A FRAMEWORK FOR KNOWLEDGE ACQUSITION SYSTEM IN PERSPECTIVE VIEW OF DIAGNOSTIC OF ROTATING MACHINERY*

Dominik WACHLA

Silesian University of Technology, Department of Fundamentals of Machinery Design
Konarskiego 18A str., 44-100 Gliwice, Poland
e-mail: dominik.wachla@polsl.pl

Summary

A concept of knowledge acquisition system for the needs of diagnostic of rotor machines was presented in the article. The concept was developed on assumption that knowledge would be acquired inductively through analysis of measure and simulative data. The founding of the system was considered, The architecture was particularly described and example of its application was provided, as well.

Keywords: knowledge acquisition, diagnostic of rotating machinery, databases, support vector machines.

SYSTEM POZYSKIWANIA WIEDZY Z PERSPEKTYWY DIAGNOSTYKI MASZYN WIRNIKOWYCH

Streszczenie

W artykule przedstawiono koncepcję systemu pozyskiwania wiedzy dla potrzeb diagnostyki maszyn wirnikowych. Koncepcję opracowano przyjmując założenie, że wiedza będzie pozyskiwana w sposób indukcyjny poprzez analizę danych pomiarowych lub symulacyjnych. Omówiono genezę powstania systemu. Szczegółowo opisano architekturę oraz pokazano przykład zastosowania.

Słowa kluczowe: pozyskiwanie wiedzy, diagnostyka maszyn wirnikowych, bazy danych, metoda wektorów wspomagających.

1. INTRODUCTION

Knowledge in technical diagnostics is a knowledge which corresponds to relations occurring between observed symptoms and determined classes of technical state of a machine. Currently conducted research concentrates on acquisition of the knowledge with application of adequate numerical models of machines. Diagnostic relations (*technical state→ symptom*) are the result of analysis of a model reaction to given changes of values of control parameters. Provision of a large amount of data is the basic feature of knowledge acquisition based on a model approach. It results in an adequately considerable number of diagnostic relations of which only few have general meaning. As a consequence, all defined relations have to be considered in the process of diagnosing a machine, which may decrease the efficiency and effectiveness of the process. Due to the above, another concise form of description of these relations is necessary. The relations may for instance be represented in a form of decision tables or decision trees, neural models, neuro-fuzzy models, and others. Obtaining a description in one of the listed forms requires application of adequate methods of their identification, which, in turn, results in preparation and use of appropriate tools.

2. A FRAMEWORK OF THE SYSTEM

The architecture of the developed system of diagnostic knowledge acquisition was based on a two-fold model. The first layer is the base layer. Its purpose is to ensure realization of tasks concerning low-level computations and data management. The base layer consists of MATLAB environment and MySQL Database Management System. The two mentioned elements were chosen due to criteria formulated at the stage of the system requirements specification.

The database system supports management of big sets of data and constitutes the second element of the base layer. Due to its popularity and availability, MySQL system was applied in the proposed architectural solution. The communication and the exchange of data between MATLAB and MySQL is ensured by SQL interface.

A layer of implementation of algorithms constitutes the second layer of the system architecture. This layer aims at providing functionality which facilitates realization of the process of database knowledge acquisition. According to the CRISP-DM methodology [6], the process of database knowledge acquisition was divided into a number of stages e.g: data preprocessing, feature extraction and selection, data modeling, etc. Having taken the above information into account, 4 routine toolboxes were distinguished in the layer of algorithm implementation. The

* This work was supported by the polish Ministry of Science and Higher Education (grant No. PBZ-KBN-105/T10/2003)

routine toolboxes provide functionality of the system within: data management, data preprocessing, feature/attribute selection, data modeling.

A number of supporting procedures was implemented within the toolboxes of data management. The procedures involved:

- entering source data from the files generated by the NLDW-MESWIR system [5]
- entering and saving data in the binary files in the internal format of the MATLAB environment (*mat-files*),
- entering and saving data from/into the files in the XML format,
- communication and exchange of data with the MySQL database server.

