Study on Data Selection Method of Historical Operation Data for Large Scale Power System

Hongyang Dai, Ying Lv, Zhihong Yu, Guangming Lu, Chang Xie and Jinxiu Hou
Power System Department, China Electric Power Research Institute, Beijing, China

Abstract. A data selection method based on similarity measurement and support vector machine (SVM) is proposed. At first, the critical clearing time (CCT) is used as the class label, and features which are strongly correlated with the class label will be extracted. Secondly, a SVM classifier is trained on the initial training instances with extracted features, and the instance which is misclassified will be removed. Thirdly, the concept of the most similar instance pair is proposed, which two instances with the minimum distance are selected, and then removes the eligible instances which is noisy and redundant instances. The proposed method which can simultaneously prune data in horizontal and vertical directions is tested by online historical data of an actual large scale power system. Experimental results demonstrate that more than 70% features and 30% instances are reduced, and the accuracy and storage reduction are also improved. This method can be well used with the good performance in large scale power system.

1. Introduction
With the rapid development of DC and AC hybrid grid, the power system becomes larger and more complicated, which brings profound influence to the operation stability of power grid [1]. It is necessary to continuously expand the performance of online security analysis, in order to meet the requirements of dispatching operation work adjusting to large scale power system. Recently, the big data technology has developed, which provides more means to solve the problems in many technical fields [2]-[4]. At present, a lot of dispatching operation data had been accumulated day by day in dispatching institutions. For example, the size of calculated data and result data which produced by the online security analysis can be amounted to 1GB in every 15 minutes. Such huge amount of data must be dealt with in learning process, which will cost massive computational resources and have long runtimes. Thus, data selection can be applied for reducing the data to a manageable size, leading to a reduction of the useless, erroneous or noisy data, before applying learning algorithms [5]-[8].

Data selection is one of the important pre-processing steps that can be applied in many data mining tasks [9]. This step aims at two aspects: (1) the data size can be reduced, leading to a reduction of the training time as long as the improvement of learning efficiency; (2) the noisy or erroneous data can be removed, leading to the improvement of the accuracy in classification problems. There are two common methods of data selection: feature selection and instance selection. Feature selection aims at reducing the number of power grid characteristics from longitudinal dimension of dataset; and instance selection aims at reducing the number of instances from latitudinal dimension of dataset, choosing a subset of the total available data to achieve the original purpose of the data mining application as if the whole data had been used.

In the last years, several approaches for data selection have been proposed in power system area. Most of methods are proposed to preserve the boundaries between different classes in the dataset, because of the relevant information provided by border instances for supporting discrimination between classes,
which destroy the structure of the data set and is bad for further analysis[10]-[11]. According to [12], [13], SVM (support vector machine) is usually used for classification with good performance. A segmentation method based on a two-level SVM is proposed [14], which reduce the negative effects of the mistaking instance. But the most useless and reductant instances are preserved with low reduction efficiency. Hybrid methods (e.g. RIPE) use ENN (Edited Nearest Neighbour) [15] to removing border points that are noisy aiming at smoothing the boundary, and use the concept of the nearest similar pair with the same class of datasets to removing redundant instances. The algorithm make better compromise in the classification accuracy and the storage compression ratio, but ignoring the problem of the nearest similar pair with the different class.

This paper is organized as follows: Section II presents the method of feature selection based on similarity measurement; Section III presents the method of instance selection combining the repetitive screening, SVM and similarity measurement; Section IV presents the experimental results, including the evaluation of the method and the comparison of three similarity measurement; finally, section V presents main conclusions of our work.

