

User Behavior Profiles Establishment in Electric Power Industry

Wei Song¹, Gang Liu¹, Di Luo^{2(✉)}, and Di Gao¹

¹ State Grid Jibei Electric Power Science Research Institute, Nanjing, China

² School of Computer and Communication Engineering, University of Science
& Technology Beijing, Beijing, China
brave1d@163.com

Abstract. The big data of electric power industry contains the information of users' values, credits, and behavior preferences. On the basis of the data, electric power companies can provide personal services as well as increasing profits. In this paper, we proposed a labeling system based on the clustering algorithms and Gradient Boost Decision Tree (GBDT) algorithm to establish user behavior profiles for State Grid Group of China, including basic information labels, behavior labels, behavior description labels, behavior prediction, and user classification. The experimental results showed that the approach can describe the behavior features of users in the electric power industry effectively.

Keywords: Behavior profile · Behavior prediction · Clustering · Gradient boost decision tree · Electric power industry

1 Introduction

There are many kinds of behaviors in electric power industry including electricity consuming, paying, capacity changing, illegally electricity using, etc. Although the electric power companies have a lot of users' behavior data, they are numerical and hard to understand directly. Some rules are needed to turn the initial numerical data to semantic labels which could easily illustrate the users' feature. Here are some of the labels: 'Zhangjiakou', 'Low monthly consumption', 'capacity change frequently', 'Recover from capacity suspend every 7 months', 'High risk of arrearage', 'Low user value'. These labels are called user profiles. Profiling the users is the basis of providing the personal services, and its applications in the electric power industry are various. First, companies can select users by labels, which are more detailed and semantic. Also, it's easy to recognize the occurrence of the user behaviors so that companies can easily monitor the user behaviors and give advise. At the same time, as the companies handle the feature of the user behaviors, precision marketing will be reasonable and the companies can arrange business better. What the most important is, according to the user profiles of electric power industry, the companies can get the semantic credit report of the users. User profile is commonly used in IT industry, but we can not build the user profiles in the electricity industry using the same approach in the IT industry

due to the great differences between the two industries. The electricity industry needs its own labeling systems and new methods to generate the semantic labels. The essence of user profiles is to label the user features, which depends on label generation techniques.

2 Related Work

User profiles are very useful in the IT, telecom, financial industry. Commonly applied techniques such as frequent pattern mining, probability matrix, etc., are used to build the mobile user behavior profiles based on the data of the users' frequent behaviors, data of regular stops and moving speeds [1]. Some labels are built base on the expert experience. These labels represent the long and short term preference of users, which can be used in the collaborative filtering algorithm to recommend web page [2]. Collaborative filtering matrix is an effective method to create user label from the recording history. Ali et al. built a user profile system named "TiVo" based on the collaborative filter matrix technique [3]. Dynamic Bayes algorithm was employed to build a user's profile by online education resource [4]. However, there is few related work to build the user profile of electricity users. Sparse encoding model was applied to create user label of abnormal electricity consumption [5]. Logistic regression algorithm was well employed to create the label showing the risk of arrearage in electric power industry [6]. Back propagation neural network model was used to predict the possibility of stealing electricity [7].

3 User Label System of Electric Power Industry

Electric user behavior profile contains static basic information and dynamic behavior information. The static basic information is stable, while dynamic behavior information is changing over time. This paper proposed a new hierarchical label system for electric power industry, including basic information label, behavior label, behavior description label, behavior prediction and user classification label.

3.1 Behavior Label

Most of the data are numerical. To show what a user did directly, the semantic label can be an effective way. The paper combines the expert experience and the clustering algorithm to achieve the discretization of the data, dividing the data into several comprehensible labels. That is, turn the numerical data to semantic label T .

