ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 1, 2025

Data Mining Techniques and Applications

*V Chandraprakash[1], Dr. P V Kumar[2], S Srinivas[3], K Anjaneulu[4]

[1] Assistant Professor, St. Martin's Engineering College, Secunderabad, Telangana – 500100

[2]Professor, Department of CSE, Osmania Univesrsity - HyderabadTelangana – 500007

[3,4]Assistant Professor, SICET Engineering College - Hyderabad Telangana – 501510

*Corresponding Author

Email: vchandraprakashit@smec.ac.in

ABSTRACT:

In the Information Technology era information plays vital role in every sphere of the human life. It is very important to gather data from different data sources, store and maintain the data, generate information, generate knowledge and disseminate data, information and knowledge to every stakeholder. Due to vastuse of computers and electronics devices and tremendous growth in computing power and storage capacity, there is explosive growth in data collection. The storing of the data in data warehouse enables entire enterprise to access a reliable current database. To analyze this vast amount of data and drawing fruitful conclusions and inferences it needs the special tools called data mining tools. This paper gives overview of the data mining systems and some of its applications.

Keywords:

Data mining system architecture , Data mining application

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 automatic summarization of data, extraction of the essence of information 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 to help organizations focus on the most important information in their data warehouses [1,2,3,4]. Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions[2].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.

Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases[3,5]. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process[1,3,5].

2 THE 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[1,2]:

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: This 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.

3. TYPES OF DATA MINING SYSTEMS:

Data mining systems can be categorized according to various criteria the classification is as follows[3]:

- Classification of data mining systems according to the type of data source mined: This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

- Classification of data mining systems according to the data model: This classification based on the data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc.

- Classification of data mining systems according to the kind of knowledge discovered: This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

- Classification of data mining systems according to mining techniques used: This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.

The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

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.

2Data 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 DATA MINING MODELS:

The data mining models are of two types[1,2,6,45]:Predictive and Descriptive.

The predictive model makes prediction about unknown data values by using the known values. Ex. Classification, Regression, Time series analysis, Prediction etc.

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc.

Many of the data mining applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of itsfuturestate.Classificationisatechniqueofmappingthetargetdatatothepredefinedgroupsor classes, this is a supervise learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that map data item to real valued prediction variable. In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone. It is also referred to as unsupervised learning or segmentation. It is the partitioning or segmentation of the data in to groups or clusters. The clusters are defined by studying the behavior of the data by the domain experts. The term segmentation is used in very specific context; it is a process of partitioning of database into disjoint grouping of similar tuples. Summarization is the technique of presenting the summarize information from the data. The association rule finds the association between the different attributes. Association rule mining is a two-step process: Finding all frequent item sets, Generating strong association rules from the frequent item sets. Sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend.

6. 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:[1,3]

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.

7. DATA MINING METHODS:

The data mining methods are broadly categories as: On-Line Analytical Processing (OLAP), Classification, Clustering, Association Rule Mining, Temporal Data Mining, Time Series Analysis, Spatial Mining,Web Mining etc.These methods use different

Types of algorithms and data.The data source can be data warehouse, database, flat file or text file. The algorithms may be Statistical Algorithms, Decision Tree based, Nearest Neighbor, Neural Network based, Genetic Algorithms based, Ruled based, Support Vector Machine etc. The selection of data mining algorithm is mainly depends on thetype of data used for mining and the expected outcome of the mining process. The domain experts play a significant role in the selection of algorithm for data mining.

A knowledge discovery (KD) process involves preprocessing data, choosing a data- mining algorithm, and post processing the mining results. There are very many choices for each of these stages, and non-trivial interactions between them. Therefore both novices and data-mining specialists need assistance in knowledge discovery processes.

The Intelligent Discovery Assistants [7] (IDA), helps users in applying valid knowledge discovery processes. The IDA can provide users with three benefits:

1. A system at ice numeration of valid knowledged is covery processes;
2. Effectiverankingsofvalidprocessesbydifferentcriteria,whichhelptochoose between the options;
3. An infrastructure for sharing knowledge, which lead stonet work externalities.

