A model for personnel selection with a data mining approach: A case study in a commercial bank

Authors:
Adel Azar[1]
Mohammad Vahid Sebt[1]
Parviz Ahmadi[1]
Abdolreza Rajaeian[1]

Affiliations:
[1]Department of Management, Tarbiat Modares University, Iran

Correspondence to:
Adel Azar

Email:
azara@modares.ac.ir

Postal address:
PO Box 14115-111, Tehran, Iran

Dates:
Received: 25 Nov. 2011
Accepted: 17 Oct. 2012
Published: 15 Apr. 2013

How to cite this article:
Azar, A., Sebt, M.V., Ahmadi, P., & Rajaeian, A. (2013). A model for personnel selection with a data mining approach: A case study in a commercial bank. *SA Journal of Human Resource Management/SA Tydskrif vir Menslikehulpbronbestuur*, *11*(1), Art. #449, 10 pages. http://dx.doi.org/10.4102/sajhrm.v11i1.449

Orientation: The success or failure of an organisation has a direct relationship with how its human resources are employed and retained.

Research purpose: In this paper, a decision-making tool is provided for managers to use during the recruitment process. The effective factors in employees' performance will be identified by discovering covert patterns of the relationship between employees' test scores and their performance at work.

Motivation for the study: Large amounts of information and data on entrance evaluations and processes have been kept in organisations. There is a need to discover the pattern in the relationship between employee's test scores and their performance at work as a tool for use during the recruitment process.

Research design, approach and method: The data mining technique that was used in this project serves as the decision tree. Rules derivation was accomplished by the Quick Unbiased and Efficient Statistical Tree (QUEST), Chi-squared Automatic Interaction Detector (CHAID), C5.0 and Classification And Regression Tree (CART) algorithm. The objective and the appropriate algorithm were determined based on seemingly 'irrelevant' components, which the Commerce Bank Human Resources management's experts describe.

Main finding: It was found that the 'performance assessment' variable was not considered as the objective. Also, it was concluded that out of 26 effective variables only five variables, such as province of employment, education level, exam score, interview score and work experience, had the most effect on the 'promotion score' target.

Practical/managerial implication: The database and personnel information of the Commerce Bank of Iran (in 2005 and 2006) was studied and analysed as a case study in order to identify the labour factors that are effective in job performance. Appropriate and scientific employment of staff that were selected from the entrance exams of companies and organisations were of crucial importance.

Contribution/value-add: It is of great importance that an extensive use of data mining techniques be applied in other management areas. Whilst this is a low-cost technique, it can help managers to discover covert knowledge in their organisations.

Introduction

Human capital is amongst the most valuable assets of an organisation. In recent years, human resources have been the centre of attention and have drawn a remarkable part of the time and assets of progressive companies. Success or failure of a company, amongst other factors, is related to how its workforce is recruited and retained (Jazani, 2000).

Suitable candidate selection for each job is also one of the most crucial issues of management decision-making. Likewise, data in hand that may be available publicly does not have great value. However, when the data is processed and becomes part of information and knowledge, it may create much value. Therefore, organisations should develop and acquire strong skills in data-processing techniques such as data mining (Khaef, Ahmad, Mottaghi & Sebt, 2007).

The existence of many databases in organisations (especially in human resources management departments), has led to data mining techniques being applied in this research as a tool to embark on one of the most important issues of management in human resources management, namely, workforce recruitment. In other words, the effective factors will be identified in employees' performance by discovering covert patterns of the relationship between employees' test scores and their performance at work. This will provide a decision-making tool for managers to use during the recruitment process.

Key concepts

What is data mining?

Data mining technique was first introduced and used in the late 1980s. Significant developments were made in the 1990s and continued through the 21st century (Hand, Mannila & Smyth, 2001). Data mining is the process of discovering the pattern of internal relation and structure of data (Berry & Linoff, 2004). In fact, data mining discovers the interesting, unexpected and valuable constructions from inside of large data and is an activity corresponding to statistics and analysis of data (Hand *et al*., 2001).

It is predicted that data searching will have revolutionary development in the next decade (Daniel & Larose, 2006). It is called one of the top ten superior technologies, which will have striking effects in world evolution by the Massachusetts Institute of Technology (MIT) (Daniel & Larose, 2005).

Knowledge discovery

In the late 1980s, new terms of knowledge discovery were applied to databases and the artificial intelligence specialists used the word 'machine learning' for the knowledge-extracting process, from the first step to the end.

This may cause a misunderstanding because when data mining tools are used in databases the results appear surprising. However, the general process of scientific discovery is a bit far removed from automation and even requires human interference. For this reason, many authors consider the process of scienctific discovery as a combination of art and science (Weiss & Indurkhya, 1998).

The process of knowledge discovery includes data mining techniques that are targeted at solving specific problems and making decisions by using mathematical models and computers, and applying data analysis on a big database. Discovered patterns and structures in the analysis will lead a researcher to find a solution, which could then be applied to the problem.

A model can qualify as knowledge when:

- it is understandable by average people
- it passes a validity test, to a certain degree
- it is functional
- it presents new information of which users had no prior knowledge (Mendonca & Sunderhaft, 1999).

Recruitment and selection of employees

Beardwell and Wright (2004) state that recruitment and selection are two processes, the focus of which are to identify and secure appropriate people to meet the human resource needs of an organisation. The two terms 'recruitment is over' and 'selection started' are often used as a subject of study (Anderson, 1994).

