Recent Developments in Damage Identification of Structures Using Data Mining

Abstract

Civil structures are usually prone to damage during their service life and it leads them to loss their serviceability and safety. Thus, damage assessment can guarantee the integrity of structures. As a result, a structural damage detection approach including two main components, a set of accelerometers to record the response data and a data mining (DM) procedure, is widely used to extract the infor-mation on the structural health condition. In the last decades, DM has provided numerous solutions to structural health monitoring (SHM) problems as an all-inclusive technique due to its powerful computational ability. This paper presents the first attempt to illustrate the data mining techniques (DMTs) applications in SHM through an intensive review of those articles dealing with the use of DMTs aimed for classification-, prediction- and optimization-based data mining methods. According to this categorization, applications of DMTs with respect to SHM research area are classified and it is concluded that, applications of DMTs in the SHM domain have increasingly been implemented, in the last decade and the most popular techniques in the area were artificial neural network (ANN), principal component analysis (PCA) and genetic algorithm (GA), respectively.

Keywords

Structural damage detection, data mining technique, artificial neu-ral network, genetic algorithm, principal component analysis

1 INTRODUCTION

Structural systems in civil engineering such as tall buildings, long hydraulic structures, and long span bridges are damage-prone under different loadings such as fatigue, aging, overloading, earth-quakes and other natural disasters during their service life (Duan and Zhang 2006; Hakim and Razak 2014a; Khanzaei et al. 2015; Ghaedi et al. 2016, 2017a, b). Existence of damage can disturb functionality and safety of the structure. Therefore, damage detection is one of the most important

factors in order to guarantee the integrity and safety of civil structures (Kanwar et al. 2007; Xu et al. 2012; Hakim and Razak 2014b; Hanif et al. 2016).

Many damage identification techniques have been applied to civil structures. For instance, visu-al inspections are the most common damage detection techniques. However, they have been used in detecting damage of many structures; they are time consuming and costly. These techniques cannot also be used for continuous monitoring of structures (Razak and Choi 2001; He 2008).

With the development of data inverse analysis methods, it is required to combine problem-based internal mechanism analysis and DM technology to facilitate the application of the measured data analyses in complex engineering system modeling (Zhang et al. 2013). Based on this strategy, DMTs have, recently, been employed in SHM area due to their powerful computational ability to detect damage in structural systems, where the main components of any structural damage detec-tion approach consist of a set of accelerometers and a DM procedure. In the monitoring process, the network of accelerometers is utilized to create a database using response data collection. The DM approach is used to extract information on the structural health condition from the database and to obtain the relationship between data in the form of patterns (Edeki 2012; Hou et al. 2013).

In this paper, various DMTs, which are applicable to SHM, have thoroughly reviewed. A brief background and different tasks of DMTs are described in Section 2. Classification, prediction and optimization-based data mining methods and their applications in SHM are presented in Sections 3 and 4. Trend and comparison of DM applications in SHM as well as the most used DMTs with highest application rate are discussed in Section 5. Here, capabilities and limitations of the most applicable data mining-based methods in SHM are presented. Finally, due to the lack of specific straightforward data mining-based SHM as well as requiring more attention to improve the health monitoring of structures, a flow chart based on the DM steps along with prediction and optimiza-tion-based algorithms is proposed for SHM assessment. In Section 6, the important conclusions are drawn. To the best of our knowledge, current research is the first attempt to illustrate the DMTs applications in SHM.

2 DATA MINING (DM)

DM is one of the key steps in Knowledge Discovery in Databases (KDD) process. This process is used to identify valid, valuable and understandable forms of data (Buchheit et al. 2000; Miranda et al. 2011). In general, DM can be categorized into two groups: descriptive mining and predictive mining. Each of these groups has its specific tasks (Pang-Ning et al. 2006). For instance, some DM tasks consist of clustering, prediction, classification, exploration and association. The purpose of clustering is to divide the samples into groups with related behavior. The numerical prediction ac-tivity determines patterns, rules or models to predict continuous or discrete target values which can also be used for other functions. Classification is used to recognize several rules which can be ap-plied in future work to determine whether a previously unknown item belongs to a known class. Exploration is used to find out dimensionality of an input data and, eventually, the association activity is used to frequently detect occurring related objects. Based on their particular utilizations in consequence of their assumptions and drawbacks, one or combination of some of these tasks can be used to find the hidden information (Obenshain 2004; Liao et al. 2012; Chen and Huang 2013).

