The Benefits of Using Data Mining Approach in Business Intelligence for Healthcare Organizations

Hisham S, Katoua

Management Information Systems Dept.
Faculty of Economics & Administration
King Abdulaziz University ,Jeddah, Kingdom of Saudi Arabia

dr_ hisham_sa@hotmail.com

Abstract

Healthcare-based business intelligence systems are complex to build, maintain and face the knowledge engineering and computing difficulties. Intelligent data mining techniques provide an effective computational methods and robust environment for business intelligence in the healthcare decision making systems. This paper discusses the benefits of using data mining and knowledge discovery methodology in business intelligence for healthcare.

Keywords: *Knowledge Engineering, Intelligent Decision Making, Business Intelligence, Healthcare, Data Mining.*

1. Introduction

Intelligent data mining (IDM) approach aims to extract useful knowledge and discover some hidden patterns from huge amount of databases, which statistical approaches cannot discover. IDM and knowledge discovery (KD) is not a coherent field, it is a dwells upon already well-established technologies including data cleaning, data preprocessing, machine learning, pattern recognition, statistics, neural networks, fuzzy sets, rough sets, clustering, etc [1,2].Recently, researchers have begun to investigate various data mining knowledge discovery methods to improve the efficiency of healthcare organizations through developing a new generation of business intelligence systems[3].

Healthcare organization uses the business intelligence (BI) solution not only for analysis but also to change business processes and drive toward the value-driven healthcare vision. BI provides an integrated view of data that can be used to monitor key performance indicators, identify hidden patterns in diagnosis, illuminate anomalies in processes, and identify variations in cost factors, all of which facilitate accountability and visibility and can drive an organization towards efficiency [4, 5].In the context of BI, intelligent technologies can be seen as enabler for managing, storing, analyzing, visualizing, and giving access to a great amount of data. For this purpose, a wide range of intelligent technologies (e.g. expert systems, online analytical processing, data mining and knowledge discovery, grid computing, cloud computing) are used in developing of a BI systems.

In this paper we focus our discussion around the data mining and knowledge discovery process in business intelligence for healthcare organizations. Section 2 presents the concept of the BI. Section 3 presents the goals of the healthcare organizations in the context of BI. Section 4 explains the data mining tasks and the knowledge discovery process for business

intelligence healthcare. Section 5 gives a brief overview of the data mining techniques in healthcare. Section 6 discusses the benefits of data mining technology from the BI point of view in healthcare sector. The final section contains the conclusion.

2. Business Intelligence Concept/Systems

Recently two distinct understandings of the term BI (respectively BI system) exist – a ***data-centric*** and a ***process-centric***. The ***data-centric*** position uses BI systems to combine operational data with analytical tools to present complex and competitive information to planners and decision makers. The objective is to improve the timeliness and quality of inputs to the decision process [4]. BI is therefore mainly used to understand the capabilities available in the organization[5]. The ***process-centric*** position notes a major shortcoming in this inherent data-centricity. Because the collection, transformation, and integration of data as well as information supply and analysis are commonly isolated from business process execution, a great part of the information that intrinsically exits within an organization remains either unused or is at most partially used but deprived of its interpretation context [6]. As they see an organization as a set of well-integrated processes [7]. BI therefore should be used to integrate the information world with the process world in order to facilitate decision making with an all-embracing information basis.

3. Healthcare Organizations In the Context of BI

3.1 Goals of the Healthcare Organization

In the context of business intelligence, the main goal of healthcare organization (HCO) is to reduce operating costs while maintaining a consistently acceptable level of patient treatment. Reduce operating costs at all levels: (a) Cost of healthcare professionals, (b) Cost of lab equipment and consumables, (c) Cost of pharmaceuticals / medical material, (d) Cost of a treatment per diagnosis related grouping, (e) Cost per type of medical intervention (e.g. specific medical operation). Moreover, goals of each HCO are:

1. Reduction of medical errors and exposure of the patient to medical hazards.
2. Support medical research with patient and treatment data.
3. Participate and support a larger healthcare system, with the exchange of medical information on a patient, as well as statistics on population morbidity and mortality.

