Liliana BYCZKOWSKA-LIPIŃSKA [1], Agnieszka WOSIAK[2]

University of Computer Sciences and Skills (1)
Institute of Information Technology, Lodz University of Technology (2)

doi:10.15199/48.2015.01.39

Feature Selection and Classification Techniques in the Assessment of the State for Large Power Transformers

Abstract. *The use of automatic feature selection and state classification of power transformer has been proposed in this research. The experimental studies were carried out on real data. Feature selection was performed using the CFS, InfoGain and ReliefF methods. The classification was carried out using C4.5, k-NN, SMO and AdaBoost algorithms. Experimental studies have proved that the automatic classification allows to obtain results comparable to the classification carried out by experts - regardless of the method the number of cases correctly classified was above 90%.*

Streszczenie*. W ramach niniejszej pracy zaproponowane zostało zastosowanie selekcji cech i automatycznej klasyfikacji stanu transformatora energetycznego. Badania przeprowadzone zostały na rzeczywistych danych. Selekcję cech przeprowadzono z zastosowaniem metod CFS InfoGain i ReliefF. Proces klasyfikacji przeprowadzony został za pomocą algorytmów C4.5, k-NN, SMO i AdaBoost. Badania eksperymentalne wykazały, że automatyczna klasyfikacja danych pozwala na uzyskanie rezultatów porównywalnych do klasyfikacji przeprowadzonej przez ekspertów - niezależnie od zastosowanej metody liczba przypadków zaklasyfikowanych poprawnie wyniosła powyżej 90%. (**Techniki selekcji cech i automatycznej klasyfikacji w procesie oceny stanu transformatora energetycznego**).*

Słowa kluczowe: eksploracyjna analiza danych, selekcja cech, klasyfikacja, transformator energetyczny
Keywords: data mining analysis, feature selection, classification, large power transformer

Introduction

The failure of power transformer causes the sequence of events within the device. These events are in turn associated with a number of the signals collected from the measuring devices. As a result we obtain a considerable amount of measurement data that need to be analyzed in order to assess the type, location and degree of abnormal operation of the transformer.

Most of the interpretative analysis is performed based on the knowledge of experts whose experience points how changes in the measurement parameters affect the state of the transformer. The most important and most difficult step in the implementation of these solutions is to transform the knowledge of experts into the appropriate decision rules and to define the membership function. The diagnosis posed as a result of expert systems require completeness of representation of the knowledge gathered by the system.

For the assessment of the device state, the analysis covers a range of measured values derived from sensors. The total number of parameters usually exceeds 100, often even 200. The reduction of dimensionality of the data allows for a faster assessment of the condition and diagnosis for the analyzed object. The reduction can be carried out both by the techniques of data extraction and feature selection [1].

Feature extraction is the transformation process that projects the set of all input features onto a space of lower dimension. Feature selection means choosing a subset of the original feature set based on a metric that usually refers to the quality of data. Selecting a subset of all available features before applying learning algorithm, is a common technique for simplifying or speeding up computations [2].

The use of the classification in the assessment of power transformer state allows you to automate the process of monitoring its condition. It provides support for the transformer station service stuff, as it allows for the concise results of the classification status of the device without the need for analyzing large amounts of measurement data and associated messages.

In this paper the use of automated classification of a power transformer state has been proposed as an alternative to previously implemented expert system. An important initial step in the analysis is the process of automated feature selection in order to reduce the dimension of the input data. The tests were carried out on real data gathered from the transformer station in Piotrkow.
The remainder of this paper is organized as follows. Section 2 presents literature review concerning data mining techniques applied in the analysis of the states of power transformers. Next section concerns the description of the proposed methodology. In Section 4 we describe the studies that were conducted. We introduce data collected for this application and discuss the results. Finally, in Section 5 we draw the conclusions.

Related Works

The assessment of the state of power transformer is a very important issue, affecting daily functioning of every person and all branches of the economy. Therefore many researches are taken to improve the functionality of the systems supporting service transformer stations and facilitating the process of analyzing signals from various devices.

In [3] the state-of-the-art methods of diagnostics of power transformers and fault detection were presented. The research reviewed computational intelligence (CI) approaches for oil-immersed power transformer maintenance reported in international journals including automatic classification methods with particular emphasis on support vector machine classifier (SVM) and k nearest neighbor (kNN) classification.

The research described in [4] investigated the benefits of using feature selection based on mutual information in power system state classification with machine learning. In the paper, the AdaBoost algorithm was used for classification based on large training datasets and feature selection was applied in order to reduce their dimensionality. The feature selection was implemented as a filter in the pre-processing stage of AdaBoost and used genetic algorithms to perform the search with the fitness function computed based on mutual information. The number of features chosen in the selection process was about 60% of the initial set of all attributes, which in turn reduced the calculation time by 22%.