3 CLASSIFICATION AND PREDICTION-BASED DATA MINING IN STRUCTURAL DAMAGE DE-TECTION

Classification and prediction-based DM algorithms can be divided into supervised and unsupervised learning broads. In supervised learning techniques (e.g. ANN, regression, support vector machine, decision tree, Bayesian analysis, etc.), targets are known, nevertheless, in unsupervised learning (e.g. clustering and principal component analysis) target known values do not exist (Liao et al. 2012). Subsequent categorization attempts to summarize the classification and prediction-based DMTs and their applications in SHM. To aid the aim, the applications of DMTs in SHM area are separately presented in a tabular form.

3.1 Artificial Neural Network (ANN)

ANN approach was first proposed in 1980s. It is a self-organizing computational technique and it can solve many functions through pattern recognition (Shahriar and Nehdi 2011; Ahmed et al. 2015). ANN can effectively be used to reconstruct nonlinear relationship learning from training (Ali et al. 2013).

A basic biological neuron, as shown in Figure 1(a), consists of a cell body, axons, dendrites and synapses. Signals (input connections) are carried by the dendrites into the cell body. Axons transfer the signals (output connections) from one neuron to others, whereas synapses are the point contacts between dendrites of one cell and axon of another cell (Karacı and Arıcı 2014). The main elements of an artificial neuron are indicated in Figure 1(b) which includes weights, bias and activation func-tion (Azimzadegan et al. 2012). A typical artificial neural network model has two parts; processing units (neurons) and connections between elements (Ali et al. 2013), in which neurons are located in layers of network. A layered ANN structure, called multilayer perceptron (MLP), is one of the most widespread ANN methods especially in structural identification (He and Yan 2007) (see Figure 1(c)). In general, a conventional ANN has three layers which are input layer, hidden layer and out-put layer (Mert et al. 2015). ANNs can be categorized by their network topology such as feed for-ward and feedback or by their learning algorithms such as supervised learning and unsupervised learning (Azimzadegan et al. 2012). Outputs (Yi) of ANN can be defined by following equation:

(1)

where x_i indicates the input values, w_{ij} illustrates the connection weight values between input, hid-den and output layers, b_i is the bias, and f_i shows the transfer function (Shah and Ghazali 2011; Chu et al. 2014).

Strain-based emulator neural network (SENN), parametric evaluation neural network (PENN), wavelet neural network (WNN), Bayesian neural network, fuzzy wavelet neural networks (FWNN), neural network-based damage classification, auto-associative neural network (AANN) and displace-ment-based neural network emulator (DNNE) are some of the important applications of neural networks in SHM. The detailed neural network technique applications are presented in Table 1.

.

3.2 Fuzzy Logic Technique

Fuzzy logic was proposed by Lotfi Zadeh for the first time in 1965. It has been employed in differ-ent applications such as pattern recognition, classification, decision making, etc. (Rutkowski 2004). As Figure 2 indicates, the basic configuration of a fuzzy technique consists of four important com-ponents, which are fuzzification, fuzzy rule-base, fuzzy inference and defuzzification. Fuzzification is a mapping from a crisp input to fuzzy membership sets. The fuzzy rule-base has set rules of fuzzy variables described by membership functions. Fuzzy inference is a decision making mechanism of the fuzzy system. The defuzzifier changes the fuzzy consequences from different rules into crisp val-ues (Nyongesa 1998). Fuzzy is a model free technique for structural system identification, where the

most important advantages of fuzzy systems are their high parallel implementation, nonlinearity and being capable of adapting (Nerves and Krishnan 1995). Applications of Fuzzy logic in SHM are detailed in Table 2.

3.3 Classification

Classification, which is based on the machine learning, is used to classify databases into several groups. This technique is employed to indicate not only a form of classification rules but also math-ematical formulae from a function of individual classes of data in a database. Therefore, these rules are capable of organizing upcoming data. Consequently, classification method can present a better view of the database contents (Han et al. 2006).Hence, refer to Table 3, key point of this technique is to predict a categorical feature based on other features of known data.

3.4 Support Vector Machine (SVM)

Support vector machine (SVM) was first introduced by Vapnik and his assistants in 1963 (Vapnik 1995; Statnikov et al. 2011). SVM works based on statistical learning theory and because of its high

accuracy and good generalization capability, it has the potential to produce high quality predictions in numerous tasks. Therefore, SVM has various applications which can be found in several areas such as machine learning, data classification and pattern recognition (Cristianini and Shawe-Taylor 2000b; Samanta et al. 2003; He and Yan 2007; Tinoco et al. 2014). Basic models of SVM are linear SVM with linear functions and nonlinear SVM with kernel functions (Mita and Hagiwara 2003). Moreover, aim of SVM classifier is to determine a separating hyperplane to divide the given data into two classes (i.e. positive class and negative class) in the optimal form. Therefore, as shown in Figure 3, the optimal separating hyperplane is determined by solving an optimization problem, de-fined as:

where w, y_i, and x_i are a m-dimensional vector, class label and given data, respectively. N is the number of samples, and b is a scalar (Mita and Hagiwara 2003; Kishore et al. 2014).

