www.ioirp.com
International Journal of Innovative Research in Computer Science and Engineering (IJIRCSE)
ISSN: 2394-6364, Volume – 3, Issue – 2. April 2018

DATA MINING TASKS AND THEIR LIFE CYCLES USING KDD PROCESS

1. Mr.S.Vetrivel , Asst Professor

2. Mrs.P.Vidhya Devi, Asst Professor

Department of Information Technology, AJK College of Arts and Science, Coimbatore

Abstract

Data mining is the process of releasing concealed information from a large set of database and it can help researchers gain both narrative and deep insights of exceptional understanding of large biomedical datasets. Data mining can exhibit new biomedical and healthcare knowledge for clinical decision making. Medical assessment is very important but complicated problem that should be performed efficiently and accurately. The goal of this paper is to discuss the research contributions of data mining to solve the complex problem of Medical diagnosis prediction. This paper also reviews the various techniques along with their pros and cons. Among various data mining techniques, evaluation of classification is widely adopted for supporting medical diagnostic decisions.

Keywords: Data Mining, Tasks, Life Cycles, Knowledge Discovery Process

1. Introduction

To generate information it requires massive collection of data. The data can be simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. To take complete advantage of data; the data retrieval is simply not enough, it requires a tool for automatisummarization of data, extraction of the essenceinformation stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all above is 'Data Mining'. Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential help organizations focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were too time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as:

1. **Exploratory Data Analysis**: It is simply exploring the data without any clear ideas of what we are looking for. These techniques are interactive and visual.

2. **Descriptive Modeling**: It describe all the data, It includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

3. **Predictive Modeling: T**his model permits the value of one variable to be predicted from the known values of other variables.

4. **Discovering Patterns and Rules:** It concern with pattern detection, the aim is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest.

5. Retrieval by Content: It is finding pattern similar to the pattern of interest in the data set. This task is most commonly used for text and image data sets.

2. Literature Survey

Author	**Year**	Heart Disease Prediction	**Knowledge resource**	**DM techniques / applications**
Jyoti Soni et.al	2011		Decision tree outperforms and sometimes Bayesian classification 's having similar accuracy as Decision tree.	Classification: Clustering Bayesian classification, Neural Networks Decision Tree, KNN
Samar AlQarzaie et.a	2011	Breast Cancer Disease	WEKA tool is used to give 93.4675% accuracy in testing set and in the training set it yields 96.8% accuracy	Classification: Decision Tree
Arvind Sharma et.al	2012	Blood Donors	By using WEKA tool, J48 decision tree acquires 89.99% accuracy	Classification: J48 Decision Tree
Shweta Kharya	2012	breast cancer	Decision tree is	Classification: Neural

		diagnosis and prognosis	a best predictor with 93.62% accuracy	Network, Association, Naive.Bayes, C4.5 decision tree algorithm
Dr. Bushra M. Hussan	2012	Prediction of medical data by K means Clustering	On changing the instances it shows 97% of accuracy.	Classification: K-means, Clustering

Table 1: literature Review

3. Data Mining Tasks

The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as:

Exploratory Data Analysis: It is simply exploring the data without any clear ideas of what we are looking for. These techniques are interactive and visual.

Descriptive Modeling: It describe all the data, It includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

Predictive Modeling: This model permits the value of one variable to be predicted from the known values of other variables.

Discovering Patterns and Rules: It concern with pattern detection, the aim is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest.

Retrieval by Content:

It is finding pattern similar to the pattern of interest in the data set. This task is most commonly used for text and image data sets.

4. Data Mining Life Cycle:

The life cycle of a data mining project consists of six phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The main phases are:

1. Business Understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

2 Data Understanding: It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

3. Data Preparation: It covers all activities to construct the final dataset from the initial raw data.

4. Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

5. Evaluation: In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be

certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

6. Deployment: The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

5. The Knowledge Discovery Process:

Data mining is one of the tasks in the process of knowledge discovery from the database. The steps in the KDD process contains:

1. Data cleaning: It is also known as data cleansing; in this phase noise data and irrelevant data are removed from the collection.

2. Data integration: In this stage, multiple data sources, often heterogeneous, are combined in a common source.

3. Data selection: The data relevant to the analysis is decided on and retrieved from the data collection.

4. Data transformation: It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure.

5. Data mining: It is the crucial step in which clever techniques are applied to extract potentially useful patterns.

6. Pattern evaluation: In this step, interesting patterns representing knowledge are identified based on given measures.

7. Knowledge representation: It is the final phase in which the discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

Conclusion

Most of the previous studies on data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. Thus, for every domain the domain expert's assistant is mandatory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. The domain experts are required to determine the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generation.

References:

[1] Lavrac, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M. &Kobler, A. (2007).Data mining and visualization for decision support and modeling of public health-care

resources.Journal of Biomedical Informatics, 40, 438-447. doi:10.1016/j.jbi.2006.10.003

[2] Li, X., Zhu, Z. & Pan, X. (2010). Knowledge cultivating for intelligent decision making in small & middle businesses.Procedia Computer Science, 1(1), 2479-2488. doi:10.1016/j.procs.2010.04.280

[3] Li, Y., Kramer, M.R., Beulens, A.J.M., Van Der Vorst, J.G.A.J. (2010). A framework for early warning and proactive control systems in food supply chain networks. Computers in Industry, 61, 852–862. Doi:101.016/j.compind.2010.07.010

[4] Liao, S.H., Chen, C.M., Wu, C.H. (2008). Mining customer knowledge for product line and brand extension in retailing.Expert Systems with Applications, 34(3), 1763-1776. doi:10.1016/j.eswa.2007.01.036

[5] Liao, S. (2003). Knowledge management technologies and applications-literature review from 1995 to 2002. Expert Systems with Applications, 25, 155-164. doi:10.1016/S0957-4174(03)00043-5

[6] Liu, D.R. & Lai, C.H. (2011). Mining group-based knowledge flows for sharing task knowledge. Decision Support Systems,50(2), 370-386. doi:10.1016/j.dss.2010.09.004

[7] Lee, M.R. & Chen, T.T. (2011). Revealing research themes and trends in knowledge management: From 1995 to 2010. Knowledge-Based Systems.doi:10.1016/j.knosys.2011.11.016

[8] McInerney, C.R. & Koenig, M.E. (2011). Knowledge Management (KM) Processes in Organizations: Theoretical Foundations and Practice. USA: Morgan & Claypool Publishers. doi: 10.2200 /S00323 ED1V01Y201012ICR018

[9] McInerney, C. (2002). Knowledge Management and the Dynamic Nature of Knowledge.Journal of the American Society for Information Science and Technology, 53(12), 1009-1018. doi:10.1002/asi.10109

[10] Ngai, E., Xiu, L. &Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36, 2592- 2602. doi:10.1016/j.eswa.2008.02.021