

2019 International Conference on
Computational Intelligence and Security (CIS'2019)
December 13-16, 2019, Macau, China

Dear Mr. LI Helong,

We are pleased to inform you that your paper entitled Intelligent energy meter fault prediction based on machine learning(ID:AP19920004)submitted to 2019 International Conference on Computational Intelligence and Security (CIS'2019) has been accepted for oral presentation at the conference and for publication in the conference proceedings, published by the Conference Publishing Services (CPS) of IEEE. However, you have to carefully study the reviewers' comments/suggestions and make a significant revision accordingly.

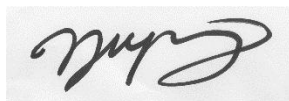
In order to successfully include your paper into the conference proceedings, it is important that you should closely follow the registration instruction and camera-ready paper submission instruction on CIS official website (URL: <http://www.cis-lab.org/>). Camera-ready paper of each accepted paper should not exceed 5 pages. Additional page(s) will be charged by the IEEE publisher. Please submit your revised paper before the camera-ready paper deadline.

Since we will select some high quality accepted papers to be included in SCI and EI journals this year after extension and revision, please make sure that the English writing of your camera-ready paper is good and has no grammar errors. For those papers with grammar errors, we will not select them to these journals.

Your kind cooperation will be greatly appreciated.

Should you have any further enquiries, please contact the CIS'2019 secretariat via e-mail:cis.conf.xd@gmail.com. Thanks a lot for your participation in CIS'2019! We are looking forward to meeting you in Macau, China, on December 13-16, 2019.

Yours sincerely,



On behalf of Programming Committee of CIS'2019

历届会议 EI 检索证明

CIS2016 (第 12 届)

Proceedings - 12th International Conference on Computational Intelligence and Security, CIS 2016

Source: Proceedings - 12th International Conference on Computational Intelligence and Security, CIS 2016, January 17, 2017, Proceedings - 12th International Conference on Computational Intelligence and Security, CIS 2016

Database: Compendex

Document type: Conference proceeding (CP)

Detailed [Hide preview](#) [Full Text Links](#)

The proceedings contain 159 papers. The topics discussed include: hadoop-based dynamic load balance scheduling algorithm of logistics inventory; divisible-load scheduling for network-based computing systems with processor startup overheads and release times; Bayesian quantile regression and variable selection for count data with an application to youth fitness survey; a new evolutionary algorithm based on decomposition for multi-objective optimization problems;

CIS2017 (第 13 届)

Proceedings - 13th International Conference on Computational Intelligence and Security, CIS 2017

Source: Proceedings - 13th International Conference on Computational Intelligence and Security, CIS 2017, v 2018-January, July 2, 2017, Proceedings - 13th International Conference on Computational Intelligence and Security, CIS 2017

Database: Compendex

Document type: Conference proceeding (CP)

Detailed [Hide preview](#) [Full Text Links](#)

The proceedings contain 121 papers. The topics discussed include: an alpha-dominance expansion based algorithm for many-objective optimization; using memetic algorithm for matching process models; an effective solution to nonlinear bilevel programming problems using improved particle swarm optimization algorithm; improving hybrid gravitational search algorithm for adaptive adjustment of parameters; a novel multi-objective evolutionary algorithm based on a further decomposition strategy;...(1 more search term)

CIS2018 (第 14 届)

Proceedings - 14th International Conference on Computational Intelligence and Security, CIS 2018

Source: Proceedings - 14th International Conference on Computational Intelligence and Security, CIS 2018, December 5, 2018, Proceedings - 14th International Conference on Computational Intelligence and Security, CIS 2018

Database: Compendex

Document type: Conference proceeding (CP)

Detailed [Hide preview](#) [Full Text Links](#)

The proceedings contain 74 papers. The topics discussed include: computationally efficient low power neuron model for digital brain; speech recognition method based on normalized simplified artificial fish swarm algorithm; an improvement evolutionary algorithm based on grid-based Pareto dominance for many-objective optimization; an encoding algorithm based on the shortest path problem; a genetic algorithm for solving linear integer;...(2 more search terms)

CIS2019 (第 15 届)



The 15th International Conference on Computational Intelligence and Security

Macau, China

December 13-16, 2019

You are here: [About CIS](#)

About CIS

Dates

Important Dates

Papers

Registration

Paper Submission

Final Paper

Keynote Speakers

About CIS

International Conference on Computational Intelligence and Security (CIS) is a major annual international conference to bring together researchers engineers, developers and practitioners from academia and industry working in all areas of two crucial fields in information processing: computational intelligence (CI) and information security (IS), to share the experience exchange and cross-fertilize ideas. In particular, the series of CIS conference provides an ideal platform to explore the potential applications of CI models algorithms and technologies to IS.

