Expert Systems with Applications 36 (2009) 7199–7209

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Planning of educational training courses by data mining: Using China Motor Corporation as an example

Chiao-Tzu Huang [a], Wen-Tsann Lin [a], Shen-Tsu Wang [b,*], Wen-Shan Wang [a]

[a] Industrial Engineering and Management Department, National Chin-Yi University of Technology, Taiwan, ROC
[b] Industrial Engineering and Engineering Management Department, National Tsing-Hau University, Taiwan, ROC

ARTICLE INFO

Keywords:
Educational training
Cluster analysis
Decision tree algorithm
Back-propagation neural networks

ABSTRACT

In Taiwan, most industries are of small and medium scale, and there are limited resources for educational training. Increasing the quality of personnel by cultivating talents for the future becomes an extremely important issue. With the growth of firms and the increase in their needs, the database is also growing. We should therefore determine how to recognize and extract the useful information contained in this database in order to apply it in such a way that assists companies in meeting their increasing and changing needs. This research collects data of personnel educational training in China Motor Corporation by cluster analysis, decision tree algorithm and back-propagation neural networks for mining analysis and classification. Based on the algorithm classification result, we finally propose the demand model suitable for educational training in other related industries. The research is expected to explore how to maximize results through planning the courses and the personnel's participation in the training. We try to determine the key factors essential to the success of educational training. Once identified, this information can then serve as the basis for other firms' future planning of educational training strategies with regard to innovation and breakthrough.

1. Introduction

Georgenson (1982) noted that although the funds of corporate educational training continued growing, only 10% of the training was effective. For innovation and breakthrough, this paper conducted a random sampling of 38,000 pieces of data in the personnel educational training database of China Motor Corporation according to employees' positions and departments. This was done through the decision tree algorithm developed upon the principle of Data Mining for study, mining and analysis in order to find the relation between the classification of educational training courses and the classification of the employees' occupation, the courses favored by the employees of different occupations and their accomplishment probability. However, in the educational training, the employees' characteristics or positions will influence the time spent on providing employee education and the funds invested in education. Thus, the educational training should be precise. Since educational training tends to be conducted by the human resources department, the cognitive difference between the employees' characteristics and the human resources department might lead to insufficient funds or the waste of the investment.

It will take more time to adjust the administration's strategies when there is an imbalance of manpower supply and demand. Thus, manpower planning and the precision of prediction will influence corporate efficiency and quality. Excess manpower supply will lead to idleness and inefficient use of personnel and indirectly result in the increase of the cost and the waste of resources; on the contrary, insufficient manpower will lead to imbalanced distribution of resources and reduce the corporate quality. Thus, with complicated training courses and plenty of employees, we should determine how to classify educational training precisely and meet the manpower demand to increase educational training quality and reduce excess expenses. This research expects to effectively analyze and evaluate by database in order to modify the original educational training model, and to construct an educational training system through manpower forecasting. It will also construct an evaluation model to formulate the most accurate forecasting models as the basis for designing employee educational training and to increase the precision of employee classification and reduce the waste of human resources.

More than 95% of the industries in Taiwan are of small and medium scale, and the firms have limited resources for educational training (Farh, 1995). Nonetheless, the funds for corporate educational training continue growing, and in order to improve human resources quality, we should cultivate employees' talents for the future. This research follows the "Input-Process-Output" (IPO)

* Corresponding author. Tel.: +886 2 22519906; fax: +886 2 22552129.
E-mail addresses: d917812@oz.nthu.edu.tw, wsz1103@pchome.com.tw (S.-T. Wang).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.09.009

proposed by Bushnell (1990) as the basic framework of the educational training model, as shown in Fig. 1.

The research purposes are reorganized below:

(1) We construct a "training input, practice training process and training effect" assessment model to explore an overview of the training effect.
(2) We explore the possible mediating effect between training input and effect during the practice.
(3) We validate the feasibility of "training input, practice training process and training effect" assessment model in small- and medium-sized enterprises within the manufacturing industry in Taiwan.
(4) The educational training fund invested in the firm is one of the critical indexes for assessing "training input". We try to find the key indexes of training funds invested in the firms to explore the relationship among training funds, the training process and the training effect.
(5) We plan educational training strategies specially adapted to meet the specific needs of employees in different positions.

2. Literature review

With the advancement of Internet and database techniques, data saving in different industries becomes diverse yet simple. In order to combine the applications of the Internet, the firms promote e-data. Plenty of data collection techniques, high-efficacy, multiple-function computers and the maturity of the Data Mining algorithm are the most critical factors with respect to the rise of Data Mining, and they are generally applied in varied industries. This research therefore applies Data Mining to the analysis of educational training data. The firms should try to extract the key information from sufficient data and effectively use this information to increase the efficacy of management, improve the quality of personnel and control costs of human resources in addition to determining the employees' needs in order to design proper service strategies. So far most of the enterprises have been unable to achieve these goals and, even if they have tried, they are hindered by having improperly related activities.

This research intends to treat cluster analysis, classification technique and neural algorithm as the initial analytical tools of educational training and human resources, and designs the process by Six Sigma to continue improving the management cycle. We also analyze and predict educational training. Finally, we compare the consistency and precision of the conclusions by performance matrix. The literature reviews the available literatures related to human resources, educational training, the management of Six Sigma and data mining.

2.1. Human resources

In knowledge economy time, knowledge becomes the main component essential to increased productivity and economic growth. Of most significant value to corporate operations are the employees' wisdom and knowledge. Manpower capital is the most unique of invisible assets having core value (Cascio, 1991). Some scholars call employees "the heart, wisdom and soul of the organization" and "the only drive for the actions in the organization" (Edvinsson and Malone, 1997).

Most scholars regard educational training as the most direct investment tool for human resources. In order to avoid the waste of the resources and to improve the organizational performance of these resources by investment in training and increased productivity, the construction of effective "educational training" becomes the critical issue concerning many firms (Carlson et al., 2000).