Wavelet support vector machine (WSVM), nonlinear support vector machines, support vector regression (SVR), multiclass support vector machine and multiclass nonlinear relevance vector ma-chine (MNRVM) are some of the important applications of SVMs in SHM. Table 4 illustrates the different use of SVM applications in detail for SHM area.

3.5 Regression Analysis

Regression analysis is one of the statistical methods that is used for prediction of different functions through different algorithms such as linear regression, nonlinear regression, logistic regression, and stepwise regression. The method is used to find the mathematical relationship between one or more independent and dependent variables (Jeon et al. 2014), as shown in Figure 4. This relationship can be defined as:

where x_i indicates the input and y_i illustrates the output; f(x) represents the linear regression function and e shows the independent random error (Liu and Li 2009).

Support vector regression (SVR), multivariate linear regression (MLR), nonlinear regression and robust regression analysis (RRA) are some of the important applications of regression analysis in SHM. Table 5 shows the applications of regression analysis in SHM, thoroughly.

3.6 Principal Component Analysis (PCA)

PCA is a well-known method for data analysis, which is used as a dimensional reduction tool (Lautour and Omenzetter 2010; Kwak 2014). PCA also is employed in exploratory data analysis and pattern recognition (Yiqiu et al. 2012). The main concept behind PCA is to transform high-dimensional correlated variables into low-dimensional uncorrelated variables by an orthogonal pro-jection, in which the new low-dimensional uncorrelated variables are known as principal compo-nents, as shown in Table 6. Main steps to have the principal components of a database is to (1)

construct a matrix from data, (2) construct the normalized matrix, (3) calculate the covariance matrix, and (4) calculate the largest eigenvalue and eigenvector. Architecture of a two-dimensional PCA is shown in Figure 5, where it can principally be defined as:

in which Y shows principal components, W indicates eigenvector and X represents the input data in the new coordinates (Ku and Waszczyszyn 2006; Hua et al. 2007; Hao et al. 2011).

3.7 Bayesian Analysis

Bayesian analysis is considered as a statistical method, which is applied to pattern recognition (Cristianini and Shawe-Taylor 2000a) and it is a classification method based on Bayes's theorem. Bayesian method has been used for addressing data uncertainty and modeling errors in structural damage detection (Jiang and Mahadevan 2008b). In recent years, a new type of Bayesian analysis has been derived from neural networks (Arangio and Beck 2012). For instance, a four-phase struc-tural damage identification approach (see Figure 6) can be used by means of Bayesian inference and fuzzy ~~wavelet neu~~ral network (WNN) model (Jiang and Mahadevan 2008a). Bayes's theorem can be defined as follows:

where P(A) is prior probability of A with no knowledge of observation B; P(B) is prior probability of B with no knowledge of A; P(A|B) is posterior probability of A after observation B; and P(B|A) is probability of observation B given that A is true (Dawsey et al. 2006; Yuen 2010). Applications of Bayesian analysis are shown in Table 7.

3.8 Clustering

Clustering is an unsupervised statistical data analysis technique, which is used in pattern recogni-tion, image analysis and bioinformatics (Park et al. 2007). This method is employed to divide da-tasets into separated similar subsets (clusters) according to typical patterns identified in the cluster-ing analysis (Ghaedi and Ibrahim 2017). Figure 7 indicates the basic concept of clustering. As shown in Figure

7(a), in order to have a successful clustering, maximum intra-cluster similarity as well as minimum inter-cluster similarity is required. Moreover, Figure 7(b) presents a schematic plot of K-means algorithm, which divides the space into three clusters (C1, C2, and C3). The K-means is one of the most descriptive partitioning clustering algorithms with a quite reliable effec-tiveness at local optimum. However, it can be employed only to numerical datasets. Furthermore, K-means has poor handling for data prone to noise and outliers (Symeonidis and Mitkas 2005). Clustering can also help to decrease the distance between datasets and improve the similarity of datasets in each cluster (Saitta et al. 2008; Yu et al. 2011; Chuang et al. 2011; Chen and Huang 2013; Xiao and Fan 2014). Table 8 illustrates the applications of clustering analysis in SHM system.

