

Intelligent Decision Support System with Embedded OLAP
Technology for the Insurance Industry

Ref. No. PQ-4555

By

Dr. Vincent Cho and Dr. Eric Ngai

Department of Management

Intelligent Decision Support System with Embedded OLAP Technology for the Insurance Industry

1. Introduction

Decision-support systems are increasingly important applications of databases. Corporations like banking, insurance, retail and health-care are beginning to use the accumulated operational data to help understand and run their business. Towards this purpose, data from the different operations of a corporation are reconciled and stored in a central database commonly called a “data warehouse”. Analysts use the data warehouse to extract the business information that enables better decision making. This interactive decision support process is called OLAP (On Line Analytical Processing).

Traditionally regular reports are compiled and printed during the non-office hours. This is to release the burden of handling daily transactions. Moreover the structure of a traditional database is mainly tailored for daily operations. Usually the reporting procedure and complicated ad-hoc queries from senior management such as cross tabulation require retrieval and merging of historical transaction records. This process does disturb the normal daily operation especially for some operational intensive business such as banking, or airlines. Thus these ad-hoc queries are almost impossible to be carried out on-line. Delay till non-office hour is necessary, this will somehow hinder managers in making instant decision and hence induce a less effective performance within a firm. In order to compliment these difficulties, some common statistics such as sales summaries across different regions, different product catagories, and different customer segments are updated when a transaction is recorded into a database. These common statistics representing the multidimensional view of the database and allow some typical analysis to be done on-line with less interruption to the daily transaction operations. This is the first step towards On Line Analytical Processing, OLAP. Academically this layout our new step on information management. However, there are various issues, which we need to overcome, to be described as follows.

The analyses of these multi-dimensional data pose new challenges. Besides statistical summaries from different views of data, data mining techniques would also be applied in conjunction with OLAP to form an integrated business solution. Among forthcoming evolutions in this domain, data mining is expected to be the most important. However, the current data mining techniques are not much connected with existing OLAP techniques. Secondly, most data mining techniques like C4.5, CN2, ITRules are stand-alone software with many parameters. Also the interfaces to these techniques are not user-friendly and required an expert to drive the software. Out of these problems, the most significant one is that these techniques are usually applied to traditional databases. One academic challenge is to research the current data mining techniques on how they can suit the new multi-dimensional structure which is adopted in the data warehouse architecture. Another academic challenge is how to design an effective interface for senior executives to drive all those sophisticated techniques. This is related to ergonomics. The main goal of this paper is to illustrate data mining techniques to the multi-dimensional data so as to build an intelligent decision support system for the insurance industry.

Insurance, as an example, taking an important role in today society and business, offers a comprehensive range of products like life insurance, saving plan, investment-linked assurance, children's education fund, hospital care, disability income protection, and unit trusts to meet different needs. The sizes of insurance or related financial companies in Hong Kong may vary from several hundreds to a few thousand employees. The funds under management for these corporate are usually in the order of billion US dollars. The captured data in an insurance company, which includes insurance premium, claims, customer profiles, and staff profiles, will definitely need a data warehouse to manage. Moreover, most insurance companies pay attention to their customers' needs. The current problem in the industry is that not many corporate have incurred data mining techniques in their data warehouse. A study has been done by Brockett et al. related to investigating neural network to predict casualty insurance [2]. However, it is again working on traditional database rather than data warehouse. This drives us to investigate our IDSS framework for an insurance company. This includes studying the characteristics of the

data warehouse in an insurance industry and researches on appropriate data mining techniques for the industry.

We try to research on three data mining functions: statistical discriminant analysis, decision trees (DT) classification and artificial neural networks (ANN), and investigate how these could be embedded into OLAP techniques for the new multi-dimensional structure as adopted by the data warehouse. Somehow our research can help an insurance or financial business management to provide support for their market strategy. Some potential applications, taking insurance as an example, are listed as follows.