2.2. Educational training

This research combines "education" and "training" into "educational training" and defines it as the training to cultivate the employees' knowledge, skills, habits and problem-solving capacities and to stimulate the employees' maximum potential to meet the demands of their present or future position (Charles, 1998). In terms of practicability, when the results of training or advanced studies cannot be transformed into a certain degree of performance in the related work, they will represent a kind of waste. Thus, the training should match the related stimulation measures to lead to the real outcome.

Coffey et al., (2003) describe a unique approach to the creation of an expert system to provide performance support and training for electronics technicians. The starting point for development of the system was the creation of a semantically rich knowledge model comprising Concept Maps and other digital media. The knowledge model was used to create the inference part of the system, and then retained to furnish an explanation of the inference component's behaviors, and to provide content for training. The findings of Ng, (2005) are based on a survey conducted in China; the present job characteristics and firm background were found to play key roles in determining training provisions. Workers who received off-the-job training were less likely to receive on-the-job training, while those who received on-the-job training were neither more nor less likely to have received off-the-job training. Unlike in developed countries, training in China was

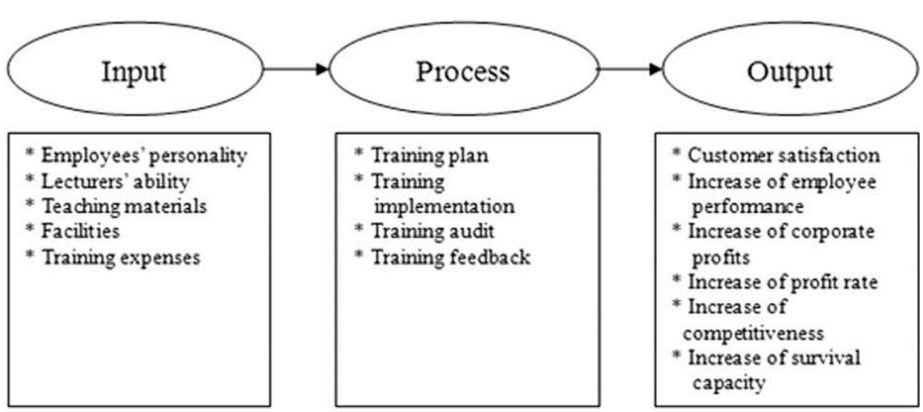

Fig. 1. Basic framework of the educational training system model source: this research.

usually intended to remedy skill deficiencies rather than enhance productivity.

2.3. Six Sigma

The managerial philosophy of Six Sigma is to thoroughly analyze the key points and analyze how they influence the input and output of the process, and deal with the problems with structural measures. The serial components of the structural dealing process are called Six Sigma management action steps. They are based on continuous improvement and feedback. The above 6σ action steps and DMAIC supported by the related statistical tools are reorganized in Fig. 2, which elaborates the statistical tool used for analysis in different steps. Su and Chou, (2008) used Six Sigma methodology to help a company achieve its expected goal through continuous project improvement. Some challenges, however, have emerged with the execution of the Six Sigma. For example, how are feasible projects generated? How are critical Six Sigma projects selected, given the finite resources of the organization? This study aims to develop a novel approach to creating critical Six Sigma projects and identifying the priority of these projects. Firstly, the projects are created from two aspects, namely, the organization's business strategic policies and the voice of the customer. Secondly, an analytic hierarchy process (AHP) model is implemented to evaluate the benefits of each project; a hierarchical failure mode effects analysis (FMEA) is also developed to evaluate the risk of each project; the priority of Six Sigma projects can also be determined from this analysis.

2.4. Data mining

Data mining, literally speaking, means to find useful knowledge from plentiful data. It resembles mine digging and extracts the useful knowledge from a complicated database. During the process, we might acquire inexplicable data or a cause-and-effect relationship outside our expectation.

2.4.1. Cluster analysis

Cluster analysis means to cluster in pairs all items or data according to similarities (or differences). Items in the same cluster will consequently be extremely similar, whereas items in another cluster will be markedly different from those in the first cluster—and those differences will be made even more apparent by the juxtaposition of those clusters. Thus, cluster analysis aims to show the similarity of data in the same cluster, where the similarity is made even more apparent by the obvious differences between and among other clusters. Lee et al. (2008) propose a method for proactive detection of Distributed Denial of Service (DDoS) attack by exploiting its architecture, which consists of the selection of handlers and agents, communication and compromise, and attack. We look into the procedures of DDoS attack and then select variables based on these features. After that, we perform cluster analysis for proactive detection of the attack. We experiment with the 2000 DARPA Intrusion Detection Scenario Specific Data Set in order to evaluate our method. The results show that each phase of the attack scenario is well partitioned, and we can detect precursors of DDoS attack as well as the attack itself.

2.4.2. Theory of decision trees

Decision Trees refers to the rules of tree structure, which further leads to classification models. In order to classify data input, we treat each node on the decision tree as a judgment formula that judges whether data input is more or less than a certain value with regard to a certain variable. With each node, we can divide data input into several classifications. Yang et al. (2005) propose an expert system called VIBEX (VIBration EXpert) to aid plant operators in diagnosing the cause of abnormal vibration for rotating machinery. In order to automate the diagnosis, a decision table based on the