3.2 Behavior Description Label

The behavior description label shows the temporal features of the user behavior and user preference. Each behavior can be described from its occurrence frequency, average variation, coverage rate, standard deviation, average time interval, periodicity, period preference. create Ratio is the coverage rate of the behavior at of the user u , and we use

$sum(at_j, u)_{ET-ST}$ as the sum of the behavior in the period T. Therefore, the formula is as follow:

$$\text{CreateRatio} = \frac{sum(at_j, u)_{ET-ST}}{\sum_{i=0}^n sum(at_i, u)_{ET-ST}} \quad (1)$$

Periodicity is for estimating if this behavior has periodic regular. If not, we use 0 to represent it, otherwise, we use d to show the periodic time interval. We use $sum()$ as the sum of the occurrence of d_j , and the $period(at, u)$ is calculated as follow:

$$period(at, u) = \begin{cases} d_i & \exists d_i \in (d_1, d_2, \dots, d_n), sum(d_i, u) \geq 0.6 \sum_{i=0}^n sum(d_i, u) \\ 0 & \forall d_i \in (d_1, d_2, \dots, d_n), sum(d_i, u) \leq 0.6 \sum_{i=0}^n sum(d_i, u) \end{cases} \quad (2)$$

The paper define TF as the period preference. If the number of the occurrence of the behavior in the one period of time is more than 60% of the total number of the occurrence of the behavior, we will label the user has periodical preference, and the label name will be the period.

Distribute the one dimension temporal feature such as occurrence frequency into several levels to describe the different frequency of the behaviors. Besides, distribute occurrence frequency, coverage rate, average time interval into several levels, and let the experts name each level so we can get the labels of the behavior preference. The labels of the period preference can be created as the name of the time period.

3.3 Behavior Prediction and User Classification Label

The electricity company will be able to arrange the product plan more reasonable by predicting the future behaviors of users. The user classification label is for dividing users into different groups, likes ‘high-value user’, ‘low-value user’, ‘high credit’, ‘bad credit’. The clustering algorithm is always used to do that, and it performs well. The behavior prediction labels are based on various classification algorithm and GBDT algorithm is the first choice. Using the clustering algorithm and GBDT algorithm, the paper classifies users into different groups which represent different risk of being behind in payment and labels them.

4 Label Generation Technique

4.1 Label Generation Technique Based on Clustering Algorithm

k-means cluster algorithm can’t be used directly in the situation [8]. While generating the label, we have to deal with the outliers instead of just ridding off them. The algorithm based on the idea of the Alex and Alessandro can perform well in such situation. The algorithm has its basis in the assumptions that cluster centers are surrounded by neighbors with lower local density and that they are at a relatively large distance from any points with a higher local density [9].

We use S to represent the N size dataset that waiting to be classified, X for each data point in this dataset, and it defined as $S = \{X_i\}_{i=1}^N$. Distances between the data point X_i and X_j are defined as δ . For each X , calculate its density ρ_i and distance δ_i . First, we have to define a cutoff distance d_c , which is given by experience. ρ_i is defined as follow:

$$\rho_i = \sum_{i,j \in I_S} \phi(d_{ij} - d_c) \quad (3)$$

where

$$\phi(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases} \quad (4)$$

Assume that $\{q_i\}_{i=1}^N$ is the index of $\{p_i\}_{i=1}^N$ in descending order, which is $\rho_{q1} \geq \rho_{q2} \geq \dots \geq \rho_{qN}$. So δ_{qi} is defined as:

$$\delta_{qi} = \begin{cases} \min(d_{qiqj}), & i \geq 2, j < i \\ \max(\delta_{qj}), & i = 1, j \geq 2 \end{cases} \quad (5)$$

Assume that there are n_c ($n_c > 0$) cluster centers in the dataset S . $\{m_i\}_{i=1}^N$ is the serial number of the data point corresponding to each of the cluster center. $d_{\max} = \max\{d_{ij}\}$ is the biggest distance between two data point in S . $\{n_i\}_{i=1}^N$ represents the number of the data point whose density is bigger than X_i and is the closest with X_i . The steps of the algorithm are as follow:

Step one: Initialization and pretreatment

- (1) Given t to determine the cutoff distance, where $t \in (0, 1)$;
- (2) Compute the distance d_{ij} between each two data points, where $i > j$, $i, j \in I_S$;
- (3) According to 2), M is the number of the distances between each two data points in dataset S and $M = 0.5(N-1)N$. Sort them in ascending order, making $d_1 \leq d_2 \leq \dots \leq d_M$. So we can get $d_c = d_{f(Mt)}$, $f(Mt)$ is the function that rounds the Mt ;
- (4) According to the formula (1), compute the density of each data point to get $\{p_i\}_{i=1}^N$, and sort them in descending order to get the index $\{p_i\}_{i=1}^N$;
- (5) According to the formula (3), compute the distance between data points, and get $\{n_i\}_{i=1}^N$.