Intelligent energy meter fault prediction based on machine learning

LI Helong

China Electric Power Research
Institute

Beijing, China

HLLi0526@126.com

YU Haibo

China Electric Power Research
Institute

Beijing, China

buaayhb@163.com

YUAN Jinshuai

School of Reliability and Systems
Engineering, Beihang University

Beijing, China

yuanjinshuai@buaa.edu.cn

Abstract—With the improvement of the intelligence and popularity of smart meters, the coverage of power information collection systems has also expanded, and the failure of smart meters is increasingly characterized by suddenness, complexity and multifaceted. Only through the maintenance personnel found that the failure of the meter is bound to have uneven personnel input and failure to timely repair. This paper analyzed the relationship between fault types and the phenomena and causes of faults based on statistics, preprocessing, clustering and prediction methods, and establishes a smart meter fault prediction. When it is effective to predict the type of failure, and accordingly equipped with the corresponding maintenance personnel, thereby reducing the cost of human resources and saving maintenance time.

Keywords- smart meters; data analysis; failure prediction model; machine learning

I. INTRODUCTION

With the continuous development of smart grids, the number of power users is increasing. The electricity information collection system is an important part in the construction of smart grids. The smart energy meter is the most basic component of the electricity collection system. Therefore, the operating state of the smart energy meter directly affects the stability, safety and economy of the acquisition system and the smart grid. The stable and reliable operation of the smart energy meter is not only related to the construction of the power grid, but also to the safe use of electricity by thousands of households. Therefore, it is very important to evaluate the operating status of smart meters [1-3].

In recent years, with the rapid advancement of communication technology and information technology, many smart energy meters have been deployed and applied, so the power company can obtain the measured data at the end of the distribution network with high frequency and wide coverage. In addition to the consumption data of the user segment. Some important operating parameters such as power, voltage, current and power factor of each test point can be obtained from current smart energy meter. With the wide application of smart energy meters in recent years, the data centers of power companies in various provinces have

accumulated a considerable amount of measurement data, and the measurement data of these energy meters can provide strong support for the evaluation of the operating state of the energy meter [4-10].

In view of this, this study proposes a method based on mathematical statistical analysis to explore the relationship between state influence and operational risk and realize the risk warning and state assessment of smart meters.

II. FAULT PREDICTION MODEL

There are many methods to solve classification problems. The single classification methods mainly included, decision tree, Bayes, artificial neural network, K-nearest neighbor, support vector machine and classification based on association rules.

A. Support Vector Machine

Support vector machine a two-category classification model (or classifier). The figures are as follows:

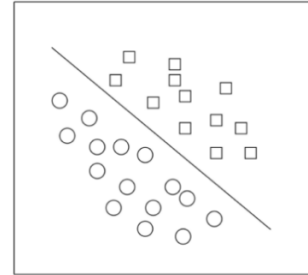


Figure 1 Support vector machine schematic

The idea of SVM is that given a set of samples containing positive and negative examples, the purpose of the SVM is to find a hyperplane to segment the sample according to positive and negative examples. SVM exhibits many unique advantages in solving small sample, nonlinear and high dimensional pattern recognition, and can be applied to other machine learning problems such as function fitting.

In the sample space, the division of the hyperplane can be described by the following linear equation:

$$g(x) = w^T x + b = 0 \quad (1)$$

If it has completed the separation of the samples, and the labels of the two samples are $\{+1, -1\}$, then for a classifier, $g(x) > 0$ and $g(x) < 0$ can be respectively Represents two different categories, $+1$ and -1 .

The core idea of the SVM is to do its utmost to maximize the separation between the two separate categories, which makes the separation more credible. And it has good classification and prediction ability for unknown new samples.

