Radio Anomaly Signal Recognition Methods Based on Clustering

Shengyang Li¹, Jianxi Lin¹, Bin Tian¹, Li Shen¹, Guorang Cui² and Jing He²*

¹Information Department, Radio Monitoring Center, Kunming, Yunnan, 650228, China
²National Pilot School of Software, Yunnan University, Kunming, Yunnan, 650091, China
*Corresponding author’s e-mail: hejing@ynu.edu.cn

Abstract. Identifying radio anomalies is one of the main purposes of radio monitoring. The current radio anomaly signal identification is mainly finished manually by the radio operators, using professional radio knowledge and their work experience. However, because the anomaly signal is hidden in the “massive” data, accompanied by a large amount of noise, and also data imbalance, the anomaly signal is difficult to find. In this paper, we combine the data unevenness processing method SMOTE and (guorong: cluster detail) to identify the anomaly signal during radio monitoring. Experimental results show that our method can improve the efficiency of existing radio anomaly signal recognition. Moreover, our experiments also show that data imbalance processing plays a key role in anomaly signal recognition.

1. Introduction
The radio spectrum resources are invisible. But it is closely related to human life. For example, people listen to the radio, answer the call, and so on. All these depend on the communications capabilities provided by the radio spectrum. With the rapid development of human demand for radiofrequency resources, radio spectrum resources become a limited and scarce resource. To highlight the importance of radio spectrum management, the Ministry of Industry and Information Technology issued the 'National Radio Plan 2016-2020' [1], which is the first time to introduce that radio spectrum resources have been placed in a strategically important position. Among the task of radio spectrum management, anomaly signal identification is one of the most critical tasks. Anomaly signal identification aims to discover that unlicensed black stations occupy public radio communication channels. Figure 1 shows the radio monitoring data at a certain time point, which its X-axis is the channel, and the Y-axis is the field strength value. Every wave that exceeds low noise is considered a signal. If a signal appears on a spectrum that has not been assigned, we consider the signal to be anomaly. For example, see Figure 1, the signal encircled by the red box is a anomaly signal if the spectrum occupied by the signal has not been allocated. Finding these anomaly signals is of considerable significance since people demand radio frequency resources is intensifying, but the radio spectrum is not inexhaustible. Currently, in the process of radio spectrum monitoring, the identification of anomalous signals is mainly made by professionally trained staff. In this way, the results of the monitoring are subjective, and manual monitoring is time-consuming. Therefore, it is necessary to improve the efficiency of anomaly signal recognition with new technologies.
The availability of big data and the rapid advance of AI technique provide unique opportunities to change the existing radio anomaly signal recognition method. For instance, a variety of machine models such as SVM[2], Gradient-Boost-Tree[3], Random Forest [4] are widely used in Recommend System [5], Pattern Recognizing [6], etc..

Since anomaly signal recognition can be regarded as a dichotomy problem (normal signal or abnormal signal), the critical idea of design a machine-based anomaly recognition method is to choose a model trained by the historical data and then to use this model to predict. However, there are remaining challenges since the real world business is complicated. For example, data imbalance and data can be dirty, etc..

In this paper, we propose an anomaly signal to recognize the method to effectively improve existing methods of radio anomaly recognition efficiency for Yunnan Radio Detection Centre. In our design, spectral field strength data Figure 1 is features extracted to reduce the effects of noise, Unbalanced processed to solve data imbalance problems, and used to train multiple machine learning models.

The main contributions of this paper are as follows:
- We investigated the problems encountered by the Yunnan Radio Monitoring Center and explored the opportunities of improving the efficiency of existing anomaly signal identification methods by using machine learning models. Moreover, in order to train our model, we collect a lot of radio monitoring data.
- We combine feature extraction method, data imbalance method (SMOTE), and multiple cluster method such as Kmeans DBSCAN, SOM etc. to provide a novel anomaly signal recognize method.
- We evaluate the performance of our method by using 4Tb Radio monitoring data. Experiment results show that. In contrast to existing manual identification by skew form Yunnan radio monitoring center method our method provides

The rest of the paper is organized as follows: The anomaly signal recognize method is presented in Sec 3. The experimental evaluation results are reported in Sec 4. we conclude the paper in Sec 5.

