전력배전 시스템에서의 취약 선로 분류를 위한 출현 패턴 마이닝

Khalid E.K.Saeed [*], Minghao Piao[*], 이헌규[*], 신진호[**], 류근호[*]
[*]충북대학교 데이터베이스/바이오인포매틱스 연구실
[**]한국전력연구원 전력 정보 기술 그룹
e-mail : [*]{abolkog, bluemhp, hglee, khryu}@dblab.chungbuk.ac.kr
[**]jinho@kepri.re.kr

Emerging Patterns Mining for Classifying Non-Safe Electrical Sections in Power Distribution System

Khalid E.K.Saeed [*], Minghao Piao[*], Heon Gyu Lee[*], Jin-Ho Shin[**], Keun Ho Ryu[*]
*Database/Bioinformatics Laboratory, Chungbuk National University
** Power Information Technology Group, Korea Electric Power Research Institute

Abstract

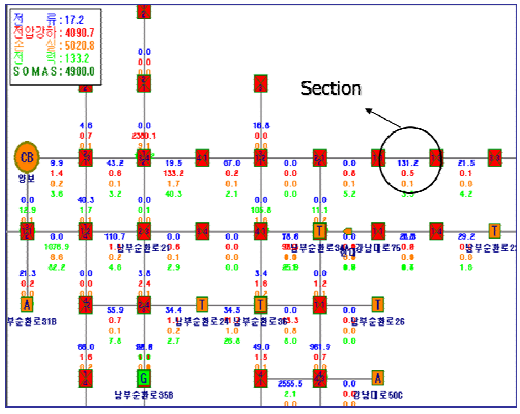
In electrical industry, classification methodology has been an important issue for analyzing power consumption patterns. It has many applications including decisions on energy purchasing, load switching as well as helping in infrastructure development. Our aim in this work is to classify the electrical section and find potentially non-safe electrical sections. For this purpose, we use Emerging Patterns based classification. The classification method uses the aggregate score of emerging patterns to build classifier. The proposed methodology was applied to a set of electrical section data of the Korea power. The test data and relational electricity information and knowledge are supported by *Korea Electric Power Research Institute (KEPRI)*.

1. Introduction

Classification methodology has been an important issue in the power industry, recently. Classification helps an electric utility making important decisions such as decisions on energy purchasing, load switching, and also helps to develop the infrastructure. It is extremely important for electric energy generation and transmission, distribution and electrical markets. In power system, data mining [1, 2, 3] is the most commonly used method to determinate load profiles and extract regularities in load data and thus has been the target of some investigations for its used in classification. Classification using data mining is usually made by building models on relative information.

In our work, we deal with the load features of electricity sections. The experts defined the sections into two parts: (I) non-safe (weak) sections and (II) safety (strong/Firm) sections. *Figure 1* shows a power distribution system, the small boxes represents electricity switches or routers, the section as illustrated in *Figure 1* is the area between two switches (routers). Unlike safety sections, non-safe electrical sections are usually the causes of serious problems such as electricity cuts, fires or damages of the infrastructure.

In this study, we use the idea of CAEP [4] to perform the classification. For classification, single Emerging Pattern can only predict the class membership for a small number of instances and not on all instances. For better overall accuracy, CAEP uses aggregate score for the distinction power and the indication of class membership for classifier. in order to keep the electricity safe, we apply the aggregate score on the data set to build classifier and try to find potentially non-safe electrical sections to help the planning and organization of electric utilities.



(Figure 1) Outline map of Korean Power Distribution System

2. Classification of Non-safe Electrical Sections

In this section, we will give a description of Emerging Patterns and the CAEP algorithm. CAEP classifies the instances by deals with the aggregate scores of Emerging

Patterns.