To maximize the spacing between different categories of data, let the data points closest to the separation surface have the largest distance. In order to describe the data points closest to the separating hyperplane, it is necessary to find two hyperplanes parallel and equidistant from the hyperplane: $H_1: y = w^T x + b = +1$ and $H_2: y = w^T x + b = -1$ the sample points on the two hyperplanes are theoretically closest to the separating hyperplane, It is their existence that determines the H_1 and H_2 position which supports the dividing line. With these two hyperplanes H_1 and H_2 , the interval mentioned above can be defined. The distance equation between the two parallel lines $ax + by = c_1$ and $ax + by = c_2$:

$$\frac{|c_2 - c_1|}{\sqrt{a^2 + b^2}} \quad (2)$$

The interval between the two hyperplanes H_1 and H_2 can be derivation is $2/\|w\|$, that means the goal now is to maximize this interval. For the convenience of subsequent derivation and calculation, further equivalent can be minimized as follows:

$$\frac{1}{\|w\|} \quad (3)$$

Assuming the hyperplane can correctly classify the samples, then

$$\begin{aligned} w^T x_i + b &\geq +1, y_i = +1 \\ w^T x_i + b &\leq -1, y_i = -1 \end{aligned} \quad (4)$$

Combine two above equations to get

$$y_i (w^T x_i + b) \geq 1 \quad (5)$$

This is constrained condition of objective function and optimized problem is

$$\begin{aligned} \min & \frac{1}{2} \|w\|^2 \\ s.t. & y_i (w^T x_i + b) \geq 1, i = 0, 1, 2, \dots \end{aligned} \quad (6)$$

B. Native Bayes

Native Bayes model predict the category to which the sample belongs by predicting probability of specific sample belongs to specific category $P(y_i | x)$ which is

$$y = \max P(y_i | x) \quad (7)$$

$P(y_i | x)$ can be written as

$$P(y_i | x) = \frac{P(x | y_i)P(y_i)}{P(x)} \quad (8)$$

Among them, $x = (x_1, x_2, \dots, x_n)$ is feature vectors of sample, $P(x)$ is priori probability. For specific sample x and category y_i , all the value of $P(x)$ are equal, and which cannot influence the relative size of $P(y_i | x)$ and can be ignored. Assuming that features x_1, x_2, \dots, x_n are independent and the following equation can be obtained:

$$\begin{aligned} P(y_i | x) &\propto P(x | y_i)P(y_i) \\ &= P(x_1 | y_i)P(y_i)P(x_2 | y_i) \dots P(x_n | y_i)P(y_i) \end{aligned} \quad (9)$$

Which, $P(x_1 | y_i)$, $P(x_2 | y_i) \dots P(x_n | y_i)$ and $P(y_i)$ can be obtained by training sample.

C. Logistic Regression

First, multinomial logistic regression can be derived by basic binary classification logistic regression. The binary classification with Logistic Regression can be calculated using the following formula:

$$P(y = 1 | x) = \frac{1}{1 + e^{-\theta^T x}} \quad (10)$$

We can obtain the following formula by rearranging the above formula

□

$$\log \frac{p}{1-p} = \theta^T x \quad (11)$$

where $p = P(y = 1 | x)$ which is probability of predicting positive sample if you input x . The best parameters of logistic regression can be learned by following maximum likelihood function.

$$L(q) = \prod_{i=1}^N P(y_i | x_i; q) = \prod_{i=1}^N (p(x_i))^{y_i} (1 - p(x_i))^{1-y_i} \quad (12)$$

Next, the multi-classification problem could to be solved. If every sample corresponds only one label, we can assume the probability every sample belongs to different label obeys geometric distribution, we can use softmax regression to classify.

$$h_q(x) = \begin{bmatrix} p(y=1|x;q) \\ p(y=2|x;q) \\ \dots \\ p(y=k|x;q) \end{bmatrix} = \frac{1}{\sum_{j=1}^k e^{q_j^T x}} \begin{bmatrix} e^{q_1^T x} \\ e^{q_2^T x} \\ \dots \\ e^{q_k^T x} \end{bmatrix} \quad (13)$$

where $\theta_1, \theta_2, \dots, \theta_k \in \mathbb{R}^n$ are parameters of model, and $1 / \sum_{j=1}^k e^{\theta_j^T x}$ can be seen as normalization of probability. For convenience, those k column vectors $\{\theta_1, \theta_2, \dots, \theta_k\}$ are sorted as $n \times k$ dimensional matrix, which is written as θ . Multi-classification logistic regression has characteristic of parameter redundancy, which means the prediction could not be changed if $\theta_1, \theta_2, \dots, \theta_k$ adds a vector at the same time. Specifically, when there are two categories,

$$h_q(x) = \frac{1}{e^{q_1^T x} + e^{q_2^T x}} \begin{bmatrix} e^{q_1^T x} \\ e^{q_2^T x} \end{bmatrix} \quad (14)$$

By using characteristic of parameter redundancy, all parameter minus θ_1 , the above formula is changed to be as follows

$$h_q(x) = \frac{1}{e^{0 \cdot x} + e^{(q_2^T - q_1^T) \cdot x}} \begin{bmatrix} e^{0 \cdot x} \\ e^{(q_2^T - q_1^T) \cdot x} \end{bmatrix} = \begin{bmatrix} \frac{1}{1 + e^{q^T x}} \\ 1 - \frac{1}{1 + e^{q^T x}} \end{bmatrix} \quad (15)$$

Where $\theta = \theta_2 - \theta_1$ and arranged formula is same as binary classification formula. Therefore, multi-classification logistic regression is an extension of binary logistic regression.