- Claims analysis - determine which medical procedures are claimed together
- Predict which customers will buy new policies
- Identify behaviour pattern of risky customers
- Identify fraudulent behaviour
- Identify a right person as employee or sales agent
- Forecasting amount of claims and cost of medical coverage; classifying most important elements that effect medical coverage; predicting which customers will buy new policies

2. Framework and Methodology

In recent years, there has been an explosive growth in the use of databases for decision support. This phenomenon is a result of the increasing availability of new technologies to support efficient storage and retrieval of large volumes of data, namely data warehousing. A data warehouse can be defined as an on-line repository of historical enterprise data that is used to support decision-making [10]. Along with the data warehouse, On-Line Analytical Processing (OLAP) is a technique for the analysis and navigation of data stored in multidimensional data warehouses [11]. It was introduced by E.F. Codd [6] with the aim of providing point-and-click simplicity in decision support systems. It particularly gained popularity as a building block of modern executive information systems.

The current challenges associated with enterprise OLAP are how to link these techniques with appropriate user-friendly interface to a decision support system and how to enhance the OLAP techniques so as to incorporate data mining. The architecture for the decision support system should be flexible enough to drive the changing OLAP techniques that are applied to a growing and huge data warehouse. The OLAP techniques, besides aggregating data from various point of views, should include latest data mining techniques. Our initiative in this research is to fill these gaps. Considering the following framework as shown in Figure 1, we are going to build an intelligent decision support system using OLAP techniques extracting the information from the data warehouse in an insurance company. The extracted data, called the data mart, is isolated from data warehouse so as to reduce the interruption on daily operations. Data mart is somehow reflecting the data related to a particular application. The update frequency of a data mart will depend on its application as determined by the senior management.

A well designed interface for the IDSS will allow senior executives to drive the OLAP techniques easily. This interface will cater for the control process that a senior manager will only need to click a few keys as defined in an application. The output is also designed to be highly interpretable and be relevant to a particular decision making.

Three common data mining techniques – discriminant analysis, decision tree, and neural network are studied to adopt the multi-dimensional feature of a data warehouse in an insurance industry. Their basic features as well as corresponding enhancement for the data warehouse adoption will be explained in the next sections. In particular we will concentrate on human resource management hoping to enhance the insurance agent selection process. We have three data mining applications that try to predict 1) the length of stay of a new insurance agent, 2) the expected sales premium to be brought into a company, 3) the persistency of those sales premium under the new insurance agent monitoring.

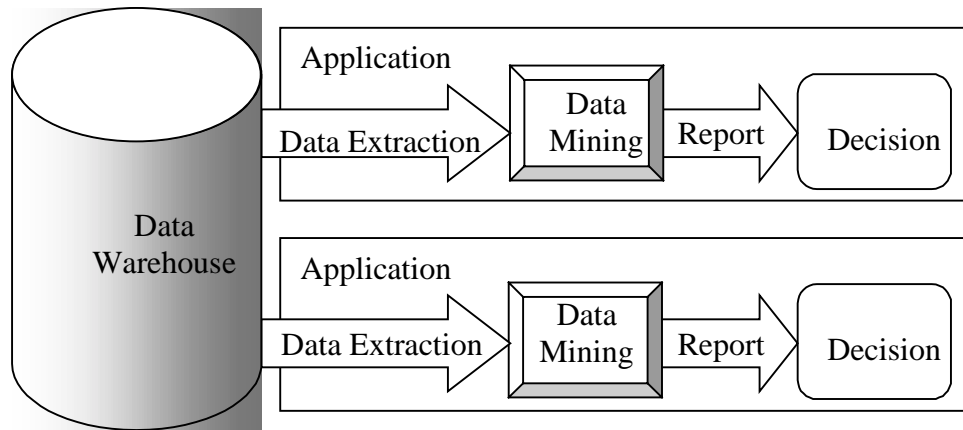


Figure 1: Framework for the Intelligent Decision Support System

2.1 Discriminant Analysis

Discriminant analysis is a statistical method using linear functions to distinguish groups. Linear discriminant functions work best and achieve an optimal solution if the underlying distributions are Gaussians and having equal covariance. When some of the underlying assumptions are not fulfilled, its classification power may decrease. However, people still like to use discriminant analysis for classification due to its simplicity. In identifying a linear discriminant function, the training error, which specifies the average loss incurred in classifying a set of training examples, is minimised. This statistical method also illustrates a number of very important principles that can be used more fully in neural networks [7].

The multidimensional data, which are the aggregated data from individual records on different aspect, will be associated with those training examples. For instance, the total sales premium, persistency index of those sales premiums, and the total on-the-job training hours will be associated with individual insurance agent record. Upon the training process, corresponding discriminant functions will be stored back to the data warehouse for future decision support. The discriminant functions will be updated periodically by retraining on latest data and outdated discriminant functions will be kept for comparison and trend analysis.

