Journal of Computers Vol. 28, No. 6, 2017, pp. 216-224
doi:10.3966/199115992017122806019

IPTV User's Complaint Prediction Based on the Gaussian Mixture Model for Imbalanced Dataset

Xin Wei[1*], Zhilin Li[1], Ronghua Liu[1] and Liang Zhou[1]

[1] College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications,
Nanjing, China, 210003
{xwei, 1015010607, 1015010608, liang.zhou}@njupt.edu.cn

Received 22 December 2016; Revised 14 January 2017; Accepted 14 February 2017

Abstract. IPTV user's experience is vital for operators to continually enhance their quality of content service and transmission. User's complaint is closely related to user's quality of experience (QoE). Predicting user's potential complaint in time is necessary. However, the happening probability of complaint is far less than that of the normal circumstances, leading to the imbalanced dataset. In order to handle this issue, an over-sampling method based on the Gaussian mixture model (GMM) is proposed. Specifically, GMM is adopted to describe the distribution of limited complaint samples. After estimated the parameters in this model, new minority class samples can be generated, which is more representative than the traditional Synthetic Minority Oversampling Technique (SMOTE). Then the Naïve Bayes classifier is used for finishing classification and prediction. Experimental results show that the proposed algorithm performs better than the competing algorithms in predicting user's complaint.

Keywords: GMM, imbalanced IPTV dataset, Naïve Bayes classifier, over-sampling

1 Introduction

Nowadays, more and more users enjoy the Internet Protocol Television (IPTV) service at home [1-2]. The IPTV operators try their best to support high-quality programs and smooth video streaming transmission in order to guarantee the excellent user's experience. To handle this issue, a large number of researchers are committed to improving user's feelings and satisfaction by finding the critical factors impacting the Quality of Experience (QoE) [3]. It does not only consider the traditional Quality of Service (QoS) parameters, such as bandwidth, packet loss, delay, but also includes factors reflecting user's viewing behaviors, such as number of switching programs, viewing duration [4]. Lin, Hu and Kong [5] provide a model of QoE for video streaming, propose an evaluation method of QoE based on stochastic model and indicate the directions of future research. Several factors influencing QoE are discussed in details [6-7]. The QoE-QoS relation has been modeled and Mean Opinion Score has been calculated to evaluate the IPTV service quality[8]. Truong, Hung and Thanh analyze the effects of multiple QoE indicators: packet reorder, jitter and packet loss.

As we known, when a user's experience is bad or declines, he may fill a complaint to the IPTV operators [5]. In other word, user's complaint is closely related to the QoE. If IPTV operators accurately and efficiently predict user's complaining behaviors in advance, they can promptly take measures to find the faults in the IPTV networks. Therefore, user's complaint prediction is vital to IPTV operators. Existing studies mainly utilize the collected Key Performance Indicators (KPI) to predict whether the user will make complaint [10-12].

In reality, users that file complaints account for a relatively small proportion of overall users. Therefore, one of the key problems in the user's complaint prediction is that the dataset inevitably becomes imbalanced. Imbalanced dataset refers to the dataset that one class of the data is represented by

* Corresponding Author

significantly more number of samples than the others [13]. Here the number of user's complaint is far less than that of non-complaint. For this binary classification problem, the class with more samples, non-complaint circumstance, is called the majority class while the other class, user's complaint, is called the minority class [14]. When traditional classification algorithms are used for processing the imbalanced dataset, it is common to obtain a biased classifier which has the high correct rate for the majority class and the low correct rate for the minority class [15]. The biased classifier has poor overall performance. Methods adopted to process the imbalanced dataset can be categorized into two main types: sampling-based methods by reconstructing the distributions of dataset from imbalanced into balanced and cost-sensitive based algorithms by changing the costs of misclassified minority examples. We propose an integrated algorithm based on Kmeans-synthetic minority over-sampling (SMOTE) technique to balance the distribution of dataset and build QoE model [11]. We use the cost-sensitive methods to build QoE model in imbalanced dataset [12]. We improve the Adaboost by adding cost coefficients for raising the cost of error classification on minority class.

