User Abnormal Viewing Behavior Detection Method Based On Canopy-FCM and Improved Isolation Forest Algorithm

Yu Wu¹*, Huiyong Liu¹

¹School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing, 100876, China
*Corresponding author’s e-mail: nothingtosay@bupt.edu.cn

Abstract. In order to analyze the users' viewing behavior more accurately, allow IPTV service providers to provide better services to users, and improve the user's viewing experience, this paper proposes a detection method of abnormal user viewing behavior based on Canopy-FCM and improved isolation forest algorithm. First, the collected viewership data is preprocessed and stored with the user viewing behavior matrix, then the Canopy-FCM clustering algorithm is used to classify users with the same rating behavior. Then, an improved isolated forest algorithm is used to perform anomaly detection and analysis on the users classified above, and finally screen out abnormal user viewing behaviors. Finally, NMI indicators and ROC curves are used to evaluate the performance of the proposed algorithm to verify the feasibility of the proposed anomaly detection method. The results show that the method proposed in this paper has a high precision and recall rate for detecting users with abnormal ratings.

1. Research Background

With the rapid development of IPTV services, the importance of accurately understanding users' viewing behaviors and in-depth exploration of users' potential needs has become increasingly prominent. Conducting behavior analysis on viewing users will not only help service providers adjust IPTV service scheduling arrangements and improve user experience, but also provide service providers with decision support, such as advertising pushes and video pushes. Therefore, it is of great significance to accurately analyze the viewing behavior of users. However, due to the huge amount of viewership data collected every day, abnormal viewership data will inevitably occur from the generation to the collection of the viewership data. Therefore, in order to analyze users’ viewing behaviors more accurately, allow IPTV service providers to provide better services to users, and improve users’ viewing experience, the accuracy, reliability and integrity of the viewership data should be guaranteed. Anomaly detection is particularly important[1][2].

Anomalous data detection is an important data mining technology, which mainly obtains hidden valuable information by mining a small number of abnormal points in the data set. The use of an effective abnormality detection method can detect and eliminate abnormal viewing data in the user viewership data in time, thereby improving the accuracy of the viewership data[3][4].

Based on the above problems, this paper proposes a user abnormal viewing behavior detection method based on Canopy-FCM and improved isolated forest algorithm. Firstly, the collected viewership data is preprocessed and stored using the viewing behavior matrix, then the fuzzy clustering algorithm is used to classify users with the same viewing behavior. Then the isolated forest algorithm is used to perform anomaly detection and analysis on the above-mentioned viewership data.
for each category, and finally the ROC curve is used to do a comparative test on the proposed algorithm to test the accuracy and superiority of the detection method proposed in this paper.

2. Data preprocessing

2.1. User Viewing Behavior Matrix
Regarding the viewing situation of a certain user in one day, the following 7 indicators are comprehensively considered: the number of minutes of live streaming per hour, the number of minutes of on-demand viewing per hour, the number of minutes of reviewing per hour, the number of minutes of viewing ads per hour, the number of channel changes per hour, the number of exits in advertisements per hour and the number of exits in normal programs per hour, using 7 rows and 24 columns of ratings behavior matrix for storage, as shown in Figure 1.

24 time periods

\[
\begin{bmatrix}
X_{1,1} & X_{1,2} & X_{1,3} & \ldots & X_{1,24} \\
X_{2,1} & X_{2,2} & X_{2,3} & \ldots & X_{2,24} \\
X_{3,1} & X_{3,2} & X_{3,3} & \ldots & X_{3,24} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
X_{7,1} & X_{7,2} & X_{7,3} & \ldots & X_{7,24}
\end{bmatrix}
\]

7 indicators

Figure 1. User rating behavior matrix

The distance between the two viewing behavior matrices is defined as equation (1).

\[
d(x, y) = \sum_{i=1}^{C} \sum_{j=1}^{\epsilon} \sqrt{(x_{ij} - y_{ij})^2}
\]  

(1)

2.2. Data standardization
Considering that each indicator of the ratings data has a different order of magnitude, in order to avoid the larger order of magnitude being dominant, this paper will standardize the data stored in the rating behavior matrix, and the processed data will be between 0 and 1. The specific formula is shown as equation (2)[5].

\[
x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}
\]  

(2)

Where \(x'\) is the standardized viewing behavior index data, \(x\) is the original viewing behavior index data, \(x_{\max}\) and \(x_{\min}\) are the maximum and minimum values of the original viewing behavior index data before standardization.