For the purpose of analysis, we need to make a distinction between the two terms. However, according to most definitions, 'recruitment' refers to the first half of the hiring process and 'selection' to the second half. Recruitment aims at identifying suitable candidates and selection focuses on choosing the best of them. Selection presents the final step of decision-making in a hiring process (Cowling & Mailer, 1990). Recruitment and selection enable organisations to employ their staff and establish their human resource base. An increasingly competitive and global business environment, along with ever-improving quality and customer service, has highlighted the importance of recruitment and selection of appropriate employees (Porter, Smith & Fagg, 2006).

Decision tree

Decision tree is a strong and widely used tool for clustering and forecasting. Unlike the neural networks, decision tree generates rules. It means that decision tree describes its prediction in the form of rules, whereas in neural networks only the final prediction is expressed, and the fact of the matter is hidden inside the net. Also in decision tree, the data set does not necessarily need to be a numeric one.

Decision tree enables us to present our predictions in the form of rules, without any need for complicated calculations for data grouping. It can be used for different kinds of data, that is, continuous or discrete, and can identify the variables with most impact in prediction and grouping.

Decision trees are also useful for exploring data to gain insight into the relationships of a large number of candidates' input variables to a target variable. Considering that decision trees combine both data exploration and modeling, they are a powerful first step in the modeling process, even when the final model is built by using some other technique.

Decision-tree methods have wide applicability for data exploration, classification and scoring. Decision trees are also a natural choice when the goal is to generate understandable and explainable rules. The ability of decision trees to generate rules that can be translated into comprehensible natural language or Structured Query Language (SQL) is one of the greatest strengths of the technique. Decision trees require less data preparation than many other techniques because they 'are equally adept at handling continuous and categorical variables' (Bekhor *et al*., p. 9). 'Categorical variables, which pose problems for neural networks and statistical techniques' (Bekhor *et al*., p. 9), are split by forming groups of classes. For this reason, decision trees are often used to pick a good set of variables that can be used as inputs in another modeling technique (Berry & Linoff, 2004).

The algorithm of a decision tree starts with developing a test, which can best divide groups. The most important objective of grouping is to obtain a model for prediction. A data set called 'training data', which includes individual variables and records, are applied for this purpose. In the steps that follow, the same procedure is performed for lower nodes with fewer numbers of data in order to obtain the best rules. Eventually, the tree becomes so large that there remains no room for further separation of node data. At this stage, the effectiveness of the created tree should be measured. For this purpose, a set of data or records, which are different from those used to create the tree, will be used. The standard which is being measured consists of correctly classified data and the predicted class is the same as the real one. In general, the advantages of using decision tree above other data mining techniques are that:

- calculations are quicker
- accuracy level is higher
- it is easier to learn
- rules make better sense.

Tests in each algorithm of the decision trees are different and selection of trees will also be completed differently.

Quick Unbiased and Efficient Statistical Tree algorithm

Quick Unbiased and Efficient Statistical Tree (QUEST) algorithm is a binary tree growth algorithm (Loh & Shih, 1997). This algorithm chooses the branch field and branch point separately. A single branching variable in QUEST chooses an unbiased variable.

Chi-squared Automatic Interaction Detector algorithm

Chi-squared Automatic Interaction Detector (CHAID) means to discover the automatic X2 challenges. This method, which was developed by Kass (Kass, 1980), uses the results of a statistical test as a very effective statistical technique for segmentation or tree growth. CHAID evaluates all amounts of a prediction field.

C 5.0 algorithm

Expansion of an ID3 algorithm is a C4.5 algorithm and its developed version is C 5.0. ID3 algorithm starts with all training prototypes in the main node of the tree. Thereafter, one specification is chosen for dividing these prototypes. For each degree of characteristic, a branch is created. A subset in proportion to the degree of the characteristic will be determined by branch and moves towards child nodes. This algorithm is applied for every child node so that all samples in one node are from the same class. This algorithm develops the range of classes from pages to numeric features. As a result, estimates are related to features that come from dividing data to subsets.

Classification and Regression Tree algorithm

CART algorithm, which is one of the most famous methods in decision tree, was introduced by Breiman and his colleagues in 1984. CART method creates binary branches based on one field. It means that each group is divided into two other groups. The norm which is used to evaluate branches is ‘Diversity’:

- Min[P(c1),P(c2)]
- 2P(c1)P(c2)
- [P(c1). log P(c2)] + [P(c2). log P(c1)].

Background of data mining application in human resources management

Data mining can be used to identify the causes of some human resource problems in organisations. Schroeck (1999), Howeedy (2002) and Olson (2002) have presented some reports about application of data mining to human resources management:

- A consultant of Metlife Insurance has created ‘If … then’ rules concluded from historical information and has used them in investigating the inspection departments. Typical results from this research (Denker, 2000) related to human resources management is as follows: discovering excessive wages by monitoring deviation from average wage in each income bracket.
- Identify employees who commute a long distance to work. It is very likely that these employees will quit their jobs if an employment opportunity close to their residence arises.
- Identify employees who do not benefit from the increasing profit of the company.
- Identify employees who do not benefit from the company’s bonus plans and health care programmes.
- Identify employees who make better effort than their peers.

As mentioned, there are high-level data mining tools such as neural networks. However, software tools developed to make ‘If … then’ rules are less costly than data mining tools.

Furthermore, data mining is used to reduce time and costs and increase access to candidates with higher quality (SAS Institute Inc. 2004).