2.1 Emerging Patterns

Emerging Patterns are those whose frequencies change significantly from one dataset to another. Each Emerging Pattern has big difference between its supports in the opposing classes and represents strong contrast knowledge. So, it can sharply differentiate the class relationship of input instances containing the Emerging Patterns. Emerging Patterns have been shown to be successful for constructing accurate classifiers. The task of mining Emerging Patterns is computationally expensive for large, dense and high-dimensional datasets. Emerging Patterns are defined as following:

***Definition 1*:** Given two different class of dataset D_1 and D_2, the Growth-Rate of an itemset X from D_1 to D_2 is defined as

$$GR(x)=\begin{cases}0 & \text{if } \mathrm{supp}_1(x)=0 \text{ and } \mathrm{supp}_2(x)=0.\\ \dfrac{\mathrm{supp}_1(x)}{\mathrm{supp}_2(x)} & \mathrm{supp}_1(x) \text{ for class1}.\mathrm{supp}_2(x) \text{ for class2.}\\ \infty & \text{if } \mathrm{support}_1(x)>0 \text{ and } \mathrm{supp}_2(x)=0.\end{cases}$$

and given a threshold ρ.

Emerging Patterns are those itemsets with large Growth-Rate from D_1 to D_2.

***Definition 2*:** Given a Growth-Rate threshold $\rho > 1$, and itemset X is said to be a Emerging Pattern from a background dataset D_1 to a target dataset D_2 if Growth-Rate $\geq \rho$.

An Emerging Pattern with high support in its home class and low support in the opposing class can be seen as a strong signal indicating the class of a test instance containing it. The score of such a signal is expressed by its supports in both classes and its Growth-Rate (GR).

2.2 Classification methods based on Emerging Patterns

Algorithms for mining Emerging Patterns have been widely studied. The mostly used approaches are Border-based approach [5] and Constraint-based approach [6]. CAEP, JEPC [7] and DeEPs [8] are classifiers using Emerging Patterns.

In this study, we use the idea of CAEP to perform the classification. For classification, single Emerging Pattern can only predict the class membership for a small number of instances and not on all instances. For better overall accuracy, CAEP uses aggregate score for the distinction power and the indication of class membership for classifier. In our work, we apply the aggregate score on the data set and the aggregate Score is defined as below.

Given an instance T and a set $E(C_i)$ of Emerging Patterns of data class C_i, discovered from the training data, the aggregate score of T for the class C_i is:

$$score(T,C_i)=\begin{cases}\displaystyle\sum_{X\subseteq T, X\in E(C_i)} \frac{GR(X)}{GR(X)+1}\times \mathrm{supc}_i(X) & , \ GR>\rho.\\ \displaystyle\sum_{X\subseteq T, X\in E(C_i)} \mathrm{supc}_i(X) & , \ GrowthRate=\infty.\\ 0 & , \ GR=0 \text{ or } GR<\rho.\end{cases}$$

3. Experiments and results

Our test data that we considered in this paper was collected by KEPRI (Korea Electric Power Research Institute), the data set features are as follow: Several electrical section codes and the features derived from each section data. The derived features are: maximum value, minimum value and average value. Since the extracted features and some electrical section codes contain continuous variables, Entropy-based discretization [9] has been used to cut up the features into several intervals, so that transferred features could be used for classification.

Let an object set *S* be composed of K classes ($d_1, d_2, ..., d_k$), having probabilities $p_1, p_2, ..., p_k$ respectively, then the entropy of *S* is defined as

$$E(S)=\sum_{i=1}^{k} p_i \log_2\left(\frac{1}{p_i}\right) \qquad (1)$$

Let an attribute *A* divide *S* into *n* disjoint subsets $S_1, S_2, ..., S_n$ then the entropy *E (A, S)* of S partitioned by A is defined as

$$E(A,S)=\sum \frac{|S_i|}{|S|}E(S_i) \qquad (2)$$

where |*X*| denotes the cardinality of the set *X*.

Information Gain (*IG*), Gain Ratio (*GR*) and Normalized Gain (*NG*) are three frequently used entropy-based criteria for evaluating the importance of an attribute on classification. In our work, the Gain Ratio has been used as evaluation criterion.

$$IG(A,S)=E(S)-E(A,S) \qquad (3)$$

$$GR(A,S)=\frac{IG(A,S)}{\displaystyle\sum_{i=1}^{n}\frac{|S_i|}{|S|}\log\frac{|S|}{|S_i|}} \qquad (4)$$

$$NG(A,S)=\frac{IG(A,S)}{\log_2 n} \qquad (5)$$

Figure 2 shows the data set features and its corresponding types before the preprocessing process and after it. The upper part of the figure show the dataset features with its corresponding data types before the preprocessing step and the lower part shows the preprocessed dataset.