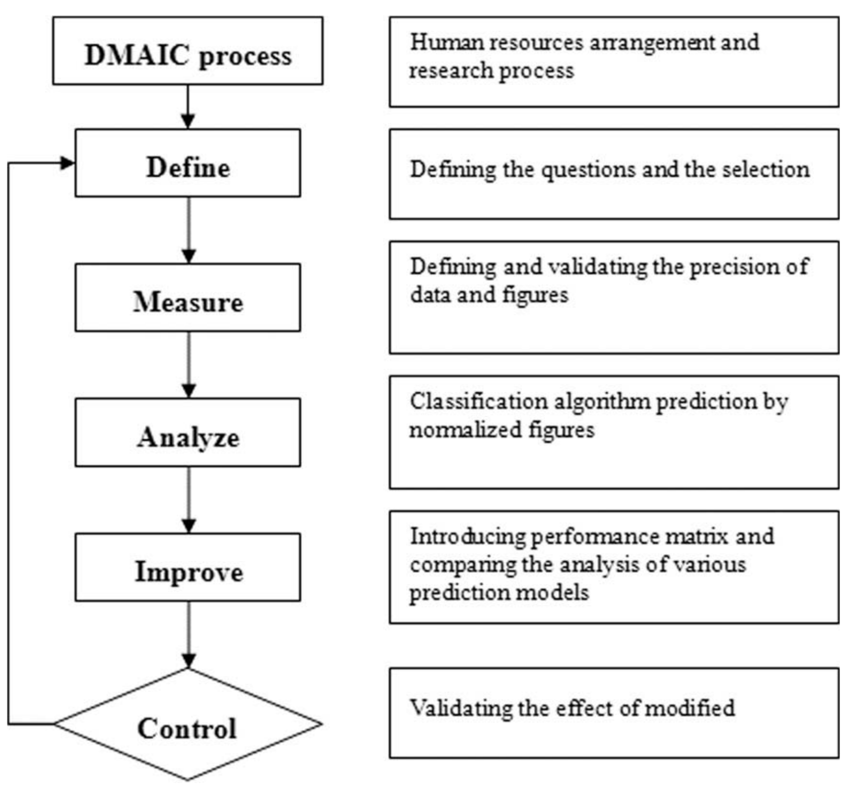

Fig. 2. Design of Six Sigma study process.

cause-symptom matrix is used as a probabilistic method for diagnosing abnormal vibration. Also, a decision tree, used in the acquisition of concept-based structured knowledge is introduced to build a knowledge base indispensable to vibration expert systems. In Piramuthu, (2008), the process of constructing these decision trees assumes no distributional patterns in the data (non-parametric); characteristics of the input data are usually not given much attention. We consider some characteristics of input data and their effect on the learning performance of decision trees. Preliminary results indicate that the performance of decision trees can be improved with minor modifications of input data.

2.4.3. Neural network

Neural network is the information processing system that simulates biological neural networks. It imitates biological neural networks by using plenty of simple and connected artificial neurons. It learns from input variables and values and constructs the patterns by continuously changing the parameters through the knowledge acquired in the learning experience. During the process, neural network requires a set of weights to output the optimized outcome. The most commonly used training is called back-propagation, which is applied to compare the input results and the acquired correct results. After each comparison, a set of modified weights is created. New input will be compared with the original value. After the repetitive process, neural networks are trained to lead to precise prediction. Schiller (2007) shows NN emulation of the forward model allows us to calculate the Jacobian of the forward model efficiently; thus the Levenberg–Marquardt optimization scheme can be used to determine the parameters of interest best fitting the measurements.

Zhao and Wang (2005) presented a feed-forward neural network model based on the LM algorithm (put forward by Levenberg and Marquardt), which is established to realize real-time identification of material properties and friction coefficient for deep drawing of an axisymmetric workpiece. Compared with the previous BP model (neural network based on back propagation algorithm) and GA-ENN (evolutionary neural network based on genetic algorithm) model, the error goal of parameter identification by the LM model is stepped downward to a new level.

3. Method

3.1. Research design

The analytical framework of this research is based on Six Sigma management and the action steps of Six Sigma when introducing the projects. After reading DMAIC measurement in Michael's (2002) Lean Six Sigma, we explain how data mining in this research is introduced for continuous improvement as the reference for other companies when introducing this research process in the future. Below we will describe data definition, data collection criterion, prediction model of data mining, improvement and control and modified models.

As for the precision analysis of educational training, we discuss with the human resources supervisors of the case firm and reorganize the actual educational training process of the said firm in Fig. 2, which includes the process of the employees' participation in the human resources training center.

3.2. Data definition

Based on the educational training system of our target firm, we explore the factors influencing the precision of the training system to find the possible reasons for the firm modifying the precision of the training. We reorganize the conclusions through the study of the training system's precision.

3.3. Data collection and measurement

After confirming our research subject, for the fit between the research samples and varied algorithms of data mining, we construct the following principles with regard to sample collection:

(1) *Elimination of noise signals:* In cluster analysis and classification algorithm, extreme values tend to interfere with the information. It becomes difficult to elaborate the analytical result. Thus, we eliminate the extreme values more than standard deviation by three times when collecting data in the variables.
(2) *Missing value:* To ensure the precision and completeness of the data, we eliminate the employees who enroll in educational training but do not actually participate or whose information or training data is incomplete.
(3) *Scope of data study:* Full-time employees in five departments of the target company from 2005 to 2007.
(4) *Modification of educational classification cases:* This refers to the employees with inferior performance after educational training in the target company from 2005 to 2006.

3.4. Analysis of demand model

The analysis of this research is based on the database and the discussion with human resources supervisors. This research collects data of personnel educational training in China Motor Corporation by cluster analysis, decision tree algorithm and back-propagation neural networks for mining analysis and classification. The content of the data includes related training conference record, training planning and written materials of training practice. "Basic information of educational training" can be found in the database of the human resources department. In terms of corporate characteristics (organizational scale, capital and establishment years), the larger-scale companies have sufficient capital and manpower and more complete systems. Therefore, they are more willing to provide the training and acquire better training effects. In terms of employee characteristics (educational level, age, work years, positions and gender), the will of the firms or the employees will influence the amount of invested training and the future effects, as shown in Fig. 3.