The entering and reading of data from/into the XML files from the level of the MATLAB environment is realized by means of praser's XML language. The SQL interface allows for direct communication and exchange of data between MATLAB and the database. The interface functionality facilitates data management through formulation of questions which are compatible with semantics and specificity of the SQL language.

A toolbox of data preprocessing includes, in the first place, routines which support preparation of the data acquired from the NDLW-MESWIR system. In particular, these are the routines of computation of Mean Absolute Value (AVE), Room Mean Square Value (RMS), amplitude -phase spectrum, etc., and routines of three- and five-point differentiation of time series. In addition, the toolbox renders available a procedure of discretization of continuous attributes according to the algorithm which was introduced by Fayyad and Irani [2]. This procedure is used among others by algorithms of feature selection [4].

A metaprocedure constituting implementation of the CFS algorithm (Correlation-base Feature Selection) [4] is the basic element of a toolbox of features/attributes selection. The functioning of the CFS algorithm is based on the search of feature space along with consideration of criterial function which provides quantitative information about the importance of the chosen subsets of features. The operation of the CFS algorithm requires implementation of at least one method of the search of state space. Due to this requirement, the feature selection toolbox provides implementation of two algorithms of the search of state space, i.e. the *Best First* algorithm [7] and the S*imple Genetic Algorithm* [3].

A toolbox of algorithms of data modeling is a next element of the implementation layer. The procedures implement algorithms of the Support Vector Machines (SVM) method [11], i.e. C-SVM, ν-SVM, ε-SVR and ν-SVR [9]. The C-SVM and ν-SVM algorithms are applied in construction of classifying models, whereas the ε-SVR and ν-SVR algorithms are used in modeling of regressive problems. For the needs of the system, the *LibSVM* [1] toolbox was adapted; the toolbox includes implementations of the mentioned SVM algorithms. The other unit of data modeling supports validation of generated models through: routines which determine values of model statistics such as *classifier performance* [10], and a set of routines which implement some techniques of model evaluation e.g. *k-fold cross-validation* [8].

3. EXAMPLE OF SYSTEM APPLICATION

To show capabilities of the system, a problem of detecting and locating cracks in the shafts of a high power turbine unit was considered. Moreover, it was assumed that the constructed classifier should locate the cracks in the shafts with accuracy of a selected stage of the turbine set.

The solution to the established task required realization of a number of ordered activities which resulted from the methodology of data knowledge acquisition. In particular:

- acquisition of a set of source data,
- definition and determination of features from the set of source data,
- construction of a set of learning examples,
- identification and validation of the classifier.

3.1. The Source Data

The needed data were obtained from a number of simulations conducted with an application of a numerical model of a 200 MW turbine set. The model was constructed at IMP PAN in Gdańsk. It was built as a FEM model of a TK 7 turbine unit at Kozienice S.A power plant. The model was adjusted to a real object on the basis of data gathered by the DT200-1 diagnostic system [5]. In order to construct a model of a turbine set, the NLDW-MESWIR system was used [5].

On the basis of the objectives of the research, a plan of generating learning data was prepared. A set of 92 cases which include a process of cracking in the turbine set in 4 locations was taken into account. The 4 locations include: a high pressure stage (HP), an intermediate pressure stage (IP), a low pressure stage (LP) and a generator (GEN). The set was supplemented with a base case which was interpreted as a reference state in which no defects occur.

A detailed plan of generating learning data was presented in the Tab. 1. A simulative experiment was carried out for each planned case. The results of calculations conducted with the use of the NLDW-MESWIR system were copied into text files.

