A NEURO FUZZY SYSTEM FOR KNOWLEDGE DISCOVERY OF INCOMPLETE CONSTRUCTION DATA

Wen-der Yu and Han-wen Lin

Institute of Construction Management, Chung Hua University, Taiwan

Abstract: This paper tackles problems encountered in mining of incomplete data for knowledge discovery of construction databases. As historical construction data are expensive to collect, any waste of incomplete data means not only loss of knowledge but also increase of costs for knowledge discovery of construction engineering. Unfortunately, incompleteness is commonplace in the existing construction databases. This paper proposes a VaFALCON (Variable-Attribute Fuzzy Adaptive Logic Control Network) neuro-fuzzy system that is equipped with the power for mining incomplete historical data. The proposed VaFALCON is shown to successfully mining of construction data with various percentages of missing attribute values.

Keywords: Construction; KDD; Data mining; Neuro-fuzzy; Incomplete data.

1. INTRODUCTION

The incompleteness, impurity, and scarcity of data are key problems confronting knowledge discovery in databases (KDD) of construction engineering and management. As historical data are valuable sources for KDD of construction knowledge, any waste of historical data can not only cause tremendous loss of valuable knowledge but also significant increase of costs for KDD of construction engineering. This problem is more serious while compared with similar problems in other industries due to two reasons: (1) the costs of construction projects are usually huge, thus it is expensive (if not impossible) to regenerate such data; (2) construction projects are unique in its nature, accumulating sufficient data for KDD has been a challenge in construction industry. Moreover, previous survey found that valuable historical data are not completely collected due to the uncertain nature of construction operations, the harsh environment of construction fields, and the attitude of construction staffs [1]. The incompleteness of data is commonplace in the historical construction databases. Unfortunately, rare traditional AI techniques, including numeric and symbolic reasoning schemes, are able to handle the incomplete data while performing data mining. Therefore, when some attributes information of historical construction data were missed, they are usually discarded. This causes another problem—data scarcity. The key reason is due to the disability of traditional data mining techniques in handling incomplete data.

This paper aims at developing a neuro-fuzzy system, based on the Fuzzy Adaptive Learning Control Network (FALCON) method [2], which is modified and improved so that it is able to handle historical data with missing attribute values. The proposed method is based on neuro-fuzzy system, so it is equipped with both learning and reasoning capabilities to mine construction knowledge from historical data. It also provides explanation of the reasoning process for system user to develop improvement strategy. Moreover, the proposed method modifies the learning mechanisms of the original FALCON with a special setup of data handling scheme. As a result, an variable-attribute learning scheme is developed. They can be used during the application. This is very useful for real-time decision making when the complete information cannot be acquired or when it is too expensive to collect.

This paper is presented in the following manner: at first, the FALCON neuro-fuzzy system is reviewed as a basis for methodology development in Section 2; the problem of data incompleteness in KDD is defined in the Section 3; the traditional approaches for handling incomplete data are described in Section 4; in Section 5, the proposed VaFALCON is introduced in details; in Section 6, two real world cases of incomplete data mining are demonstrated with the proposed VaFALCON method; finally, the conclusions and future works are discussed Section 7.

2. NEURO-FUZZY DATA MINING FOR CONSTRUCTION ENGINEERING

Construction engineering was conceived as an experienced-based discipline [3];, knowledge acquired from previous works plays a key role for

successful performance of the new projects. Not only the construction know-how of the contractors, but also the design capabilities of the design firms and the management skills of CM consultants rely heavily on such knowledge.

Today information technologies provides easy means to capture and store digitized data. With the computerization of construction industry, more and more of data have been collected and stored in databases. Even though the databases are promising sources for useful and valuable construction knowledge [4], the raw data are usually rarely of direct benefit. The true value of data resides in the ability to extract useful knowledge for decision-makers. When the scale of data manipulation, exploration, and inferencing exceed human capacities, it naturally prompts the need for intelligent data analysis methodologies, which could discover useful knowledge from data [5]. The emerging of KDD and Data mining (DM) techniques provides solutions for such need. The term KDD refers to the overall process of knowledge discovery in databases. Another related term, DM, is a particular step in the KDD process, which involves the application of specific algorithms for extracting patterns, models, or rules from data [6]. Among the many existing data mining algorithms, soft computing techniques (such as neuro-fuzzy systems) are most widely applied for a KDD process [6].