On the other hand, an acceptable level of patient treatment involves:

1. Evidence based medicine, accurate diagnosis and efficient treatment,
2. On time admittance in the hospital and healthcare treatment,
3. Treatment with respect for the patient- analysis of options,
4. Reduction of risks during treatment, and
5. Capture of medical history of the patient in order to support evidence based medicine.

3.2 Business Intelligence Healthcare Processes

HCOs typically prescribe how their processes have to be performed; especially those processes that represent complex routine work, that involve many persons and organizational units and that are in general frequently performed [8]. In the context of BI, processes can be seen as primary trigger for information and data collection, processing and distribution and

are therefore very important. A differentiation in medical, business and support processes is shown in figure 1. ***Medical processes*** are those activities and work practices within a health care organization which are mainly focused on the health services delivery (e.g. nursing, medical treatment). ***Business processes*** comprise activities that are needed to effectively run the health care organization and may not be, or only partially sector specific (e.g. financial accounting). ***Support processes*** are used from both kinds of processes but only have an indirect impact on medical and business activities (e.g. supply of materials).

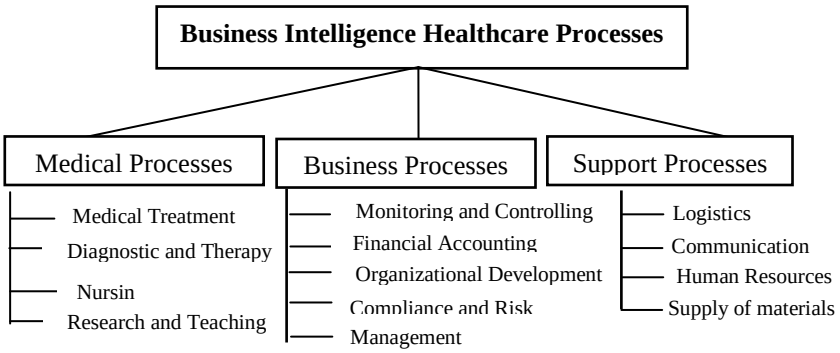

Figure1: Healthcare Process of the Context of Business

4. Data Mining for Business Intelligence in Healthcare

Fig. 2 shows the main functional phases of the knowledge discovery process. The preprocessing phase is often referred to as data cleaning. The cleaned data is stored in the warehouse. This is followed by data mining phase and its results are provided to an output generator (visualization) producing reports, action lists, or monitor reports. Each phase is supported by different methodologies. Data mining itself exhibits a plethora of algorithmic tools such as statistics, regression models, neural networks, fuzzy sets and evolutionary models.

The knowledge discovery process is arranged into five steps:

1. understanding the domain in which the discovery will be carried out,
2. forming the data set, its cleaning, and warehousing,
3. extracting patterns, this is essence of data mining,
4. post-processing of the discovered knowledge,
5. putting the results of knowledge discovery into use.

Fundamental issues in knowledge discovery arise from the very nature of databases and the objects (data) they deal with. They are characterized as follows: (a) huge amounts of data, (b) dynamic nature of data, (c) incomplete or imprecise data, (d) noisy data, (e) missing attribute values, (f) redundant or insignificant data. The knowledge discovery process is dynamic, highly interactive, iterative, and fully visualizable. Its main goals are to: (a) extract useful medical and administrative reports,(b) spot interesting events and trends,(c),support decision-making processes, and (d) exploit the data to achieve scientific, business, or operational goals.

Data mining is supported by a host that captures the character of data in several different ways.

1-Clustering: The key objective is to find natural groupings (clusters) in highly dimensional data. Clustering is an example of unsupervised learning, and it is a part of pattern recognition.