Meanwhile, the best research using data mining for personnel selection was titled ‘Data mining to improve personnel selection and enhance human capital’ authored by Chien and Chen (2007). They have presented a model for personnel selection in the electronics industry. Whilst our approach shares some similarities with Chien and Chen’s research, it has some differences as follows:

- The knowledge discovered by Chien and Chen was based on parameters such as age, gender, marital status, education and experience whereas in our research the relations of those parameters were with entrance exam marks. Obviously limited parameters will bring limited knowledge too.
- Their case study was based on a capital-intensive industry, whereas our research has focused on a service industry, in which human capital plays a major role.
- The strength of Chien and Chen’s research is the number of data records in excess of 3000.
- In their research there are three target variables, namely, job performance, retention and turn-over reasons. However, in the current research, target variables are ‘job performance’ and ‘promotion scores’. This is because the latter variables are related not only to personnel but also to the organisation.

- The algorithm used in Chien and Chen's research was the CHAID method using only one tree, but in the current research an attempt was made to allocate an appropriate algorithm using logical reasons and to mention the level of error and also eliminate the limitation of using only one tree.

In addition, and from results of previous studies, the technique used by researchers in the data mining area is just 'Decision Tree'. However, different algorithms are applied in the data mining area.

Research design

Research approach

The success or failure of an organisation has a direct relationship with how its human resources are employed and retained. It is the case that organisations keep large amounts of information and data on entrance evaluations and processes. This information, however, is often left unutilised, or at best, analyzed through rudimentary statistical methods. In the current era, new intelligent technology has created tools to aid humankind in transforming large volumes of data into knowledge and information. Data mining is considered a solution for analysing this data. The present thesis takes this approach further by avoiding opinion-based methods that are traditionally used in the selection of new employees.

Research method

In this project, the status of employees' upgrade is recognised by using data mining techniques and rules, and the relation between entrance exam marks and staff situation with job performance. Firstly, job performance, the private information of employees in commercial banks and the human indices, were effective in performance or upgrade, were identified by studying exam databases. In the next step, a suitable model was designed to address the shortcomings of previous researches by learning from previous studies about data mining and human source performance.

Thereafter, a data warehouse was created. With data mining techniques, the data set was cleansed. In other words, the unsuitable and redundant variables were removed. After data cleansing, the records with empty fields were deleted and suitable amounts were replaced for lost data.

Research participants

In order to test the designed model, the sample set of candidates who passed the entrance exam were applied and were hired in 2004 and 2005 in the commercial bank.

This was part of a subset that consisted of over 30 000 participants in the entrance exam. However, because it is necessary to evaluate the relation between the test marks and the participants' performance and promotion, only the part of the data sample associated with candidates who were hired by the bank and had worked there for at least one year was chosen.

The number of accepted and hired individuals in 2004 and 2005 was over 940. However, we could find only 711 individuals with complete records. The records consisted of 26 information fields about personnel information, hiring status, exam marks and job performance. They were chosen as the final data set.

Access to this database requires some data preparation such as data conversion, review of missing data, reduction of data size, data enrichment, data repair and data cleansing. This is completely explained in detail in the case study.

Many researchers (Breiman, Friedman, Olshen & Stone, 1984; Chien & Chen, 2007; Hartigan, 1975; Quinlan, 1986) in human resources management have used the decision tree technique in data mining. The decision tree was chosen for this purpose. The algorithms that will be used in this technique are QUEST, CHAID, C5.0 and CART. In order to choose the best algorithm, an approach that is explained in the next section was taken.

Measuring instruments

In the next step, using decision tree techniques, data was classified and analysed and the primary results were obtained. Finally, the orders which were not obvious were explained. In this research the data set were prepared by MS Excel and rules were drawn using SPSS Clementine software (version 12.0).

Finally, validity and reliability of the rules were examined and then the model for knowledge discovery and application for personnel selection was presented.

Conceptual model

Based on defined variables and existing databases, the conceptual model for discovering knowledge from entrance exam results was designed according to Figure 1. As illustrated in the model, in the first step, the variable groups were defined and the most effective variables were identified. Thereafter, the data mining process will be implemented by entering these variables as well as target variables for each algorithm. Results will be used to predict employees' job performance and their promotion in the future. The final results will be used as organisational knowledge to improve efficiency in the organisation.

Case study

Data preparation

Data preparation includes all the steps that prepare records and variables to create the mode l and the tree. This step is usually the longest and most important in data mining and knowledge discovery processes. The better the quality of the data, the more precise the modeling and results will be. This process has been carried out in six steps, explained as follows:

- **Step 1:** Several meetings were arranged with the managers of human resources to brief them on the objectives of the project and also to evaluate the data records of the personnel.

- **Step 2:** The type of variables in four categories (personal variables, examination variables, job variables and performance variables) were defined according to existing fields in the records. This was done in collaboration with human resource associates of the bank.
- **Step 3:** According to the bank's human resource records, out of all the applicants who wrote the entrance exam in 2004 and 2005, 970 individuals were admitted and employed. Because of the dispersion that existed in the exam database and the employees' records, data records were reduced to 717 records with 37 fields by using MS Excel tools.
- **Step 4:** Considering that each field represents only one variable, immediately after creating the database, all redundant or incomplete variables or fields are deleted, combined or converted. During this stage, the initial 37 fields were reduced to 26.
- **Step 5:** The quantitative and qualitative values of each variable were reviewed. Some variables were concatenated. All records then were evaluated and six records eliminated. Therefore, the number of complete records was reduced to 711.