As for the characteristics of corporate educational training, we show the actual input by actual training expenses (dollars), total training person-times (time), total training hours (hours), average training expense (each person), average training time (time, each person), average training times (times, each person) and average training hours (hours per person), as shown in Table 1–3.

4. Results

4.1. Cluster analysis

The difference between cluster analysis, prediction and classification is that training data is not required for the former. In cluster analysis, similar data will be allocated in the same cluster. However, the clusters are extremely different and each cluster can be treated at the same level.

According to classification methods, cluster analysis can be divided into Hierarchical Cluster Analysis and Non-Hierarchical Cluster Analysis. Hierarchical Cluster Analysis is suitable for the situation with less observation and in which we select the number of clusters by Tree Diagram. It includes the Linkage Method and the Variance Method. This research applies the Linkage Method and

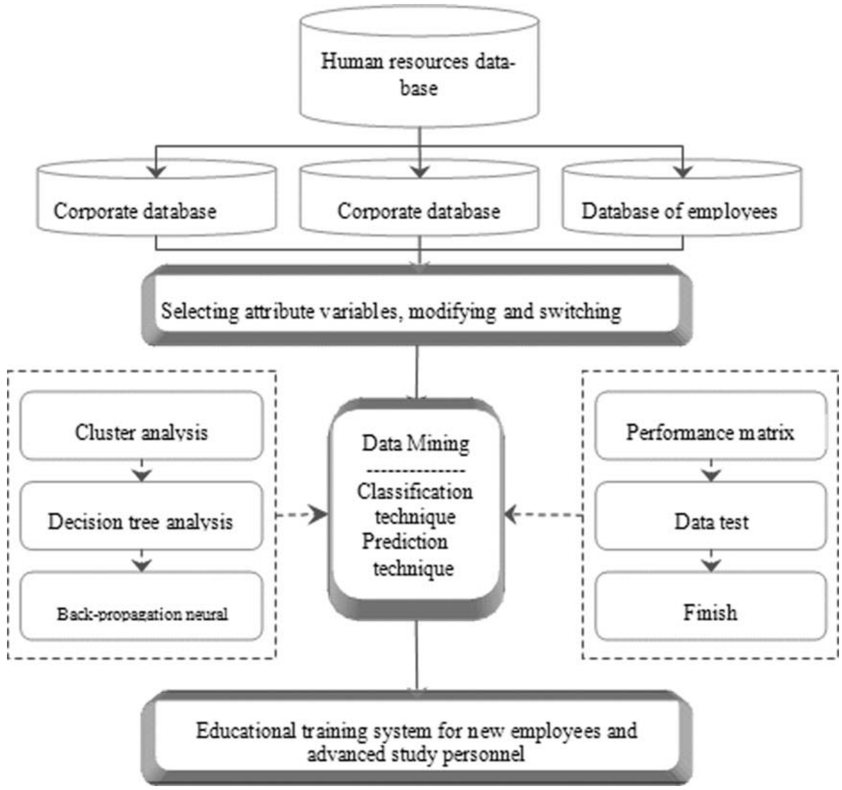

Fig. 3. Framework of data mining process.

Table 1
Corporate characteristics

Number of employees (people)	Capital (10,000 dollars)	Establishment years (year)
1–200 (including)	Less than 80 million (including)	1–10 years (including)
200–400	More than 80 million	10–20 years
More than 400		More than 20 years

Table 2
Employee characteristics

Educational level	Age	Work years
Others	Less than 20 years (including)	Less than 1 year (including)
Junior high school	20–30 years (including)	1–5 years (including)
Senior high school (vocational school)	30–40 years (including)	5–10 years (including)
College	40–50 years (including)	10–15 years (including)
University	50–60 years (including)	15–20 years (including)
Master or doctor	More then 60 years	More than 20 years
Position	Department	Sex
Administrator, on-site personnel, service personnel	Production	Male
Department manager or department supervisor	Marketing	Female
Director, general manager or manager	Human resources	
	Finance	
	R&D	

conducts the clustering from the view of more and less variance among the clusters.

In Table 4, step 1 is the combination of observation 11 and observation 24. The difference factor after combination is 202. The new observation number (new cluster) is 11 (present numbers include 11 and 24). We then move to step 4 (or phase 4).

In step 4, observation 18 and 11 are combined. The difference factor after combination is 441. Observation is combined for the first time. The previous combination is in phase 0. Observation 11 (new observation number refers to the lesser one) in new cluster No. 11 includes three observations, No. 11, 18 and 24. The new cluster No. 11 will be combined in step 13.

In step 13, observation 11 and 6 are combined and the difference factor after combination is 13157.107. The previous combination of observation 11 is in step 1. The previous combination with observation 6 is in step 0. The new number after the combination

Table 3
Characteristics of corporate educational training

Actual training expense (dollars)	Total training person-time (times)	Total training hours (hours)
(Including) Less than 10	(Including) Less than 100	(Including) Less than 1000
10–30 (including)	100–500 (including)	1000–5000 (including)
30–50 (including)	500–1000 (including)	5000–10,000 (including)
50–70 (including)	1000–1500 (including)	10,000–15,000 (including)
Above 70	Above 1500	Above 15,000
Average actual training expense/person (dollars)	Average training times/person (time)	Average training hours/person (hours)
(Including) Less than 1000	(Including) Less than 5	(Including) Less than 50
1000–5000 (including)	5–10 (including)	50–100 (including)
5000–10,000 (including)	10–15 (including)	100–150 (including)
10,000–15,000 (including)	15–20 (including)	150–200 (including)
Above 15,000	Above 20	Above 200

of two observations is 6. New cluster No. 6 includes 4 observations, No. 6, 11, 18 and 24. New cluster No. 6 will be combined in step 14.

In step 14, observations 6 and 7 are combined. The difference factor after combination is 17626.875. The previous combination with observation No. 6 is in step 10, and that of observation 7 is in step 0. The new number of two combined observations is cluster No. 6. The next combination will be in step 19.

