The Impact of Data Mining on System Analysis Process

تأثير تقنية تعدين البيانات على اساليب تحليل النظم

Dr. Yahya M. Hadi Al_mayali
Computers Department, Girls Education College,
Kuffa University.
e-mail: Yahya_mayali@yahoo.com
Mobile: 07901733096

ABSTRACT
With the advent of the computer age, people have begun using computers to automate the data gathering process and store the information in databases. Computers are so well suited to this task that huge databases with terabytes of information have been generated. It is well beyond the scope of the human mind to sort through all of this data and find any useful patterns for predicting future events. The mining has been invented as one technique of the machine learning field to deal with this new problem by using computers to automate the process of searching data in huge databases for useful patterns, which is can used to build a new system.
In this paper we need to show how data mining techniques can help the systems analysts people for studying and extract facts for building new systems.
Keywords
Machine Learning, System Analysis, Data Mining, Knowledge Discovery from Databases.

المستخلص
أدى التطور المتصاعد في تكنولوجيا الحاسوب (كيانات مادية وبرامج) وكذلك التطور في نظم الأتصالات والأنترنيت وأعتماد الحاسوب في العمل اليومي الى بناء العديد من نظم قواعد بيانات لمختلف المجالات الأقتصادية والأجتماعية والعلمية. أن الأعتماد اليومي للحاسوب ادى الى نمو متصاعد في حجم المعلومات المخزونة في نظم قواعد البيانات وان دراسة هذه الأنظمة لغرض تطويرها.والتدقيق فيها والبحث في هذا الكم الضخم من البياتات بأستخدام الطرق التقليدية اصبح اكبر من طاقة العقل الأنساني. لذا فمن الأجدر ان يوظف الباحثون الحواسيب لاغراض اتمتة عملية التحليل واستخراج الأنماط المخفية للمتغيرات داخل نظم قواعد البيانات من خلال ايجاد طرق واساليب تتلائم والمرحلة الحالية،واستخدام النتائج في بناء وتطوير الأنظمة.أوجد الباحثون تقنية التعدين كواحد من اساليب تعلم الماكنة لكي تمثل حلا لمشكلة البحث في البيانات الكبيرة الحجم من خلال إستعمال الحواسيب لأتمتة عمليةالبحث في البيانات واستنباط المؤشرات المفيدة التي يمكن ان تستخدم لآغراض تطوير النظم الحالية اوبناء نظم جديدة.
هذا البحث يبين تأثيرتقنية تعدين البيانات على اساليب تحليل ونصميم النظم واعتمادها كأحد طرق ايجاد الحقائق وأكتشاف الأنماط المخفية داخل قواعد البيانات حيث تعتبر من الأساليب المهة والناجحة للتعامل مع الكم الهائل في المعلومات.

1. INTRODUCTION

Few years ago, systems analysts use the fact finding methods to search over hundreds, maybe even thousands of bits of data looking for underlying patterns that would reliably predict system future outcomes. Through the internet and other private networks can generate hundred megabytes of data per second, which is may be reach more than four terabytes a day (terabytes = 2^{40} Bytes). For comparison, four terabytes is enough space to store fifteen hundred copies of the thirty two-volume text of any large Encyclopedia such as Encyclopedia of Britain [1].

Figure (1) depict the huge amount of are automatically collected. So once any organization has all of this data, they will need some way to analyze it to produce useful conclusions, otherwise the analysis of this huge volume of data may be larger than the ability of any human being. .

Therefore we need a new technique to be able for examining such a volume of information today. One of such new technique is data mining.