FALCON (Fuzzy Adaptive Learning Control Network) is one of the most successful development of soft computing techniques. It was first proposed by Lin and Lee for the purpose of automatic control [2]. It's then modified and enhanced by Yu and Skibniewski to automatic constructability knowledge acquisition [7]. Their research found FALCON provides many features desirable for construction knowledge acquisition including: (1) capability of handling uncertain information; (2) functions of learning; (3) explicit knowledge representation; and (4) trace-back functions for problem solving. An FALCON network consists of five layers of neurons: (1) input layer—taking input information; (2) input linguistic layer—transforming input information into fuzzy linguistic terms; (3) fuzzy rule layer—representing fuzzy IF-THEN rule base; (4) output linguistic layer—performing fuzzy inference process of fuzzy decision rules; (5) output layer—transforming fuzzy information to crisp decision information. The original FALCON model is shown in Figure 1. Details of FALCON computational algorithms are referred to Lin and Lee [2].

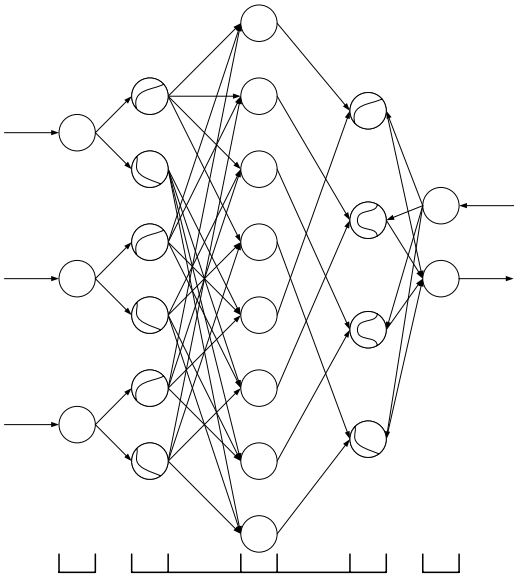

Figure 1. FALCON model [2]

3. DEFINTIONS OF INCOMLETENESS FOR CONSTRUCTION DATABASES

Incomplete data are commonly found in the traditional construction databases. Two categories of incompleteness are defined in this paper: (1) missing data—incomplete coverage of data in some intervals of the universe of discourse; (2) missing values—incomplete information in some interesting attributes. Following describes the problems of the two types of data incompleteness.

3.1 Missing data

The first type of data incompleteness problem is lack of data sets in some intervals of the universe of discourse for a specific attribute. This type of data incompleteness can be further classified into two categories: (1) interpolation type; (2) extrapolation type. For the interpolation problems, the data sets are missing between two clusters of data in the universe of discourse, see Figure 2. Thus, the missing data are usually recovered by interpolation schemes.

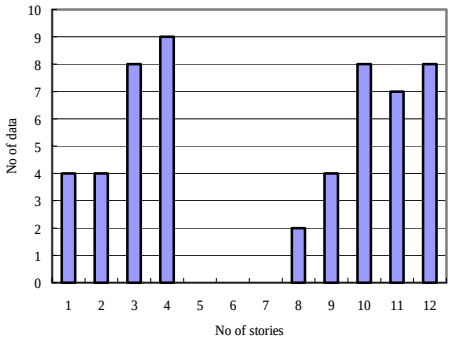

Figure 2. Interpolation type of missing data

On the other hand, for the extrapolation problems, the data sets are missing at extremes parts of the universe of discourse. Thus, extrapolation schemes are adopted for data recovery. Figure 3 shows an example of extrapolation type of missing data.