3. Construction of anomalous viewing behavior detection model
The specific architecture of user abnormal viewing behavior detection based on Canopy-FCM and improved isolated forest algorithm is shown in Figure 2. The entire model includes three parts: viewing data preprocessing, viewing data cluster analysis, and viewing data abnormality detection.
Collect Viewership Data → Data Preprocessing → Viewing Behavior Matrix → Data Standardization

Canopy Algorithm → FCM Algorithm → Get the Classification Results of Viewership Data

Isolated Forest Algorithm Training → Isolated Forest Algorithm Testing → Anomaly Scores and Analyze Anomalous

Figure 2. Specific architecture for user abnormal viewing behavior detection based on Canopy-FCM and improved isolated forest algorithm

4. Cluster analysis of viewership data

4.1. FCM clustering algorithm

Because different types of viewing users have different viewing behaviors, and abnormal viewing behaviors in one category may be similar to normal viewing behaviors in another category. In order to avoid this situation and improve the accuracy of anomaly detection, considering that FCM is a flexible fuzzy division, this paper uses FCM clustering algorithm to classify users with the same viewing behavior. [5] As a kind of unsupervised algorithm based on partition, FCM algorithm is an improvement of hard partition C-means algorithm. Its core idea is to continuously update the membership matrix and clustering center until the objective function becomes stable, and output the final clustering result.[6]

Where the cluster center update formula is shown as equation (2).

$$v_i = \frac{\sum_{j=1}^{n} (u_{ij})^m x_j}{\sum_{j=1}^{n} (u_{ij})^m}, 1 \leq i \leq c$$

Where $$u_{ij}$$ represents the membership degree of $$x_i$$ belonging to the cluster center $$c_j$$, $$x_i$$ represents the $$i$$-th sample, $$c_j$$ represents the center of the $$j$$-th cluster, $$m$$ is the fuzzy weighted index and $$m \geq 1$$.

The update formula of the membership matrix is shown as equation (4).

$$u_{ij} = \left[ \sum_{k=1}^{c} \left( \frac{\| x_j - v_i \|^2}{\| x_j - v_k \|^2} \right)^{m-1} \right]^{-1}, 1 \leq i \leq c, 1 \leq j \leq n$$

The objective function is defined as equation (5).
4.2. Canopy-FCM algorithm

Since the search direction of the gradient method used by the FCM algorithm is always along the direction of energy reduction, the algorithm itself is sensitive to the initial value. When the initial value is selected near a minimum point of the objective function, the algorithm is easy to fall into the clustering result represented by the minimum point and generate a local optimal solution. Therefore, in view of the above problems of the FCM algorithm, the Canopy algorithm and the FCM algorithm are combined. First, an approximate number of clusters and initial cluster centers are obtained through Canopy clustering, and the obtained results are passed to the FCM algorithm for more efficient clustering. The fusion algorithm combines the advantages of the Canopy and FCM algorithms, and makes up for the shortcomings between each other, can obtain more accurate clustering results in a shorter time, and reduce the number of iterations of the algorithm.[7][8]

The specific process of canopy algorithm is as follows.

Input: data set D, parameters T1, T2 (T2<T1)
Output: cluster number c, cluster center set

a. A data object P is randomly deleted from the data set D as a new canopy initial center
b. For the remaining objects in D, if the distance from P is less than T1, it is assigned to canopy where P is located; if the distance from point P is less than T2, it is deleted from D (calculate canopy)
c. If D is not empty, jump to a
d. Output current canopy number and canopy initial center set

In addition, in order to improve the stability and accuracy of the FCM algorithm in the next step, it is necessary to detect the number of data objects in each canopy cluster. If a canopy cluster has only one or few data objects, it is considered that the canopy cluster is isolated. They should be deleted from the total canopy cluster to get the C value and initial cluster center of the FCM clustering process.