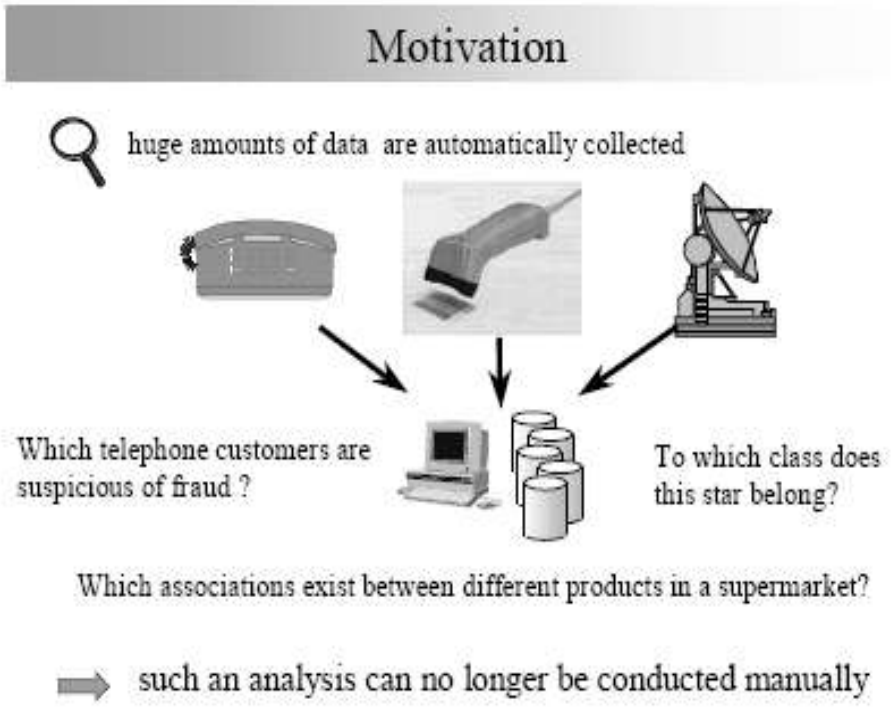

Figure (1) Resources of data to update the databases

Data mining, also known as knowledge discovery from databases, is the analysis of data to search for underlying patterns that will hold for all occurrences of the data source. These patterns can then be used to predict future events with a fair degree of certainty. These predications can be used to find measures to prevent undesired events and to promote desirable trends. Data mining has numerous applications in science, security, and business [2].

2. THE CURRENT ISSUE IN DATA MINING

With real case studies of organizations deploying data mining as a catalyst for enhancing and reengineering their business processes, data mining is now entering mainstream IT as a mature and tested technology. With this phase of evolution data mining has moved beyond the debate on algorithms and into the debate on usability. There are three main issues which should be considered by any organization considering the introduction of data mining:

1.Methodology,
2.Ease of use and
3.Performance / scalability.

Methodology

For data mining to gain wide acceptance, it is important to have a step by step methodology for a data mining project. This ensures that the benefits reported by seasoned data miners are repeatable by other people in various business sectors. This can help dispel the belief that data mining is a kind of 'black art' which can only be practiced by specialist. Such a methodology is beginning to emerge and there is certainly wide agreement on the main steps of such a methodology.

Figure (2) illustrate and summarized data mining methodology steps.

Ease of Use

Data mining tools are increasingly used by computer literate business users. This requires these tools to be no more difficult to use than a spreadsheet program. Furthermore the data mining tool needs to support all the steps of a data mining methodology. Finally, because of the nature of data mining, the tool has to support extensive data and patterns reporting and visualization.

Performance and Scalability

With the decreasing costs of data processing and storage comes the data rich organization. It is now common place for small and medium sized organizations to hold gigabytes of data relating to a business process. It is therefore essential that data mining tools can deliver acceptable performance on large volumes of data regardless of the computing platform / architecture being used. There are a number of computing architectures for data mining