Tab. 1. The plan of generating source data: α_p-angular position of the crack, Wp - non-dimensional coefficient of the depth of the crack

Defect name	Element No.	α_p	W_p	Example No.	Class label
Base case	—	—	—	1	*PBARTMAX*
Cracks in HP stage	31	270	{0.050, 0.075, 0.100, …, 0.650}	2÷26	*CRHB*
Cracks in IP stage	59	0	{0.050, 0.075, 0.100, …, 0.650}	27÷51	*CRIDA*
Cracks in LP stage	103	90	{0.050, 0.075, 0.100, …, 0.475}	52÷69	*CRLFB*
Cracks in generator	125	125	{0.050, 0.075, 0.100, …, 0.600}	70÷92	*CRGBC*

3.2. Learning Data

In diagnostic of rotating machinery, the fundamental evaluation of a technical state is based on analysis of chosen frequency components of amplitude-phase spectra of mechanical vibrations [5]. The frequency components are connected with rotation velocity of a rotor. In particular, the elements 0.25X, 0.33X, 0.5X, 1X, 2X, 3X and 4X are evaluated; where X in [Hz] denotes a nominal frequency of rotation of a rotor.

Having taken the above into account, it was assumed that the learning examples would be constructed as sets of a selected frequency components of amplitude-phase spectra defined for vibration signals which were obtained at the stage of generating source data.

Two sets of learning examples were established. A division criterion was applied in relation to the signal category. The first set of learning examples was defined for the velocity of absolute vibrations fixed in the *x* and *y* directions. The second set was established for relative vibrations fixed in the x and y directions, as well. In each sets of learning examples, 196 features were marked.

The research objectives, i.e. classifiers induction, require a definition of a set of adequate categories for the considered problem. On the basis of the plan of generating learning data (tab. 1), 5 categories were defined and labeled as: *CRHB*, *CRIDA*, *CRLFB*, *CRGBC* and *PBARTMAX*. Then two complete sets of learning examples were obtained.

3.3. Feature Selection

According to the methodology of knowledge discovery from databases, a stage of identifying and verifying the classifier should be preceded by a stage of the relevant feature selection. Such a procedure aims at acquiring models which are characteristic of decreased complexity as well as increased generalizing capabilities. For this reason, the process of relevant feature selection was conducted using CFS algorithm [4]. The scheme of feature selection consisted of 11 experiments where the following algorithms were used for searching of feature space:

- The *Best First* algorithm [7] – once,
- The simple genetic algorithm [3]– ten times.

Within the *Best First* algorithm, a strategy of *bi–directional* search was used. In turn, for the simple genetic algorithm, values of parameters recommended in the literature [3] were assumed.

The entire number of features in both sets of learning examples was reduced from 169 to:

- 69, for the set of learning examples constructed on the velocity of absolute vibrations,
- 42, for the set of learning examples constructed on the relative vibrations.

The details of the obtained results are shown in [12].

3.4. Identification and Validation of a Classifier

Taking into account the objectives of the conducted investigations, the ν-SVC algorithm was considered to be applied in building a classifier for detection and localization of cracks in shafts of the turbine set. The identification of SVM classifiers was conducted for:

- sets of learning examples containing a complete set of features,
- sets of learning examples containing only the features marked in the selection process.

The following identifiers distinguishing these sets were accepted in order to differentiate between the particular learning sets: *FS* – an identifier for learning sets with a complete set of features, *BF* – an identifier for learning sets in which features were selected by means of the *Best First* algorithm, *GA1÷GA10* – an identifier for learning sets in which features were selected as a result of a tenfold activation of the genetic algorithm.

Tab. 2. The SVM classifiers identification scheme

Model ID	Kernel function	ν	C	γ
SVM-LIN	Linear	0.1	1.0	—
SVM-RBF	RBF	0.1	1.0	0.1

The accepted categories of determined SVM models and the applied learning parameters of the ν-SVC algorithm were presented in the Tab.2. The values of the ν and C metaparameters were selected on the basis of our own knowledge and on the pre-experiments. The method of *k-fold cross validation* with a set of learning examples divided into 3 subsets was applied in the process of the classifier identification. The quality of the identified classifiers was measured by calculation of, among

others, a classifier performance. The obtained results were presented in Fig. 1.