2. Related Work
Anomaly signal recognition is one of the most important studied field on radio detection. In this section, we identify the most relevant research thread in this area:

Radio anomaly signal recognition has been done manually by humans for a long time. With artificial intelligence shines in all walks of life, there are a lot of efforts on trying to use machine learning models to improve efficiency of abnormal signal recognition. Bo Li and et al. [7] propose a wireless intelligent analysis method based on fuzzy pattern recognition. The fuzzy clustering analysis method is used to classify and identify C-band anomaly signals, but the calculation is complicated and takes a long time. Yuequn Xiao and el al. [8] applied cyclic bispectrum to signal feature extraction, which effectively improved the recognition rate of radar emitter signals. Y. Lin and J. C. Li [9] propose a neural network classification method for identifying radar signals, which has a good recognition rate. Qiang Li and et al. [10] used the improved BP neural network to identify the radio
anomaly signal and obtained good experimental results. However, for small sample classification, there are often over-fitting and over-learning phenomena. Bo Feng and et al. [11] propose a method for effectively identifying the type of radio anomaly signal, which is based on support vector machine for classification. Xuan Zhang [12] used genetic algorithm to select features and BP neural network to predict detected signals. Her experimental results show that this method has related good recognition effect.

Different from the above-mentioned researches, we consider the data imbalance problem and integrate the SMOTE method into the identification method. Moreover, we tried a variety of machine learning models such as SVM, GDBT, etc. and verified the performance of our recognition method.

3. Identification Methodology of Anomaly Signal
In this section, we discuss the details of our anomaly signal recognize methods design. As shown in Algorithm 1, the identification method of radio anomaly signal needs to go through the following steps: Feature extraction of training data and Clustering methods selection, which will be discussed in Sec 3.1 and Sec 3.2

**Algorithm 1 Radio Anomaly Signal Recognition Methods Based on Clustering**

**Input:** Radio monitoring data (raw)
1. Feature extraction of radio monitoring data (raw);
2. The characteristic data is processed unbalance;
3. Select and evaluate clustering models
   Output analysis results

3.1 Feature extraction of training data
Feature extraction of training data. Since the field strength value belongs to high-dimensional data, it is difficult to be used. And also, there are a lot of noises in the raw data, by mapping the raw data to the feature space, the training data can train the model more effectively. Therefore, reduction the dimensionality of data is necessary. The 16 features and the calculation formula of some features are as follows:
   - Mean Value (Mv)
   - Variance (Var): the degree to which the instantaneous amplitude value of the frequency point deviates from its mean value.
   - The first peak (Fp)
   - The second peak (Sp)
   - The third peak (Tp)
   - The number of continuous points whose amplitude is greater than the bottom noise (Cpn)
   - The average interval of a point whose amplitude which is greater than bottom noise (Cpi)
   - The ratio of the number of peaks to the total number of points (Pr)
   - The variance of the peak point after filtering the threshold (Pev)
   - Frequency point occupancy (Fbo)
   - Zero crossing rate (Zcr): the number of times the signal spectrum graph crosses the threshold line
   - Normalize the mean variance of the absolute value of the instantaneous amplitude (Msd)
   - The kurtosis of instantaneous amplitude is normalized (Kur): a statistical measure used to describe the distribution of instantaneous amplitude of frequency points around its mean value.
   - The standard deviation of the absolute value of the instantaneous amplitude (Sd)

3.2 Clustering methods selection
The following representative machine learning algorithms are selected:
• K-means clustering algorithm is a clustering analysis algorithm that is solved iteratively. Its step is to randomly select k objects as the initial clustering center, then calculate the distance between each object and each seed clustering center, and assign each object to the nearest clustering center. The cluster center and the objects assigned to it represent a cluster. Each time a sample is allocated, the cluster center of the cluster is recalculated according to the existing objects in the cluster. This process is repeated until a termination condition is met.

• DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a representative Density-Based Clustering algorithm. Different from partition and hierarchical clustering method, it defines cluster as the maximum set of points connected with density, and can divide regions with high enough density into clusters, and can find clusters of arbitrary shape in the spatial database of noise.

• Hierarchical Clustering (Hierarchical) is a type of Clustering algorithm that creates a Hierarchical nested cluster tree by calculating the similarity among data points of different categories. Hierarchical methods first calculate the distance between samples. Merge the closest points into the same class each time. Then, the distance between the classes is calculated, and the nearest class is merged into a large class. Keep merging until a class is synthesized.

• The self-organizing mapping (SOM) network clustering the input vectors through the competitive learning of neighborhood elements and weight adjustment. Compared with other clustering algorithms, this algorithm does not need to initialize the clustering center and cluster number.

• Spectral clustering (Spectral) is a kind of graph theory based clustering method. By clustering the eigenvectors of the Laplace matrix of sample data, the original clustering of sample data can be achieved. Spectral clustering can be understood as mapping the data of high-dimensional space to low-dimensional space, and then clustering in low-dimensional space with other clustering algorithms (such as KMeans).

4. Experiments
In this section, we evaluate the identification performance of different clustering methods.

4.1 Setup
Before giving an analysis detail, we firstly describe the experimental setup, including the data sets and metrics used throughout the measurements.

4.1.1 Data sets

Figure 2. fixed-monitoring-station

Figure 3. Moving-monitoring-station

A 4TB data set collected from different kinds of radio monitoring equipment (see Figure 4) is used in our experiments. This data set consists of a series of records, and each record is the intensity of the field at a time point (see Figure 4, the Y-axis is the field intensity values, the X-axis is the Frequency spectrum, and the Z-axis is the time points).
As we mentioned before, the radio anomaly signal identification mainly determines whether there is a black station illegally occupying the common channel at the current time. Since the normal signal will only transmit signals in the frequency band assigned to it for a specified time, the normal signals have a very high similarity (see Figure 4). We manually label each record as a training data training classifier based on historical radio monitoring data. The monitoring data of a certain day is selected as the test data. The effectiveness is measured by predict accuracy. Therefore, the high accuracy is the more effective the classifier is.

4.1.2 Measurement metrics
The effectiveness of models is measured by the Purity, Silhouette Coefficient (SC), Calinski Harabasz (CH). Formally, these metrics can be calculated as follows:

- Purity:
\[ \text{Purity} = \sum_{m} \frac{m}{m} P_i \]  

(1)

Where, \( P_i = \max(P_j) \) (\( P_j \) refers to the probability that members in cluster \( i \) belong to class \( j \)), \( K \) is the number of clusters, and \( m \) is the number of members involved in the whole cluster division.

- Silhouette Coefficient (SC):

\[ SC(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \]  

(2)

Where, \( a(i) \) is the average distance between sample \( i \) and other samples in the same cluster, and \( b(i) \) is the minimum of the average distance between sample and all samples in other clusters.

- Calinski Harabasz (CH)

By comparing the traces of the gaps between classes and the traces of the gaps within classes, the judgment algorithm was found that the larger the CH was, the closer the class was, and the dispersion between classes was. The formula was as follows:

\[ CH(k) = \frac{\text{tr}(B_k)m - k}{\text{tr}(W_k)k - 1} \]  

(3)

Where \( m \) is the number of training samples, \( k \) is the number of categories, \( B_k \) is the covariance matrix between classes, \( W_k \) is the covariance matrix of data within categories, and \( \text{tr}(\cdot) \) is the trace of the matrix.