The authors of [5] used Supporting Vector Machine (SVM) to solve the problem of multi-classification for small number of samples and non-linear data in transformer oil chromatography fault diagnosis. They proposed to use genetic algorithm (GA) to select SVM.

Methodology

The power transformer state assessment is carried out based on a series of measurements, which are interconnected with particular dependencies and have to give a snapshot of the diagnosed object. The device state monitoring systems are mainly based on the expert knowledge, which allows for building the appropriate diagnostic rules. The project and implementation of such a system were the subject of our previous work described in [6, 7], and its logical structure is shown in Fig. 1.

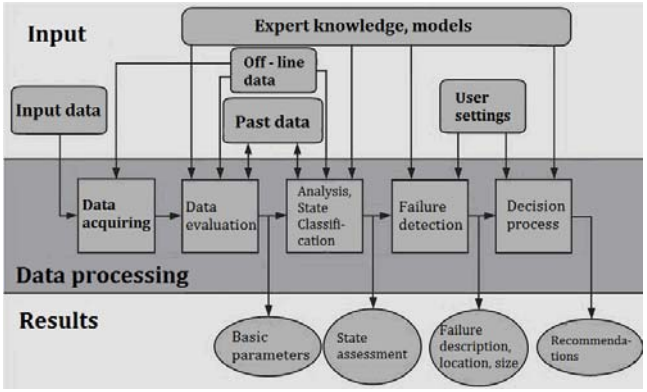

Fig. 1. The logical structure of the transformer state monitoring system [7].

The data mining techniques enable to automate some steps of data analysis or allow them to improve. The data mining can be defined as the process of selection, screening and modeling of large data sets in order to provide useful results of the final analysis [8]. Depending on the purpose of the exploration process, it is necessary to choose the appropriate method of exploratory data analysis.

In this paper we propose the use of automatic classification as an alternative to expert rules. In order to improve the classification, feature selection is performed as a preliminary step of analysis. The diagram of the proposed monitoring system incorporating feature selection and classification is shown in Fig. 2.

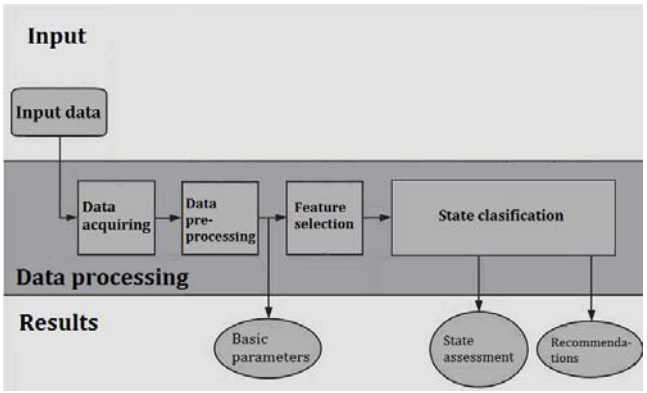

Fig. 2. The diagram of the monitoring system incorporating feature selection and classification [Source: Own work].

According to the proposed methodology (Fig. 2), the first step in the process of the transformer state assessing is the data preprocessing by the reduction of the input data using extraction (transformation). It involves the projection of input data onto the space of lower dimension, thereby speeding up the process of analysis. The most commonly used feature extraction methods include the Principal Component Analysis (PCA) and the Independent Component Analysis (ICA) [1]. In our proposed solution the goal of the transformation is to remove these features from an input set of all parameters that are not informative for the transformer state changes in the training set.

In the next step (Fig. 2) feature selection is performed. Feature selection process is to extract such a subset S of all available parameters D, which best represents the entire data collection and do not adversely affect the informativity of data and the accuracy of the calculations. Choosing the right method of feature selection is a crucial problem for the further analysis. The literature sources describe various algorithms and do not allow to draw conclusions for the ranking of solutions. It is also not possible to determine which method is best to keep in a specific case. Therefore it is necessary to compare the methods in particular applications, since different methods give good results for some data sets, but for other data may be much less suitable [9].

The classification of a transformer state - the third step (Fig. 2) - may relate to distinguish two classes: safe / unsafe or more classes, for example, extraction four states: normal, warning, alarm and emergency. Regardless of the number of these classes, the task of automatic classification techniques is to convert the input data, to summarize their characteristics, and identify key information related to operating conditions [10, 11]. The choice of an appropriate automatic classification algorithm, as in the case of automatic selection of data, usually results from empirical studies [4].

Experimental Analysis and Results

The experimental studies were conducted for the data collected from the diagnostic for the state of power transformers operating at the station in Piotrków [6]. In the previous solution the assessment of the state was based on expert knowledge and the process of the analysis took into account all the parameters obtained from the sensors and dissolved gas analysis (DGA).