2-Regression Models: These originate from standard regression analysis and its applied part known as system identification. The underlying idea is to construct a linear or nonlinear function

3-Classification: This concerns learning that classifies data into the predetermined categories. The term originates from pattern recognition, in which a vast number of classifiers have been developed.

4-Summarization: This is an approach towards characterizing data via small number of features/attributes. In the simplest scenario one can think of a mean and standard deviations as two extremely compact descriptors of the data. This technique is often applied in an interactive exploratory data analysis and automated report generation.

5-Link analysis: It is concerned with determination of relationships (dependencies) between fields in a database. In a particular case we may be interested in the determination of the correlation between the variables.

6-Sequence Analysis: This type of analysis is geared toward problems of modeling sequential data. Pertinent models embrace time series analysis, time series models, and temporal neural networks.

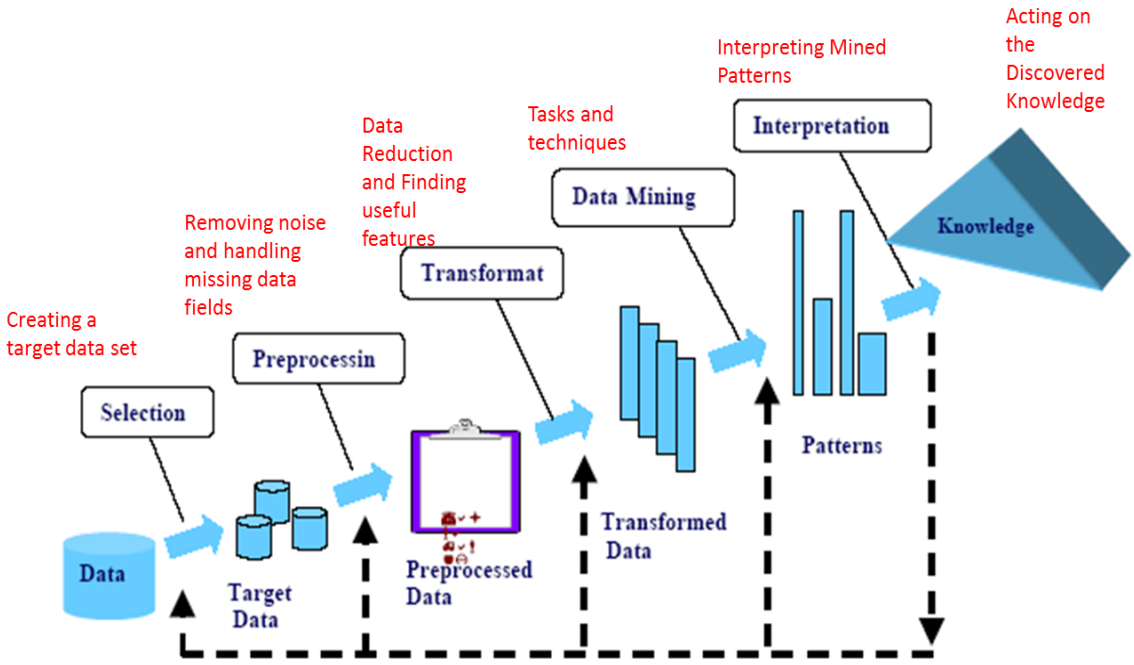

Fig. 2 Producing Knowledge with Intelligent Data mining approach

5. Data Mining Techniques in Healthcare

There are various data mining techniques available with their suitability dependent on the domain application (for more information, see [1,2]). Statistics provide a powerful and fundamental tool for quantification and evaluation of results. However, machine learning algorithms based on statistics need to be modified and scaled before they are applied to data mining. This section presents a brief account about the well known data mining techniques.

*1 Neural Networks (NN) :*NN are inspired in biological models of brain functioning. They are capable of learning by examples and generalizing the acquired knowledge. Due to these abilities the neural networks are widely used to find out nonlinear relations which otherwise could not be unveiled due to analytical constraints. The learned knowledge is hidden in their structure thus it is not possibly to be easily extracted and interpreted. NN is one of many data mining analytical tools that can be utilized to make predictions on key healthcare indicator such as cost or facility utilization. NN are known to produce highly accurate results and in medical applications can lead to appreciate decisions.