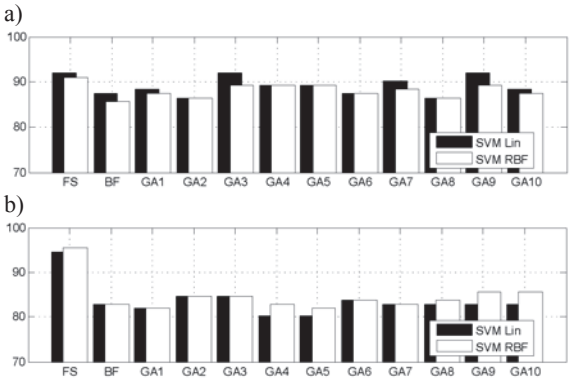

Fig. 1. Classifiers performance: a) absolute vibrations b) relative vibrations

The results (Fig. 1) prove high efficiency of the procured SMV classifiers not only for the complete sets of learning examples (FS), but also for the sets of learning data created on the basis of selected sets of relevant features (BF, GA1-GA10). It needs to be noted that the highest efficiency of the SVM classifiers is observed in the case of the complete sets of learning data; the efficiency slightly decreases along with the decrease in the number of features in the learning sets.

Such a phenomenon may result, among others, from the manner in which the SVM method functions or from the properties of the CFS algorithm. Due to the limited scope of the conducted research, the obtained results cannot constitute an explicit recommendation to omit the stage of the feature selection in the process of knowledge acquisition. Nonetheless, they prove that the prepared system requires further study.

4. SUMMARY

While preparing the system, it was taken into account that the system is meant to acquire diagnostic knowledge with the use of learning data generated by the NLDW–MESWIR software on the basis of the model of a power plant turbine set. The system verification was conducted along with realization of a practical task concerning model (knowledge) acquisition for a hypothetical problem of identification of a classifier which detects and locates cracks in the shaft of the turbine set. Such an approach facilitated examination of the system as far as the implementation and the correctness of the project assumptions are concerned; the project assumptions are connected with the system functionality. The acquired results prove validity of the assumptions and the practical usefulness of the system. Additionally, data concerning the system areas requiring supplementation and corrections were obtained.

REFERENCES

[1] Chang C. C., Lin C. J. *LIBSVM: a library for support vector machines*, 2001. http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[2] Fayyad U. M., Irani K. B.: *Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In 13'th International Joint Conference on Uncertainly in Artificial Intelligence (IJCAI93)*, pages 1022–1029, Chambery, France, 1993.

[3] Goldberg D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, 1989.

[4] Hall M. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, Waikato University, Department of Computer Science, Hamilton, NZ, 1998.

[5] Kiciński J. et al.: *Modelowanie i diagnostyka oddziaływań mechanicznych, aerodynamicznych i magnetycznych w turbozespołach energetycznych (in Polish)*. IMP PAN, Gdańsk,

[6] Larose D. T.: *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience, 2004.

[7] Michalewicz Z., Fogel D. B.: *How to Solve It: Modern Heuristics*. Springer-Verlag, Berlin Heidelberg, 2004.

[8] Moczulski W.: *Metody pozyskiwania wiedzy dla potrzeb diagnostyki maszyn (in Polish)*. Zeszyt 130, Politechnika Śląska, Gliwice, 1997.

[9] Schölkopf B., Smola A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[10] Witten I. H., Frank E.: *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2 edition, 2005.

[11] Vapnik V.: *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[12] Wachla D. System Verification. In Ciupke K., Moczulski W. eds., *Knowledge Acqusition For Hybrid Systems of Risk Assessment And Critical Machinery Diagnosis*. ITE, Radom, 2008.

Dominik WACHLA (PhD Eng.) is an assistant professor at the Faculty of Mechanical Engineering at Silesian University of Technology in Gliwice. His research is focused on the application of methods of artificial intelligence in the technical diagnostics of machinery and industrial processes.