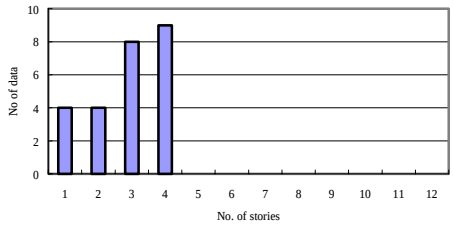

Figure 3. Exterpolation type of missing data

3.2 Missing values

In the second type of data incompleteness problem, some attributes of interest may not be available due to numbers of reasons: (1) the attributes are not considered important at time of data entry; (2) the mistakes made by collector; and (3) equipment malfunctions. This type of problem is more severe while merging databases from different sources. For example, in Table 1, data sources of firm A, B, and C provide inconsistent data format. The resulted database shows a typical example of missing-value type of data incompleteness, where the missing attribute values are depicted with shadowed cells.

Table1 Missing values in heterogeneous databases

Firm	weath.	Temp.	Humid.	item	Prod.
A	cloud	18			high
B		23	90%		low
C			60%	Form.	Ave.

3.3 Definitions of data incompleteness in this paper

This paper will tackle problems regarding to "missing values" rather than "missing data". That is, the incompleteness is defined as the percentage of unavailable attribute values. The missing data problem is not considered in this paper. In order to evaluate the degrees of incompleteness for missing values, two measures of data incompleteness are defined: (1) percentage of incomplete attributes (PIA)—measuring the number of unavailable attributes, which consist at least one missing value, over the number of total attributes; (2) percentage of overall incompleteness (POI)—measuring the total number of data set with at least one missing value over the number of total data set. In calculation of the above two measures of data incompleteness, only the precondition attributes characterizing a data set are considered, the consequence part of a data set, such as the last column (productivity) in Table 1, is not included.

4. TRADITIONAL METHODS FOR DATA CLEANING OF MISSING VALUES

There has been several methods proposed for filling in missing attribute values. Such methods are named "data cleaning". Han and Kamber proposed six methods for cleaning data with missing values [8]:

(1) Ignore the tuple—this method simply discards the tuple for all data. This is not a very effective, unless the tuple contains several attributes with missing values.
(2) Fill in the missing value manually—the method employs a domain expert to fill in the missing values based on his/her personal judgment. This method is quite straightforward and easy to implement. However, it is time consuming and sometimes error prone with bias of the domain expert.
(3) Use a global constant to fill in the missing value—this method replaces all missing attribute values with the same constant, e.g., "unknown". When "unknown" is adopted as the global constant, the problem is still unsolved for FALCON, since the "unknown" data is not acceptable for FALCON.
(4) Use the attribute mean to fill in the missing value—this method is similar to the previous one except that the global constant is replaced by global mean of the considered attribute. This method is a safer way for data cleaning. However, it is still biased.
(5) Use attribute mean for all samples belonging the same class as the given tuple—this method requires the user to classify the data first. However, this is sometimes impossible.
(6) Use the most probable value to fill in the missing value—this method employs some inferencing methods to recover the missing values, such as statistic regression, Bayesian formalism, decision tree induction, etc. This is possible the best, among the existing data cleaning approaches, for recovering missing attribute values. However, while the PIA or POI is high, this method may fail.

Considering the data mining process of FALCON, the traditional data cleaning methods are either limited in their capability or biased in their nature while handling missing attribute values. Moreover, when the ratio of missing values is high, all of the existing data cleaning methods tend to fail. It is therefore very desirable to develop a data mining technique that is able to handle data incompleteness problems regardless of the availability of domain experts and the severity of incompleteness.

5. THE PROPOSED VaFALCON APPROACH

This paper proposes a modified FALCON, namely Variable-Attribute Fuzzy Adaptive Logic Control Network (VaFALCON), for mining of incomplete construction data. The data incompleteness is defined previously as PIA (percentage of incomplete attributes) and POI (percentage of overall incomplete values). In order to improve the drawbacks of traditional data cleaning methods, the proposed VaFALCON aims at handling incomplete construction without restrictions on PIA and POI. Moreover, the proposed VaFALCON is designed to take incomplete construction data directly without data cleaning. Following describes the details of the proposed VaFALCON.