From above analysis, most existing algorithms handling imbalanced dataset by generating new minority class samples directly from existing samples. They don't grasp the property of the minority class samples. If we can effectively obtain the distribution of minority class samples, the generated new samples will be more representative and can more appropriate for describing the property of minority class. It may finally be convenient for the subsequent classifiers to predict user's complaint. Gaussian mixture model (GMM) is a statistical tool for modeling probability distribution of real data sets. Due to its benefits from analytical tractability and universal approximate capacity for continuous probability density functions, it has been widely used in several domains [16-18]. Here we use the GMM to describe the distributions of the minority class samples. After performing the proposed estimation and generation algorithms, the property of original minority class samples is learned and the new minority class samples can be obtained.

The rest of the paper is organized as follows. In Section II, we provide a review of the synthetic oversampling method, SMOTE. In Section III, we propose our algorithm based on GMM to debate the problem of the imbalanced dataset. In Section IV, we provide and discuss experimental results. In Section V, we give conclusion.

2 The Synthetic Minority Over-sampling Technique

SMOTE is a powerful method that has been successfully performed in IPTV user's complaint prediction. First of all, we define subsets $S_{\min} \subset S$ and $S_{maj} \subset S$, where $S_{\min}$ is the set of user's complaint (minority class) examples in the imbalanced dataset S, and S_{maj} is the set of non-complaint (majority class) examples in S. $S_{\min} \cap S_{maj} = \{\phi\}$ and $S_{\min} \cup S_{maj} = \{S\}$. Specifically, consider the K-nearest neighbors for each minority class example. The K-nearest neighbors are defined as the K elements of $S_{\min}$ whose Euclidian distance between itself and x_i under consideration exhibits the smallest magnitude along the n-dimensions of feature space. To create a synthetic sample, we randomly select one of the K-nearest neighbors, then multiply the corresponding feature vector difference with a random number between [0,1], and finally, add this vector to x_i

$$x_{new} = x_i + \left(\tilde{x}_i - x_i\right) \times \delta \tag{1}$$

where $x_i \in S_{\min}$ is the minority instance under consideration, $\tilde{x}_i$ is one of the K-nearest neighbors for $x_i : \tilde{x}_i \in S_{\min}$, and $\delta \in [0,1]$ is a random number. Therefore, the resulting synthetic instance according to (1) is a point along the line segment joining x_i under consideration and the randomly selected K-nearest neighbor $\tilde{x}_i$.

3 The Proposed GMM-Naïve Bayes Algorithm

Different from the traditional SMOTE algorithm, here GMM is selected to describe the distribution of minority class samples. After estimation, we can use the learned GMM to generate new minority class

samples. These procedures change the dataset from imbalanced to balanced. Finally, the Naïve Bayes classifier is used to process the balanced dataset, finishing training and prediction.

3.1 Over-sampling Algorithm Based on GMM

Select GMM for describing distribution of minority class samples. Gaussian mixture model (GMM) can be seen as a linear superposition of multiple Gaussian components, which can provide a better probability model than Gaussian distribution alone. The probability density function of the GMM can be expressed by (2):

$$p(x_i) = \sum_{k=1}^{K} p(x_i, z_i = k) = \sum_{k=1}^{K} \pi_k N(x_i \mid \mu_k, \Sigma_k) . \quad (2)$$

GMM is composed of K Gaussian distribution, each Gaussian distribution called a "Component". These components are superimposed together to form a mixed Gaussian model. In order to make the expression and the subsequent estimation more convenient, a random binary variable z_i is introduced, which is a 1-of-K representation. $z_i = k$ denotes x_i coming from the kth component. Moreover, the joint probability distribution $p(x_i, z_i)$ is defined according to the edge probability distribution $p(z_i)$ of z_i and the conditional probability distribution $p(x_i \mid z_i)$. π_k in (2) can be expressed as:

$$p(z_i = k) = \pi_k . \quad (3)$$

where the parameter π_k must satisfy the condition: $0 \le \pi_k \le 1$ and $\sum_{k=1}^{K} \pi_k = 1$.