2.2 Neural Networks

Neural network is a computing technique inspired by the function of nerve cells in the brain. They are composed of many parallel, interconnected computing units. Each of these performs a few simple operations and communicates results to its neighbouring units. Typically the nodes of a neural network are organised into layers with each node in one layer having a connection to each node in the next layer as shown in Figure 2. Associated with each connection is a weight and each node has an activation threshold. During pattern recognition, each node operates as a simple threshold device. A node sums all the weighted inputs (multiplying the connection weight by the state of the previous layer node). The weighted sum is then fed into an (typically non-linear) activation function for the node's output.

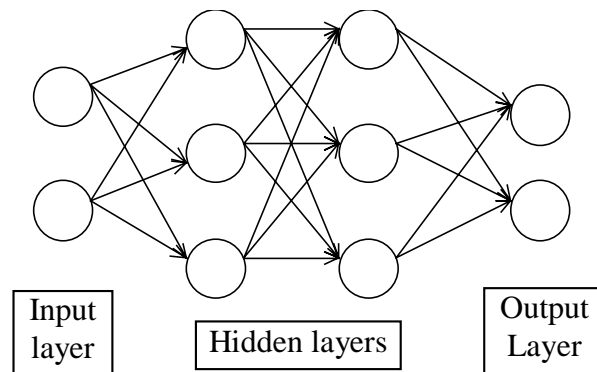


Figure 2: Neural Network

It is the connection weights and the topology that determine ability of the network and its the application. A learning algorithm is a procedure used to find the values of these weights for a given application. A popular neural network learning algorithm is the back-propagation algorithm [1, 17]. The deviation of output from its correct value for a case is back-propagated through the network; all relevant connection weights and unit biases are adjusted, using gradient descent, to make the actual output closer to the target. Training continues until the weights and biases stabilise. However, in order to avoid overfitting all neural networks should be thoroughly validated on testing data - data outside the training set [18].

In contrast with current statistical methods, which are usually linear-based, neural networks would cater for a non-linear system. Neural networks are also very effective in learning cases that contain noisy, incomplete or even contradictory data. The ability to learn and the capability to handle imprecise data make them very effective in handling financial and business information. A main limitation of neural networks is that they lack of explanation capabilities. They do not provide users with details of how they reason with data to arrive at particular conclusions. Neural nets are black-boxes which accept input and the user has to believe in the correctness of the corresponding output. Another limitation is the relative slowness of the training process. It typically takes order of magnitudes longer to train a neural net than to build a tree or generates a regression equation.

In case of handling the training data from the data warehouse, we associate various aggregated summaries or performance indices to a training record and we keep the trained neural network for future use. Regularly, we update the neural network by retraining on the latest data mart.

2.3 *Decision Trees*

A decision tree is a set of conditions that used to classify an object. Decision trees are usually represented by a graph of leaves and stems. A leaf in a decision tree identified by class, while a stem represented a condition of an attribute of a measured object. Decision tree is popular for description of a discrete class dependent on a logical expression about the attributes' value. Moreover, it can provide explanatory classification and its learning algorithms are fast.

There are several popular induction techniques such as ID3, C4.5, CART, and CHAID for building a decision tree. Most of the decision tree systems use an overfit-and-simplify learning strategy to handle noisy data: a hypothesis is formed by firstly growing a complex tree which "overfits" the data, and then simplifying or pruning the complex tree. The process of growing is usually governed by some heuristic functions such as binary split which recursively partition the data. CART uses the Gini index to measure the impurity at a node, and then chooses the split to maximize the reduction in impurity; ID3

and C4.5 use a measure of information gain; and the CHAID decision tree algorithm uses a splitting criterion based on chi-square test statistic.

On the pruning process, there are two approaches depending on whether or not a testing data set is used to estimate the error rate of a decision tree. Cost-complexity pruning and reduced-error pruning use separate testing cases to access the error rates. The approach taken in C4.5 belongs to the second family of techniques that use only the training set from which the tree is built. The method named pessimistic pruning is inspired by a statistical correction based on prior probability distribution of error cases.