Several other attempts have been made to automate this process and design of a generalized data mining tool that posses intelligence to select the data and data mining algorithms and up to some extent the knowledge discovery.

8. 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]Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, *Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.*

[2] Larose,D.T.,"DiscoveringKnowledgeinData:AnIntroductiontoDataMining",*ISBN0-471-66657-2, John Wiley & Sons, Inc, 2005.*

[3] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", *Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.*

[4].Chapman, P.Clinton,J., Kerber, R.,Khabaza, T.,Reinartz,T., Shearer,C.andWirth,R.."CRISP-DM 1.0 : Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (The Netherlands), 2000".

[5]. Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," *AI Magazine, American Association for Artificial Intelligence, 1996.*

[6]. Tan Pang-Ning, Steinbach, M., Vipin Kumar. "Introduction to Data Mining", *Pearson Education, New Delhi, ISBN: 978-81-317-1472-0, 3rd Edition, 2009.*

[7]. Bernstein, A. and Provost, F., "An Intelligent Assistant for the Knowledge Discovery Process",

WorkingPaperoftheCenterforDigitalEconomyResearch,NewYorkUniversityandalsopresentedatthe IJCAI 2001 Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases.

[8]. Baazaoui, Z., H., Faiz, S., and Ben Ghezala, H., "A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data, volume 5, ISSN 1307-6884," *Proceedings of World Academy of Science, Engineering and Technology, April 2005.*

[9]. Rantzau, R. and Schwarz, H., "A Multi-Tier Architecture for High-Performance Data Mining**,** A Technical Project Report of ESPRIT project, The consortium of CRITIKAL project, Attar Software Ltd. (UK), Gehe AG (Denmark); Lloyds TSB Group (UK), Parallel Applications Centre, University of Southampton (UK), BWI, University of Stuttgart (Denmark), IPVR, University of Stuttgart (Denmark)".

[10]. Botia, J. A., Garijo, M. y Velasco, J. R., Skarmeta, A. F., "A Generic Data mining System basicdesign and implementation guidelines*", A Technical Project Report ofCYCYT project of Spanish Government. 1998. Web Site: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.1935*

[11]Campos,M.M.,Stengard,P.J.,Boriana,L.M.,"Data-CentricAutomatedDataMining",,WebSite.: www.oracle.com/technology/products/bi/odm/pdf/automated_data_mining_paper_1205.pdf

[12].Sirgo, J., Lopez, A., Janez, R., Blanco, R., Abajo, N., Tarrio, M., Perez, R., "A Data Mining Engine based on Internet, Emerging Technologies and Factory Automation," *Proceedings ETFA '03, IEEE Conference, 16-19 Sept. 2003. Web Site* :www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.8955

[13]BiancaV.D.,PhilippeBouladeMareüiland MartineAdda-Decker,"Identificationofforeign-accented French using data mining techniques, Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI)". Web Site *:www.limsi.fr/Individu/bianca/article/Vieru&Boula&Madda_ParaLing07.pdf*

[14]Halteren, H. van, "Linguistic Profiling for Author Recognition and Verification", *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics USA**,** Barcelona, Spain, Article No. 199, Year of Publication: 2004.*

[15]. Halteren, H. V., Oostdijk N., "Linguistic profiling of texts for the purpose of language verification, The ILK research group, Tilburg centre for Creative Computing and the Department of Communication and Information Sciences of the Faculty of Humanities, Tilburg University, The Netherlands." WebSite: *www.ilk.uvt.nl/~antalb/textmining/LingProfColingDef.pdf*

[16]. Antonie, M. L., Zaiane, O. R.,Coman, A., "Application of Data Mining Techniques for Medical Image Classification", *Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD 2001) in conjunction with ACM SIGKDD conference, San Francisco, August 26, 2001.*

[17].Kusiak, A., Kernstine, K.H., Kern, J.A., McLaughlin, K.A., and Tseng, T.L., "Data Mining: Medical and Engineering Case Studies". *Proceedings of the Industrial Engineering Research 2000 Conference, Cleveland, Ohio, pp. 1-7,May 21-23, 2000.*

[18].Luis, R., Redol, J., Simoes, D., Horta, N., "Data Warehousing and Data Mining System Applied to E-Learning**,** Proceedings of the II International Conference on Multimedia and Information &Communication Technologies in Education, Badajoz, Spain, December 3-6th 2003.

