBD | DPHS-MDS | DPHS-bus-MDS |
|---------------|-----------|-------------|---------------|------------------|------|---------|----------|-------------|
| vehicle1      | 0.5370    | 0.5980      | 0.7179        | 0.7459           | 0.6455 | 0.7208  | 0.7185   | **0.7787**   |
| ecoli1        | 0.8148    | 0.6915      | 0.8680        | 0.9050           | 0.8566 | 0.8654  | 0.8905   | **0.9056**   |
| ecoli2        | 0.3905    | 0.3426      | 0.8578        | 0.8769           | 0.8743 | 0.8814  | 0.8864   | **0.9121**   |
| yeast3        | 0.2649    | 0.2693      | 0.9023        | 0.8794           | 0.8310 | 0.8644  | 0.8857   | **0.9129**   |
| ecoli3        | 0.7157    | 0.6577      | 0.8129        | 0.8597           | 0.7186 | 0.8141  | 0.7622   | **0.8811**   |
| vowel0        | 0.8042    | 0.8745      | 0.9556        | 0.9598           | 0.9694 | 0.9580  | 0.9689   | **0.9892**   |
| yeast-1vs7    | 0.6397    | 0.4530      | 0.6318        | 0.7094           | 0.5079 | 0.5874  | 0.4517   | **0.7148**   |
| ecoli-0-1-3-7vs2-6 | 0.6857 | 0.7293   | **0.8128**   | 0.7410           | 0.5401 | 0.7264  | 0.5401   | 0.7405      |
| Average       | 0.6066    | 0.5770      | 0.8199        | 0.8346           | 0.7429 | 0.8022  | 0.7630   | **0.8544**   |
| Friedman rank | 6.75      | 7.00        | 4.00          | 2.75             | 5.56  | 4.25    | 4.44     | **1.25**     |

3.2. Business Anomaly Detection in the Power Dispatching Automation System

The effectiveness of the proposed method is verified by combining three kinds of business abnormal data of the power dispatching, which include application disconnection (AD), data jump (DJ) and telemetry table not refreshing (TTNR) [12]. The experimental results are shown in table 3. The proposed method achieves the best results in F-measure and G-mean, which shows that the proposed method has advantages in business anomaly detection of the power dispatching automation system.
Table 3. Experimental results of DPHS-cnus-MDS and related methods on business abnormal datasets.

| Business abnormal datasets | F-measure | G-mean |
|----------------------------|-----------|--------|
|                           | RUS boost | SMOTE boost | Easy Ensemble | Balanced Bagging | BRAF SBD | DTE-SBD | DPHS-MDS | DPHS-bus-MDS |
| AD                        | 0.9045    | 0.9645  | 0.9634      | 0.9695          | 0.967    | 0.9795  | 0.9799  | **0.9996** |
| DJ                        | 0.9721    | 0.9721  | 0.9712      | 0.9757          | 0.9692   | 0.9829  | 0.9847  | **0.9999** |
| TTNR                      | 0.9745    | 0.9721  | 0.9834      | 0.981           | 0.9803   | 0.9848  | 0.9864  | **1**    |

4. Conclusion

In order to solve the problem of the overlap of samples in the boundary region, an under-sampling method by removing majority clustering noises in the boundary region is proposed. Combined with DPHS-MDS, the proposed method is compared with relevant methods. The experimental results on 8 imbalanced datasets show that the proposed method performs better on F-measure and G-mean, which are improved by 5.4% and 9.1% respectively compared with DPHS-MDS. Friedman rank achieved 1.13 and 1.25 on F-measure and G-mean respectively, ranking first. Finally, the proposed method is carried out in the business anomaly datasets of the power dispatching automation system, and the better detection effect is achieved. The F-measure and G-mean are all above 0.99.

To sum up, the proposed method provides a new idea of balancing dataset, which can be applied to more hybrid sampling methods, and combined with different classifiers to verify its robustness, which will be further studied in the future.

Acknowledgements

This work was supported by State Grid Science & Technology Project of SGCC—Intelligent Management and Control of Integrated Operation of Secondary Power System (5700-2020040366A-0-0-00).

References

[1] Habeeb R, Nasaruddin F, Gani A, et al. 2019 Real-time big data processing for anomaly detection: A survey International Journal of Information Management 45 289-307.

[2] Seiffert C, Khoshgoftaar T M, et al. 2008 RUSBoost: Improving classification performance when training data is skewed IEEE International Conference on Pattern Recognition pp 1-4.

[3] Chawla N V, Lazarevic A, Hall L O, et al. 2003 SMOTEBoost: Improving prediction of the minority class in boosting European Conference on Principles of Data Mining and Knowledge Discovery (Berlin Heidelberg: Springer) pp 107-119.

[4] Liu X Y, Wu J and Zhou Z H 2009 Exploratory undersampling for class-imbalance learning IEEE Transactions on Systems Man & Cybernetics Part B 39 (2) 539-550.

[5] Lango M and Stefanowski J 2018 Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data Journal of Intelligent Information Systems 50 (1) 97-127.

[6] Bader-El-Den M, Teitei E and Perry T 2018 Biased random forest for dealing with the class imbalance problem IEEE Transactions on Neural Networks and Learning Systems 30 (7) 2163-2172.
[7] Sun J, Lang J, Fujita H, et al. 2018 Imbalanced enterprise credit evaluation with DTESBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates *Information Sciences* **425** 76-91.

[8] Gao X, et al. 2020 An ensemble imbalanced classification method based on model dynamic selection driven by data partition hybrid sampling *Expert Systems with Applications* **160** 113660.

[9] Ester M, et al. 1996 A density-based algorithm for discovering clusters in large spatial databases with noise *Proc. Int. Conf. Knowledg. Discovery & Data Mining.*

[10] Cover T M, Hart P E, et al. 1967 Nearest neighbour pattern classification *IEEE Transactions in Information Theory* **13** (1) 21-27.

[11] García S, Fernández A, Luengo J, et al. 2010 Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining *Experimental Analysis of Power Information Sciences* **180** (10) 2044-2064.

[12] Li X P, Gao X, Yan B, Chen C X, Chen B, Li J L and Xu J H 2019 An approach of data anomaly detection in power dispatching streaming data based on isolation forest algorithm *Power System Technology* **43** (04) 1447-1456.