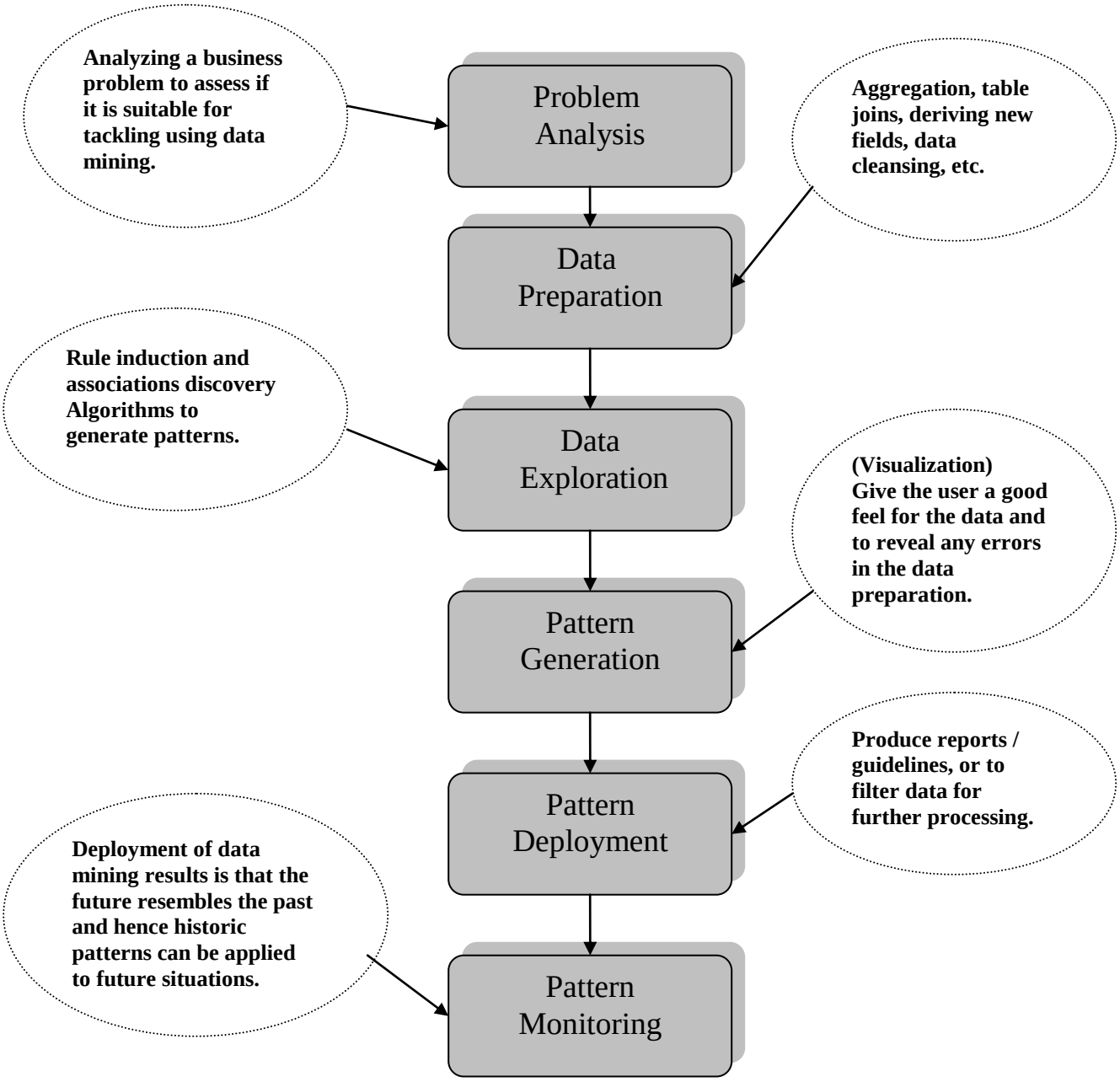

Figure (2) Data Mining Methodology Steps

3. THE BASIC DATA MINING PROCESS

Before any data mining systems can be used on a set of data, automated data reduction techniques often need to be used to summarize the data to a manageable size. These data reduction techniques often include cataloging, classification, clustering, segmentation, and partitioning, as well as other forms of sorting [2, 5, 6]. This tends to be a tedious and time-consuming part of the process because tremendous

amounts of data need to be manipulated. Data cleaning should also occur at this stage, meaning that incorrect, inconsistent, incomplete, or missing data needs to be accounted for. Unfortunately, most current data mining tools do not deal with data cleaning very well.