2.Support Vector Machines(SVM): SVM approach has its roots in statistical learning theory (SLT) and provides a way to build "optimum classifiers" according to some optimality criterion that is referred to as the maximal margin criterion. SVM have demonstrated significant efficiency when compared with neural networks. Their main advantage lies in the structure of the learning algorithm which consists of a constrained quadratic optimization problem, thus avoiding the local minima drawback of NN. An interesting development in SLT is the introduction of the Vapnik- Chervonenkis dimension, which is a measure of the complexity of the model.SVM treat both the problem of how to minimize complexity in the course of learning and how high generalization might be attained[9].

3. Clustering: Clustering techniques apply when the instances of data are to be divided into natural groups. The classical clustering technique is k-means where clusters are specified in advance prior to application of the algorithm. This corresponds to parameter k. Then k points are chosen at random as clusters centers. All instances are assigned to their closest cluster center according to the Euclidian distance metric. Next the centroid, or mean, of each cluster center is calculated. These centroids are taken to be the new cluster centers for their respective clusters. The whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive runs. At this point the cluster centers have stabilized and will remain the same [2]. There are many variants of clustering even for the kmeans algorithm depending upon the method of choosing the initial centers.

4. *Association Rule Mining (ARM):* ARM is one of the most well studied data mining tasks. It discovers relationships among attributes in databases, producing if-then statements concerning attribute-values [10]. An association rule $X \Rightarrow Y$ expresses that in those transactions in the database where X occurs; there is a high probability of having Y as well. X and Y are called respectively the antecedent and consequent of the rule. The strength of such a rule is measured by its support and confidence. The confidence of the rule is the percentage of transactions with X in the database that contain the consequent Y also. The support of the rule is the percentage of transactions in the database that contain both the antecedent and the consequent.

5. Rough Sets Theory (RST): RST was proposed as a new approach to vague concept description from incomplete data. The RST is one of the most useful techniques in many real life applications such as medicine, pharmacology, engineering, banking and market analysis. This theory provides a powerful foundation to reveal and discover important structures in data and to classify complex objects. One of the main advantages of rough set theory is that it does not need any preliminary or additional information about data. Information about rough sets software for data analysis was given in [3].

*6.Genetic Algorithms (GA):*Many classifications models have been proposed in the literature, such as distributed algorithms, restricted search, data reduction algorithms, parallel algorithms, neural networks and decision trees, genetic algorithms[11]. These approaches either cause loss of accuracy or cannot effectively uncover the data structure.GA provide an approach to learning that based loosely on simulated evolution. The GA methodology hinges on a population of potential solutions, and as such exploits the mechanisms of natural selection well known in evolution. Rather than searching from general to specific hypothesis or from simple to complex GA generates successive hypotheses by repeatedly mutating and recombining parts of the best currently known hypotheses. The GA operates by iteratively updating a poll of hypotheses (population). One each iteration, old members of the population are evaluated according a fitness function. A new generation is then generated by probabilistically selecting the fittest individuals form the current population. Some of these selected individuals are carried forward into the next generation population others are used as the bases for creating new offspring individuals by applying genetic operations such as crossover and mutation.

6. Benefits of Data Mining and Knowledge Discovery for Healthcare

From the above discussion, it can be concluded that, based on the healthcare task, both healthcare knowledge engineer (HKE) and the experts can determine the data mining technique which appropriate to this task (Table 1 shows some examples).In addition, a HCO may implement data mining and knowledge discovery methodology (KDM) with the help of a skilled employee who has good understanding of health care industry. And, data mining methodology can be effective at working with large volume of data to determine meaningful pattern and to develop strategic solutions. Moreover, decision makers can learn lessons from the use of KDM in other domains and apply KDM to problems of health care industry (Physicians and Pharmaceutical companies, Hospitals, Insurance companies, etc). Following are some of the important areas where IDM techniques can be of tremendous use in health care management.