5.1 Mechanism of variable-attribute procession

Most of the current data mining techniques, including traditional FALCON and artificial neural networks, take only complete data without missing values as their inputs. While the input data are incomplete, data cleaning is performed to fill out the missing values. Therefore, the first step to develop VaFALCON is to establish a mechanism for processing variable numbers of attributes. Using the FALCON model in Figure 1 as a basis for discussion, there are three inputs (X_1, X_2, and X_3) and one single output (Y). Referring to Table 2, data set *A* is a complete data that consist of input and output pairs as ([a,b,c], D), where [a,b,c] is the vector of inputs and D is the value of output. The other incomplete data set $\bar{A}$ contains a missing attribute X_1 whose value is unknown and shown as "nan" meaning "not a number".

Table 2 Complete vs. incomplete data set

Attribute	X_1	X_2	X_3	Y
Data set *A*	a	b	c	D
Data set $\bar{A}$	nan*	b	c	D

*Not a number

The propagation of a complete data set in FALCON is shown in Figure 4, where the interconnections between the first two nodes in Layer 2 (pre-condition fuzzy linguistic terms) and Layer 3 (rule nodes) are shown as solid links, which means physical connections. As the first input (X_1) is missing, the resulted FALCON is shown in Figure 5. In Figure 5, the fuzzy linguistic term nodes of the first input (X_1) are disconnected (shown as dashed line) with the rule nodes in the following layer. The signals are not propagated via dashed-line links. In the traditional FALCON, the network of Figure 5 can not learn due to the undetermined links between Layer 3 and Layer 4.

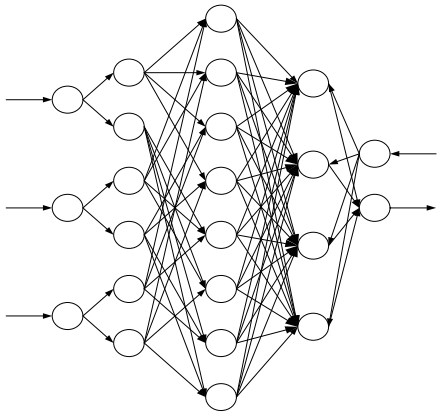

Figure 4. Connections of FALCON for complete data

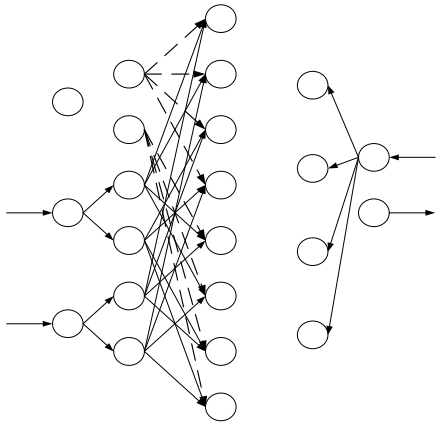

Figure 5. Connections of FALCON for incomplete data

Using the first principle of data cleaning described in Section 4, the first tuple (X_1) can be ignored while the information of the other two tuples (X_2 and X_3) are conserved. To make the best use of the data information, the FALCON is degraded as shown in Figure 6, where the first input is omitted. With the two residual tuples, the FALCON of Figure 6 can still learn from example data.

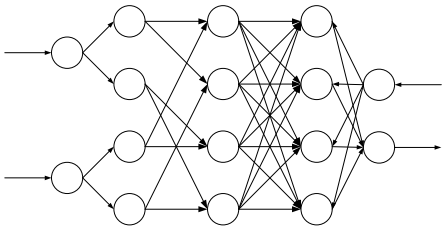

Figure 6. Degraded FALCON for incomplete data

5.2 Modified learning algorithms for VaFALCON

In order to implement the mechanism of variable-attribute procession for VaFALCON described in the previous section, the learning algorithms of traditional FALCON are modified as shown in the following.

(1) Modification of Kohonen learning rule

In order to take incomplete input data, Kohonen learning rule is modified as follows:

For means of membership functions—

if $x \neq nan$

$$\bar{w}_i^{k+1} = \bar{w}_i^k + \eta^k \left(x - \bar{w}_i^k \right)$$

$$\bar{w}_j^k = \bar{w}_j^k \quad , \quad \text{for } j = 1,2,...,n \quad j \neq i \qquad (1)$$

end

For spreads of membership functions—

if $x \neq nan$

$$\sigma_i = \frac{\left| m_i - m_{nearest} \right|}{\gamma} \qquad (2)$$

end

In both equation (1) and (2), "*nan*" (not a number) means missing of attribute value. The logic judgment in the first line of equation (1) and (2) represents that the modification is performed only when the attribute is not empty.