The experimental studies aimed at:

- verifying the efficacy of the use of automatic classification for a state of a transformer as an alternative to the previously implemented expert system,
- extracting such subset of attributes from the whole set of gathered parameters that identifies a group of devices in a particular state,
- identifying the best pairwise combination of algorithms for feature selection and classification dedicated to the data from the real-source transformer station.

The input set of data consisted of 2172 cases described using 172 attributes. As a result of preprocessing, the attributes that were not related to the classification of a state of a transformer have been removed. Consequently the data set was limited to 82 attributes describing the state of the transformer.

The feature selection was carried out using the following methods:

- correlation- based feature selection (CFSSubsetEval - CFS and CorrelationAttributeEval - CAE algorithms),
- information-gain (InfoGain),
- ReliefF.

The classification process was performed using four specified algorithms:

- J48 (C4.5),
- K-nearest neighbors (IBk),
- sequential minimal optimization (SMO),
- AdaBoost.

The goal of the classification was to distinguish one of the four states of power transformer condition: normal, warning, alarm and emergency.

All experiments were conducted using Weka data mining tools by Machine Learning Group at the University of Waikato [12, 13].

The evaluation of the efficacy of each pairwise combination of feature selection algorithms and classification algorithms was based on the following criteria: accuracy, sensitivity (recall = true positive rate), specificity, false positive rate (FP rate), precision, root mean square error and number of features. Most of these factors are dependent on the ratio between correctly classified cases in comparison to the model results.

These cases can constitute four types of collections:

- TP (True Positives) - cases correctly classified by the algorithm as positive results,
- FP (False Positives) - cases incorrectly classified by the algorithm as positive results,
- TN (True Negatives) - cases correctly classified by the algorithm as negative results,
- FN (False Negatives) - cases incorrectly classified by the algorithm as negative results.

As a preliminary step of the analysis and the point of reference for later use of feature selection, we carried out the classification using the whole set of attributes, without excluding any of them from the process of analysis. The results are summarized in Table 1.

Table 1. The results of classification methods using all features.

Method	Precision	Recall	RMS Error	Time to build (s)
C4.5	0.888	0.923	0.1765	0.87
k-NN	0.904	0.914	0.145	0.01
SMO	0,858	0,905	0.3255	9.49
Ada Boost	0.811	0.901	0.2275	0.38

To summarize the values of precision values (column 2.), recall (column 3.) and the mean square error (column 4.) shown in Table 1, we can conclude that the best results were obtained for classification using k-NN. Moreover, this method required a very short calculation time - in comparison to the SMO method, the difference was significant (0.01 vs. 9.49).

The results of classifications using different methods, taking into account the preceding feature selection are presented in Table 2. (CFS selection), Table 3. (CAE selection), Table 4. (Infogain algorithm) and Table 5. (ReliefF algorithm). The tables consists of five columns which represent: the name of the method of classification, accuracy, recall, mean square error and computation time.

It is worth to remark that the subsets of features formed by the different feature selection algorithms were partially disjoint, that means methods indicated various attributes, both in terms of their importance, as well as their number: from 9 attributes for CFS feature selection, by 11 attributes for CAE algorithm and 18 parameters for InfoGain method up to 37 features selected by ReliefF algorithm.

Table 2. The results of classification methods after CFS feature selection (No of features = 9).

Method	Precision	Recall	RMS Error	Time to build
C4.5	0,870	0,909	0.1961	0.07
k-NN	0,901	0,912	0.1448	< 0.01
SMO	0,811	0,901	0.3266	0.66
Ada Boost	0,811	0,901	0.2275	0.06

Table 3. The results of classification methods after CAE feature selection (No of features = 11).

Method	Precision	Recall	RMS Error	Time to build
C4.5	0,879	0,917	0.1848	0.11
k-NN	0,900	0,911	0.1456	< 0.01
SMO	0,864	0,906	0.3257	0.66
Ada Boost	0,811	0,901	0.2166	0.06

Table 4. The results of classification methods after InfoGain feature selection (No of reatures = 18).

Method	Precision	Recall	RMS Error	Time to build
C4.5	0,870	0,912	0.1945	0.23
k-NN	0,899	0,911	0.1479	< 0.01
SMO	0,811	0,901	0.3266	1.43
Ada Boost	0,811	0,901	0.228	0.07

Table 5. The results of classification methods after ReliefF feature selection (No of features = 37).

Method	Precision	Recall	RMS Error	Time to build
C4.5	0,883	0,914	0.1854	0.24
k-NN	0,902	0,913	0.1444	< 0.01
SMO	0,827	0,899	0.3261	4.78
Ada Boost	0,811	0,901	0.2275	0.1

The case study results shown in Tables 2.-5. in respect to the objectives proved that the automatic classification of data allows to obtain results comparable to the classification carried out by experts - regardless of the method the number of cases classified correctly was above 90%. Use of data selection as a preliminary stage of the classification process has not affected the accuracy of classification. Still more than 90% of the cases was classified in accordance with the guidelines of the experts. Slightly lower (89.90%) results were obtained only for a pairwise combination of selection performed by ReliefF and SMO classification algorithm. It is also worth noting that regardless of the feature selection algorithm, the best results were obtained using the k-NN method of classification (IBk algorithm).