Data mining programs can use any of a number of algorithms including advanced data visualization, a multitude of statistical procedures, tree-based modeling and segmentation, genetic algorithms, machine learning, inductive reasoning and association, neural networks, and pattern/image recognition algorithms. [4] Whatever method the data-mining program uses, it is executed on a sample of the reduced and cleaned data set. The trends and rules the program generates are then tested against another segment of the data set to see if they still hold true, and therefore represent true patterns rather than coincidences specific to the data sample. At this point, human analysts examine the patterns found by the data mining to determine which can be useful or should be explored more intricately. Truly useful and important patterns are often repeatedly tested with new data sets to confirm their reliability [7].

4. A BASIC DATA MINING ALGORITHM

It necessary to make one more assertion regarding the difference between data mining and data modeling. Data mining is about discovering understandable patterns (trees, rules or associations) in data.Data modeling is about discovering a model that fits the data, regardless of whether the model is understandable - (e.g. tree or rules) or a black box (e.g. neural network). Based on this assertion, we restrict the main data mining technologies to induction and the discovery of associations and clusters.

Rule Induction

Rule or decision tree induction is the most established and effective data mining technologies in use today. It is what can be termed 'goal driven' data mining in that a business goal is defined and rule induction is used to generate patterns that relate to that business goal. Rule induction will generate patterns relating the business goal to other data fields (attributes). Inferring rudimentary rules is one method to learn a 1- level decision tree (1R) in the following a Pseudo code for 1R.

For each attribute,

For each value of the attribute, make a rule as follow:
Count how often each class appears
Find the most frequent class
Make the rule assign that class to this attribute-value
Calculate the error rate of the rules
Choose the rules with the smallest error rate
Note: "missing" is treated as a separate attribute value

The resulting patterns are typically generated as a tree with splits on data fields and terminal points (leafs) showing the propensity or magnitude of the business event of interest.
Decision tree learning is one of the most widely used and practical methods for data mining. It is a method for approximating discrete-valued functions that is robust to noisy data and capable of learning disjunctive expressions. I will describe the basic decision tree algorithm ID3 [6]. Decision trees are popular because they are robust against errors in the data and missing attributes.

Learned function is represented as a decision tree. Learned trees can also be re-represented as sets of *if-then* rules to improve human readability.
Table(1) below represent some training data extracted from whether database case be used to illustrate how the decision tree build and the Entropy formula can be used.

Table (1) Example of training data from whether database

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No -
D2	Sunny	Hot	High	Strong	No -
D3	Overcast	Hot	High	Weak	Yes +
D4	Rain	Mild	High	Weak	Yes +
D5	Rain	Cool	Normal	Weak	Yes +
D6	Rain	Cool	Normal	Strong	Yes +
D7	Overcast	Cool	Normal	Strong	Yes +
D8	Sunny	Mild	High	Weak	No -
D9	Sunny	Cool	Normal	Weak	Yes +
D10	Rain	Mild	Normal	Weak	Yes +
D11	Sunny	Mild	Normal	Strong	Yes +
D12	Overcast	Mild	High	Strong	Yes +
D13	Overcast	Hot	Normal	Weak	Yes +
D14	Rain	Mild	High	Strong	No -

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute in the given example. This process is then repeated for the sub tree rooted at the new node [6].