(2) Modification of *fuzzy AND* inference

The *fuzzy AND* inference of rule nodes in the third layer of FALCON performs t-norm computation, i.e., minimization or intersection. In VaFALCON, the goal is to avoid the situation that the memberships of the missing attributes become the outputs (i.e., minimum) of *fuzzy AND* operations. To achieve this goal, the memberships of missing attributes are replaced with the constant value "1.1" as shown in Table 3 and Table 4.

Table 3 Original outputs of Layer 2 in FALCON

Node	1	2	3	4	5	6
Output	nan	nan	0.825	0.236	0.148	0.567

Table 4 Modified input of Layer 3 in FALCON

Node	1	2	3	4	5	6
Output	1.1	1.1	0.825	0.236	0.148	0.567

After the modification described above, it is guaranteed that no missing attribute will become the output of *fuzzy AND* operations.

5.3 Learning process of VaFALCON

The learning process of VaFALCON consists of two phases: (1) Self-organizing phase—including modified Kohonen learning and reinforcement competitive learning to construct the primitive FALCON structure; (2) Back-propagation phase—applying back-propagation learning rule to fine-tune the network. The learning process ends when the expected error rate is achieved or the maximum number of learning cycles is reached.

6. SYSTEM TESTING

In order to demonstrate the power of the proposed VaFALCON, system testing is performed for two cases of real world construction problems: (1) mining data of building construction costs; (2) mining of duration data for slurry wall construction. The experiment is designed to compare the estimation accuracy of three scenarios: (1) learning of complete data set; (2) learning of data sets with various degrees of data incompleteness in terms of PIA and POI; (3) learning of incomplete data that discards the data sets with missing values. In the second scenario, the POI is set for 5%, 10%, 25% and 100%. The missing attributes are selected by a random process. The PIA for all cases in Scenario 2 is set to be 100%, that is at least one missing value is found in every tuple.

(1) Case I—mining data of building construction costs

In Case I, the historical data of building construction costs was collect from a research by Yu [8]. Totally 25 data sets were collected. Among those, 22 data are used for training and the rest 3 data are used as testing set. For Scenario 1 (complete data), the system accuracy is 94.66%. The accuracy of Scenario 2 and 3 for various degrees of POI are shown in Figure 7.

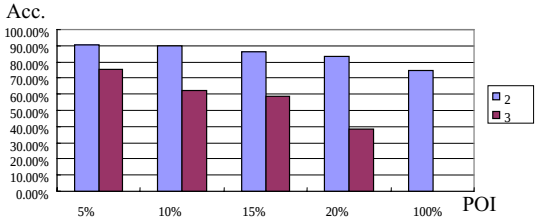

Figure 7. Comparison of accuracy between Scenario 2 and Scenario 3 for Case I

It is found from Figure 7 that the system accuracy of Scenario 3 decreases dramatically as the POI increase from 5% to 100%, while the system accuracy of Scenario 2 (with the proposed VaFALCON) decays moderately for the same degree of POI. As POI is 100% (i.e., every data in the training set consists at least one missing attribute), the proposed VaFALCON can still achieve system accuracy as high as 74.41%, while the Scenario 3 (discarding all data) cannot learn at all (system accuracy is 0%).

(2) Case II—mining of duration data for slurry wall construction

In Case II, the historical data of building construction costs was collected from a research by Yang [9]. Totally 27 data sets

were collected. 24 data are used for training and the rest 3 data are used as testing set. For Scenario 1, the system accuracy is 94.62%. The accuracy of Scenario 2 and 3 for various degrees of POI are shown in Figure 8.