Finally, the database of 711 records with 26 variables was prepared in an Excel worksheet to build the model and the tree. These variables were grouped in four categories as follows:

- Personal variables (including age, Grade Point Average [GPA], professional experience, marital status, gender, education, major, name of school).
- Exam variables (including marks for various subjects, overall score, interview score).
- Job variables (including years at job, employment status, position).
- Performance related variables (including promotion score, performance evaluation and total score).

Target variables definition

Target variables such as promotion score and performance assessment are determined in this research with respect to the conceptual model.

Promotion score: This score is defined as the spread between the scores of the position he or she was hired for, and his or her current position. If there is no change this score will be 0. The maximum score that was reached was no greater than 1150. Considering this variable is one of the target variables, the scores are grouped in four levels as shown in Table 1.

Performance assessment: This variable will be determined based on employee's performance evaluation, which was conducted at least twice per year for each employee. The maximum score is 100. All the quantitative scores are grouped in seven qualitative categories as shown in Table 2.

TABLE 1: Data classification of preferment score.

Scale	Label	Preferment Score
Excellent promotion	A	700–1150
Good promotion	B	300–500
Bad promotion	C	50–200
No Promotion	D	0

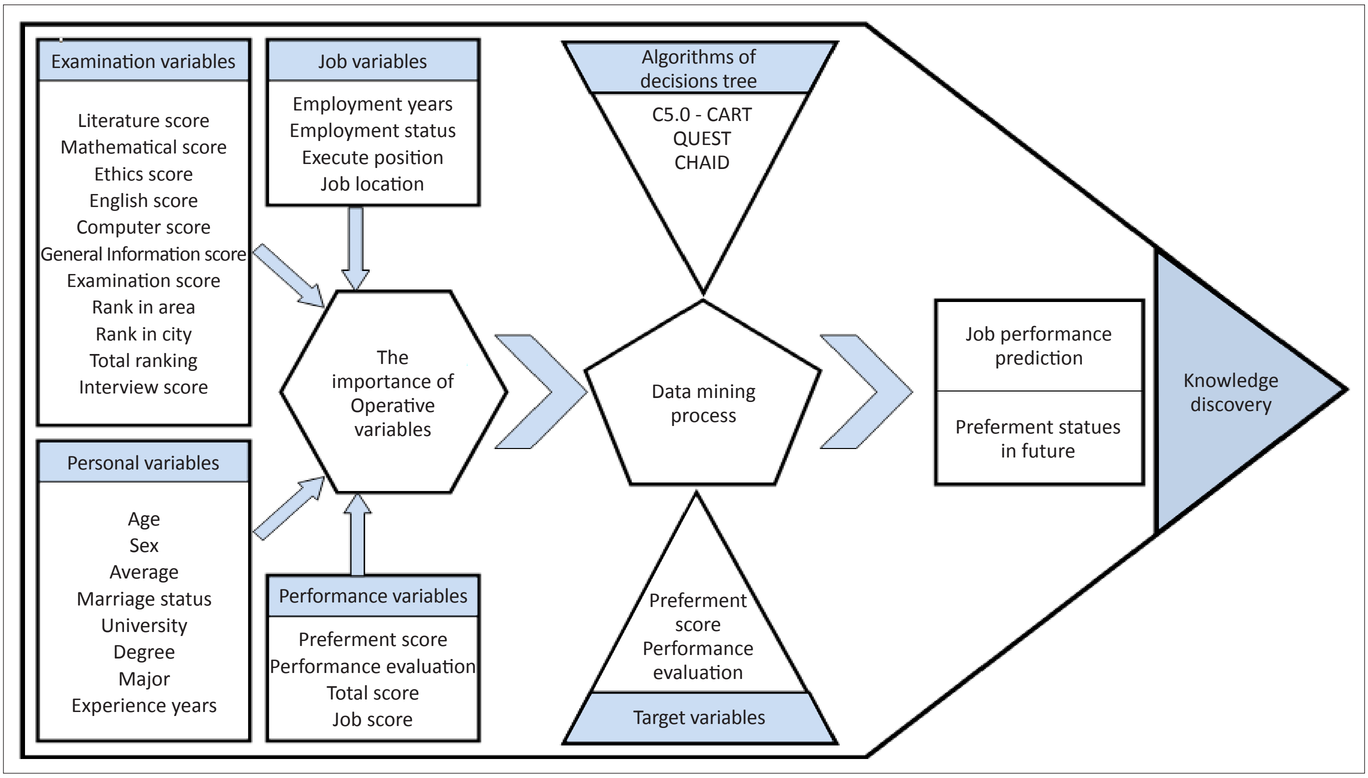

FIGURE 1: Knowledge discovery conceptual model from database of entrance examinations.

Trees making

To build the tree we need to find answers to the following questions:

- Should all variables be simultaneously incorporated in building the tree?
- Does the accuracy of the tree depend on the number of records or variables?
- What combination of variables makes the best tree?
- Does the selection of target variables have significance in building the tree?
- Which data mining algorithm will reach the best conclusion?

For answering above questions, the used process for trees making are as follows:

- **Step 1:** Considering four identified groups are variables and two target variables, including several algorithms to build the decision tree, all applicable models will be evaluated. The existence of four independent variables, two target variables, and six feasible algorithms to build the decision tree (QUEST, CHAID, C5.0, CART in Towing state and CART in Ordered state) will result in 180 alternative ways to build the tree:

  $(2^x – 1) * y * A$ [Eqn 1]

  Where, '*x*' is the number of groups of independent variables, '*y*' is the number of target variables, and 'A' is the number of algorithms. Therefore:

  $(2^4 – 1) * 2 * 6 = 180$ [Eqn 2]

  550 training records were randomly selected, for which each of the identified 180 models were trained with use of SPSS Clementine software and the accuracy of each was calculated and registered. In the seven detected states with QUEST algorithm, there was no possibility of building a tree. The permissible number of levels is limited to four, because going beyond the fourth level would make drawn rules too long and therefore useless.
- **Step 2:** With respect to the obtained results and considering the most effective variables amongst all the trees (173 trees), 12 trees were constructed. The sums of errors that were obtained from each algorithm for all 185 available trees are presented in Table 3.
- **Step 3:** Finally, the accuracy for each target variable and algorithm type were estimated, the result of which is shown in Table 4.