5. Anomaly detection of viewership data

5.1. Isolated Forest Algorithm

The isolated forest[9] algorithm is an unsupervised learning algorithm proposed by Fei Tony Liu and Zhou Zhihua in 2008 to determine whether there are abnormal points in the data set. Due to the high dimensionality and the large amount of viewership data, if an anomaly detection method based on density or distance is used, it will cause problems such as a large amount of calculation and long running time. The metric used by the isolation forest algorithm does not depend on the calculation of distance and density. It is an anomaly detection algorithm based on partition and integrated learning. The design of the isolated forest algorithm uses two notable characteristics of abnormal data: few and special.

The anomaly detection using the isolated forest algorithm is divided into two phases: training phase and testing phase. The training phase uses the sub-samples of the training set to construct the isolation tree, and the test phase passes the test instances through the isolation tree to obtain the anomaly score of each test sample.

In the training phase, the algorithm flow of using the sub-samples of the training set to construct the isolation tree is as follows.

Input: Training data set D
Output: An isolated tree

a. Randomly select \( \Psi \) points from the training data set as sub-samples and put them into the root node of
an isolated tree
b. Randomly generate a cutting point p (the cutting point is generated between the maximum and minimum values in the current node data)
c. A hyperplane is formed from the tangent point, and the current data space is divided into 2 subspaces. Points smaller than p are placed on the left branch of the current node, and points larger than p are placed on the right branch of the current node.
d. Repeat b and c to construct new leaf nodes until the data can no longer be split or the number of splits reaches \( \log_2 \psi \), then stop splitting.

Since the isolated tree and the binary search tree have an equivalent structure, the isolated forest algorithm uses the expected average length of the binary search tree to estimate the average length of the isolated tree. As shown in equation (6).

\[
C(n) = 2H(n-1) - \frac{2(n-1)}{n}
\]  

(6)

Among them, \( H(i) \) is the harmonic number, \( n \) is the number of leaf nodes, and \( C(n) \) is the average value of the path length when \( n \) is given, so the isolated forest algorithm uses it to normalize each path length.

The anomaly score is defined as equation (7).

\[
S(x, n) = 2 \frac{E(h(x))}{C(n)}
\]  

(7)

Where \( E(h(x)) \) is the average value of the output distance of each isolated tree in the isolated forest. Analyzing by formula, the following conditions can be used for abnormal analysis:

1) If S is very close to 1, the sample is very likely to be abnormal
2) If the abnormality analysis is far less than 0.5, the possibility that the abnormality is abnormal is very small

5.2. Selection of cut point for isolated forest

The key to detection in the isolated forest algorithm is the construction of the separation tree, and the key to the construction of the separation tree is the choice of cutting points. The cut point of the separation tree in the original isolated forest algorithm is selected by determining the range of the data, and randomly selecting a data value within the range as the cut point, so as to determine the left and right subtrees of the separation tree and complete the tree construction, which will affect the accuracy of anomaly detection. The choice of the cutting point in the separation tree determines the position of the target point on the separation tree. Therefore, this paper no longer chooses the random value method to determine the cutting point vaguely. Instead, through filtering, a better cut point is selected to construct a separation tree.

First propose the evaluation index of the separation tree cutting point. As shown in equation (8)

\[
ass(X, v) = \frac{dist(X, v)}{disp(X, v)}
\]  

(8)

Where X is the selected sub-sample set for constructing the separation tree, and v is the cutting point.

\( dist(X, v) \) indicates the degree of data difference on both sides of the cut point. The larger the value, the greater the data difference. It is defined as equation (9).

\[
dist(X, v) = |E(x \mid x \in X, x < v) - E(x \mid x \in X, x > v)|
\]  

(9)

\( disp(X, v) \) indicates the degree of dispersion of the data on both sides of the tangent point, the larger the value, the more concentrated. It is defined as equation (10).

\[
disp(X, v) = D(x \mid x \in X, x < v) + D(x \mid x \in X, x > v)
\]  

(10)

Therefore, the larger the value \( ass(X, v) \), the better choice of cut point.
The cutting point selection algorithm flow is as follows.