In step 19, observation 6 and 13 are combined, and the difference factor after combination is 58594.298. The previous combination with observation No. 6 is in step 14. The previous combination with observation 13 is step 15. New number of two combined observations is cluster No. 6. The next combination will be in step 23.

In step 23, the difference factor of observation No. 1 and 6 abruptly becomes 213318.310. The difference between the two is significant. Therefore, they should not be combined. In addition, in step 24, the increase of the difference factor between observation No.1 and No. 6 is significant. The factor value is 213318.310. Thus, they should also not be combined. Therefore, it is proper to allocate 25 observations into 3 clusters:

(1) *No. 1 cluster:* including observation 1, observation 2, observation 3, observation 4, observation 5, observation 8, observation 10, observation 12 and observation 17.
(2) *No. 6 cluster:* including observation 6, observation 7, observation 11, observation 13, observation 14, observation 16, observation 18, observation 19, observation 20, observation 21, observation 23 and observation 24.
(3) *No. 9 cluster:* including observation 9, observation 15, observation 22 and observation 25.

In Fig. 4, we draw the Tree Diagram with the result of the agglomeration process and the convergence order above to allow the users to have a clearer analysis.

The analyzed data is reorganized into Table 5. The departments include five constructs: production, marketing, human resources, R&D and finance. There are the following three clusters:

(1) Cluster 1 is "general courses" and the participation rate in different departments is high.
(2) Cluster 2 is the "core course" of automobile manufacturing. The participation rate in the production and R&D departments is high, whereas the rate in the marketing, human resources and finance departments is low. Thus, it is the professional course of automobile manufacturing.

Table 4
Cluster agglomeration process

Phases	Combination cluster		Coefficient	Phase cluster appears first		Next phase
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	11	24	202.000	0	0	4
2	18	20	441.000	0	0	4
3	4	12	854.500	0	0	9
4	11	18	1457.000	1	2	13
5	13	19	2061.500	0	0	6
6	13	16	2699.000	5	0	15
7	6	23	3378.500	0	0	10
8	3	5	4406.000	0	0	18
9	4	10	5449.167	3	0	11
10	6	21	6737.000	7	0	13
11	1	4	8158.083	0	9	16
12	9	22	10463.083	0	0	17
13	6	11	13175.107	10	4	14
14	6	7	17626.875	13	0	19
15	13	14	23144.375	6	0	19
16	1	2	28670.225	11	0	18
17	9	25	34751.892	12	0	21
18	1	3	43412.506	16	8	20
19	6	13	58594.298	14	15	23
20	1	17	75078.208	18	0	22
21	9	15	96696.042	17	0	24
22	1	8	123574.861	20	0	23
23	1	6	213318.310	22	19	24
24	1	9	437068.640	23	21	0

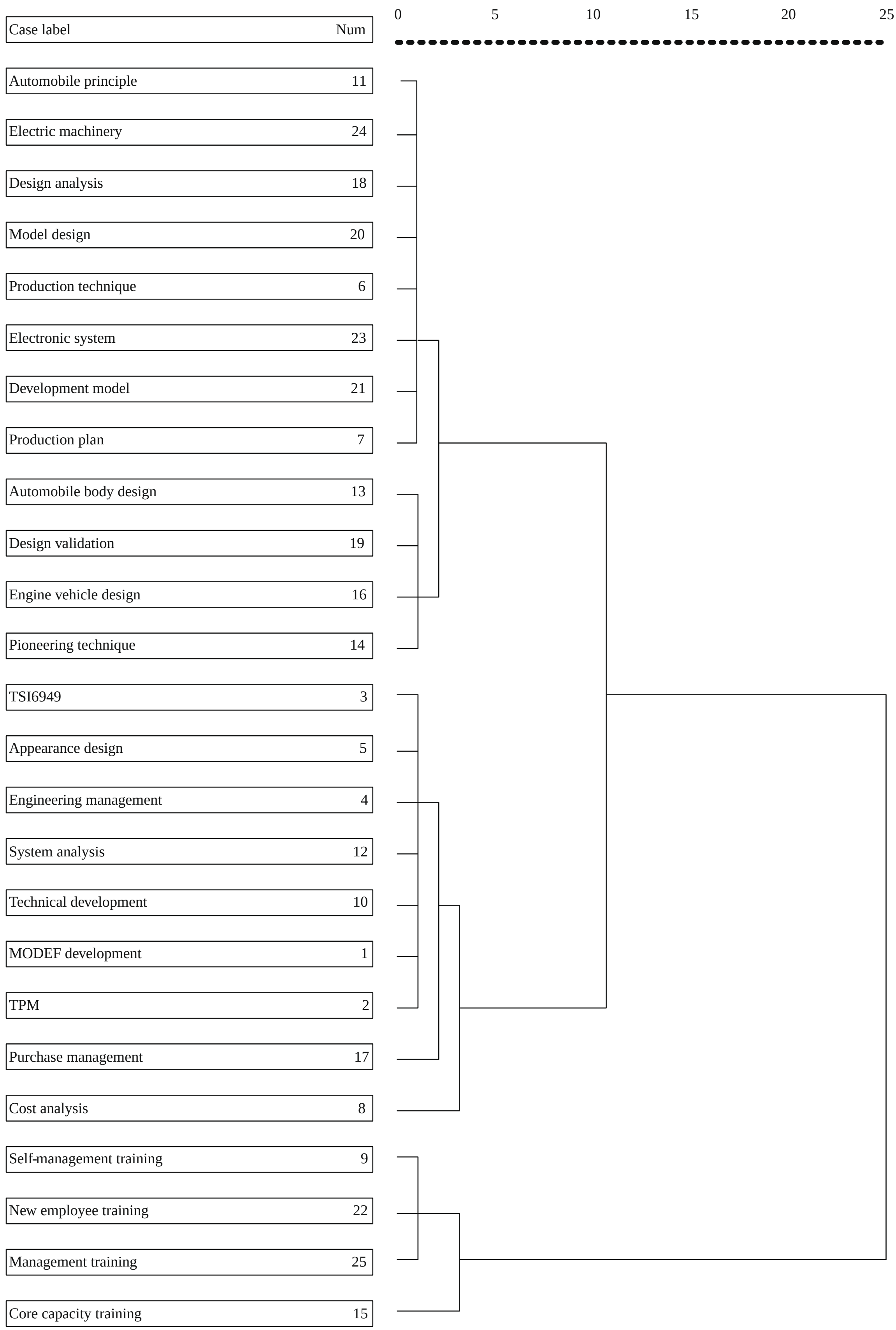

Fig. 4. Tree Diagram of classification result of Hierarchical Cluster Analysis.