TABLE 2: Data classification of performance assessment score.

Scale	Performance assessment score
Excellent	100
Very good	99–100
Good	98–99
Normal	96–98
Weak	93–96
Very weak	90–93
Nonperformance	< 90

TABLE 3: Algorithms accuracy from 185 perquisite trees.

Algorithm name	Algorithm accuracy average (%)	Total Algorithms accuracy average (%)
QUEST	63.49	67.29
CHAID	60.33	-
C 5.0	80.43	-
CART	65.11	-
CART (Towing)	66.78	-
CART (Ordered)	66.78	-

QUEST, Quick Unbiased and Efficient Statistical Tree; CHAID, Chi-squared Automatic Interaction Detector; CART, Classification And Regression Tree.

Initial selection of suitable trees for drawing rules

Given the number of produced trees, it is obvious that many rules will be drawn from these trees. In order to draw conclusions from the rules, it is necessary to separate good trees from other produced models. In doing so, several criteria were defined for choosing trees, as follows:

- trees with over 70% accuracy
- trees with one or more levels (meaning a rule has been drawn)
- trees that includes at least one of the individual variables or exam variables
- trees whose effective variables are amongst either in the individual variable or exam variable groups.

By applying the above criteria, 17 trees were selected in the initial step to go for exam step, amongst which 3 trees were selected. All three had performance evaluation as their target variable and the algorithm used for all was C5.0.

Examining produced trees

Experimental data was used to examine the 17 selected models for the purpose of creating rules. From the final data,

TABLE 4: Target variables accuracy from 185 perquisite trees.

Target variable	Algorithm name	Algorithm accuracy average (%)	Comments	Total Algorithms accuracy average (%)
Performance assessment	QUEST	30.88	Average from 12 trees	43.11
	CHAID	30.94	Average from 16 trees	-
	C 5.0	68.20	Average from 16 trees	-
	CART	39.99	Average from 16 trees	-
	CART (Towing)	42.78	Average from 16 trees	-
	CART (Ordered)	42.78	Average from 16 trees	-
Preferment Score	QUEST	93.60	Average from 13 trees	91.21
	CHAID	89.73	Average from 16 trees	-
	C 5.0	92.65	Average from 16 trees	-
	CART	90.23	Average from 16 trees	-
	CART (Towing)	90.77	Average from 16 trees	-
	CART (Ordered)	90.77	Average from 16 trees	-

QUEST, Quick Unbiased and Efficient Statistical Tree; CHAID, Chi-squared Automatic Interaction Detector; CART, Classification And Regression Tree.

161 records were randomly selected as experimental data to be used for this step. In other words, all 161 selected records were used for training the tree and making the model.

In this step, each of the 17 selected trees in the former step was examined by inputting the experimental data and measuring their accuracy. Because the accuracy of three trees, whose target variable was 'performance evaluation' had significantly declined, those trees as well as the 'performance evaluation' variable were eliminated before entering the rule creation step. The accuracy levels for the eliminated trees are as show in Table 5.

Therefore the number of selected models for creating rules was reduced to 14. Considering that the models and the rules generated by CART (Ordered) and CART (Towing) algorithms were identical, only one of these was chosen. The test results of the algorithms are presented in Table 6.

It can be concluded that the 'promotion score' variable and CART algorithm have resulted in the highest accuracy. This occurred whilst the 'performance evaluation' variable was not observed in any of the final models and the models built based on that had resulted in a high error level.

Create rules

After the final selection of trees, rules and trees need to be created. The purpose of creating rules is to evaluate each tree and to translate each rule from mathematical and logic language to plain language comprehensible by experts.

TABLE 5: Test accuracy of trees with 'performance assessment' target variable.

Tree number	Accuracy with training data (%)	Accuracy with experience data (%)	Difference excess (%)
C5-A-2	60.91	23.60	37.31
C5-A-5	97.64	21.12	76.52
C5-A-11	84.00	21.74	62.26

TABLE 6: Algorithms accuracy of 14 chosen trees.

Algorithm name	Accuracy with training data (%)	Accuracy with experience data (%)	Difference excess (%)
QUEST	69.09	70.19	-1.10
CHAID	71.82	77.02	-5.20
C 5.0	72.32	66.46	5.86
CART	81.74	79.33	2.41
CART (Ordered & Towing)	72.61	70.52	2.09
Average	75.48	72.95	2.53

QUEST, Quick Unbiased and Efficient Statistical Tree; CHAID, Chi-squared Automatic Interaction Detector; CART, Classification And Regression Tree.

It is worth noting that each rule is associated with an occurrence probability that can be estimated based on the number of the records associated with the defined rule:

$$P = nc \ / n \qquad \text{[Eqn 3]}$$

Where P is the probability of the rule occurrence, n is the number of all considered records and nc is the number of coincided records with rule.

It is obvious that the deeper (belonging to the lower level of trees) the branch is the lower the value of n will be.