Input: Training data set D
Output: A cut point

1. Determine the median \( x_m \) of the data in the subsample, calculate \( ass(X, x_m) \)
2. Calculate the intermediate number \( x_l \) from \( x_i \) to \( x_m \), and the intermediate number \( x_r \) from \( x_i \) to \( x_m \), calculate \( ass(X, x_l) \) and \( ass(X, x_r) \)
3. If \( ass(X, x_l) > ass(X, x_r) \), let \( x_r = x_m \); otherwise let \( x_l = x_m \)
4. Repeat 1-3 until one of the following is met, output \( x_m \)
   a. no more data points between \( x_l \) and \( x_r \)
   b. both \( ass(X, x_l) \) and \( ass(X, x_r) \) are less than \( ass(X, x_m) \)

6. Result analysis

6.1. Analysis of cluster analysis algorithm results

In order to prove the stability and effectiveness of the Canopy-FCM algorithm proposed in this paper, this paper selects several public data sets from the UCI machine learning library for experiments, and selects the FCM algorithm and the Canopy-FCM algorithm proposed in this paper for comparative experiments. The description of the data set used is shown in Table 1.

| Data set name | Number of samples | Number of attributes | Number of clusters |
|---------------|-------------------|----------------------|-------------------|
| Iris          | 150               | 4                    | 3                 |
| Wine          | 178               | 13                   | 3                 |
| Seeds         | 210               | 7                    | 3                 |
| Wave          | 5000              | 21                   | 3                 |

In order to evaluate the clustering effect of the clustering algorithm, this paper uses the two most commonly used evaluation indicators—accuracy and Normalized Mutual Information[10] to compare the experimental results.

The above data sets are distributed on the traditional FCM algorithm, K-means algorithm and Canopy-FCM algorithm in this paper to conduct experiments, and analyze the clustering accuracy of these algorithms on different data sets and the results of NMI indicators.

The experimental results of the Canopy-FCM algorithm and the comparison algorithm on each data set are shown in Table 2.

| Data set name | FCM     | K-means | Canopy-FCM |
|---------------|---------|---------|------------|
| Iris          | 0.8912  | 0.8671  | 0.8923     |
| Wine          | 0.7645  | 0.7765  | 0.8021     |
| Seeds         | 0.9021  | 0.8927  | 0.8872     |
| Wave          | 0.4521  | 0.4762  | 0.6124     |

Analysis and comparison show that in the selected data sets, the Canopy-FCM algorithm proposed in this paper has a higher accuracy rate than other traditional FCM algorithms and K-means algorithms.

In this paper, the experimental results of the Canopy-FCM algorithm and the comparison algorithm on the NMI indicators on each data set are shown in Table 3.
Table 3. NMI indicator comparison

| Data set name | FCM   | K-means | Canopy-FCM |
|---------------|-------|---------|------------|
| Iris          | 0.7321| 0.7641  | 0.7762     |
| Wine          | 0.4215| 0.4182  | 0.4781     |
| Seeds         | 0.6719| 0.7139  | 0.6652     |
| Wave          | 0.3124| 0.3372  | 0.3651     |

It can be seen from the analysis that in the selected data, in addition to the Seeds data set, the Canopy-FCM algorithm proposed in this paper has a higher NMI compared to other comparison algorithms. On the Seeds dataset, the NMI value of this algorithm is slightly worse than other algorithms.

Run 10 experiments on the data set in the Canopy-FCM algorithm and other algorithms in this paper, and compare the average number of iterations. The experimental results are shown in Table 4.

Table 4. Comparison of average number of iterations

| Data set name | FCM   | K-means | Canopy-FCM |
|---------------|-------|---------|------------|
| Iris          | 32    | 34      | 27         |
| Wine          | 38    | 40      | 29         |
| Seeds         | 44    | 49      | 33         |
| Wave          | 38    | 40      | 31         |

It can be seen from the table that the Canopy-FCM algorithm in this paper effectively reduces the number of iterations of the algorithm compared with the traditional FCM algorithm and K-means algorithm, and has been well verified in the data set selected in this paper.