Table 5
Classification result

Cluster	Corporate business Elective courses	Automobile manufacturing professional core courses	General courses
Training courses	• TS16949 • Appearance design • Engineering management • System analysis • Technical development • MODEF development • TPM • Purchase management • Cost analysis	• Automobile principle • Electric machinery • Design analysis • Model design • Production technique • Electronic system • Development model • Production plan • Automobile body design • Design validation • Engine vehicle design • Pioneering technique	• Core capacity training • New employee training • Management training • Self-management training

(3) Cluster 3 is "advanced study course of corporate business". The participation rate of the marketing, human resources and finance departments is high, and the qualification rate is high.

4.2. CHAID mining of decision tree algorithm

By decision tree algorithm, we can divide and classify data according to the significant standard set and construct of a tree with the classified events. On the tree, we can generalize some event rules and find the relationship among the events on which to base predictions.

In the tree diagram, we can test the significant standard fixed for certain attributes of the event by each internal node and arrange the branches according to the significant standard. Each branch is the probable value of the said attribute or the set of several probable values. In other words, there might be more branches under the branches of each internal node, and they can be the internal node of the next branch for the test on the significant standard of another attribute. When the internal node does not reach the significant standard, there will be no more branches. At the point, the internal node is called a leaf node. We can explain the reasons and results of the attributes in the event through inferring the information showed on each leaf node along the path upward. After mining data with CHAID in the decision tree algorithm, the tree diagram of target variables and different prediction variables.

In the tree diagram in Fig. 5, we can find the selection of the courses of five departments from the employees participating in the training. There is significant difference between high-rank, medium-rank supervisors and basic employees, which refers to two nodes; the accomplishment rate of high-rank supervisors (the first node) participating in the training is up to 86.29%.The accomplishment rate of medium-rank supervisors and basic employees (the second node) participating in training is up to 92.75%, which is much higher than the first node. Since p-value of the first node is less than the 0.05 significant standard set by this paper, we will have more branches for the first node. With regard to the first node, we find that we can allocate the occupations of the participants into human resources, finance, production and R&D and marketing (2 nodes). In the human resources, finance, production and R&D departments, there are 88.64% participants enrolling in the training (the third node); in the marketing department, the accomplishment rate of the participants joining in training is 77.00% (the fifth node). Since p-value of the second node (medium-rank supervisors and basic employees) is more than the 0.05 significant standard set by this paper, we will have no more branches. According to Table 6, we find that the precision rate of analytical result of the data is up to 99.89%.

4.3. Back-propagation neural networks

4.3.1. Construction of neural networks

With optimized algorithms with different standards, the back-propagation algorithm leads to varied algorithms. Generally speaking, with regard to Function Approximation and the network with plenty of weights, Levenberg Marquardt back-propagation (LMBP) reveals considerably precise training. Compared with other algorithms, LMBP acquires lower mean square error. Although the saving demand of memory of LMBP is much more than other algorithms, this research still selects it as the learning principle of the network with regard to the above consideration.

After confirming the learning algorithm of back-propagation neural networks, this research classifies the units of input layer, hidden layer and output layer. In the input layer, there are 6 variables (educational level, age, work year, position, department and gender) that can influence training classification. As to the output layer, there are three levels of target value of training classification: first, second and third levels.

In the hidden layer, one to two layers will usually lead to the optimized convergent effect. Excess or few layers will influence the convergent effect of the network since one or two hidden layers can properly reflect the interaction among input units, and excess hidden layers will result in complicated network and local absolute maximum. Generally speaking, in reality, one hidden layer and the proper number of artificial neurons can solve complicated function reflection. Two hidden layers are only needed when there are more training samples or when non-linear correlation is significant. Based upon the above literatures, this research operates back-propagation neural networks by one hidden layer; the structure of neural networks is drawn in Fig. 6 (Zhao & Wang, 2005).

This research constructs the basic framework of neural networks. However, the number of artificial neurons in the hidden layer and transfer functions in the hidden layer and the output layer are not set. In the following, we will elaborate on the optimization of parameters in back-propagation neural networks.

Example of input layer:

Data of the employee no. 1234 = (6, 2, 1, 1, 3, 1).

In output layer, there are 4 levels of target value of training classification. Thus, this research changes every piece of data in the output layer into vectors. The solution of output vectors is whether it is in the training of level 1. If the answer is "yes", the network output will be 1. If the answer is "no", the network output will be 0.

Example of output layer:

- Training of the first level = (1, 0, 0)
- Training of the second level = (0, 1, 0)
- Training of the third level = (0, 0, 1).