Similarly, for a given value of z_i , the conditional probability distribution of x_i can be expressed as a Gaussian distribution:

$$p(x_i \mid z_i = k) = N(x_i \mid \mu_k, \Sigma_k) . \quad (4)$$

Let us use $\gamma(i,k)$ denote the probability of z_k for a given x_i, $p(z_i = k \mid x_i)$. Its value can be obtained from the Bayes theorem:

$$\gamma(i,k) = \frac{p(z_i = k) p(x_i \mid z_i = k)}{\sum_{k'=1}^{K} p(z_i = k') p(x_i \mid z_i = k')} = \frac{\pi_k N(x_i \mid \mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} N(x_i \mid \mu_{k'}, \Sigma_{k'})} . \quad (5)$$

$\gamma(i,k)$ can be seen as the probability of the data x_i in the component k, so we can divide the x_i into the category with the greatest probability according to the probability of x_i in K components.

Estimate parameters in GMM from the original minority class samples. Here, suppose $x_i \in S_{\min}$ obey a mixed Gaussian distribution (denoted by $p(\mathrm{X})$). We need to determine the mixing coefficient π_k, the mean μ_k and the covariance Σ_k of each component. We need to find a set of parameters so that the probability of the model to generate these given data points is maximal, which we can express as $\prod_{i=1}^{N} p(x_i)$, where N denotes the number of samples in $S_{\min}$. From (2), the log-likelihood function of the GMM is shown in the following:

$$\ln P(\mathrm{X}) = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_i \mid \mu_k, \Sigma_k) \right\} . \quad (6)$$

A good tool for maximum likelihood estimation of mixed models is the Expectation Maximization (EM) algorithm [18]. It is an iterative algorithm for maximum likelihood estimation when the observed data are incomplete data, which greatly reduces the computational complexity of maximum likelihood estimation.

For the Gaussian Mixture Model, in order to solve the mean μ_k of the Gaussian component, we let the partial derivative of (6) for μ_k equal to zero, yielding the following equation:

$$\sum_{i=1}^{N} \frac{\pi_k N(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i \mid \mu_j, \Sigma_j)} \Sigma_k^{-1}(x_i - \mu_k) = 0 . \tag{7}$$

Therefore,

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k) x_i . \tag{8}$$

where $N_k = \sum_{i=1}^{N} \gamma(i,k)$, we can regard N_k as the number of data x_i is assigned to component k.
Similarly, in order to solve the covariance Σ_k, we let the partial derivative of (5) for Σ_k equal to zero, yielding the covariance:

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k)(x_i - \mu_k)(x_i - \mu_k)^T . \tag{9}$$

Finally, after solving the mean and covariance, we need to calculate the mixing coefficient π_k, taking $\sum_{k=1}^{K} \pi_k = 1$ into account, requiring that the sum of the mixing coefficients is equal to 1. So we use the Lagrange multiplier method to maximize the following expression:

$$\ln p(\mathrm{X}) + \lambda(\sum_{k=1}^{K} \pi_k - 1) . \tag{10}$$

Seeking partial derivative of the above formula for π_k:

$$\sum_{i=1}^{N} \frac{N(x_i \mid \mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_j N(x_i \mid \mu_{k'}, \Sigma_{k'})} + \lambda = 0 . \tag{11}$$

$$\pi_k = \frac{N_k}{N} . \tag{12}$$

The procedure of parameter estimation by the EM algorithm is summarized in Table 1.

Table 1. GMM parameter estimation algorithm from the original minority class samples

Require: $\{x_i\}_{i=1}^{N} \in S_{\min}$, initial values $\Theta^0 = \{\pi_k^0, \mu_k^0, \Sigma_k^0\}_{k=1}^{K}$;
1: E-step: Calculate $\gamma(i,k)$ according to the current parameter Θ^{t-1} by (5), where the superscipt "t-1" is the iternation number.
2: M-step: Estimate Θ^t by (8), (9) and (12), respectively.
3: Repeat E-step and M-step until the log-likelihood function converges, $\ln P(\mathrm{X})^t - \ln P(\mathrm{X})^{t-1} \le \varepsilon$.
Return: $\Theta^T = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$.