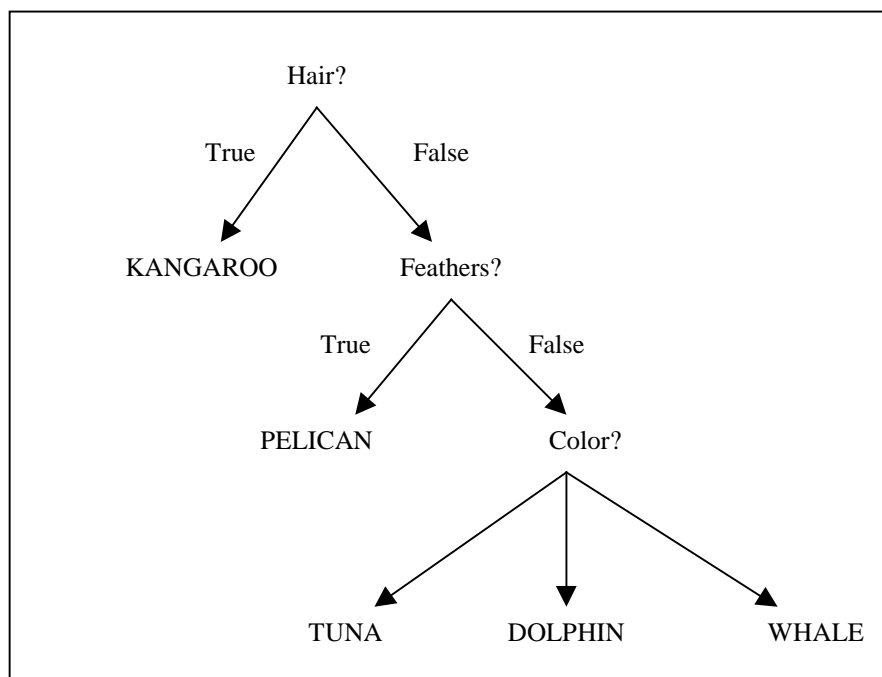


Figure 3: A Simple Decision Tree (Adapted from Sestito and Dillon, 1994)

In case of using a decision tree in a data warehouse environment, its leaves will take various aggregated attributes. Also the tree is updated periodically with its history being stored up for trend analysis.

3. Data Collection

Having contacted most of the insurance corporate in Hong Kong, we come up with an eminent insurance operator, which agrees to provide us their agency performance data for

our analysis and design. The data set includes records of more than 3000 agents over its last three years of operations in Hong Kong. The attributes to be considered includes sex, date of birth, nationality, academic level, number of dependents, job position, working experience, management experience, amount of insurance brought, eligibility of selling particular products, hire date, termination date (if any), previous job nature, and previous annual income.

4. Experiment

An experiment was performed by applying the three data mining methods in predicting 1) the length of service, 2) sales premium and 3) persistency index of insurance agents. These three predicted variables would definitely help a manager to select an insurance agent in a recruitment process. Within the three years period from 1998 to 2001, 3053 records of agents are used for training purpose while another 500 randomly selected records are used for testing. In the experiment, Neural networks used are single hidden layer feed-forward model, Decision Trees are built by C4.5 and the Discriminant analysis is making use of the Fisher's linear discriminant function. Nominal attribute such as sex is represented by dummy variables such as sex1, sex2 in Discriminant and ANN models. In this case the dummy variables of sex1 and sex2 are binary and exclusive.

4.1 Predicting Length of Service

Length of service was classified into two classes, less than one year (class 1) and more than or equal to one year (class 2). Obviously, we intent to employ someone who will work for more that one year (class 2). The result shows that the accuracy predicted by Discriminant for training and testing data are 60.07% and 57.20% respectively; the accuracy of ANN for training and testing data are 69.28% and 61.40% respectively; the accuracy of Decision Tree for training and testing data are 75.30% and 59.40% respectively. Table 1 shows the classification matrix of testing data predicted by the three methods. ANN gives the highest accuracy, $(37.6 + 23.8)\% = 61.4\%$, for unseen data. To interpret the percentages, take the figures of ANN model for illustration, if an agent candidate has been predicted as class 1, there is $37.6 / (37.6+21.8) = 63.3\%$ that she

actually is in class 1. On the other hand, if the candidate has been predicted as class 2, there is 58.6% that she actually is in the class 2. That is, predicting a candidate as class 1 would be more accurate than class 2. Comparing the results in Table 1, The ANN model gives the highest conditional accuracy for both classes 1 and 2 in predicting length of service.