III. EXPERIMENT

Before The goal of experiment is to evaluate precision and effectiveness of three classification algorithm. The experiment was completed with Python 3.6 under Window 10 system. The experimental computer configuration is: Intel core i5-7300U@2.6GHz CPU, DDR3 1866 MHz, 8GB Memory, 128G SSD solid hard disk. The experimental data is failure data in Chongqing from 2015 to 2018. The flowchart of the algorithm is as follows:

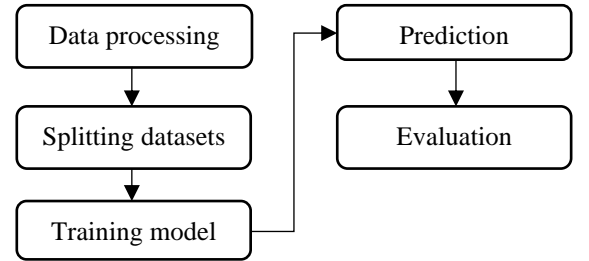


Fig.2 Algorithm flow

A. Units

Smart meter operation data for fault prediction is usually collected under unspecified conditions, so these low-quality data are difficult to use directly for fault prediction. The data set has a total of 20 attributes, representing different information (see Table 1).

Tab.1 Data set attribute

Char	Type of char
Factory info	CHAR
Batch info	CHAR
Inventory time	INTEGER

Installed time	DATE
Operation time	INTEGER
Num of Failure	INTEGER
Failure time	DATE
Full inspection	BOOLEAN
Sampling inspection	BOOLEAN
Running Sampling	BOOLEAN
location	CHAR
Failure causes	CHAR
Failure position	CHAR
Abnormal category	CHAR
Abnormal phenomenon	CHAR
Abnormal time	DATE
Accuracy level	CHAR
Average error	FLOAT
On-site inspection time	DATE
Basic error	FLOAT

B. Data processing

Since it is not possible to collect appropriate data for the operation of the smart meter information collection system in a short period of time, it is necessary to improve the smart meter data by means of preprocessing. Data cleaning needs to be performed before data preprocessing. Data cleaning mainly includes data filtering, data deduplication and accuracy improvement. Data preprocessing mainly missing value processing, continuous data normalization, discrete data encoding, feature selection.

Feature selection is usually also part of the pre-processing. Feature selection is independent of any machine learning algorithm. Features are selected based on correlations between features determined by scores in various statistical tests. Different types of data correlation coefficients can be defined according to the following table.

Tab.2 Feature selection method

Feature	Continuous feature	Discrete feature
Continuous feature	Pearson correlation coefficient	LDA
Discrete feature	ANOVA	chi-squared test

The fault type is classified as discrete data, and the remaining attributes can be divided into discrete data (inventory time, running time, average error, basic error, etc.) and continuous data (factory information, batch information, installation time, full inspection, random inspection, running sampling, location, phenomena and causes of problems, abnormal categories), so the selection of feature attributes is mainly divided into discrete types of faults and other discrete attributes and discrete types of faults and remaining discrete attributes Correlation.

The correlation between the discrete fault type and the remaining discrete attributes is selected by linear discriminant analysis. Analyze using linear discriminant analysis and summarize the results in the table below.

Tab.3 Vector after dimensionality reduction

Feature	Vector in dimension reduction direction
Inventory time	9.73990218e+09
Operation time	-1.43223707e+10
Average error	-1.31394626e+05
Basic error	5.68744983e+03

According to the principle that the corresponding feature vector has the best eigenvector segmentation performance, the four continuous attributes are highly correlated with the fault location. The correlation between the discrete fault type and the remaining discrete attributes is selected by chi-square analysis.

After calculating correlation between attributes and fault location the full inspection and the Sample inspection are not related to the fault location, because the full inspection and the sampling inspection are all qualified in the record system.