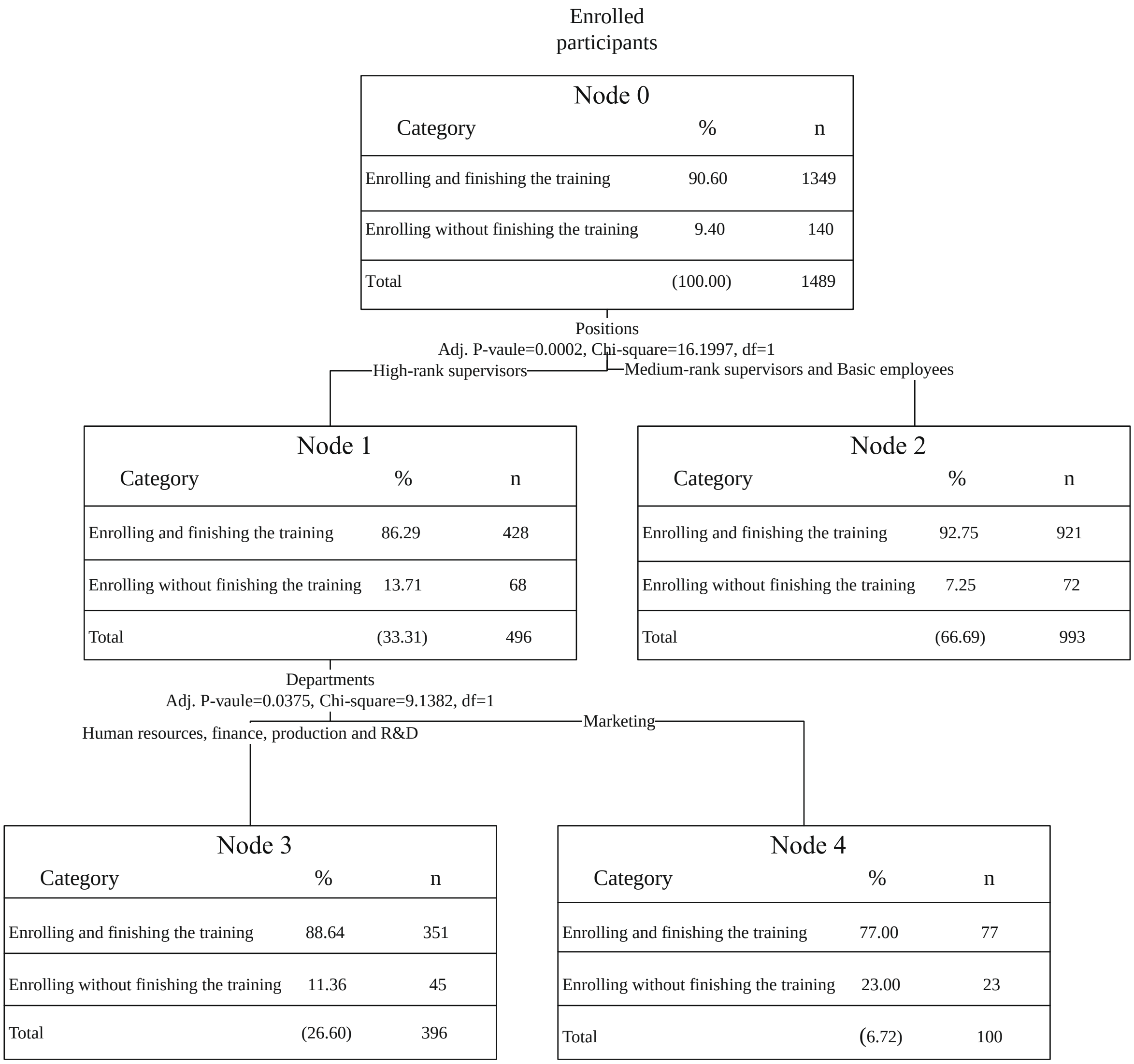

Fig. 5. Tree Diagram of employees who have participated in educational training courses.

Table 6
Precision rate of analysis data in the Tree Diagram

	Risk statistics
Risk estimate	0.0021363
SE of risk estimate	0.000180357

4.3.2. Optimized number of artificial neurons

The number of artificial neurons in the hidden layer is significantly influenced by the network quality. When there are more artificial neurons, the convergence will be slower, a situation that cannot reduce the error. On the contrary, when there are fewer artificial neurons, the correlation between input and output cannot

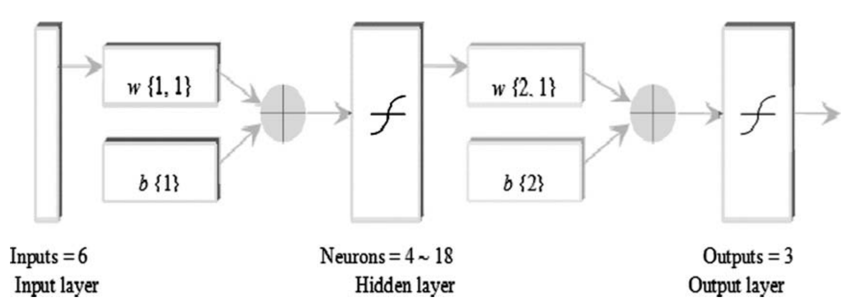

Fig. 6. Structure of neural networks.

Table 7
Precision rate of training classification (back-propagation neural networks)

		Results of classification and prediction (training)				
		Training of the first level	Training of the second level	Training of the third level	Training of the fourth level	Total
Number	Training of the first level	405	45	0	0	450
	Training of the second level	12	389	46	3	450
	Training of the third level	0	6	421	23	450
Percentage (%)	Training of the first level	90.00	10.00	0.00	0.00	100
	Training of the second level	2.67	84.45	10.23	0.67	100
	Training of the third level	0.00	1.33	93.56	5.11	100
Total precision rate of training: 91.4%						

be properly reflected. Generally speaking, the number of artificial neurons in the hidden layer can be based on those in the input layer (6) and the output layer (4) (Yeh, 1999):

(1) Simple question: number of artificial neurons in hidden layer = (number of input artificial neurons + number of output artificial neurons) ÷ 2 = 4.
(2) General question: number of artificial neurons in hidden layer = (number of input artificial neurons + number of output artificial neurons) = 9.
(3) Difficult question: number of artificial neurons in hidden layer = (number of artificial neurons + number of output artificial neurons) × 2 = 18.

