International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 1, January - 2014

Data Mining Issues and Challenges in Healthcare Domian

Dr. C. Sunil Kumar[1] , Dr. A. Govardhan[2], B. Sunil Srinivas[3]
[1]*Professor in CSE, MREC, Hyderabad, A.P., India*
[2]*Professor of CSE & SIT Director, JNTUH, Hyderabad, A.P, India*
[3]*Associate Professor in CSE, VVIT, Hyderabad, AP, India*

Abstract

Data mining (DM) has become important tool in business and related areas and its task in the healthcare field is still being explored. Currently, most applications of DM in healthcare can be classified into two areas: decision support (DS) for clinical practice, and policy development. However, it is difficult to find experimental literature in this area since a considerable amount of existing work in DM for healthcare is theoretical in nature. In this paper, the challenges that limit the progress made in this area and present considerations for the future of DM in healthcare are reviewed.

1. Introduction

Healthcare associations today are capable of generating and gathering huge amounts of data **[1]**. This raise in volume of data requires usual way for these data to be extracted when needed. With the use of DM methods it is possible to dig out interesting and useful knowledge and regularities. Knowledge acquired in this method, can be used in suitable area to progress work efficiency and improve quality of decision making procedure **[2, 3, 4 and 5]**. In the above stated points there is a great necessitate for new generation of computer theories and tools to help public with extracting useful data from constantly rising volume of digital information. Information technologies (ITs) are being increasingly implemented in healthcare associations in order to respond to the needs of physicians in their daily decision making behaviors. DM tools can be very useful to control restrictions of public such as subjectivity or error due to exhaustion, and to provide indications for the decision-making processes. The fundamental nature of DM is the recognition of relations, models that provide support for forecasting and of decision making process for diagnosis and treatment plan. These models can be called analytical, and they are being integrated into information systems (ISs) of hospitals as models for decision making, dropping the subjectivity and decision making time. In addition, the use of IT in healthcare enables complete management of medical knowledge and its secure exchange among receivers and providers of healthcare examinations. Widespread use of information technology enables the elimination of manual tasks of data extraction from charts or filling of specialized questionnaires, extraction of data directly from electronic records, transfer on secure electronic system of medical records that will save lives and reduce the cost of healthcare, timely detection of infectious diseases with advanced collection of data etc. Retrieval of data with the help of computers can improve the quality of decision making and avoids human or medical errors.

When there is a huge volume of data that needs to be classified, decision making by people is typically poor. DM signifies the process of analyzing raw data with the help of computer and extraction of their semantics. It is frequently defined as discovering previously unknown and potentially useful information from large volume of unstructured data. Thanks to this technique, it is possible to predict trends and customer behavior and thus provide the organization's business success. This is accomplished by data analysis from various perspectives and finding the connections and relations among jointly unconnected data. In the process of DM previously unidentified trends and models from a database (DB) of historical information are being revealed and that information is being converted into considerable business solutions.

2. Application of Data Mining in Healthcare

In modern period many important changes are brought, and ITs have found wide application in the domains of human activities, as well as in the healthcare.

The development and implementation of new ITs that allow global networking; give recent medicine the description of informatical medicine. ITs gradually provide the help in system approach of solving medicinal problems. Disposition of the right data enables the preparation of accurate reports, for e.g., usage of hospital facilities, or number of occupied beds. At the same time it is easier to examine treatment and to check the information exchange. Use of ITs enables change of the healthcare system - how to improve community health, the healthcare of the system users, reduce costs, and save time. It is well known that healthcare is a difficult area where new knowledge is being gathered daily in a growing rate. Big part of this knowledge is in the form of paper work, resulting from an investigations conducted on data and information collected from the patient's medical records. There is a big inclination today to make this information available in E-Form, converting information to knowledge, which is not an easy task.

.