In this part, the smart meter data is preprocessed. Firstly, the whole of the smart meter is statistically analyzed. The frequency of different fault types is statistically analyzed. Then the data screening, missing value processing and continuous data normalization are done. And the normalization of discrete data, etc., to facilitate the subsequent data input into the machine learning algorithm. Finally, according to the characteristics of attribute categories in the classification attribute of fault data, this paper proposes a feature

selection method based on linear discriminant analysis and chi-square test. It selects the factory information, batch information, installation time, sampling, location, and appearance. The main attributes of the problem phenomenon and cause, abnormal category, abnormal phenomenon, abnormal time, active accuracy level, on-site inspection time, inventory time, running time, average error and basic error.

C. Evaluation indicator

The evaluation index is a quantitative indicator for the same data, inputting different algorithms, or inputting the same algorithm but different parameters to give the algorithm or parameters good or bad. By comparing the fault type of the test set with the matching rate of the fault type of the training set, the accuracy of the algorithm model prediction can be obtained. In addition, precision, recall and F1 score are also common methods for evaluating prediction results.

D. Experimental results and analysis

The date in paper is real failure date in Chongqing from 2016 to 2017 whose featured can be seen as Table 5. There are 24586 data after processing including factory info, batch info, bar code number, inventory time, operation time, running time, number of failure, failure time, location, failure causes, abnormal category, abnormal phenomenon, abnormal time, accuracy level, basic error etc. Data is split 75-25 into train and test datasets respectively in predictive modeling.

Tab.4 Intelligent Meter Fault Type Prediction Based on Three Classification Algorithms

Model	Precision	Recall	F1score	Sample size
SVM	0.96	0.96	0.94	6147
Native Bayes	0.98	0.98	0.98	6147
Logistic Regression	0.84	0.90	0.86	6147

The above are the results of three classification algorithms, in which the logistic regression performance is poor, while the support vector and naive Bayes perform better. From the point of view of accuracy, Naive Bayes

has the best effect, followed by support vector machine, but the difference is not much different from Naive Bayes, and the effect of logistic regression is the worst. Support vector machine and native Bayes have higher than logistic regression where the recall rates are same. Combined with the definition of F1 score, the classification method with high accuracy and recall rate, the score of F1 is also high. The results in Table 10 can also confirm this. Therefore, naive Bayes classification works best when the samples tested are the same.

IV. CONCLUSION

Combining the method of machine learning to analyze the fault data of smart meter can effectively predict the fault type of smart meter. The performance of different algorithms is different in the experiment. After experiment, naive Bayes is better as a traditional machine learning algorithm than the other two machine learning algorithms with an accuracy rate and a recall rate of 0.98.

ACKNOWLEDGMENT

This work is supported by the China State Grid Technology Project under Grant No. 5442JL170021

- [1] Wang Dewen, Yang Liping. Stream Processing Method and Condition Monitoring Anomaly Detection for Big Data in Smart Grid[J]. Automation of Electric Power Systems, 2016,40(14):122-128
- [2] Yan Yingjie, Sheng Gehao, Chen Yufeng etc. A Method for Anomaly Detection of State Information of Power Equipment Based on Big Data Analysis, 2015,35(1):52-59
- [3] Miranda ,G. Castro, S. Lima. Diagnosing Faults in Power Transformers With Auto associative Neural Networks and Mean Shift[J]. IEEE Trans on Power Delivery,2012,27(3):1350-1357
- [4] Luo Gang, Shi Dongyuan, Chen Jinfu, Duan Xianzhong. Automatic identification of transmission sections based on complex network theory[J].IET Generation Transmission & Distribution, 2013, 8(7): 1203-1210.1
- [5] Capgemini Consulting. Big data black out: are utilities powering up their data analytics [EB /OL],2016.
- [6] GTM Research. The soft grid 2013—2020: big data & utility analytics for smart grid [EB /OL],2016.
- [7] Yokoyama E, Uchimura M. Variable number of tandem repeats and pulsed-field gel electrophoresis cluster analysis of enterohemorrhagic Escherichia coli serovar O157 strains[J]. Journal of food protection, 2007, 70(11): 2583-2588.
- [8] Fan Shengping, Cao Shunan, Zhang Jin. Application of Fuzzy Math in Fault Diagnosis for Condenser [J]. Automation of Electric Power Systems, 2016,40(14):122-128
- [9] Jiang Tao. Reliability evaluation and selective maintenance for complex systems with inspection data[D]. University of Electronic Science and Technology of China,2017.
- [10] Guo C, Chai Y, Wang C. Multi-source heterogeneous data recognition based on linguistic labels[C]Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2016 International Conference on. IEEE, 2016: 278-285