Thus, in order to test the optimized number of artificial neurons, this research increases the number of artificial neurons from 4 to 18 in the hidden layer. As to the transfer function of the hidden layer and the output layer during the test, this research replaces it with hyper-tangent function (Tansig). After confirming the optimized number of artificial neurons, this research tests the optimized combination of transfer function of the hidden layer and the output layer.

4.3.3. Classification result

With the result, we conduct the classification and acquire the classification precision rate of training (1800 samples) (see Table 7). After reorganization, 405 out of 1800 samples are classified and predicted as the first level; 389 samples are classified and predicted as the second level; 421 samples are classified and predicted as the third level; 431 samples are classified and predicted as the fourth level. A total of 1646 samples are classified precisely; the precision rate is 91.40%.

According to back-propagation neural networks in this research, we find that the classification precision rate of back-propagation neural networks on training is up to 91.40%.

5. Discussion and conclusion

Based on the results of cluster analysis and decision tree algorithm, this paper draws the following conclusions. First of all, according to cluster analysis, we find that there are more production and R&D courses in the China Motor Corporation and fewer marketing, human resources and finance courses. We suggest that the company increase the number of courses in the latter category. Employees in all departments can participate in the planning of general courses, and the advanced study and core courses can be much better organized. Employees may not be satisfied with their present work and therefore intend to learn other skills to change their occupation. Provided they are aware of this situation, supervisors can focus on meeting the career-building needs of their workers, with the goal of retaining their employees. With cluster analysis, we can indicate the similarity of certain courses with the goal of arranging a series of courses in order to provide effective educational training, encourage the employees to systematically join in the courses, increase the qualification rate, match the work content with the employees' profession and aspirations and increase work efficiency. Finally, according to decision tree algorithm, we find that human resources, finance, production and R&D courses in an educational training center are advantageous. The participants in the educational training are the high-ranking supervisors. Based on the above two kinds of algorithms, we find that different algorithms show the information of different constructs. The decision makers can thus control the current situation of employee performance through continuing education. It also shows the precision of the attributes among the variables in this paper.

We apply data mining and try to find the correlation between the course classification of educational training and the positions of those employees participating in educational training from the database of the China Motor Training Center. We first recognize the cluster relation among the above three variables by cluster analysis. According to the Tree Diagram, we identify the similarity of certain courses in the China Motor Training Center in order to develop a series of training courses; we further find the decision tree diagram of three variables by CHAID in the decision tree algorithm. According to the tree diagram, we find the most favored, prestigious courses in the China Motor Training Center's training and the preferred training courses of the employees in different positions. We also realize that the qualification rate of those joining in training courses of different classifications is higher. We indicate the correlation between the classification of training courses and the employees' positions, and between the favorite course classification of the employees with different positions and their accomplishment rate.

Based on the results of the two kinds of algorithms, China Motor Training Center can plan its training courses according to its employees' different positions and can construct prestigious, effective educational training courses with high efficiency.

References

Bushnell, D. S. (1990). Input, process, output: A model for evaluating. *Training and Development Journal*, 41–43.

Cascio, W. F. (1991). *Costing human resources: The financial impact of behavior in organization*. Boston: PWS-Kent.

Charles, F. W. (1998). *Management development through training*. Reading Massachusetts: Addison Wesley publishing Company.

Carlson, D. S., Bozeman, D. P., Kacmar, K. M., Wright, P. M., & McMahon, G. C. (2000). Training motivation in organizations: An analysis of individual-level antecedents. *Journal of Managerial Issues, 12*, 271–287.

Coffey, J. W., Can~asa, A. J., Roger Carffa, G. H., Reichherzera, T., & Suria, N. (2003). Knowledge modeling and the creation of El-Tech: A performance support and training system for electronic technicians. *Expert Systems with Applications, 25*, 483–492.

Edvinsson, L., & Malone, M. S. (1997). *Intellectual capital: The realizing your companys true value by its hidden roots*. New York: Harper Business.

Farh, J. L. (1995). Human resource management in Taiwan, The Republic of Chain. In L. F. Moore & P. D. Jennings (Eds.), *Human resource management on the Pacific Rim: Institutions, practices*. Berlin: Walter de Gruyter.

Georgenson, D. L. (1982). The problem of transfer calls for partnership. *Training and Development Journal, 36*, 75–78.
Lee, K., Kim, J., Kwon, K. H., Han, Y., & Kim, S. (2008). DDoS attack detection method using cluster analysis. *Expert Systems with Applications, 34*, 1659–1665.
Michael, L. G. (2002). *Lean Six Sigma for Service*. New York: McGraw-Hill.
Ng, Y. C. (2005). Training determinants and productivity impact of training in China: A case of Shanghai. *Economics of Education Review, 24*, 275–295.
Piramuthu, S. (2008). Input data for decision trees. *Expert Systems with Applications, 34*, 1220–1226.
Schiller, H. (2007). Model inversion by parameter fit using NN emulating the forward model: Evaluation of indirect measurements. *Neural Networks, 20*, 479–483.
Su, C. T., & Chou, C. J. (2008). A systematic methodology for the creation of Six Sigma projects: A case study of semiconductor foundry. *Expert Systems with Applications, 34*, 2693–2703.
Yang, B. S., Lim, D. S., & Tan Andy, C. C. (2005). VIBEX: An expert system for vibration fault diagnosis of rotating machinery using decision tree and decision table. *Expert Systems with Applications, 28*, 735–742.
Yeh Y. C. (1999). Application and practice of neural network model. Scholars Books, Taiwan.
Zhao, J., & Wang, F. (2005). Parameter identification by neural network for intelligent deep drawing of axisymmetric workpieces. *Journal of Materials Processing Technology, 166*, 387–391.