Table 1: Predict length of service

		ANN Predicts		Discriminant Predicts		DT Predicts	
		Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Actual	Class 1	37.6 %	16.8 %	28.8 %	25.6 %	42.6 %	11.8 %
	Class 2	21.8 %	23.8 %	17.2 %	28.4 %	28.8 %	16.8 %

Figure 4 shows a decision tree that predicting length of service. In the generated tree, “job position” and “insurance brought” are the two most important factors that are used to determine the length of service.

```

POSITION CODE in {EC,ME}:
| LIFE INSURANCE BROUGHT (SUM INSURED) <= 464700 : < 1 year (84.0/13.9)
| LIFE INSURANCE BROUGHT (SUM INSURED) > 464700 :
| | PREVIOUS JOB NATURE = C: >= 1 year (1.0/0.8)
| | PREVIOUS JOB NATURE = S: < 1 year (3.0/2.1)
| | PREVIOUS JOB NATURE = P: < 1 year (1.0/0.8)
| | PREVIOUS JOB NATURE in {M,B}:
| | | Age <= 38.6301 : >= 1 year (8.0/2.4)
| | | Age > 38.6301 :
| | | | SEX = F: < 1 year (4.0/1.2)
| | | | SEX = M:
| | | | | Age <= 43.663 : < 1 year (7.0/2.4)
| | | | | Age > 43.663 : >= 1 year (2.0/1.0)
POSITION CODE = IC:
| LIFE INSURANCE BROUGHT (SUM INSURED) > 0 : >= 1 year (70.0/12.7)
| LIFE INSURANCE BROUGHT (SUM INSURED) <= 0 :
| | MARITAL STATUS = S:
| | | PAST ANNUAL INCOME <= 97201 :
| | | | EMPLOYMENT STATUS = PT: >= 1 year (15.0/8.4)
| | | | EMPLOYMENT STATUS in {FT,SE}: < 1 year (234.0/72.0)
| | | | EMPLOYMENT STATUS in {UE,PF}: < 1 year (0.0)
| | | PAST ANNUAL INCOME > 97201 :
| | | | EMPLOYMENT STATUS in {UE,PF}: >= 1 year (2.0/1.0)
| | | | EMPLOYMENT STATUS = PT: < 1 year (1.0/0.8)
| | | | EMPLOYMENT STATUS in {FT,SE}:
...
...

```

Figure 4: Part of Tree that determining Length of Service

The standardized coefficients of Discriminant analysis predicting length of service is shown in Table 2, which measure the relative importance of independent variables. It can be seen that the coefficients of “job position” and “marital status” are the highest, which indicates that they are the most important factors in predicting length of service.

Table 2: Standardized Canonical Discriminant Function Coefficients

	Std. Coeff.
Position	0.754
Sex	0.335
Age	0.040
Martial Status	
Single	0.671
Married	0.312
Divorced	0.337
Number of Dependents	0.127
Academic Level	-0.098
Previous Job Nature	
Clerical/Technician	-0.179
Sales/Servicing	0.003
Professional	-0.043
Management	-0.105
Employment Status	-0.054
Past Annual Income	-0.466
Working Experience	-0.105
Management Experience	-0.182
Insurance Brought	-0.368

As a result, if we overlap the most influential variables as indicated in different data mining techniques, 1) job position, 2) insurance brought, and 3) martial status are important related to the length of service for a new insurance agent.

4.2 Predicting Premium

Sales premium, which corresponds to the total amount of insurance policy to be handled by an insurance agent, was classified into two classes, one is less than HK\$ 1M (class 1) while the other is more than or equal to HK\$ 1M (class 2). Thus the insurance agent falling into class 2 is to be a high performer. Discriminant predicts 57.44% and 64.60% correctly in training and testing data respectively; accuracy from ANN are 67.25% and 53.20% respectively; and the accuracy from decision tree are 77.10% and 53.00%

respectively. Discriminant analysis produces the highest overall accuracy in the testing dataset. It also produces the highest conditional accuracy of 57.8% in predicting class 1, while Decision Tree gives the highest conditional accuracy of 72.0% in predicting class 2.