Conclusions

The research described in the paper proved the efficacy of the proposed methodology. The assessment of the state of the transformer carried out by automatic methods provided the correct classification results in over 90% of cases. The analyzed space of attributes was automatically reduced by feature selection methods, which had a positive effect on the time of analysis without any loss arising from lack of messages about changes in the state of the transformer.

Further work will be associated with the use of other data mining techniques aimed at more efficient analysis of the collected data, primarily related to the selection of features and classification. In particular, it is worth considering the use of genetic algorithms, which according to the literature review, can give good results [14, 15, 16].

REFERENCES

[1] FODOR I.K.: A survey of dimension reduction techniques, Technical Report, Lawrence Livermore National Laboratory, US Department of Energy, 2002

[2] GUYON I., ELISSEEFF A.: An introduction to variable and feature selection. Journal of Machine Learning Reseach, 3, 1157–1182, 2003

[3] SUN H.-Ch., HUANG Y.-Ch., HUANG Ch.-M.: Fault Diagnosis of Power Transformers Using Computational Intelligence: A Review, Energy Procedia No 14, pp. 1226 – 1231, 2012

[4] PISICA I., TAYLOR G, LIPAN L.: Feature selection filter for classification of power system operating states, Computers and Mathematics with Applications 66 (2013) 1795–1807
[5] HAN Han, WANG Hou-jun, DONG Xiucheng: Transformer Fault Dignosis Based on Feature Selection and Parameter Optimization, Energy Procedia No 12, pp. 662 – 668, 2011, DOI:10.1016/j.egypro.2011.10.090
[6] BYCZKOWSKA-LIPIŃSKA L., WOSIAK A.: System diagnostyczny do oceny stanu pracy transformatora energetycznego, Przegląd Elektrotechniczny 12/2006 pp. 137-140, 2006
[7] BYCZKOWSKA-LIPIŃSKA L., WOSIAK A., KAŹMIERSKI M., KERSZ I.: Korzyści ekonomiczne wynikające z zastosowania systemów monitoringu transformatorów energetycznych, Przegląd Elektrotechniczny 12/2007, pp.101-104, 2007
[8] GIUDICI P.: Applied Data Mining Statistical Methods for Business and Industry, Wiley & Sons, 2003
[9] LAL T. N., CHAPELLE O., WESTON J. ELISSEEFF A.: Embedded Methods, In: Feature Extraction, Foundations and Applications (Eds. Isabelle Guyon, Steve Gunn, Masoud Nikravesh, Lotfi A. Zadeh), ISBN-10 3-540-35487-5, Springer-Verlag Berlin Heidelberg 2006, pp. 137-165
[10] PISICA I., EREMIA M.: Making smart grids smarter by using machine learning, In: 46th International Universities' Power Engineering Conference, UPEC, 2011, pp. 1–5.
[11] JAZEBI S., VAHIDI B., JANNATI M.: A novel application of wavelet based SVM to transient phenomena identification of power transformers, Energy Conversion and Management 52 (2011) pp. 1354–1363, 2011
[12] HALL M., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN R., WITTEN I.H., The WEKA data mining software: an update, SIGKDD Exploration Newsletter vol. 11, 2009, pp. 10-18, DOI: 10.1145/1656274.1656278
[13] WITTEN I. H., FRANK E.: Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, Morgan Kaufmann, 2011
[14] FEI S.W., ZHANG X.B.: Fault diagnosis of power transformer based on support vector machine with genetic algorithm, Expert Systems with Applications 36 (2009) 11352–11357
[15] SHINTEMIROV A., TANG W., WU Q.H.: Power transformer fault classification based on dissolved gas analysis by implementing bootstrap and genetic programming, IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews 39 (January (1)) (2009) 69–79
[16] FEI S.W., LIU C.I., MIAO Y.B.: Support vector machine with genetic algorithm for forecasting of key-gas ratios in oil-immersed transformer, Expert Systems with Applications 36 (2009) 6326–6331.

***Authors**: prof. dr hab. inż. Liliana Byczkowska-Lipińska,, University of Computer Sciences and Skills, ul. Rzgowska 17 a, 93-008 Lodz, Poland e-mail: liliana.byczkowska-lipinska@p.lodz.pl, dr inż. Agnieszka Wosiak, Politechnika Łódzka, Instytut Informatyki, ul. Wólczańska 215, 90-924 Lodz, Poland, e-mail: agnieszka.wosiak@p.lodz.pl*