Step two: Determine the cluster centers

The cluster centers must fit two conditions: (1) cluster centers are surrounded by neighbors with lower local density; (2) cluster centers are at a relatively large distance from any points with a higher local density. For each data point, the bigger ρ_i and δ_i it has, it has more chance to be the cluster centers. Consider both of the ρ_i and δ_i , we can get γ_i defined as follow:

$$\gamma_i = \rho_i \delta_i, i \in I_S \quad (8)$$

Apparently, the bigger γ_i it has, it is much likely to be the cluster centers. Draw the figure of the γ_i in the descending order. There must be some points that have larger γ_i and can be easily identified. Make them as the cluster centers.

Step three: Classify the rest of the data points

For the rest of the data points, they are sorted in the descending order of their ρ_i . So compare each of them to the different cluster centers, and classify them into the cluster that has the smallest distance.

5 Experiment and Analysis

The paper selects the real data of large-scale industry users from State Grid Group of China, which contains 183662 monthly electricity consumption data, 23562 electricity inspection data, 13839 capacity changing data, ranging from 2014.1 to 2016.10. Besides, there are also 7008 arrearage data ranging from 2016.1 to 2016.10.

5.1 Generate the Behavior Label

Create behavior labels according to monthly consumption data, and get the ‘Extreme high electricity consumption’, ‘High electricity consumption’, ‘Medium electricity consumption’, ‘Low electricity consumption’ and ‘Extreme low electricity consumption’ labels. The monthly consumption of large-scale industry users has a wide distribution. A small number of extreme large-scale steel plants have much higher monthly consumption than the others, which become the outliers. Therefore, we rid this part of data into the ‘High electricity consumption’ level and label the behavior. The paper set $t = 0.05$ to get the best result. And Fig. 1 is what we got.

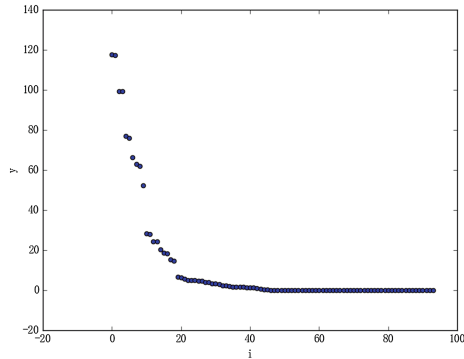


Fig. 1. γ_i in descending order of each data point

From the Fig. 1, we can see there are 20 points on the figure that are much high than the others. These points will be selected to be the cluster centers, and after that, the cluster centers will be classified into the rest of the four levels to get the other labels.

5.2 Generate the Behavior Description Label

The paper is using the monthly consuming behavior and capacity changing behavior as an example in this section. According to the formula in Sect. 3, we can get Table 1.

Table 1. The behavior feature of monthly consuming and capacity changing

Behavior	Frequency	Coverage rate	Average time interval	Standard deviation	Periodicity	Period preference
No consumption	17	0.57	1.5 month	1.36	1 month	Former half of year
Low electricity consumption	7	0.23	2.1 months	2.06	No periodicity	Latter half of year
Extreme low electricity consumption	4	0.13	2.75 months	3.05	No periodicity	0
Capacity suspend	5	0.56	6 months	3.36	No periodicity	Beginning of years
Recovery from capacity suspend	4	0.44	7 months	0	7 months	Latter half of year

From the table above, classify them with cluster algorithm to get the behavior description label: ‘No consumption frequently’, ‘Low electricity consumption’, ‘Capacity suspend happens at the beginning of the year’, ‘Recover from capacity suspend every 7 months’.