Table 3: Predict Sales Premium

		ANN Predicts		Discriminant Predicts		DT Predicts	
		Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Actual	Class 1	31.6 %	8.8 %	18.6 %	21.8 %	32.4 %	8.0 %
	Class 2	38.0 %	21.6 %	13.6 %	46.0 %	39.0 %	20.6 %

As shown in Figure 5, which is the decision tree generated by C4.5, working experience and job position are important factors in predicting an agent's sales premium.

```

WORKING EXPERIENCE <= 0.1 :
  POSITION CODE = IC: < 1000000 (455.0/56.4)
  POSITION CODE = ME: < 1000000 (0.0)
  POSITION CODE = EC:
    Age <= 23.7479 : >= 1000000 (7.0/2.4)
    Age > 23.7479 : < 1000000 (18.0/3.7)
WORKING EXPERIENCE > 0.1 :
  PAST ANNUAL INCOME <= 92043 :
    MARITAL STATUS in {S,M,D}:
      PREVIOUS JOB NATURE in {C,S,P,M}:
        POSITION CODE in {IC,EC}:
          ACADEMIC LEVEL in {US,SC,GR}: < 1000000 (645.0/195.0)
          ACADEMIC LEVEL in {PS,DP}:
            EMPLOYMENT STATUS in {UE,PF,SE}: >= 1000000 (2.0/1.0)
            EMPLOYMENT STATUS = PT:
              WORKING EXPERIENCE <= 0.6 : >= 1000000 (4.0/1.2)
              WORKING EXPERIENCE > 0.6 : < 1000000 (5.0/2.3)
            EMPLOYMENT STATUS = FT:
              PREVIOUS JOB NATURE in {C,S,P}:
                PAST ANNUAL INCOME <= 84840 : < 1000000 (32.0/11.6)
                PAST ANNUAL INCOME > 84840 : >= 1000000 (2.0/1.0)
  ...
  ...

```

Figure 5: Part of Tree Predicting Premium

In

Table 4, the standardized discriminant coefficients show that previous job, past annual income and age contribute significantly in predicting an agent's sales premium.

Table 4: Standardized Canonical Discriminant Function Coefficients

	Std. Coeff.
Position	-0.270
Sex	-0.142
Age	-0.510
Marital Status	
Single	-0.139
Married	0.041
Divorced	0.026
Number of Dependents	0.098
Academic Level	-0.081
Previous Job Nature	
Clerical/Technician	0.543
Sales/Service	0.551
Professional	0.135
Management	0.633
Employment Status	0.301
Past Annual Income	0.574
Working Experience	0.186
Management Experience	-0.066
Insurance Brought	0.265

By considering the analysis from decision tree and discriminant analysis in the prediction of sales premium, 1) previous job, 2) past annual income, 3) age, 4) working experience and 5) job position are significant factors.

4.3 Predicting Persistency

Persistency is an index that measures ratio of lapse amount relative to the total premium, it was classified into two classes, one is 0% - 75% (class 1) and the other is 75% - 100% (class 2). The higher the persistency, the less chance that the premium policy to be lapsed. Thus class 2 is regarded to be high performers. For training and testing data, the accuracy of Discriminant are 69.89% and 57.80% respectively; the accuracy of ANN are 72.56% and 58.40% respectively; the accuracy of Decision Tree are 78.20% and 58.20% respectively. The highest one is ANN model, which also produces the highest conditional accuracy of 56.9% in predicting class 2. For predicting class 1, Decision Tree produces the best conditional accuracy of 69.4%.

Table 5: Predict Persistency

		ANN Predicts		Discriminant Predicts		DT Predicts	
		Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Actual	Class 1	13.0 %	34.4 %	14.0 %	33.4 %	10.0 %	37.4 %
	Class 2	7.2 %	45.4 %	8.8 %	43.8 %	4.4 %	48.2 %