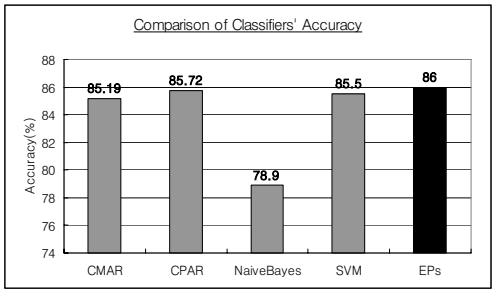
Feature	Type
Customer Electricity Use Code	Nominal
	Continuous
Maximum value of Electricity Section data	Continuous
Minimum value of Electricity Section data	Continuous
Average value of Electricity Section data	Continuous

Data Preprocessing

Feature	Type
Customer Electricity Use Code	Nominal
Maximum value of Electricity Section data	Nominal
Minimum value of Electricity Section data	Nominal
Average value of Electricity Section data	Nominal

(Figure 2) Data Preprocessing

The accuracy of the classifier was compared with several classifiers in order to know how well the classifier can classify the data. Our experiments shows good results for our classifier, comparing to other classifiers, SVM (Support Vector Machine), CPAR (Classification based on Predictive Association Rules) [10], CMAR (Classification Based on Multiple Association Rules) [11] and Naïve Bayes classifier. *Figure 3* shows the result of the experiment.



(Figure 3) Comparison of the Classifiers' accuracy

4. Conclusions

Classification methodology has been an important issue in the power industry, recently. Classification helps an electric utility making important decisions such as decisions on energy purchasing, load switching, and also helps to develop the infrastructure.

In this paper, the proposed main mining task is classification of electrical sections using emerging patterns. We tried to classify potentially non-safe in order to make the use of electricity more safe and secure. Our classification method shows good result comparing its accuracy with other classification algorithms: CPAR, CMAR, Naïve Bayes and SVM.

Acknowledgements

This work is supported by Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MOST) (R01-2007-000-10926-0).

References

[1] Heon Gyu Lee, Jin-ho Shin, Keun Ho Ryu, "Application of Calendar-Based Temporal Classification to Forecast Customer Load Patterns from Load Demand Data," IEEE CIT 2008, pp.149-154.

[2] Minghao Piao, Jin Hyoung Park, Heon Gyu Lee, Jin-Ho Shin, Keun Ho Ryu, "Power Load Pattern Classification from AMR Data," The 29th KIPS Spring Conference 2008, pp.231-234.

[3] Heon Gyu Lee, Jin-Ho Shin, Hong Kyu Park,Young-il Kim, Bong-Jae Lee, Keun Ho Ryu, "Temporal Classification Method for Forecasting Power Load Patterns From AMR Data," Korean Journal of Remote Sensing, Vol. 23, No. 5, pp.393-400, 2007.

[4] G. Dong, X. Zhang, L. Wong, and, J. Li, "CAEP: Classification by Aggregating Emerging Patterns," Proceedings of 2nd International Conference on Discovery Science, pp.30-42.

[5] J. Li, K. Ramamohanarao, and G. Dong, "The space of jumping emerging patterns and its incremental maintenance algorithms," Proceedings of the 17th International Conference on Machine Learning, pp. 551-558.

[6] X. Zhang, G. Dong, and K. Ramamohanarao, "Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets," Proceedings of the 6th ACM SIGKDD international conference on Knowledge Discovery and Data Mining , pp.310-314.

[7] J. Li, G. Dong, and K. Ramamohanarao, "Making use of the most expressive jumping emerging patterns for classification," Knowledge and Information Systems, 2001, pp.131-145.

[8] J. Li, G. Dong, K. Ramamohanarao, and L. Wong, "DeEPs: A new instance-based discovery and classification system," Machine Learning, 2004, pp. 99-124.

[9] J. Dougherty, R. Kohavi, M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features," Machine Learning: Proceeding of the 12th International Conference, Morgan Kaufmann Publishers, pp.194-202.

[10] X. Yin, Han, "CPAR: Classification based on Predictive Association Rules," In: Proceedings of SIAM International Conference on Data Mining, pp.331-33.

[11] W. Li, J. Han, Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rule," Proceedings of the 2001 IEEE International Conference on Data Mining, pp.369-376.