Out of 89 drown rules and considering frequency of rules in some trees, 68 rules were obtained from data mining. Also, the variables that had strong correlation with the target variable (promotion score) were province of employment; examination score; interview score; education degree; years of experience; mathematics score; English score; university graduated from; education discipline and general information score. Therefore only 10 variables had most effect on the 68 created rules.

Results

Selection of rules

It is fair to assume that not every created rule is reasonable. There are three main reasons for this:

1. Not all presented rules have a high probability and frequency.
2. Some rules may contradict others.
3. Models created from data do not necessarily contain knowledge, therefore, they might have been created randomly.

Considering that the analysis of all these rules could be tedious and difficult, final rules will be created based upon expert judgments according to the following principles:

- Rules that have both $n >= 100$ and $P > = 60\%$.
- Rules that have both $n > = 50$ and $P > = 70\%$.

By applying the above principles and our expert judgment only 19 rules were chosen. Some of the selected rules are presented in Table 7.

TABLE 7: The final rules sample.

Rule number	Number in branch (*n*)	Probability (*P*) (%)	Rule description
R12	228	82.00	Applicants whose examination score and interview score are less than or equal to 6449.5 and 87.5 respectively, their promotion will be 'D' after 3 years.
R15	192	63.50	Applicants whose examination scores are more than 6449.5 and interview scores are less than 84.5, their promotion will be 'D' after 3 years.
R27	117	89.70	Applicants without experience that have B.S. Degree, their preferment will be 'D' after 3 years.
R36	220	83.50	Applicants with experience whose 'examination score' is less than 6449.5, their Promotion will be 'D' after 3 years.
R40	182	62.60	Applicants whose examination scores are more than 6449.5 and whose 'interview score' is less than 84.5 their preferment will be 'D' after 3 years.
R53	138	100.00	Applicants that have a high school diploma or college diploma degree and their 'Job Score' is more than 2075, their promotion will be 'B' after 3 years.
R60	137	83.20	Applicants that have B.S. degree, their promotion will be 'D' after 3 years.

Also, the number of variables that have a strong correlation with the target variable (promotion score) were decreased to five variables as follows:

- education level or degree (repeated in 11 rules)
- examination score (repeated in 8 rules)
- interview score (repeated in 8 rules)
- province of employment or job location (repeated in 6 rules)
- years of experience (repeated in two rules).

Final research model

The summary of the foregone steps is depicted in Figure 2. It is worthwhile to note that in the final of the eleven-step approach presented earlier, stresses on the repeatability of the data mining process and the benefit of the discovered knowledge are in redoing of the steps. This way, improvements can be made in data mining results and rules creation in each iteration of the exam. Furthermore, data mining will be performed in a more guided manner.

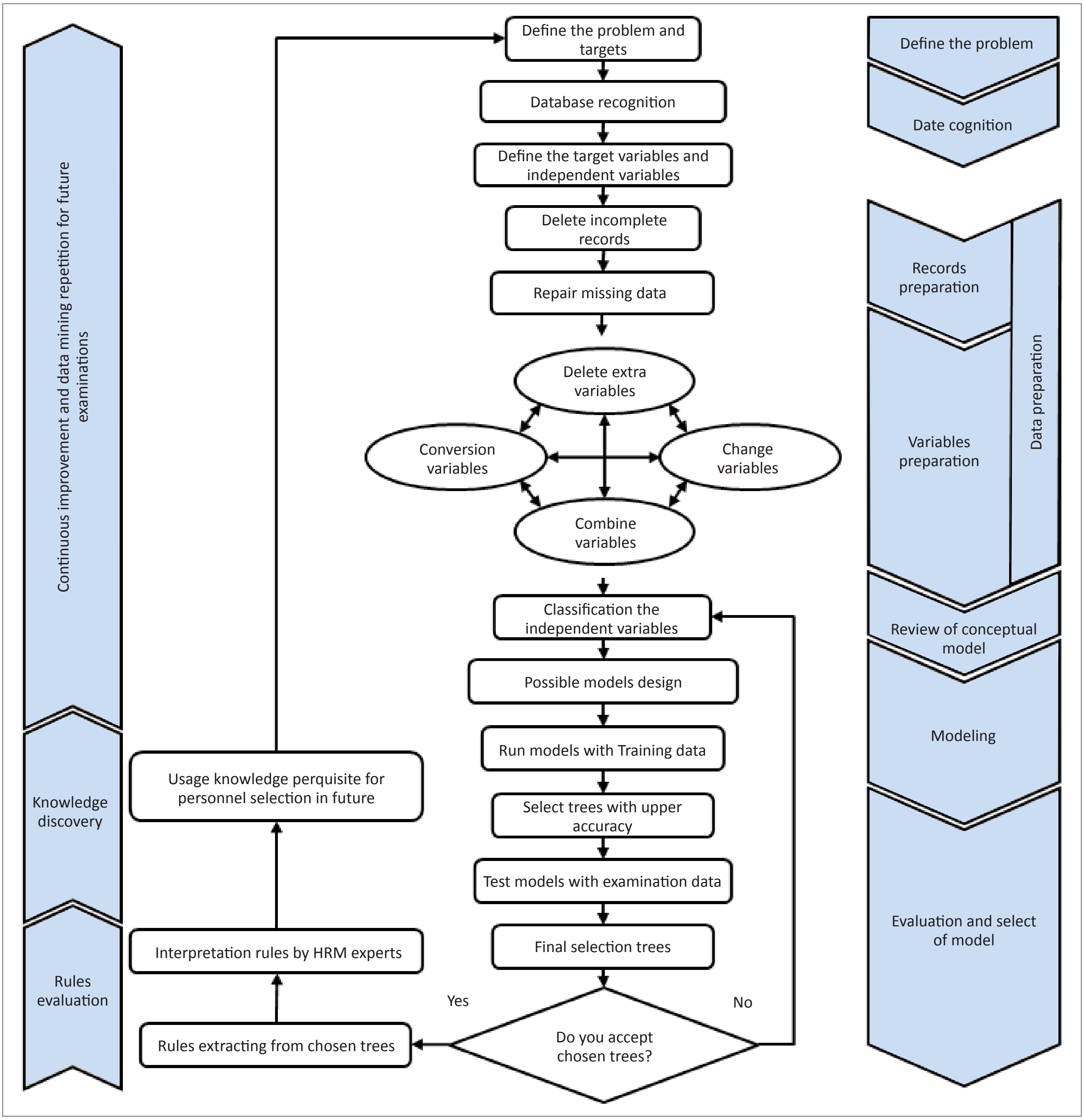
Continuous improvement and data mining repetition for future examinations
Knowledge discovery
Rules evaluation
Define the problem and targets
Database recognition
Define the target variables and independent variables
Delete incomplete records
Repair missing data
Delete extra variables
Conversion variables
Change variables
Combine variables
Classification the independent variables
Possible models design
Run models with Training data
Select trees with upper accuracy
Test models with examination data
Final selection trees
Do you accept chosen trees?
Yes
No
Rules extracting from chosen trees
Interpretation rules by HRM experts
Usage knowledge perquisite for personnel selection in future
Define the problem
Date cognition
Records preparation
Variables preparation
Data preparation
Review of conceptual model
Modeling
Evaluation and select of model