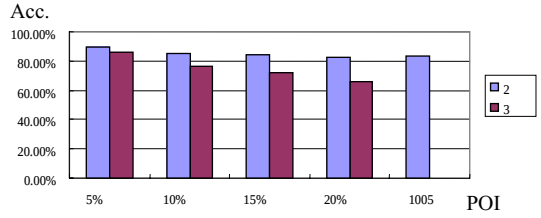

Figure 8. Comparison of accuracy between Scenario 2 and Scenario 3 for Case II

Similar results are found from Figure 8 that the system accuracy of Scenario 3 decreases dramatically in Case II, while the system accuracy of Scenario 2 is still high comparatively. As POI is 100%, the proposed VaFALCON can achieve system as high as 83.02%, while the Scenario 3 cannot learn at all (system accuracy is 0%)

From both of the above examples, it is concluded that the proposed VaFALCON can learn from incomplete construction data for various degrees of POI. It has significantly improved the traditional approaches of data cleaning in handling incomplete data. The improvement of system accuracy ranges from 14.98% (5% POI) to 74.71% (100% POI) for building construction costs and from 3.74% (5% POI) to 83.02% (100% POI) for slurry construction duration.

7. CONCLUSIONS AND FUTURE WORKS

This paper presents a proposed VaFALCON neuro-fuzzy data mining technique for knowledge discovery of historical construction databases. The proposed VaFALCON is capable of handling incomplete construction data. This feature is very desirable for construction industry, where data are scarce and data accumulation is expensive. Any waste of historical data may cause major cost increases for KDD. Moreover, due to the harsh environment of construction job site, incomplete data collection is nature in construction databases. The proposed VaFALCON can meet all of the above needs. It is found from the demonstration examples that the improvement of system accuracy range from 3.74% (5% POI) to 83.02% (100% POI), which verifies the value of the proposed VaFALCON method.

The proposed VaFALCON is able to handle incomplete data with missing attribute values, however the missing data problem discussed in Section 3.1 of this paper is still unsolved. Ambitious researchers are encouraged to pursue in that field.

ACKNOWLEDGEMENT

This paper is based on the results of a research project sponsored by the National Science Council, Taiwan, under project No. NSC92-2211-E-216-017. Sincere appreciations are given to the sponsor by the authors.

REFERENCES

[1] Wu, C. F., Yu, W. D., and Yang, J. B. "A Study on the Estimation of Realistic Construction Duration and the Schedule Compression Incentives," *Report to the Public Construction Commission*, Public Construction Commission, Executive Yuan, Taiwan Government, 2002. (in Chinese)

[2] Lin, C. T., and Lee, C. S. G., "Neural-network-based fuzzy logic control and decision system," *IEEE Transactions on Computers*, Vol. 40, No. 12, pp. 1320-1336, 1991.

[3] Ardery, E. R., "Constructability and constructability programs: White paper," *J. of Constr. Engrg. and Mgmt.*, ASCE, 117(1), pp. 67-89, 1991

[4] Yu, W. D., and Yang, J. B., "Data Mining for the Cost Estimating of Highway Bridges Construction with a Neuro-Fuzzy System", *Proceedings of 2001 Ninth National Conference on Fuzzy Theory and Its Applications*, Nov. 23~24, National Central University, Chung-li, Taiwan, pp. 437~442, 2001.

[5] Fayyad, U., and Uthurusamy, R., "Data mining and knowledge discovery in databases," *Commun. ACM*, vol. 39, pp. 24–27, 1996.

[6] Mitra, S., Pal, S. K., and Mitra, P., "Data mining in soft computing framework: A survey," IEEE Trans. Neural Networks, vol. 13, No. 1, pp. 3–14, 2002.

[7] Yu, W. D., and Skibniewski, M. J., "Quantitative constructability analysis with a neuro-fuzzy knowledge-based multi-criterion decision support system," *Automation in Construction*, Vol. 8, No. 5, pp. 553-565, 1999.

[8] Han, J., and Kamber, M., *Data Mining—Concepts and Techniques*, Morgan Kaufmann Publishers, San Diego, U.S.A., pp. 109, 2001.

[9] Yu, J. S., "Developing building cost estimating system using case-based reasoning approach," *Master Thesis*, Department of Civil Engineering, National Central University, Chungli, Taiwan, R.O.C., 2001. (in Chinese)