Generate new minority class samples by the learned GMM. After the GMM parameter estimation procedure finishing, it can be considered that the distribution property of the minority class samples can be grasped and represented by this learned model. Then, new minority class samples can be generated by the learned GMM. The procedure is described in Table 2.

Table 2. New minority class samples generation from the learned GMM

Require: $\Theta^T = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$, N';
1: Uniformlly generate a random number δ in [0,1].
2: If $\delta \in \left[\sum_{c=1}^{k-1} \pi_c, \sum_{c=1}^{k} \pi_c\right]$, generate a new sample from Gaussian distribution $\tilde{x} \sim N(\mu_k, \Sigma_k)$
3: Repeat above steps N' times.
Return: $S'_{\min}$

It is noted that N' denotes the number of new minority class samples needing to generated, which is determined by the imbalanced dataset. The final minority class set is $S_{\min}^{new} = S_{\min} + S_{\min}^{'}$, which is used for the subsequent prediction algorithm.

3.2 Complaint Prediction by Naïve Bayes Algorithm

The core of the prediction algorithm is to construct a classifier. Considering the factors of high computationally efficiency, high accuracy and solid theoretical foundation, Naïve Bayes algorithm has been widely used. Naïve Bayes classifier is based on a simple assumption: the given property values under classification characteristic conditions are independent. Based on attribute conditional independence assumptions, Naïve Bayes classifier has a simple star architecture (represented by D), as shown in Fig. 1.

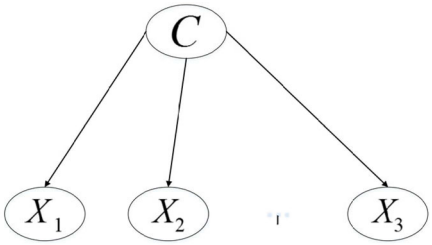

Fig. 1. The structure of Naive Bayes classifier

Concretely, joint probability distribution based on naive Bayes classifier structure is as follows:

$$p(c, x_1, \cdots, x_n) = p(c)\, p(x_1, \cdots, x_n) = p(c) \prod_{i=1}^{n} p(x_i | c). \tag{13}$$

where $p(c)$ is prior probability and $p(x_i | c)$ is class-conditional probability.

According to the joint probability distribution, the expression of Naive Bayes classifier is obtained as follows:

$$\underset{c(x_1, \cdots x_n)}{\operatorname{argmax}} \left\{ p(c) \prod_{i=1}^{n} p(x_i | c) \right\}. \tag{14}$$

Obviously, the training process of naive Bayes classifier is to estimate the prior probability of class $p(c)$ and class-conditional probability $p(x_i | c)$ for each attribute based on the training set S, $S = S_{maj} + S_{\min}^{new}$.

4 Experimental Results and Analysis

The authors may express their acknowledgement and the financial support project number here.

4.1 Dataset and Preprocessing

In our experiment, two original datasets are from Jiangsu Telecom. Dataset 1 is the IPTV alarming list from April 1st to April 10th. Dataset 2 is the user's complaint list (Telecommunications Service received the data from user complaints). The data preprocessing is as follows:

Attribute selection. As mentioned above, in dataset 1, each instance has a user id. In our study, we choose 10 attributes whose meanings are listed in Table 3.

Table 3. All attributes and meanings

Attributes	Meanings
SEVERITY	The alarm level.
ALARM_NUM	The number of alarm time.
LOSSRATE	The impact factor of packet loss, representing the impacts that network has on MOS value.
DOWN_BANDWIDH	The down bandwidth of the top-set box.
MEDIARATE	The rate of the media.
MDI_DF	The delay factor of media delivery index.
MDIMLR	The media loss rate of media delivery index.
VSTQ	The quality of network transmission, reflecting the status of the network.
MOS_VALUE	The average score of the video.
CPU_USAGE	CPU usage.