4.2 Analysis

Table 1. Evaluations

| Method   | Purity (%) | SC   | CH              |
|----------|------------|------|-----------------|
| KMeans   | 93.375     | 0.666| 25696.108       |
| DBSCAN   | 57.68      | -0.323| 9.763          |
| Hierarchical | 93.65   | 0.665| 25487.179       |
| SOM      | 93.663     | 0.743| 47598.037       |
| Spectral | 93.837     | 0.664| 25462.225       |

There are 19907 normal signal records and 2639 anomaly signal records among 22546 records of Radio monitoring data. As shown in Table 1, the cluster purity, contour coefficient and CH of DBSCAN clustering algorithm are the worst in all comparison clustering algorithms, indicating that the data is not applicable to DBSCAN algorithm. The clustering effect of self-organizing mapping network is the best in all comparison clustering algorithms. It depends on the self-organizing adaptive characteristics of SOM itself, and solves the problem that clustering algorithm is sensitive to clustering center initialization. The experimental results of the KMeans algorithm, the Spectral algorithm, and the Hierarchical algorithm are close, which is determined by the characteristics of the distribution of the data itself.

5. Conclusions

Aiming at the present radio anomaly signal recognition service, we design the radio anomaly signal recognition algorithm. The experimental results show that the algorithm can recognize most normal signals and radio anomaly signals effectively. Moreover the algorithm can effectively reduce the workload of manual identification of radio anomaly signals.

References

[1] Ministry of Industry and Information Technology of People’s Republic of China, National radio management program 2016-2020.
http://www.miit.gov.cn/n1146285/n1146352/n3054355/n3057267/n3057272/c5222986/content.html

[2] Vedran Kirinic, Ervin Ceperic, Sasa Vlahinic, and Jonatan Lerga. Support vector machine state estimation. Applied Artificial Intelligence, 33(6):517-530, 2019.

[3] Zeyi Wen, Bingsheng He, Ramamohanarao Kotagiri, Shengliang Lu, and Jiashuai Shi. Efficient gradient boosted decision tree training on gpus. In 2018 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2018, Vancouver, BC, Canada, May 21-25, 2018, pages 234-243, 2018.

[4] Hanwu Luo, Xiubao Pan, Qingshun Wang, Shasha Ye, and Ying Qian. Logistic regression and random forest for effective imbalanced classification. In 43rd IEEE Annual Computer Software and Applications Conference, COMPSAC 2019, Milwaukee, WI, USA, July 15-19, 2019, Volume 1, pages 916-917, 2019.

[5] Rafael Gomes Mantovani, Andre L. D. Rossi, Edesio Alcobaca, Joaquin Vanschoren, and Andre C. P. L. F. de Carvalho. A meta-learning recommender system for hyperparameter tuning: Predicting when tuning improves SVM classifiers. Inf. Sci., 501:193-221, 2019.

[6] Joanna Wisniewska and Marek Sawerwain. Recognizing the pattern of binary hermitian matrices by quantum knn and SVM methods. Vietnam J. Computer Science, 5(3-4):197-204, 2018.

[7] Bo Li. C-band radio signal intelligence based on fuzzy pattern recognition analysis.

[8] Yuequn Xiao, Yongjun Zhao, and Jiandong Zhu. A new method for identifying radar emitter signals.

[9] Y. Lin and J. C. Li. Radar signal recognition algorithms based on neural network and grey relation theory.

[10] Xuan Zhou Qiang Li, Zheng Pei. The research of radio signal classification based on improved bp neural network.

[11] Bo Feng, Zheng Fei, Yinliang Zhong, and Fangli Ma. C-band radio anomaly signal recognition based on support vector machine.

[12] Xuan Zhang. Feature selection of radio abnormal signal based on genetic algorithm.

[13] Yuan-Ting Yan, Ruiqing Liu, Zihan Ding, Xiuquan Du, Jie Chen, and Yanping Zhang. A parameter-free cleaning method for SMOTE in imbalanced classification. IEEE Access, 7:23537-23548, 2019.

[14] Huan Zhang, Liangxiao Jiang, and Liangjun Yu. Class-specific attribute value weighting for naive bayers. Inf. Sci., 508:260-274, 2020.

[15] Sriparna Saha, Shreyasi Datta, and Amit Konar. Decision tree based single person gesture recognition. In Computational Intelligence for Semantic Knowledge Management - New Perspectives for Designing and Organizing Information Systems, pages 71-89. 2020.