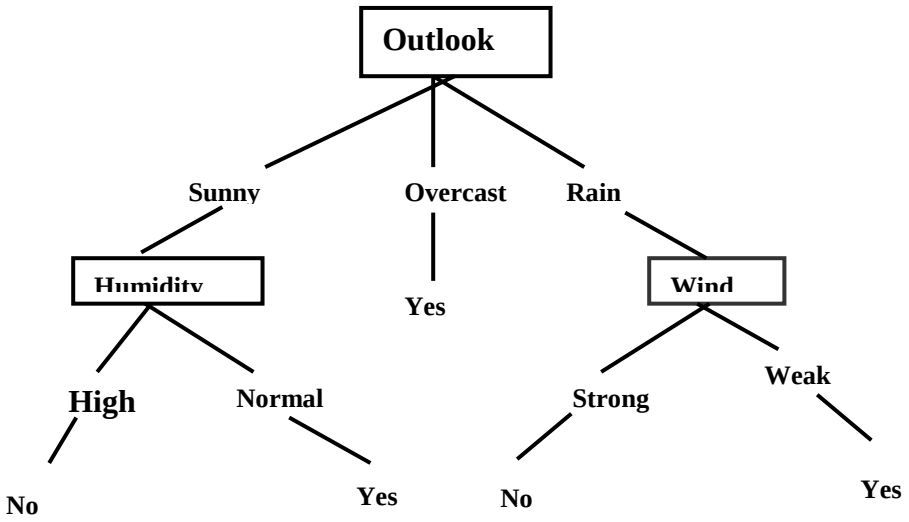

Figure (3) Basic decision tree

Figure (3) illustrates a simple learned decision tree. This decision tree classifies Saturday mornings according to whether they are suitable for playing tennis.

For example, the instance: (Outlook = Sunny, Temperature = Hot, Humidity = High, Wind = Strong) would be sorting down the leftmost branch of this decision tree and would therefore be classified as a negative instance because it leads to a .No.. Note that the Wind and Temperature attributes are ignored because they are not part of the branch that leads to the conclusion. This allows decision trees to avoid considering unnecessary information [6].

ID3 learns decision trees by constructing them top-down, beginning with the question .which attribute should be tested at the root of the tree?. To answer this question, each attribute is evaluated using a statistical test to determine how well it alone classifies the training examples. The best attribute is selected and used as the test at the root node of the tree. A descendant of the root node is then created for each possible value of this attribute, and the training examples are sorted to the appropriate descendant node. This process is repeated using the training examples associated with each descendant node to select the best attribute to test at that point in the tree. This forms a greedy search for an acceptable decision tree. Note that the algorithm never backtracks to reconsider earlier choices [6].

In order to decide which attribute to test at each node in the tree, we begin by defining entropy. Entropy characterizes the (im)purity of an arbitrary collection of examples. Given a collection S, containing positive and negative examples of some target concept, the entropy of S relative to this boolean classification is:

$$\text{Entropy}(S) = -p(+)\log_2 p(+) - p(-)\log_2 p(-) \quad \text{.....} \quad (1)$$

where p(+) is the proportion of positive examples in S and p(-) is the proportion of negative examples in S [6]. To illustrate, suppose S is a collection of 14 examples of some Boolean concept, including 9 positive and 5 negative examples (we adopt the notation [9+, 5-] to summarize such a sample of data).Then the entropy of S is:

$$\text{Entropy}([9+,5-]) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0.940 \quad \ldots\ldots\ldots\ldots \quad (2)$$

Thus far we have discussed entropy in the special case where the target classification is Boolean. For calculating entropy in a more general way, if the target attribute can take on c different values, then the entropy of S relative to this c-wise classification is defined as:

$$\text{Entropy}(S) = \sum_{i=1}^{c} -Pi \log 2 Pi \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

Where pi is the proportion of S belonging to class i. Note the logarithm is still in base 2 because entropy is a measure of the expected encoding length measured in bits [6]. Entropy is important for deciding which attribute would be the best test for a node of a decision tree because it is used to calculate the information gain. Information gain is the expected reduction in entropy caused by partitioning the examples according to this attribute. More precisely, the information gain, Gain(S, A) of an attribute A, relative to a collection of examples S, is defined as:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in Values(A)} |Sv| / |S| \, Entropy(S_v) \quad \ldots\ldots\ldots\ldots \quad (4)$$

Where values(A) is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v [6].