Data cleaning. Data cleaning is an important step to remove irrelevant and redundant information present or noisy and unreliable data. Data cleaning process can be further divided into three steps: data error cleaning, duplicate checking and labeling. Specifically, the error data that uploaded by set-top boxes should be cleared firstly. Secondly, the duplicate checking is conducted to remove the samples with same values. Processing these data clearly helps improve the classification results and avoid over generalization. Finally, traverse each sample in dataset 1 and mark the samples whose user ids also appear in the dataset 2 as minority class. The other instances are marked as majority class. In this way, dataset 1 is divided into two categories and labeled.

After these steps, the total number of the samples in our dataset now is 439050, among which 4871 samples belong to minority class and 434179 belong to majority class. It can be seen that the degree of imbalance data set is quite large. It is difficult for users to report very accurate predictions for the reporting barrier. Moreover, it is important to choose the appropriate evaluation criteria for imbalanced datasets.

4.2 Evaluation Criteria of Imbalanced Datasets

In the presence of imbalanced data, it is difficult to make relative analysis when the evaluation metrics are sensitive to data distributions. In the paper, evaluation criterion of classification models are defined based on confusion matrix shown in Table 4.

Table 4. The confusion matrix and result of the balanced cart

	True class (T)	False class(F)
Positive output(P)	TP	FP
Negative output(N)	FN	TN

Considering a basic two-class classification problem, let {T, F} be the true positive and negative class label and {P, N} be the predicted positive and negative class labels. $G-mean$ takes the classification performance both of the minority class and majority class into account, calculated by (14). When we test the model, we use the common performance indicators of the imbalanced dataset. In this paper, $G-mean$ is used for our evaluation criterion.

$$G-mean=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}}. \tag{15}$$

4.3 Experimental Results and Analysis

After data cleaning and over-sampling by using proposed algorithm in Section III, the minority class samples expands to 89 times of the original minority dataset. This processed dataset is divided into two parts. One part is used as the training dataset, the other part as the test dataset. We randomly select a subset which contains 10000 minority class samples and 10000 majority class samples from the training dataset to build the Naïve Bayes classifier.

In order to verify the model, we randomly select different proportions of imbalanced dataset from the

testing dataset. The proportions are 1:30, 1:60 and 1:89. Each proportion has five subsets which are used for testing. Each subset contains 200 minority class data. We compare the performance of the proposed algorithm (GMM-Naïve Bayes) with Borderline-SMOTE-Naïve Bayes algorithm in different proportions. The results are shown in Fig. 2 to Fig. 4.

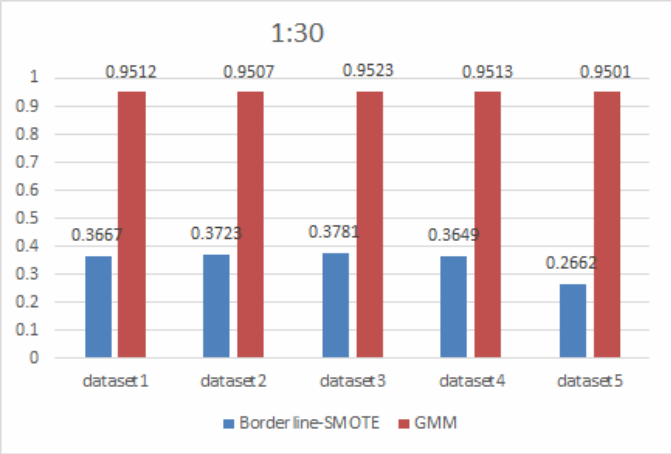

Fig. 2. G-mean comparison of Borderline-SMOTE-Naïve Bayes model and GMM-Naïve Bayes model (1:30)