2. Feature selection base on correlative coefficient

When selecting the features to be processed, it is needed to select the features that are most related with the class label. In order to find the features having high relationship with the class label, the correlation coefficient between them is calculated. In the probability theory and statistics, the correlation coefficient reflects the strength and direction of the linear relationship between two variables, and the most commonly used is Pearson correlation coefficient. Pearson Correlation coefficient calculation formula is defined as follows.

$$\text{Cor}_{AB} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{\left(\sum_{i=1}^{n}(a_i - \bar{A})^2\right)^{1/2} \left(\sum_{i=1}^{n}(b_i - \bar{B})^2\right)^{1/2}}$$

(1)

Where A and B are two linear variables, N is the number of elements in A or B, $a_i$ is the value in A, $b_i$ is the value in B, $\bar{A}$ is the average of A, $\bar{B}$ is the average of B.

The range of the correlation coefficient is [-1, +1], the correlation coefficient greater than “0” represents that two groups of variables are positive correlation, and more closer to “+1”, more stronger correlation degree; on the contrary, the correlation coefficient less than “0” represents that two groups of variables are negative correlation, and more closer to “-1”, more stronger correlation degree; “0” represents that two groups of variables are uncorrelated. The correlation degree is determined by following scope of correlation coefficient showed in Table 1.

| Correlation coefficient | Correlation degree       |
|-------------------------|--------------------------|
| 0.8-1.0                 | Very strong correlation   |
| 0.6-0.8                 | Strong correlation        |
| 0.4-0.6                 | Middle degree correlation |
| 0.2-0.4                 | Weak correlation          |
| 0.0-0.2                 | Zero correlation          |

3. Instance selection based on SVM and similarity measurement

One method of instance selection often has limitations, Combination of several methods can give play to complementary advantages and make up for the shortcomings each other. The method of instance selection proposed in this paper was the combination of the SVM and similarity measurement, which can remove the noisy and useless instances. This method was mainly divided into three steps: repetitive screening, selection based on SVM, selection based on similarity measurement
3.1. Repetitive screening
The operation of the power system is periodic. The periodicity is relatively stable in a short period of time, and with the growth of time scale, the difference of operation mode for power system will inevitably increase. Features are reduced greatly after feature selection used to the initial dataset, the difference between two instances will be disappear, that there will be two or more instances of the same characteristics. Thus, the repetitive instances need to be deal with at first, for choosing one of the same instances and deleting the remaining instances.

3.2. Instance selection based on SVM
The support vector machine (SVM) is a widely used tool in classification problems. It trains a classifier by finding an optimal separating hyperplane which maximizes the margin between two classes of data in the kernel induced feature space. In learning by SVM, SVM calculates an alignment discernment line which maximize margin. SVM is excellent in generalization capability and it can extend to nonlinear by a kernel trick [13]-[14].

The instance was trained by SVM algorithm with the kernelled decision function represented as:

\[
f(x) = \text{sgn}[(w^T \cdot x + b) + \alpha \sum_{i=1}^{m} y_i K(x_i, x) + b']
\]

Where \(\text{sgn}\) is the sign function to determine the classification of the instance (for example, the calculation result is negative or positive which represent the different classification respectively), \(K\) is linear kernel, \(\alpha\) and \(b\) are the parameters occurred in training process, \(x_i\) is support vector, \(y_i\) is the class label, \(x\) is instance to be discriminated.

The Gaussian kernel is commonly used.

\[
K(x, x_i) = \exp\left(\frac{-||x-x_i||^2}{\sigma^2}\right)
\]

Where the parameter \(\sigma > 0\). Except the parameters in kernel function, \(c\) called cost parameter need to be specified in the training process which is a positive constant. The cost parameter that denotes the penalty of slacks can enhance the generalization ability of SVM algorithm.

It need to be clear that, the parameters of decision function are default in the model training process if using the mature data analysis tools, such as R programming language or MATLAB. Generally, \(c\) is equal to reciprocal of characteristics, and \(c\) is equal to 1 in SVM decision function.

The step of instance selection base on SVM is described below.