With the increase of costs in healthcare associations and the growing need to control all the expenses, suitable analysis of medical data has become a problem of the utmost importance. All healthcare associations need an expert analysis of their medical information, the scheme that is overwhelming and costly. There is a great potential for DM application in healthcare. Healthcare associations are very oriented on use of patient's data. Ability to use a data in DBs in order to extract useful information for quality health care is a key of success of healthcare institutions. Healthcare ISs contain huge volumes of data that include information on patients, data from laboratories that are continually growing. With the use of DM techniques, useful models of information can be found in this data that will later be used for further research. A very significant issue is how to classify huge volumes of data. Automatic categorization is done based on the similarities that are present in information. This type of categorizations is useful only if the conclusion, that is drawn, is acceptable for the physician or the end user. DM provides support for recognition of reliable relations among treatment and result.

In medical investigation, DM begins with the assumptions and results are accustomed accordingly. This is different from standard DM practice, which simply begins with a set of data without obvious assumptions. While the traditional DM is focused on models and trends in datasets (DSs), DM in healthcare is more focused on minority that is not in accordance with models and trends. The reality that standard DM is more focused on describing and not explaining the patterns and developments is the one thing that deepens the difference among standard and healthcare DM. Healthcare needs these clarifications since the small difference can stand between life and death of a patient. Analytical techniques used in DM, in most cases have long been known mathematical methods and algorithms. Although DM is a little technology, the process of data investigations is nothing new. The thing that linked these methods and large DBs is a cheaper storage space and processing power. Here are some of the techniques of DM, which are successfully used in healthcare, such as artificial neural networks (ANNs), decision trees (DTs), genetic algorithms (GAs) and nearest neighbor method (NNM) **[14]**.

ANNs are analytical methods that are formed on the basis of advanced learning processes in the human brain. As the human brain is capable to, after the learning process, draw hypothesis based on previous inspections, NNs are also capable to forecast changes and events in the system after the process of learning. NNs are groups of connected I/O units where each connection has its own weight. The learning procedure is performed by balancing the net on the basis of relations that exist among elements. Based on the significance of cause and effect between certain data, stronger or weaker connections among neurons are being formed. Network formed in this way is ready for the unknown data and it will react based on formerly acquired knowledge. **ANNs** are ideal for multiprocessor systems, where a huge number of procedures are performed in parallel.

DT is a graphical expression of the relations that exist among the data in the DB. It is used for data categorization. The consequence is displayed as a tree. DTs are mainly used in the classification and forecast. It is a simple and a powerful way of representing knowledge. The patterns obtained from the DT are characterized as a tree structure. The instances are classified by sorting them down the tree from the root

node to some leaf node. The nodes are branching based on if-then condition. Tree view (TV) is a clear and easy to understand, DT algorithms are considerably faster than NNs and their learning is of shorter duration. DT is a tree where each node indicates a test or decision on the item of information that is listed for kindness. The

option of a particular trade depends on the outcome of the test. In order to classify the data, process is starting from the root node and following the argument down until it reaches the final node, at which time a verdict is made. DT can also be inferred as a special form of a rule set, which is distinguished by its hierarchical organization of rules.

GAs is based on the principle of genetic modification, alteration and natural selection. These are algorithmic optimization strategies inspired by the principles observed in natural evolution. The GA creates a number of random solutions to the difficulty. All these solutions may not be good, a group of solutions can be skipped entirely, and it can come down to the overlap ping solutions. Poor results are discarded, and the good ones preserved. A good solution is then being hybridized, and then the whole procedure is repeated. Finally, similar to the process of natural selection, only the best results remain. So, from the set of potential solutions to the problems that contends with each other, the best results are chosen and combined with each other in order to obtain a universal solution from the set of solutions that will become better and better, similar to the process of evolution of organisms. GAs is used in DM to formulate hypotheses about the dependencies among variables in the form of association rules (ARs) or other internal formalism. The difficulty of this method is that it requires an enormous amount of processing power and it is too slow for trivial issues. Since evolutionary computation is a robust and parallel search algorithm, it can be used in data mining to find interesting knowledge in noisy environment **[6, 7 and 8]**.