5.3 Generate the Behavior Prediction Label and User Classification Label

The paper uses RFM model [12] combined with expert experience to define the risk of the arrearage. The paper uses nine months arrearage data as training data ranging from 2016.1 to 2016.9 and classifies the RFM value into different clusters using the clustering algorithm above and gets 17 clusters. Giving the R, F and M the same weight, classify the 17 clusters into 4 levels according to the expert experience. Finally, we can label users with ‘High risk in arrearage’, ‘Medium risk in arrearage’, ‘Low risk in arrearage’, ‘No risk in arrearage’.

Build the predicting model for risk of the arrearage based on GBDT algorithm according to the monthly consumption record, contract breaking record, capacity changing record. The result is displayed using the confusion matrix as follow:

GBDT	No	Low	Medium	High
No	363	0	0	0
Low	0	589	126	3
Medium	0	196	588	23
High	0	3	32	70

The paper compares it with the same training and testing data using the random forest (RF) algorithm. Compare the two different algorithms with the total accuracy, the recognition rate of the ‘No risk in arrearage’, ‘Low risk in arrearage’, ‘Medium risk in arrearage’, ‘High risk in arrearage’ respectively. The comparison is showed as Table 2.

Table 2. The comparison between GBDT and RF

Algorithm	Total	No	Low	Medium	High
GBDT	0.807	1	0.820	0.729	0.667
RF	0.803	0.997	0.822	0.719	0.634

We can know that both algorithms have almost the same accuracy in predicting the risk of the arrearage. However, the GBDT has better performance in recognizing the higher risk users.

6 Conclusion

The paper proposes a method to build the user behavior profiles of the electric power industry. The user behavior profile can recognize and describe behaviors, make predictions and classify users into different groups. Through the experiments, the user behavior profile that we build can describe the user perfectly. However, there are some disadvantages. Some of the labels need human involvement and are not generated quantitatively. The predicting model of the behavior may update after a period of time. In the future, we are going to build the user profile with more detailed data to fit the need of providing personal services.

References

1. Huang, W., Xu, S., Wu, J. et al.: The profile construction of the mobile user. *J. Mod. Inf.* **10** (2016)
2. Liu, Y.: A Generation Method of Mobile Users’ Tags based on Time Features. Dalian University of Technology (2015)
3. Ali, K., Van Stam, W.: TiVo: making show recommendations using a distributed collaborative filtering architecture// Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, pp. 394–401 (Aug 2004)
4. Slimani, H., Faddouli, N.E., Bennani, S., et al.: Models of digital educational resources indexing and dynamic user Profile evolution. *Int. J. Emerg. Technol. Learn.* **11**(1), 26 (2016)

5. Li, Z., Lujun, Z., Weiguo, G.: Application of sparse coding in detection for abnormal electricity consumption behaviors. *Power Syst. Technol.* **11**, 3182–3188 (2015)
6. Xu, R., Nd, W.D.: Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **16**(3), 645–678 (2005)
7. Chen, X., Li, C., Xiaoxiao, L., et al.: Research on electricity demand forecasting based on ABC-BP neural network. *Comput. Meas. Control* **22**(3), 912–914 (2014)
8. Kojima, K.: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. *Berkeley Symposium on Mathematical Statistics & Probability* (1969)
9. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. *Science* **344**, 1492 (2014). doi:[10.1126/science.1242072](https://doi.org/10.1126/science.1242072)
10. Deng, X.: Research of O2O E-commerce Recommendation Model on Gradient Boosting Regression Trees. *Anhui University of Science & Technology* (2016)
11. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29** (5), 1189–1232 (2000)
12. Fader, P.S., Hardie, B.G.S.: RFM and CLV: Using Iso-Value Curves for Customer Base Analysis. *J. Mark. Res.* **XLII**(4), 415–430 (2005)

Advances in Intelligent Systems and Interactive
Applications

Proceedings of the 2nd International Conference on
Intelligent and Interactive Systems and Applications
(IISA2017)

Khafa, F.; Patnaik, S.; Zomaya, A.Y. (Eds.)

2018, XVIII, 895 p. 407 illus., Softcover

ISBN: 978-3-319-69095-7