- **Step 1:** Labelled training set \(S\) is trained by SVM, finding an optimal separating hyperplane and learning a classification model.
- **Step 2:** \(S\) as the test set is classified by classification model learned in step 1, obtaining the classification result \(R\).
- **Step 3:** If the instance in test set classified by SVM differs from the class in the given training data, the instance need to be removed.
- **Step 4:** Step1 ~ Step3 is repeated until there is no incorrect discernment.
- The method is a process of removing instances iteratively, it can be used to remove noise instances, make the decision boundary more clearly.

3.3. Instance selection based on similarity measurement
The method in section B can effectively remove noise instances near the boundary, but also remain all of the internal instance, leading to unsatisfactory compression ratio of capacity. The instance selection method based on similarity measurement was described in this section.
First of all, a new concept was introduced: the most similar instances pair. For an instance A in dataset, computing the most similar instance B, if A is also the most similar instance to B, it can say that A and B are the most similar instances pair.

In this paper, three methods were used to calculating the most similar instance pair: Euclidean Distance, Hausdorff Distance \cite{15} and Correlation distance.

There are two n-dimensional instances \( A = (a_1, a_2 \ldots a_n) \) and \( B = (b_1, b_2 \ldots b_n) \).

- The Euclidean Distance is defined as follows:
  \[
  D_{AB} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}
  \]  
  \( (4) \)

- The Hausdorff Distance is defined as follows:
  \[
  H_{AB} = \max(h(A,B), h(B,A))
  \]  
  \( (5) \)

Where

\[
 h(a,b) = \max_{a \in A} \min_{b \in B} ||a-b||
\]  
  \( (6) \)

And \( ||a-b|| \) is the Euclidean distance.

- The Correlation Distance is defined as follows:
  \[
  D_{AB} = 1 - Cor_{AB}
  \]  
  \( (7) \)

Where, \( Cor_{AB} \) is the correlation coefficient defined in section 2.

The step of instance selection base on similarity measurement is described below.

- Step 1: the most similar instance \( x' \) of the instance \( x \) was calculate and marked, and repeat this step until all instance’s most similar instance were calculate and marked in dataset \( S \).
- Step 2: all the most similar instance pair were extracted after traversal of the instance \( x \) and its most similar instance \( x' \).
- Step 3: the class of each instance in the most similar pair was compared. If the same class, anyone of two instances was deleted; if not the same class, the two instances were deleted at the same time.

The method is a process of removing instances either closed to the boundary or internal instance far away from the boundary, it can be used to remove either noise instances or reductant instances.

4. Experiments

4.1. Initial Dataset

State variables of power system were selected as the analysis of the attribute in dataset, including the electric variables of equipment in static state and statistics of area and power plants and stations. Selecting static variables under the static state can shorten the time of stability judgement. The speed of stability judgement was reduced if transient variables were selected, which needed a period of time in transient stability simulation process. There was no definite conclusion of the relation between static variables and power system stability, so that more static variables were choose as far as possible under the premise of computing resources. The static variables and statistics are shown in Table 2. Critical clearing time (CCT) that represent the system boundary of stable and unstable was selected as class label to analysis the dataset. And it can characterize the stability degree when specified fault occurred in power system, the longer critical clearing time, the less impact to power system, and the system is more stable. System instability will be created by the specified fault if the critical clearing time is less than the normal protection operation time.
The historical data of online security and stability analysis in June was selected to be the instance of initial dataset in which 2484 instances and 9815 attributes existed at last. The information of initial dataset is shown in Table 3.

Table 2. The state variables and statistics.

| Equipment   | Variables             |
|-------------|-----------------------|
| AC line     | Active power/PAC      |
| DC line     | Power/ PDC            |
| Generator   | Power/ PG, Voltage/VG |
| Load        | Voltage/V_L           |

| Statistics  | Variables                      |
|-------------|--------------------------------|
| Area        | The sum of generators active output/ PAG |
|             | The sum of loads / PAL          |
|             | The average of voltage/ VA      |
| Plants and stations | The sum of generators active output/ PTG |
|             | The number of operated generators/ NTG |
|             | The sum of loads/ PTL           |

Table 3. The Information Of Initial Dataset.

| Dataset | Instances | Attributes | Class |
|---------|-----------|------------|-------|
| Number  | 2484      | 9815       | 9     |

In order to verify the validity of this method, the data set was divided into training set and test set before using algorithm. So that, 80% instances were randomly selected in each class as the training set, and the remaining instances as test set. 10 kinds of random numbers were used to get the result after calculating the average as the final result of this method.