NNM is a technique that is also used for data categorization. Unlike other techniques, there is no learning process to create a pattern. The data used for learning is in fact a pattern. When the new data shows up, the algorithm analyzes all the data in the DB to find a subset of instances that are the best fit and based on that it is able to forecast the outcome. The study conducted on the application of NNM on benchmark data set (DS) to detect competence in the diagnosis of heart diseases, produced the results that application of this technique achieved an accuracy of 97.5% which is a top percentage than any other published study on the same set of data. DM in healthcare demands close cooperation between managers of quality in healthcare and DM specialists, and it is consisted of analysis driven by data and analysis driven by interest **[9, 10 and 11]**. Type of analysis driven by data is used because analysis driven by interest can predict unanticipated patterns in data. AR rule is usually used within these analyses. This approach that uses both types of analysis has a good and a bad, because the users are not thrilled with a huge number of findings that are way beyond their field of expectations, and then again, unanticipated patterns don't stay ignored.

3. The Obstacles for Data Mining in Healthcare

One of the biggest troubles in DM in medicine is that the raw health data is huge and heterogeneous **[12, 13]**. These data can be assembled from diverse sources such as from conversations with patients, laboratory results and interpretation of doctors. All these components can have a major crash on diagnosis, prognosis and treatment of the patient, and should not be ignored. The scope and difficulty of medical data is one of the obstacles to successful DM. Missing, incorrect, inconsistent or non-standard data such as pieces of information saved in dissimilar formats from different data sources create a major obstacle to successful data mining. It is very difficult for people to process gigabytes of records, although working with images is relatively easy, because physicians are being able to identify patterns, to accept the basic trends in the data, and formulate rational decisions. Stored information becomes less useful if they are not available in easily apprehensible format. The function of visualization techniques is increasing in this, as the picture are easiest for people to understand, and can provide plenty of information in a snapshot of the results.

Physicians' interpretations of images, signals, or other scientific data are written in unstructured language, so it is very hard to perform DM of such data. Even specialists in the same area cannot agree on universal terms that indicate the status of the patient. Not only those different names are being used to describe the same disease, but also tasks are getting even more complicated by using different grammatical structures to explain the relations among medical entities. Also, another barrier is that almost all diagnoses and treatments in medicine are inaccurate and subjected to error rates. Here the analysis of specificity and sensitivity are being used for the measurement of these errors. One of the unique characteristics of medical DM is that the basic data structures in medication are poorly mathematically characterized in comparison with other

areas of physics science, because the conceptual structure of medicine consists of a description in words and pictures, with very few formal limitations in the dictionary, image composition, or permissible relations among the basic concepts.

Within the issue of knowledge integrity assessment, two biggest challenges are: (1) How to expand efficient algorithms for comparing content of two knowledge versions. This challenge demands development of efficient algorithms and data structures for evaluation of knowledge integrity in the DS; and (2) How to enlarge algorithms for evaluating the influence of particular data modifications on statistical importance of individual models that are collected with the help of common classes of DM algorithm. Algorithms that measure the power that modifications of data values have on discovered statistical importance of patterns are being developed, although it would be impossible to develop a universal measure for all DM algorithms.

DM in healthcare can be restricted in data access, since the raw inputs for DM frequently exist in different settings and systems, like administrations, clinics, laboratories etc. Therefore, data must be collected and integrated before DM can take place. Building of DW ore DM begins can be a very costly and time consuming process. Healthcare associations DM must use big investment resources, especially time, effort and money. DM project can fail from various reasons, like lack of managerial support, inadequate DM expertise etc.

4. Future Issues

At present, DM is still considered to be in its early years. As new applications continue to emerge, certain issues for the DM in the healthcare field will need to be considered:

4.1 Improved Data Sharing Among Agencies – Several institutions such are overcoming privacy problems that limit information sharing by blocking out significant patient identification information such as Social Security Numbers (SSNs). For example, HCFA has established a data availability link on their web site and support data exchange for research purposes. The challenge will be to overcome propriety conditions imposed by private institutions. Researchers may want to develop contractual relationships with such institutions which may limit the publication of explicit findings but will provide the opportunity to work with real data instances. Finally, researchers will increasingly find public information becoming available on the web.

4.2 Integrated Web Mining Tools - Text Mining (TM) has recently come into focus for mining on the web. However, the web includes a large amount of non-text based data that may need to be considered in the future, especially as online telemedicine begins to prosper. Another feature that is gaining important reputation is the automation of billing transactions via the Internet. This change will provide a great opportunity to use DM techniques to detect fraudulent online transactions.