3.9 Decision Tree

One of the statistical methods in order to solve classification problems is decision tree that can widely be implemented (Gal et al. 2013). Components of a decision tree, as shown in Figure 8, con-sist of nodes and arcs. The chance nodes indicate features, while decision nodes represent pattern classes and arcs denote feature values. The main advantages of using this approach is easy solution to understand and its higher interpretable capacity compared to other statistical techniques (Wei and Hsu 2008).

Starting point to design a decision tree model is a decision node creation including several alter-natives. For example, Kim et al. (Kim et al. 2011) considered two alternatives (i.e. first inspection time (tinsp,1) after and before initial service life (t(0)life)) and four branches to create a decision tree model for prediction of lifetime of deteriorating structures with one inspection, as shown in Figure 9(a). In Figure 9(b), each branch represents different cases of decision tree including late inspection in first branch, changing node without damage detection in second branch and damage detection in third or fourth branches. The repair process can be conducted by decision maker selec-tion in the third or fourth branch. Applications of decision tree method in SHM are presented in Table 9.

4 OPTIMIZATION-BASED DATA MINING IN STRUCTURAL DAMAGE DETECTION

4.1 Genetic Algorithms (GA)

Genetic algorithm (GA) was first proposed by John Holland in 1970's. In GA, a chromosome is used to determine the solution. The chromosome includes a group of genes that optimize parame-ters. In damage identification problems, damage parameters are determined in real number of genes, which form the chromosome of an individual solution. This algorithm employs a random solution from a current population. Then, next generation will be created using crossover and mutation op-erators (Kouchmeshky et al. 2008; Aghajanloo and Sabziparvar 2012). Figure 10 illustrates a typical

---

crossover and mutation operator in GA algorithm. Accordingly, the basic steps to form a GA in-clude:

Select a population of chromosomes;
Calculate the fitness of each chromosome;
select parents chromosomes from population,
randomly; Crossover creates offspring;
Mutation forms a new generation;
Calculate the fitness of new chromosomes;
Repeat the selection of parents chromosomes to find the fittest chromosomes in the next gen-eration (Fu 2000).

Genetic algorithm is an attractive tool to optimize difficult problems because of its benefits such as parallelism, convergence to global optima, adaptation, and no need for the gradient of the objec-tive function. Considering these benefits, GA has been successfully applied in structural damage identification problems (Sohn et al. 2001; Raich and Liszkai 2003). Applications of GA in SHM are shown in Table 10.

4.2 Particle Swarm Optimization (PSO)

PSO which was first proposed by Kennedy and Eberhart (Kennedy and Eberhart 1995), is one of the population-based artificial intelligence optimization techniques. The approach was simulated by the social behavior of organisms such as bird flocking to be used a suitable tool for global optimiza-tion. In PSO, a particle represents a potential solution where each particle has two updatable fea-tures; position and velocity. The main steps of this algorithm can be presented as follows:

Particles initialization with arbitrary position and velocity;

Objective function evaluation and updating the position and velocity based on the best fitness function;

Determination of global minimum fitness value;

Velocity modification and particle movement to new position;

Process iteration to meet criterion (Talatahari et al. 2013).

Furthermore, PSO is easy to apply and has great computational capacity. In comparison with other optimization approaches, PSO is, however, more efficient and requiring fewer number of func-tion evaluations while gives better or the same quality of results, but it has some weaknesses such as trapping into local optimum in a complex search space and disability to do a good local search around a local optimum (Dimou and Koumousis 2009; Gholizadeh and Fattahi 2014; Gundogdu et al. 2015). Optimization function, as the critical attribute of PSO, is very useful to assess the struc-tural damages; however, as shown in Table 11, PSO implementation in SHM is not as much of-GAapplications.

4.3 Ant Colony Optimization (ACO)

Ant Colony Optimization (ACO) was introduced by Marco Dorigo in 1992 for combinatorial opti-mization problems (Yu and Xu 2011; Guerrero et al. 2014; Cottone et al. 2014). ACO is a probabil-istic population-based technique inspired by the ants' foraging activities and their communication system by using pheromone trail, as the most significant role to discover the direction (Amini and Ghaderi 2012; Fidanova et al. 2012; Majumdar et al. 2012). Optimization problems can be solved utilizing ACO by iterating two steps:

Construct candidate solutions in a probabilistic way by using a probability distribution over the search space;

Modify the probability distribution by candidate solutions (Yu and Xu 2011).