Note that the first term in Equation (4) is just the entropy of the original collection S, and the second term is the expected value of the entropy after S is partitioned using attribute A. For example, suppose S is a collection of training-examples described by attributes including Wind, which can have the values Weak and Strong. As before, assume S is a collection containing 14 examples, [9+, 5-]. Of these 14 examples, suppose 6 of the positive and 2 of the negative examples have Wind = Weak, and the remainder have Wind = Strong. The information gain due to sorting the original 14 examples by the attribute Wind may then be calculated as:

Values(Wind) = Weak, Strong

S = [9+, 5-]

S_{Weak} [6+, 2-] ←
S_{Strong} [3+, 3-] ←
Gain(S,Wind) = Entropy(S)- $\sum_{V \in (Weak, Strong)}$ $|S_v|/|S|$ □Entropy(S_v)
= Entropy(S) – (8/14)Entropy(S_{Weak})
. (6/14)Entropy(S_{Strong})
= 0.940 – (8/14)0.811 – (6/14)1.00
= 0.048

At each node of a decision tree ID3 calculates all the information gains of the possible attributes, then it chooses the attribute that will yield the highest information gain and makes that the test of the current node. In this way ID3 maximizes the utility of each node and generally generates shorter trees. The definition of the ID3 algorithm is ID3 (Examples, Target attribute, Attributes). Examples are the training examples [6].

Target attribute is the attribute whose values are to be predicted by the tree. Attributes is a list of other attributes that may be tested by the learned decision tree. Returns a decision tree that correctly classifies the given Examples.

- Create a Root node for the tree
- If all Examples are positive, Return the single-node tree Root, with label = +
- If all Examples are negative, Return the single-node tree Root, with label = –
- If Attributes is empty, Return the single-node tree Root, with label = most common value of Target attribute in Examples
- Otherwise Begin
- A ← the attribute from Attributes that best classifies Examples
- The decision for Root ← A
- For each possible value, vi, of A,

  - Add a new tree branch below Root, corresponding to the test A = vi

  - Let Examples vi be the subset of Examples that have value vi for A

  - If Examples vi is empty
- Then below this new branch add a leaf node with label = most common value of Target attribute in Examples
- Else below this new branch add the subtree ID3(Examples, Target attribute, Attributes- {A})
- End
- Return Root

* The best attribute is the one with the highest information gain, as defined in Equation (4).

The Discovery of Associations

This is the second most common data mining technology and involves the discovery of associations between the various data fields. One popular application of this technology is the discovery of associations between business events or transactions. For example discovering that 90% of customers that buy product A will also buy product B (basket analysis) or that in 80% of cases when fault 1 is encountered then fault 7 is also encountered. If the sequence of events is important then another data mining technology for discovering sequences can be used.A second application of association's discovery data mining is the discovery of associations between the fields of case data. Case data is data that can be structured as a flat table of cases. Records of mortgage applications are an example of case data. In such data, associations can be found between data fields; for example that 75% of all applicants that are over 45 and in managerial occupations are also earning over £40,000. Such associations can be used as a way of discovering clusters in the data.

Note that this differs from rule induction on case data in that no outcome needs to be defined for the discovery process.

Classification and clustering

Classification has been an age old problem. Many researchers tried to invent variety of methods and concepts of classifying all form of life for organizing the rich diversity in living organisms.

Today's classification systems, try to find differentiating features between classes and use them to classify unknown instances. It has been recognized as an important problem in data mining. Among other knowledge discovery tasks and has attracted great attention for the past years in the data mining community. Figure (4) depict the classification system in general and figure (5) explain the classification procedure in data mining [3,5] .