4.2. Results of Feature Selection

Features that beyond middle degree correlated to the CCT were selected for subsequent analysis by the method in section 2, and features that weak correlated and zero correlated to the CCT were removed. Then, the correlation coefficient that was equal to 0.4, 0.5, and 0.6 was selected to be threshold respectively, the number of features and accuracy after feature selection were compared in Figure 1.

![Figure 1. Comparison of different threshold.](image)
Figure 1 shows that with the increase of the correlation coefficient threshold, the number of features reduced sharply, but the change of the classification accuracy was more complex. Thus, 0.5 has been used as threshold to select features, and 574 features were selected in the end for the further analysis.

4.3. Results of instance selection
The instance selection method proposed in this paper was applied, and Table 4 shows the results. Except accuracy and reduction to evaluate the performance of instance selection method, effectiveness was introduced in this paper which has a comprehensively assessment. Thus, we consider effectiveness equals accuracy multiple reduction. In Table 4, No. represents ten random numbers; Accuracy represents the classification result of dataset without instance selection; Accuracy1 represents the classification result of dataset after using the instance selection based on SVM and Reduction1 represents its selection effect; Accuracy2 represents the classification result of dataset after using the combined instance selection based on SVM and similarity measurement and Reduction2 represents its selection effect.

| No. | Accuracy | Accuracy1 | Reduction1 | Accuracy2 | Reduction2 |
|-----|----------|-----------|------------|-----------|------------|
| 1   | 0.8864   | 0.8783    | 0.1070     | 0.8783    | 0.3842     |
| 2   | 0.8945   | 0.8986    | 0.1236     | 0.8945    | 0.4043     |
| 3   | 0.8945   | 0.9047    | 0.1135     | 0.8945    | 0.4018     |
| 4   | 0.8925   | 0.8884    | 0.1120     | 0.8864    | 0.3918     |
| 5   | 0.9006   | 0.8966    | 0.1226     | 0.9026    | 0.4028     |
| 6   | 0.8864   | 0.8844    | 0.1130     | 0.8864    | 0.3998     |
| 7   | 0.8864   | 0.8966    | 0.1215     | 0.8884    | 0.3988     |
| 8   | 0.9249   | 0.9229    | 0.1190     | 0.9128    | 0.4013     |
| 9   | 0.9067   | 0.9087    | 0.1175     | 0.9067    | 0.3968     |
| 10  | 0.9006   | 0.9026    | 0.1190     | 0.9047    | 0.3953     |
| Average | 0.8974 | 0.8982 | 0.1169 | 0.8955 | 0.3977 |

effectiveness -- 0.1050 0.3561

Table 4 shows that instance selection based on SVM has a good performance in terms of accuracy and the instance selection based on SVM and similarity measurement has a good performance in terms of reduction. Through the comprehensive comparison of effectiveness between two methods, the second method is better than the first one which prove the validity and efficiency of the method proposed in this paper.

4.4. Comparison of three similarity measurement

![Figure 2. Comparison of accuracy.](image)
To find a better similarity measurement of instance selection, the performance of three methods including Euclidean Distance, Hausdorff Distance and Correlation distance were compared considering the indexes: accuracy, reduction and effectiveness. The Figure 2, 3 and 4 show, respectively, the resulting indexes of each method.