Figure 6 shows part of decision tree that is predicting an agent's persistency. Working experience and job position are two most important factors.

```

WORKING EXPERIENCE <= 0 :
|
|   Age <= 22.7425 :
|   |   POSITION CODE = EC: > 75% (5.0/1.2)
|   |   POSITION CODE = ME: > 75% (0.0)
|   |   POSITION CODE = IC:
|   |   |   MARITAL STATUS = M: <= 75% (4.0/1.2)
|   |   |   MARITAL STATUS in {D,W}: <= 75% (0.0)
|   |   |   MARITAL STATUS = S:
|   |   |   |   Age <= 19.8301 : > 75% (69.0/25.3)
|   |   |   |   Age > 19.8301 : <= 75% (86.0/34.6)
|   |   Age > 22.7425 :
|   |   |   NUM OF DEPENDENTS <= 0 : <= 75% (295.0/44.8)
|   |   |   NUM OF DEPENDENTS > 0 :
|   |   |   |   MARITAL STATUS = M: > 75% (15.0/5.8)
|   |   |   |   MARITAL STATUS = W: <= 75% (1.0/0.8)
|   |   |   |   MARITAL STATUS in {S,D}: > 75% (0.0)
|   |   WORKING EXPERIENCE > 0 :
|   |   |   POSITION CODE in {IC,EC}: > 75% (2732.0/850.3)
|   |   |   POSITION CODE = ME:
|   |   |   |   ACADEMIC LEVEL = US:
|   |   |   |   |   PREVIOUS JOB NATURE = S: > 75% (1.2/0.9)
|   |   |   |   |   PREVIOUS JOB NATURE in {C,P}: <= 75% (0.0)
|   |   ...
|   ...

```

Figure 6: Part of Tree Predicting Persistency

In predicting an agent's persistency, discriminant analysis shows that previous job is a most important factor, the next important factors are age and working Experience.

Table 6: Standardized Canonical Discriminant Function Coefficients

	Std. Coeff.
Position	-0.150
Sex	-0.069
Age	-0.485
Martial Status	
Single	-0.252
Married	-0.141
Divorced	-0.064
Number of Dependents	0.184
Academic Level	0.132
Previous Job Nature	
Clerical/Technician	1.022
Sales/Serviceing	1.026
Professional	0.424
Management	0.858
Employment Status	0.260
Past Annual Income	0.115
Working Experience	0.376
Management Experience	0.070
Insurance Brought	0.199

Thus the combined analysis from decision tree and discriminant analysis concludes that 1) working experience, 2) job position, 3) age and 4) previous job are the most influential factors in determining the persistency of an insurance agent’s sales premium.

Table 7 shows that the three methods outperform one another in different situations. ANN performs well in predicting length of service and persistency, while in predicting premium, discriminant analysis gives the best result. Decision Tree outperforms the prediction on training data of all the three goals, this indicates that Decision Tree's ability of induction is very good, while ANN does its best job on deduction.

Table 7: The overall out-performer in different applications

	ANN	Discriminant	Decision Trees
Length of Service	✓		
Premium		✓	
Persistency	✓		

The overall result from all applications shows that 1) working experience, 2) job position, 3) age, 4) marital status, 5) previous job 6) past annual income, and 7) insurance brought are important factors in determining a new insurance agent on his/her length of stay, sales premium to be handled, and persistency in avoiding the lapsing of those premiums. Thus we can make use of those important attributes in order to improve the new insurance agent selection process.

5. Conclusion

The study demonstrates using three data mining techniques to identify agent candidates on their length of stay in a company and their sales performance. In tradition, insurance companies used to hire an agent without classification, but would terminate the contract if the agent could not make enough sales in a period. However, such strategy is not cost effective because some costs such as training cost, printing of business cards would be incurred.

Among a number of data mining techniques, although decision tree is a little bit less out-perform, it is the easiest tool for classification prediction. Unlike other techniques, result of decision tree analysis is human interpretable. The benefit of statistical linear discriminant analysis is that the calculation is relatively simple and thus the training speed is fast. This method shows its ability in predicting sales premium in our experiment. Artificial neural networks is a sophisticated method, despite of its slow training speed, its predicting ability is good in predicting two out of three goals in our experiment.

In this study, the accuracy of prediction is not high enough. This may due to the short period of data collection. Those agents worked over three years were excluded in model building. Moreover, some other useful attributes such as personality test score, were not included. This is because the company that engaged did not have the record of personality test score in their staff profile. This would limit out accuracy in predicting the performance of an insurance agent. After evaluated by managers of the engaged insurance company, it is suggested that the goals can be refined into three or four classes

rather than two. For instance, the length of stay would be classified as within six month (class 1), six month to one year (class 2), one year to two year (class 3) and more than or equal to two years (class 4). Such suggestion will be adopted in our future research.