FIGURE 2: Knowledge discovery model from database of entrance examinations with data mining.

Discussion

Obviously, many methods are introduced for recruitment in management science. Non-use of information technology in this field and using the idea of knowledge hidden in the database of current staff performance creates new issues in this field. The main objective of this study is to prevent the improper selection of employees by using statistical and commonly known methods.

The purpose of this research was to search the database for employee performance by using data mining methods in order to discover patterns of employee efficiency and effectiveness and to convert this information into useful knowledge for the organisation. After discovering these patterns, the rules related to factors affecting performance were extracted and provided to managers. This will help them select the best employees.

According to the comments of bank experts about rules, some conclusions derived are as follows:

- Deleting of target variables such as 'performance assessment' demonstrates that system performance evaluation is not implemented properly.
- According to derived rules, multiple-choice tests conducted are not standard and not appropriate criteria for use in an interview session.
- Age, gender and marital status of employees have no effect on their upgrade's status.
- Exam courses such as literature, education, mathematics and general knowledge have no effect on performance and promoting those selected.
- Some upgrades have been made when an employee submits a degree after employment.
- Employees' university of choice has an effect on the interview.
- Employees' major field of study has no effect on employment.
- With regard to the accuracy of extracted rules, higher knowledge represents superior candidates.
- The main reason for promoting people in the Northern areas is related to the extent of these areas and the fact that posts are emptied earlier in these areas.

Although this information is openly available to managers, documents can now be submitted to check this with confidence.

Finally, it is obvious that knowledge cannot necessarily be obtained from extracted rules. Other research should be done to provide additional knowledge.

The following recommendations can be submitted to improve the study:

- With regard to the removal of test variables, it is essential to change the type of selected course and its content.
- It is better to change scoring methods for performance evaluation and training tools should be used for completing information.
- It is essential that a realistic performance assessment is kept separate from performance assessment based on payment.
- It is appropriate to design and replace exams in which intelligence and capabilities such as report writing and basic banking information can be measured.
- It is proposed that training sessions should be held for interviewees before conducting interviews.
- It is appropriate to record the details of interviews and scores in separate databases.
- It is better to define a coefficient for each course of the entrance exam.
- It is essential to maintain exam information, private information, job information and performance information for more detailed assessment in the future.

The continuous update of current databases in organisations can be considered important for the data mining process. The process of providing and maintaining data can be improved by software specialists.

Conclusion

The appropriate and scientific employment of staff selected on the basis of organisations and companies entrance exams, especially in service-based organisations, are of crucial importance. The use of data mining and the discovery of covert knowledge in these organisations will be very effective.

In this article, whilst existing approaches to data mining applied in human source management were reviewed, a new model with a stress on the best feasible tree or trees selection appropriate with the data bank was presented. The status of employees' upgrades were recognised by using data mining techniques, the rules and relation between entrance exam marks and staff situation with job performance. Firstly, job performance, the private information of employees in commercial banks and the human indices which were effective in performance or upgrade were identified by studying exam databases. In the next step, a suitable model was designed to address the shortcomings of previous research by learning from these studies about data mining and human source performance. Secondly, a data warehouse was created. With data mining techniques, the data set was cleansed. In other words, the unsuitable and redundant variables were removed. After data cleansing, the records with empty fields were deleted and suitable amounts were replaced for lost data. Thirdly, with decision tree techniques, data was classified and analysed and the primary results were obtained. Finally, the orders which were not obvious were explained.

In conclusion, the validity and reliability of the rules were examined and the model for knowledge discovery and application for personnel selection was presented. As a result, in a conceptual approach to the most effective variables, as well as target variables, it was determined that the 'performance evaluation' target variable was not an appropriate choice for building a relation with other variables. Also, it was concluded that, out of 26 effective variables only five variables, such as province of employment; education level; exam score; interview score and work experience had the most effect on the 'promotion score' target.