Considering the accuracy, although there is little difference between three methods, the performance of Hausdorff Distance and Correlation distance are little better than Euclidean Distance. Once considering the reduction, the reduction of Euclidean Distance is much better than others. Thus, Figure 4 shows that, after comprehensive comparison of effectiveness, Euclidean Distance can better applied to instance selection.

5. Conclusions

In this paper, we have presented a data selection method based on similarity measurement and SVM. This methodology efficiently deals with the feature selection and instance selection. The experiments have shown that the method can remove internal reductant instances and noisy instances, not only keep the distribution of dataset, but also meet the requirement of data selecting. The comparison of three similarity measurements shows that Euclidean Distance can well applied to instance selection. In conclusion, the data selection method proposed is valid and effective, with the additional advantage of being able to scale up to large datasets.

6. References

[1] LI Mingjie, "Characteristic Analysis and Operational Control of Large-Scale Hybrid UHV AC/DC Power Grids," Power System Technology, vol. 40, pp. 985-991, Apr. 2016.
[2] XUE Yusheng, LAI Yening, “Integration of Macro Energy Thinking and Big Data Thinking Part One-Big Data and Power Big Data,” Automation of Electric Power Systems, vol. 40, pp. 1–8, Jan. 2016.
[3] XUE Yusheng, LAI Yening, “Integration of Macro Energy Thinking and Big Data Thinking Part Two-Applications and Explorations,” Automation of Electric Power Systems, vol. 40, pp. 1–13, Apr. 2016.
[4] PENG Xiaosheng, DENG Diyuan, et al, “Key Technologies of Electric Power Big Data and Its Application Prospects in Smart Grid,” Proceedings of the CSEE, vol. 35, pp. 503–511, Feb. 2015.
[5] S. Garcia, J. Luengo, and F. Herrera, Data preprocessing in data mining, Springer, 2015.
[6] Noor Izzri Abdul Wahab et al. “Transient Stability Assessment Of A Large Actual Power System Using Least Squares Support Vector Machine With Enhanced Feature Selection”, 2008 Australasian Universities Power Engineering Conference, pp. 1-6.
[7] Rui Zhang, Yan Xu, Zhao Yang Dong, David J Hill. “Feature Selection for Intelligent Stability Assessment of Power Systems”, 2012 IEEE Power and Energy Society General Meeting, pp. 1-7.
[8] Oveis Abedinia, et al. “A New Feature Selection Technique for Load and Price Forecast of Electrical Power Systems”, IEEE Transactions on Power Systems, vol. 35, pp. 1-11, Apr. 2016.
[9] E. Leyva et al. “Three New Instance Selection Methods Based on Local Sets: A Comparative Study with Several Approaches from A Bi-Objective Perspective”, *Pattern Recognition*, vol. 48, no. 4, pp. 1523-1537, 2015.

[10] K. Nikolaidis, J.Y. Goulermas, and Q. Wu. “A Class Boundary Preserving Algorithm for Data Condensation”, *Pattern Recognition*, vol. 44, no. 3, pp. 704-715, 2011.

[11] D.L. Wilson, “Asymptotic Properties of NearestNeighbor Rules Using Edited Data”, *IEEE Trans. Systems, Man, and Cybernetics*, vol. 2, no. 3, pp. 408-421, July 1972.

[12] D.R. Wilson, and T.R. Martinez, “Reduction Techniques for Instance-Based Learning Algorithms”, *Machine Learning*, vol. 38, no. 3, pp. 257-286, Mar 2000.

[13] V.N. Vapnik, *Statistical Learning Theory*. John Wiley and Sons, 1998.

[14] Nello Cristianini, John Taylor, *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.

[15] Huttenlocher D P, Klanderman G A, Rucklidge W J. Comparing images using the Hausdorff distance. *IEEE Transactions on PAMI*, vol. 9, pp. 850-863, 1993.

7. Acknowledgments

Project supported by Science and Technology Projects of SGCC: Research on online early warning and fault disposal decision technology for multiple serious faults in large power grid.