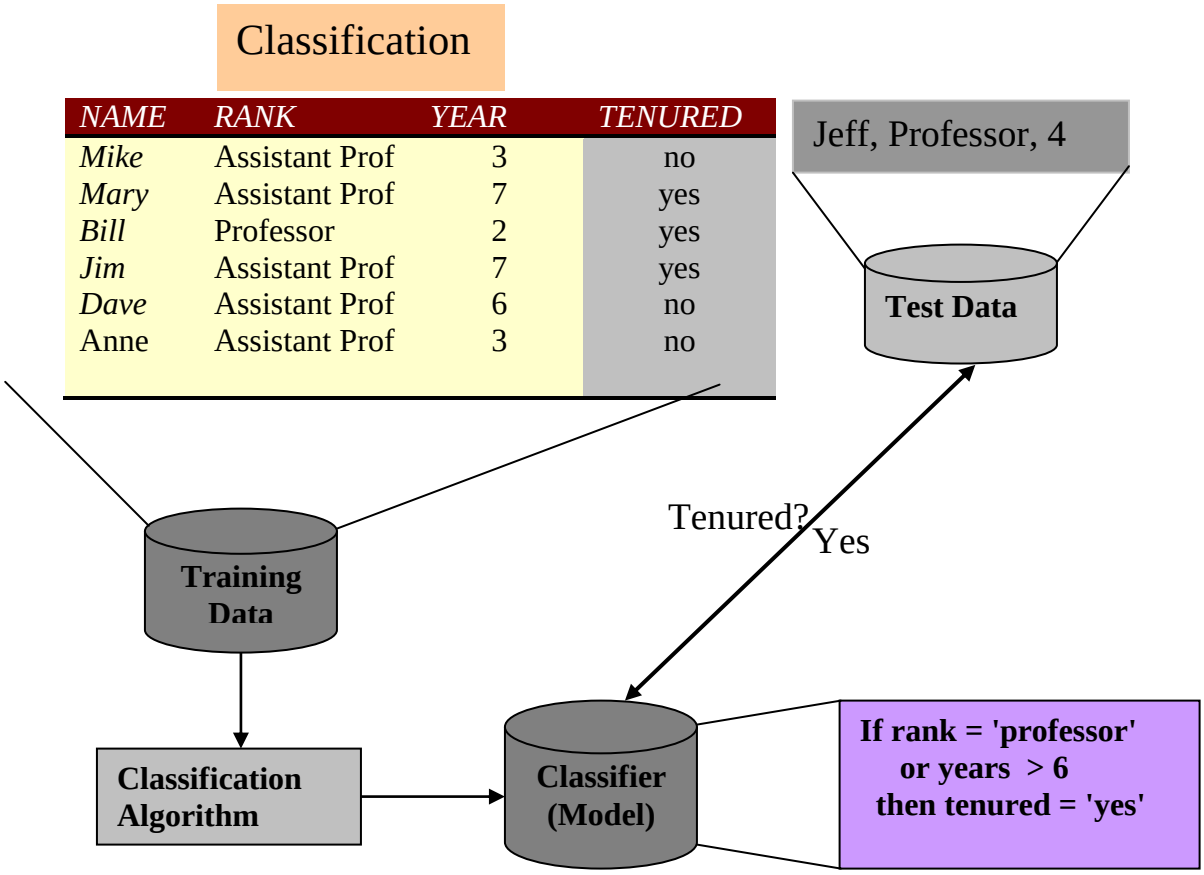

Figure (5) Classification procedure in data mining

5. DATA MINING IN BUSINESS

Most organizations can be currently labeled 'data rich', since they are collecting increasing volumes of data about business processes and resources. Typically, these data mountains are used to provide endless 'facts
and figures' such are 'there are 60 categories of occupation', '2000 customer's accounts ' etc. Such 'facts and figures' do not represent knowledge but if anything can lead to 'information overload'.
However, patterns in the data represent knowledge and most organizations nowadays can be labeled 'knowledge poor'.