6. Contributions:

Year 2000 has been indicated to be a century of information technology. There is no doubt that the world in which we live will continue to change, and that technology will be a major driving force of change. The rate of change is also likely to continue to increase. The 1980s trend toward increased numbers of mergers, alliances, acquisitions, divestitures, and partnership formations is likely to continue on both a national and global basis. As business market become more competitive, building an effective intelligent decision support system will be a key factor to success. Under these considerations, investigation on how to make use of the information from data warehouse is unprecedingly a definite task. This will not only provide a foundation for academics on how data mining techniques can be applied to a data warehouse, but also build a reference for software developers or distributors to enhance their IDSS software.

Acknowledgement

This research is funded (grant A-PB66) by The Hong Kong Polytechnic University, Hong Kong (SAR), China.

References

[1] Aleksander and Morton, *Introduction to Neural computing*, North Oxford Press. 1990.

[2] Brockett P.L., Cooper W.W., Golden L.L., and Xia X., "A Case Study in Applying Neural Networks to predicting Insolvency for Property and Casualty Insurers", *Journal of the Operational Research Society*, Vol. 48, 1153-1162, 1997.

[3] Chauduri S. and Dayal U., "An overview of data warehousing and OLAP technology", *SIGMOD Record*, 26(1):65-74, March 1997.

- [4] Cho V., Wüthrich B. and Zhang J., "Text Processing for Classification", *Journal of Computational Intelligence in Finance*, special issue on "Financial News Analysis using Distributed Data Mining", pp.26, accepted Dec 98, 1998.
- [5] Clark P. and Niblett T., "The CN2 Induction Algorithm", *Machine Learning*, Vol. 3. Pp. 261-283, 1988.
- [6] Codd E., Codd S., and Salley C., "Providing OLAAP (on-line analytical processing) to user-analysts: An IT mandate", *Technical report*, E.F. Codd & Associates, 1993.
- [7] Duda R.O., Hart P.E. and Stork D.G., *Pattern Classification*, 2nd edition, New York: John Wiley & Sons, Inc., 2001
- [8] Dzeroski S. and Lavrac N., "Inductive Learning in Deductive Databases", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5 No. 6., pp 939-949, November 1993.
- [9] Goodman R. M. and Smyth P., "The Induction of Probabilistic Rule Sets – IRule algorithm", *Proc. of the Sixth International Workshop on Machine Learning*, pp. 129-132, 1989.
- [10] Inmon W., *Building the Data Warehouse*, Wiley & Sons, Inc., Second edition, 1996.
- [11] Kimball R., *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Datawarehouses*, John Wiley & Sons. Inc., 1996.
- [12] Laks V.S. Lakshmanan and Fereidoon Sadri, "Modeling Uncertainty in Deductive Databases", *Proc. Int. Conf. on Database Expert Systems and Applications*, DEXA'94, Sept. 1994, Athens, Greece, Lecture Notes in Computer Science, #854, Springer 1994, pp. 724-733.
- [13] Michalski R.S., Mozetic I., Hong J. and Lavrac N., "The Multi-Purpose Incremental Learning System AQ15 and Its Testing Application to three Medical Domains", *Proc. AAAI-86*, pp.1041-1045, 1986.

- [14] Nunamaker, J.F., Chen, M and Purdin, D.M. (1990), "Systems Development in Information Systems Research", *Journal of Management Information Systems*, Winter, Vol. 7, 1990-91, pp. 89-106.
- [15] Sestito S., and Dillon T.S. (1994), *Automated knowledge Acquisition*, New York: Prentice Hall.
- [16] Tseng F.S., Chen A.L.P. and Yang W.P., "Refining Imprecise Data by Integrity Constraints", *Data & Knowledge Engineering Journal*, North-Holland, pp. 299-316.
- [17] Wasserman P.D., *Neural Computing: Theory and Practice*, New York: Van Nostrand Reinhold, 1989.
- [18] Weiss, S.M. and Kulikowshi, C.A., *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems*, San Mateo, CA: Morgan Kaufmann, 1991.
- [19] Wüthrich B., "Probabilistic Knowledge Bases", *IEEE Transactions on Knowledge and Data Engineering*, Vol 7(5), pp. 691-698, October 1996.