6.2. Analysis of anomaly detection algorithm results

In order to prove the stability and effectiveness of the new algorithm proposed in this paper to improve the isolated forest algorithm, this paper uses the viewership data of 9752 users in a county in August 2020 as the research object, and evaluates the anomaly detection model through the ROC curve [11] and AUC.

AUC refers to the area under the ROC curve. Under normal circumstances, the value of AUC ranges from 0.5 to 1, and the larger the value of AUC, the better the performance of the model. The ROC curve obtained is shown in Figure 3, AUC=0.8827, which is relatively close to 1, and is at a relatively ideal level. In addition, this paper uses the traditional iForest and the improved iForest to conduct multiple data mining, and compares the average AUC of the results of the two algorithms. The results are shown in Table 5. The analysis shows that the AUC value obtained by the improved iForest algorithm is larger, which can more accurately detect the abnormality of the viewers.
Table 5. Comparison of AUC values of the two algorithms

| algorithm    | AUC average |
|--------------|-------------|
| iForest      | 0.8241      |
| Improved iForest | 0.8827    |

7. Conclusion
In order to analyze the users' viewing behavior more accurately, allow IPTV service providers to provide better services to users, and improve the user's viewing experience, this paper proposes a detection method of abnormal user viewing behavior based on Canopy-FCM and improved isolation forest algorithm. The accuracy rate, Normalized Mutual Information and the number of iterations are used to evaluate the clustering algorithm. The ROC curve and AUC value are used to compare the improved iForest algorithm. The experimental results show that the method proposed in this paper has high accuracy and recall rate for detecting users with abnormal ratings, which verifies the feasibility of the detection method proposed in this paper.

Acknowledgments
This work is supported by Beijing Wisdom Information Security Software Technology Co., Ltd.

References
[1] Zhang, B. (2020) Analysis of IPTV Users' Viewing Behavior Based on Time Correlation. China Science and Technology Information, 6:62-63.
[2] Zhang, Y., Chen, W., Zha, W., Gu, X. (2015) A Time-Topic Coupled LDA Model for IPTV User Behaviors. IEEE Transactions on Broadcasting, 61:56-65.
[3] Sun, Y., Li, S.H., Cui, C., Li, B., Chen, S.S., Cui, G.Y. (2018) An improved method for detecting outliers in power user data based on Gaussian kernel function. Power System Technology, 42:1596-1604.
[4] Jing, Y., Qi, Q., Wang, J., Feng, T., Liao, J. (2019) ALSR: An Adaptive Label Screening and Relearning Approach for Anomaly Detection. In: 2019 IEEE Symposium on Computers and Communications (ISCC). Barcelona. Spain. pp. 1-6.
[5] Zhao, M., Li, Y.N., Li, C., Yang, L. (2020) Anomaly detection of power consumption data based on fuzzy clustering and isolated forest. Journal of Shanxi University of Technology(Natural Science Edition), 36:38-43
[6] Yih, J. (2010) FCM algorithm based on Normalized Mahalanobis distances in image clustering. In: 2010 International Conference on Machine Learning and Cybernetics. Qingdao. pp. 2724-2729.
[7] Yu, Q., Ding, Z. (2016) Improved Canopy-FCM algorithm based on MapReduce. In: 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Datong. pp. 1975-1979.
[8] Zhou, J.G. (2018) Research on optimization and improvement of Canopy-Kmeans clustering algorithm based on Hadoop platform. Journal of Anhui Radio and TV University, 4:117-128
[9] Zhang, X. et al. (2017) LSHiForest: A Generic Framework for Fast Tree Isolation Based Ensemble Anomaly Analysis. In: 2017 IEEE 33rd International Conference on Data Engineering (ICDE). San Diego. CA. pp. 983-994
[10] Amelio, A., Pizzuti, C. (2015) Is normalized mutual information a fair measure for comparing community detection methods? In: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Paris. pp. 1584-1585
[11] He, X., Gallas, B.D., Frey, E.C. (2010) Three-Class ROC Analysis—Toward a General Decision Theoretic Solution. IEEE Transactions on Medical Imaging, 29:206-215