Our definition of data mining is the process of discovering knowledge from data. Data mining enables complex business processes to be understood and re-engineered. This can be achieved through the discovery of patterns in data relating to the past behavior of a business process. Such patterns can be used to improve the performance of a process by exploiting favorable patterns and avoiding problematic patterns. Examples of business processes where data mining can be useful are customer response to mailing, lapsed insurance policies and energy consumption. In each of these examples, data mining can reveal what factors affect the outcome of the business event or process and the patterns relating the outcome to these factors. Such patterns increase our understanding of these processes and therefore our ability to predict and affect the outcome. Many businesses are interested in data mining because of the falling cost of data storage, the increasing ease of collecting data over networks, and the

immense computational power available at low prices. The development of robust and efficient data mining algorithms has caused most businesses to create huge databases containing as much information about their activities as possible. Already available on the market are generic multitask data mining tools to perform a variety of discovery operations.

Making data mining programs useful to businesses requires several elements. First, the problem needs to be stated in the business users. terms, including viewing the data in a business model perspective. Second, the program needs to support specific key business analyses such as segmentation, which is very important in marketing applications. Third, the results of the data mining need to be presented in a form geared to the business problem being solved. Finally, there has to be support for protracted data mining on an increasing data set, since business databases are continually growing to store increasing numbers of business transactions [1].

Data mining applications have been developed for a variety of businesses including marketing, finance, banking, manufacturing, and telecommunications [1, 6].

Many of these applications use a predictive modeling approach, but they encompass the full range of methods previously mentioned. Data mining in marketing falls into the broad area called database marketing. It consists of analysis of customer databases to select the best potential customers for a particular product.

Data mining offers a convenient way to monitor these large computer networks. By detecting anomalous activities in the logs of computers, a data mining system could flag suspicious events for later inspection by system administrators, allowing them to avoid checking all the normal daily activities. The data mining system does this by developing a profile of the typical activities of each user in the network. Deviations from the expected pattern could be harmful or abusive behavior and would therefore be flagged. The system would have to be flexible enough to compensate for normal deviations from expected behavior like users learning new programs or doing new tasks [2].

6. CONCLUSIONS

Although data mining is still limited in its functionality, its potential is nearly unlimited. Data mining expected to be a vital tool for developing the new systems or tuning the existing ones. Therefore data mining can be added to the systems analysis techniques especially for the large and huge database systems. Already business, science, and security have derived benefits from its development. Databases that used to store millions of bits of useless information can be mined for insights that can greatly profit the miners. Recently, scientific instruments and business systems have been gathering extra information that was apparently useless simply because it was so easy to do.The creation of data mining makes this excess information useful. Research to expand the types and magnitude of data that data mining systems can effectively mine is well underway. The needs of business, security, and science will provide incentive to invest time and money into such development.

Many researchers expect the data mining will advance faster than the growth of databases and allow the mining of nearly infinite databases, such as mining the entire World Wide Web.

7. REFERENCES

[1] Brachman, Ronald J., Tom Khabaza, Willi Kloesgen, Gregory Piatetsky-Shapiro, Evangelos Simoudis, .Mining Business Databases,. *Communications of the ACM*, v39, p42(7) (Nov 1996).

[2] Brodley, Carla E., Terran Lane, and Timothy M. Stough, .Knowledge Discovery and Data Mining,. *American Scientist*, v86, p54(8) (Jan-Feb 1999).

[3] Mitchell, Tom M., "Machine Learning and Data Mining.", *Communications of the ACM*, v42, p30 (Nov 1999).

[4] Fu, LiMin, .Knowledge Discovery Based on Neural Networks, *Communications of the ACM*, v42, p48 (Nov 1999).

[5] Risi Thonangi. "CLASSIFYING CATEGORICAL DATA". A MASTER OF SCIENCE BY RESEARCH (COMPUTER SCIENCES),at the International Institute of Information Technology, Hyderabad, 2005.

[6] Mitchell, Tom M., Machine Learning. McGraw-Hill Companies, 1997.

[7] Studt, Tim, .Scientific Data Miners Make use of all the Tools Available: Data Mining can Extract Previously Unknown and Potentially Useful Information from Extremely Large and Complex Data Bases,. *R & D*, v39, p62C(2) (April 1997).