4.3 DW Standardization- While uniform DW standards may take a while to appear, there needs to be a bridge to help facilitate mining of data from various sources. The development of inter-agency, flexible standards may mitigate the need for extensive cleaning tools.

4.4 DW Compression- As DWs continue to grow, the difficulties for mining a massive DS will continue. A process of compression without compromising data quality would diminish some of those issues. Scaling a wider range of accessible tools for large DSs is met with a number of difficulties including idea, data and computational difficulty, and storage requirements. This is particularly a worry if mining is going to continue to be a driving force behind desktop decision support.

4.5 Focus on representation and interpretation of findings: Early, formative research on DM and more broadly Knowledge Discovery (KD) in Databases addressed the need to conceptualize the discovery of new facts in databases as a process with DM being one important component of this. In most literature on DM focuses on the application of DM techniques for pattern extraction and reduces the significance of pattern interpretation. This difficulty becomes overstated due to lack of adequate empirical studies. Consequently, researchers have not adequately examined issues that relate to presentation and interpretation of these patterns. DM can potentially benefit from research in the area of data visualization but in healthcare that addresses data visualization and pattern interpretation techniques are yet to find a research study. In fact to understand the nature of outcomes from DM techniques, it is important to think of DM as one component in an

overall decision support environment that integrates data cleaning, data visualization, and interpretation.

5. Conclusion

Empirical research in DM for healthcare is limited. Several factors, most importantly those related to data quality and availability, have limited research in this locale. In our evaluation, addressing these two issues will give significant drive to research in this arena. In particular, healthcare institutions and governing bodies need to establish strong data quality standards before the environment can be conducive to productive research. Secondly, a positive partnership must be established among institutions maintaining DWs.

6. References

[1] Allen, H.G. "The Healthcare Data Warehouse", Data Management Review, (7:3), 1997, pp. 50-51.

[2] Brosset, S. E., Sprague, A. P., Hardin, J. M., Waites, K. B., Jones, W. T. and Moset, S. A. "Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance", Journal of the American Medical Informatics Association, (5:4),1998, pp. 373-381.

[3] Chen, L.d and Sakaguchi, T. "Data Mining: Information Storage and Retrieval Systems", Information Systems Management, (17:1),2000, pp. 65-71.

[4] Chung, H. M. and Gray, P. "Special Section: Data Mining", Journal of Management Information Systems, (16:1), 1999, pp. 11-17.

[5] Degoulet, P. and Fieschi, M. Introduction to Clinical Informatics, Springer, New York,1997. [6] Fayyad, U., Haussler, D. and Stolorz, P. "Mining Scientific Data", Communications of the ACM, (39:11),1996, pp. 55-60

[7] Goodall, C. R. "Data Mining of Massive Datasets in Healthcare", Journal of Computational & Graphical Statistics, (8:3), 1999, pp. 620-635.

[8] Lavrac, N. "Selected Techniques for Data Mining in Medicine", Artificial Intelligence in Medicine Journal, (16:1), 1999, pp. 3-23.

[9] Moser S, T., Warren, J. T. and Brossette, S. "Application of Data-mining to Intensive Care Unit Microbiologic Data", Emerging Infectious Diseases, (5:3),1999, pp. 454- 459.

[10] Canlas, R. D. (2009). "Data Mining in Healthcare: Current Applications and Issues", Carnegie Mellon University, Australia.

[11] Gupta, S., Kumar, D., & Sharma, A. (2011). "Data Mining Classification Techniques Applied for Breast Cancer Diagnosis And Prognosis", Indian Journal of Computer Science and Engineering (IJCSE) 188-195.

[12] Yang, Q., & Wu, X. (2006). "10 Challenging Problems in Data Mining Research", International Journal of Information Technology & Decision Making Vol. 5, No. 4, 597–604.

[13] Cios, K. J., & Moore, G. W. (2002), "Uniqueness of Medical Data Mining", To appear in Artificial Intelligence in Medicine Journal.

[14] Shouman, M., Turner, T., & Stocker, R. (2012). "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients". 2012 International Conference on Knowledge Discovery (ICKD 2012) IPCSIT Vol. XX. Singapore: IACSIT Press.