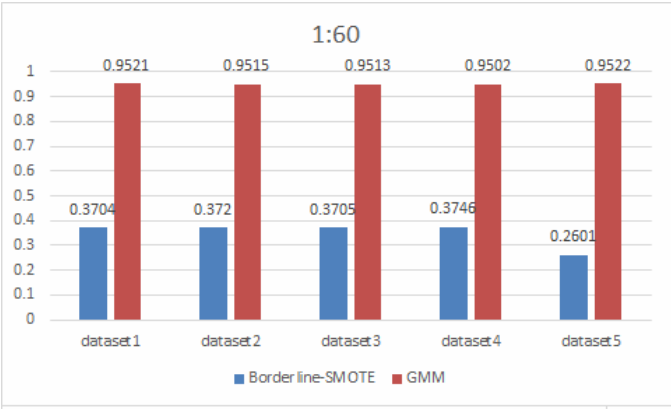

Fig. 3. G-mean comparison of Borderline-SMOTE-Naïve Bayes model and GMM-Naïve Bayes model (1:60)

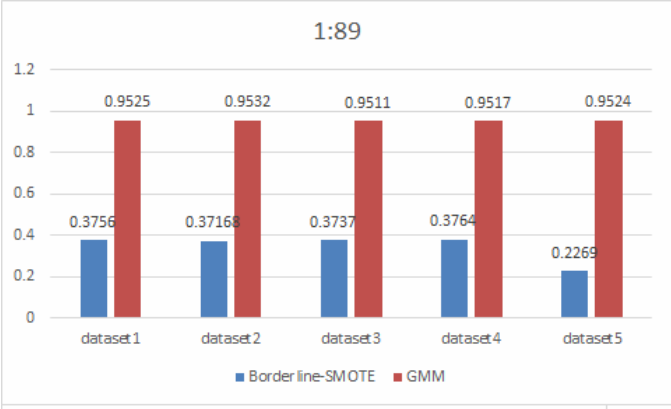

Fig. 4. G-mean comparison of Borderline-SMOTE-Naïve Bayes model and GMM-Naïve Bayes model (1:89)

From Fig. 2 to Fig. 4, we can see that the proposed GMM-Naïve Bayes algorithm has better performance than Borderline-SMOTE- Naïve Bayes algorithm in the imbalanced dataset. The reason is that the GMM can better describe the distribution property of the minority class examples. Therefore, the generated samples are also effective for training and testing.

Additionally, we also use the original data to test the performance of the GMM-Naïve Bayes algorithm. We also compare its performance with that of the Kmeans-Naïve Bayes [11], Borderline-SMOTE-Naïve Bayes and no-SMOTE algorithms in different proportions. The proportions are also 1:30, 1:60 and 1:89. The result is shown in Fig. 5.

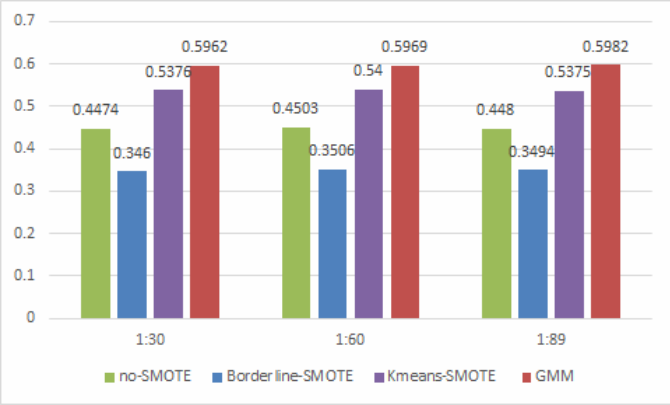

Fig. 5. G-mean comparison of no-SMOTE, Borderline-SMOTE-Naïve Bayes, Kmeans-Naïve Bayes and GMM-EM- Naïve Bayes algorithm

From Fig. 5, although the performance is not as good as above, we can see that the GMM-Naïve Bayes still has better performance than no-SMOTE, Borderline-SMOTE-Naïve Bayes and Kmeans-Naïve Bayes algorithms in the imbalanced dataset.

In summary, the above experimental results show that the GMM-Naïve Bayes algorithm can really improve prediction efficiency and accuracy of minority class. In our experiment, the samples belonging to minority class are very important. The performance of prediction is essential for operators to adjust the metrics before the users send a complaint file. Therefore, the proposed scheme is an effective tool for handling this imbalanced dataset processing problem.