a) E-governance structures in health care
b) Health insurance
c) Forecasting treatment costs and demand of resources
d) Anticipating patient’s future behavior given their history
e) Public health informatics
f) Data modeling for health care applications
g) Executive information system for health care

Table 1: Healthcare tasks and the Appropriate Data Mining Techniques

Healthcare Task	The Appropriate Data Mining Technique
Classification	Neural Networks Support Vector Machine Decision Trees Genetic Algorithms Rule induction
Clustering	K-means
Regression and prediction	Support Vector Machine Decision Trees Rule induction, NN
Association and Link Analysis (finding correlation between items in a dataset)	Association Rule Mining
Summarization	Multivariate Visualization

In addition, IDM techniques provide several benefits for the healthcare. These are for example:

a) Storing medical history in a warehouse may be useful for statistical analyses, as well as for discovering patterns and trends – used for diagnosis and reporting purposes;
b) Link analysis – may support accurate diagnosis (by showing links between symptoms and illnesses) and efficient treatment – by revealing links between illnesses and medical drugs;
c) Clustering and classification tasks are useful for statistical purposes.

7. Conclusion

Business intelligence provides an integrated view of data that can be used to monitor key performance indicators, identify hidden patterns in diagnosis, illuminate anomalies in processes, and identify variations in cost factors, all of which facilitate accountability and visibility and can drive an organization towards efficiency. From the technical point of view, healthcare based business intelligence systems are complex to build, maintain and face the knowledge-acquisition difficulty. Efficiency of such systems is determined by the efficiency of the intelligent techniques and methodologies. Intelligent data mining techniques provide an effective computational methods and robust environment for business intelligence in the healthcare decision making domain. Data mining techniques can be seen as enabler for storing, managing, representing, analyzing, visualizing, and giving efficient access to a huge amount of data.

References

[1] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski. "Data Mining Methods for Knowledge Discovery", Kluwer Academic Publishers, 1998.

[2] I. H. Witten and E. Frank, Data Mining – Practical Machine Learning Tools and Techniques, Elsevier, 2005.

[3] A. M. salem, safia A. Mahmoud., "Mining patient Data Based on Rough Set Theory to Determine Thrombosis Disease", Proceedings of First Intelligence conference on Intelligent Computing and Information Systems, pp 291-296. ICICIS 2002, Cairo, Egypt, June 24-26,2002.

[4] Negash S., Business intelligence, Communications of the Association for Information Systems , 13: 177-95, 2004.

[5] Gluchowski P. ,Business Intelligence - Konzepte, Technologien und Einsatzbereiche. HMD - Praxis der Wirtschaftsinformatik , 222: 5-15 (in German), 2001.

[6] Bucher T,and Dinter B. , Process orientation of information logistics - An empirical analysis to assessbenefits, design factors, and realization approaches, IEEE 41th Annual Hawaii Int Conf on System Sciences (HICSS-41) , Waikoloa, Big Island, Hawaii, 2008.

[7] Hammer M, and Champy J. ,Reengineering the corporation - A manifest for business revolution, NewYork: Harper Collins Publishers; 1993.

[8] Vassilacopoulos G,and Paraskevopoulou E., A process model basis for evolving hospital information systems, Journal of Medical Systems, 21 (3): 141-53, 1997.

.[9] C. Cortes and V. Vapnik, "Support vector networks", Machine Learning, vol. 20, pp. 273-297, 1995.

[10] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data, pages 207–216, Washington, D.C., May 1993.

[11] Abdel-Badeeh M.Salem and Abeer M.Mahmoud, "A Hybrid Genetic Algorithm-Decision Tree Classifier", Proc. of the 3rd Int Conf on New Trends in Intelligent Information Processing and Web Mining, Zakopane, Poland, pp. 221-232, June 2-5, 2003.