Ant colony based algorithm can be classified into three forms including ant-quantity, ant-density and ant-cycle. The main concept of ACO is the power of finding the shortest way from

ant's nest to food source, as indicated in Figure 11. ACO Applications used by researchers in SHM are shown in Table 12.

5 RESULTS AND DISCUSSIONS

5.1 Comparison of DMTs Applications in SHM

The DMT applications used in SHM area have been introduced and discussed in sections 3 and 4. As shown in Figure 12, the general development of DM applications in SHM, primarily since 2000, demonstrates that the main three DMTs used in SHM were ANN, PCA and GA, respectively. Overall, outcomes showed that, applications of statistical techniques in the SHM were less than artificial intelligence techniques. The percentage of DMT's applications in SHM is shown in Figure 13. This figure indicates that ANN and PCA were the highest application rate by conducing 30% and 20% of researches, respectively. Further, GA by 10% application rate stood at the third level. In contrary, decision tree, clustering, Bayesian, PSO, regression and ACO techniques were rarely used in SHM. Other DMTs such as SVM and fuzzy were occasionally used in SHM with 6% to 8% application rate.

5.2 Capabilities and limitations

In recent years, there has been a vast increase in the number of reverse analysis in SHM generally based on two methodologies; artificial neural networks and optimization techniques. In optimiza-tion-based data mining, meta-heuristic techniques, including (1) biology inspired intelligent global search approaches such as GA and (2) swarm-based techniques such as ACO and PSO, have several applications in SHM. Nevertheless, in general, drawback of optimization approaches is time consum-

ing. On the other hand, these methods have many differences in detail. For instance, PSO has memory and all particles retain the knowledge of good solutions, but in GA, there is no memory and when the population changes, all previous knowledge about the problem will be destroyed. From this viewpoint, ACO is not appropriate for continuous optimization problems. Despite using PCA in several SHM studies as it was able to decrease dimensions of multidimensional datasets, ANN is the most powerful DMT in classification, regression and prediction which has been broadly used in many SHM applications, because of its nonlinear learning abilities. ANN is more flexible and more accurate in comparison to other DMTs. Figure 14 illustrates the discussed DMTs trends for all applicable methods. This is also demonstrated the development of DMTs from 2000 to pre-sent.

In the present study, a flow chart based on DM steps along with prediction and optimization-based algorithms is proposed for SHM assessment and indicated in Figure 15. According to proposed flow chart, SHM assessment starts with measuring the damage level. After collecting the da-ta, preprocessing step including cleaning, construction, integration and transformation of data is carried out in order to build the database. In next step (modeling), the suitable DMTs are suggest-ed to be applied for training the data. Since various DMTs such as ANN, Fuzzy, PCA, SVM, GA, PSO, ACO, Bays, etc. exist for the same problem type, therefore, they can be employed for differ-ent purposes such as optimization, classification and perdition. Moreover, test design generation, patterns creation and their validation are also detailed in this step to determine the severity and location of damages. After pattern assessment, several rehabilitation activities (e.g. modification and strengthening the structural members, repair or replacement of the existing damaged members, and minor or major maintenance) are suggested to improve the health condition of civil structures. Based on existing problem, one or combination of these activities needs to be done to protect the structural performance.

6 CONCLUSION

DM is just one of the very important steps in the Knowledge Discovery in Databases (KDD) process to extract the models, patterns and rules from raw data in large databases. There are three main approaches, namely statistical techniques, machine learning techniques and artificial intelligence techniques to apply in structural real world problems. In this paper, various DMTs which are applicable to SHM from 2000 were widely reviewed and, consequently, their main concept, method-ologies, capabilities and drawbacks were discussed. Based on the presented information, the follow-ing conclusions can be drawn.

ANN, PCA and GA are the most common DMTs in SHM, respectively, in order to structural damage identification of various types of civil structures; buildings, bridges, reinforced concrete beams, dams, truss structures, and steel plates.

Classification and prediction based DMTs can be categorized into two different methods which are statistical methods and artificial intelligence methods.

Statistical methods such as decision tree and Bayesian have the lowest application rate in SHM due to lack of capacity, flexibility and complexity. In contrast, artificial intelligence techniques such as ANN and GA have the highest application rate in SHM due to their accu-racy, flexibility, autonomy, complexity, and optimization capability.

In optimization-based DMTs, GA is the best optimal tool to improve the efficiency of difficult problems due to its criteria, which is based on parallelism. GA can utilize a coding set of variable from a population instead of employing the variables directly from a particular solution.