5 Conclusions

In this paper, an over-sampling method based on the Gaussian mixture model (GMM) is proposed. Specifically, GMM is adopted to describe the distribution of limited complaint samples instead of the KNN in traditional SMOTE algorithm. Then the parameters in the GMM are estimated by the EM algorithm. New minority class samples can be generated by the learned model. Finally the Naïve Bayes classifier is used for finishing classification and prediction. Experimental results show that the proposed algorithm performs better than the competing algorithms in predicting user's complaint.

Acknowledgements

This work is partly supported by the State Key Development Program of Basic Research of China (2013CB329005), the National Natural Science Foundation of China (Grant No. 61322104, 61571240), the Priority Academic Program Development of Jiangsu Higher Education Institutions, the Natural Science Foundation of Jiangsu Province (Grant No. BK20161517), the Qing Lan Project, the Scientific Research Foundation of NUPT (Grant No.NY217022).

References

[1] H. Ketmaneechairat, A survey and comparison of some popular IPTV applications, in: Proc. 8th International Conference on

Computing Technology and Information Management (NCM and ICNIT), 2012.

[2] B. Rong, Y. Qian, M. Guiagoussou, M. Kadoch, Improving delay and jitter performance in wireless mesh networks for mobile IPTV service, IEEE Transactions on Broadcasting 55(3)(2009) 642-651.

[3] Y. Chen, K. Wu, Q. Zhang, From QoS to QoE: a survey and tutorial on state of art, evolution and future directions of video quality analysis, IEEE Communications Surveys and Tutorials 17(2)(2015) 1126- 1195.

[4] D. Tsolkas, E. Liotou, N. Passas, A survey on parametric QoE estimation for popular services, Journal of Network and Computer Applications 77(1)(2017) 1-17.

[5] C. Lin, J. Hu, X. Kong, Survey on models and evaluatio of Quality of Experience, Chinese Journal of Computer 35(1)(2012) 1-15.

[6] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, H. Zhang, A quest for an internet video quality-of-experience metric, Proceedings of the 11th ACM Workshop on Hot Topics in Networks 8(4)(2012) 97-102.

[7] T.D. Pessemier, K.D. Moor, W. Joseph, L.D. Marez, L. Martens, Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching, IEEE Transactions on Broadcasting 59(1)(2013) 47-61.

[8] H. Kim, S. Choi, A study on a QoS/QoE correlation model for QoE evaluation on IPTV service, in: Proc. 12th International Conference on Advanced Communication Technology (ICACT), 2010.

[9] T. Truong, N. Hung, N. Thanh, Service provisioning with quality-of-experience support in IMS-based IPTV, in: Proc. International Conference on Ubiquitous and Future Networks (ICUFN), 2012.

[10] R. Huang, X. Wei, C. Lv, X. Li, S. Zhang, Prediction model for user's QoE in imbalanced dataset, in: Proc. The First International Conference on Computational Intelligence Theory, Systems and Applications (CCITSA), 2015.

[11] T. Wang, R. Huang, X. Wei, F. Zhou, Improving user's quality of experience in imbalanced dataset, in: Proc. International Computer Symposium, 2016.

[12] L. Wang, J. Jin, R. Huang, X. Wei, J. Chen, Unbiased decision tree model for user's QoE in imbalanced dataset, in: Proc. International Conference on Cloud Computing Research and Innovations (ICCCRI), 2016.

[13] H. He, E. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21(9)(2009) 1263-1284.

[14] W. Rivera, P. Xanthopoulos, A priori synthetic over-sampling methods for increasing classification sensitivity in imbalanced data sets, Expert Systems with Applications 66(12)(2016) 124-135.

[15] Z. Sun, Q. Song, X. Zhu, A novel ensemble method for classifying imbalanced data, Pattern Recognition 48(5)(2015) 1623-1637.

[16] K. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, Cambridge, MA, 2013.

[17] X. Wei, Z. Yang, The infinite Student's t-factor mixture analyzer for robust clustering and classification, Pattern Recognition 45(12)(2012) 4346-4357.

[18] G. McLachlan, T. Krishnan, The EM Algorithm and Extensions, 2nd ed., John Wiley & Sons, New York, 2008.