It is of great importance that an extensive use of data mining techniques can be made in other management areas. Whilst this is a low-cost technique, it can help managers to discover covert knowledge in their organisations.

Limitations of the study

The limitations of the study can be classified in two separate groups as follow:

- Lack of access to data mining software in Iran.
- Lack of access to personnel data bases.

Recommendation for future research

In order to complement and enhance the results of this study in human resource management, the following recommendations can be considered for future research:

- Designing a model for recruitment of entrance exam volunteers by means of a fuzzy data mining approach.
- Investigating the difference between the result of data mining for the recruitment model of entrance exam's volunteers by means of a decision tree and neural network technique.
- Investigating the relationship between the evaluation scores of interviewees with the future performance of employed people by means of the data mining approach.
- Applying data mining techniques in other areas of management.
- Investigating the relationship of banks' employee performance in provincial branches with the index of deposits absorption.

Acknowledgements

Competing interests

The authors declare that they have no financial or personal relationship(s) which may have inappropriately influenced them in writing this article.

Authors' contributions

A.A. (Tarbiat Modares University) was the supervisor of the dissertation. M.V.S. (Tarbiat Modares University) was the main provider of dissertation and was responsible for experimental and project design and wrote the manuscript. P.A. (Tarbiat Modares University) was the advisor of dissertation. A.R. (Tarbiat Modares University) partly wrote the manuscript.

References

Anderson, A. (1994). *Effective personnel management: A skills and activity-based approach.* Oxford: Blackwell Business.

Beardwell, J., & Wright, M. (2004). Recruitment and selection. In I. Beardwell, L. Holden & T. Claydon (Eds.), *Human Resource Management: A contemporary approach* (4th edn., pp.190–229). London: Prentice Hall.

Bekhor, S., Mahalel, D., Prashker, J., Prato, C.G., Galtzur, A., & Factor, R. (2007). *Exploratory research by means of data mining for investigating the relationship between the infrastructure, the drivers and the characteristics of offences*. Paper presented to the Ran Naor Foundation for the Advancement of Road Safety Research. Retreived n.d., from http://www.oryarok.org.il/webfiles/audio_files/behor_mahalel.pdf?&lang=en_us&output=json

Berry, M.J.A., & Linoff, G.S. (2004). *Data mining techniques for marketing sales and customer relationship management.* (2nd edn.). The Hoboken, New Jersey: John Wiley & Sons Publishing Inc.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and regression trees*. New York: Chapman & Hall.

Chien, C.F., & Chen, L.F. (2007). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications, 34*, 280–290. http://dx.doi.org/10.1016/j.eswa.2006.09.003

Cowling, A., & Mailer, C. (1990). *Managing human resources.* (2nd edn.). London: Edward Arnold.

Daniel, T., & Larose, J. (2005). *Discovery knowledge in data: An introduction to data mining*. Wiley Interscience.

Daniel, T., & J. Larose. (2006). *Data mining methods and models*. The Hoboken, New Jersey: Wiley & Sons Publishing Inc.

Denker, R. (2000). *Audites for human resource applications*. Retrieved 2008, from http://www.wizsoft.com/default.asp?Win=8&winsub=49

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge: The MIT Press.

Hartigan, J.A. (1975). *Clustering Algorithms (Probability & Mathematical Statistics)*. John Wiley & Sons Inc.

Howeedy, R. (2002). *First door uses web trends to create a human resource database.* Retrieved 2008, from http://www.information-management.com/issues/20020601/5262-1.html

Jazani, N. (2000). *Human resource management*. Tehran: Ney Publishing Inc.

Kantardzic, M. (2003). *Data mining: Concept, models and algorithms*. Wiley: IEE Press. PMid:12858662

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, *29*(2), 119–127. http://dx.doi.org/10.2307/2986296

Khaef, E., Ahmad, A., Mottaghi, P., & Sebt, M.V. (2007). *Examining the influence of employment model utilization based on data mining, on the employees' replacement rate*. First Iran Data Mining Conference, Nov.

Loh, W.Y., & Shih, Y.S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*, 815–840.

Mendonca, M., & Sunderhaft, N. (1999). *Mining software engineering data: A survey*, Data & Analysis Center for Software (DACS) State-of-the-Art Report, No. DACS-SOAR-99-3.

Olson, D.L., & Jesse, S. (2012). Case study of open-source enterprise resource planning implementation in a small business. *Enterprise IS, 6*(1), 79–94. http://dx.doi.org/10.1080/17517575.2011.566697

Olson, R.K. (2002). Genetic and environmental influences on reading and related cognitive skills. *Dyslexi, 4*, 10–16.

Porter, K., Smith, P., & Fagg, R. (2006). *Leadership and management for HR professionals.* Oxford: Butterworth Heinemann.

Quinlan J.R. (1986). The effect of noise on concept learning. In Michalski *et al*., (Ed.), *Machine Learning: An Artificial Intelligence Approach*, Vol. 2. Morgan Kaufmann.

SAS Institute Inc. (2004). SAS® 9.1 SQL Procedure User's Guide. Cary, NC: SAS Institute Inc.

Schroeck, M.J. (1999). *Data warehousing in human resource management systems.* Information Management Magazine. Retrieved 2008, from http://www.information-management.com/issues/19990601/1019-1.html?zkPrintable=true

Weiss, S.M., & Indurkhya, N. (1998). *Predictive data mining: A practical guide*. San Francisco: Morgan Kaufmann.