[19].Chen, H., Chung, W., Qin, Y., Chau, M., Xu, J. J., Wang, G., Zheng, R., Atabakhsh, H., "Crime Data Mining: An Overview and Case Studies", *A project under NSF Digital Government Programme, USA, "COPLINK Center: Information and Knowledge Management for Law Enforcement,", July 2000 – June 2003*.

[20].Kay, J., Maisonneuve, N., Yacef, K., Zaiane O., "Mining patterns of events in students' teamwork data", *Proceedings of the ITS (Intelligent Tutoring Systems) 2006 Workshop on Educational Data Mining*, pages 45-52, Jhongli, Taiwan, 2006.

[21] Rao, R. B., Krishnan, S. and Niculescu, R. S., "Data Mining for Improved Cardiac Care" , *SIGKDD Explorations Volume 8, Issue 1.*

[22] Ghani,R.,Probst,K.,Liu,Y.,Krema,M.,Fano,A.,"TextMiningforProductAttributeExtraction**",** *SIGKDDExplorationsVolume8,Issue1.*

[23] DeBarr, D., Eyler-Walker, Z., "Closing the Gap: Automated Screening of Tax Returns to Identify Egregious Tax Shelters". *SIGKDD Explorations Volume 8, Issue 1.*

[24] Kanellopoulos, Y., Dimopulos, T., Tjortjis, C., Makris, C. "Mining Source Code Elements for Comprehending Object-Oriented Systems and Evaluating Their Maintainability", *SIGKDD Explorations Volume 8, Issue 1*.

[25] Schultz, M. G., Eskin, Eleazar, Zadok, Erez, and Stolfo, Salvatore, J., "Data Mining Methods for Detection of New Malicious Executables". *Proceedings of the 2001 IEEE Symposium on Security and Privacy, IEEE Computer SocietyWashington, DC, USA , ISSN:1081-6011, 2001.*

[26] Cai,W.andLiL.,"AnomalyDetectionusingTCPHeaderInformation,STAT753ClassProject Paper, May 2004.". Web Site:*http://www.scs.gmu.edu/~wcai/stat753/stat753report.pdf.*

[27].Nandi, T., Rao, C. B. and Ramchandran, S., "Comparative genomics using data mining tools, Journal of Bio-Science, Indian Academy of Sciences, Vol. 27,No. 1, Suppl. 1, page No. 15-25, February 2002".

[28] Khreisat, L., "Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study". *Proceedings of The 2006 International Conference on Data Mining, DMIN'06, pp 78-82, Las Vegas, Nevada, USA, June 26-29, 2006.*

[29] Onkamo,P.andToivonen,H.,"Asurveyofdataminingmethodsforlinkagedis-equilibrium mapping", *Henry Stewart Publications 1473 – 9542. Human Genomics. VOL 2, NO 5, Page No. 336–340, MARCH2006.*

[30] Smith,L.,Lipscomb,B.,andSimkins,A.,"DataMininginSports:PredictingCyYoungAward Winners". *Journal of Computer Science, Vol. 22, Page No. 115-121,April 2007.*

[31] Deng, B., Liu, X., "Data Mining in Quality Improvement". *Proceedings of the Twenty-Seventh Annual SAS UsersGroup InternationalConference2002bySASInstituteInc.,Cary,NC,USA.ISBN1-59047-061-3.WebSite:http://www2.sas.com/proceedings/sugi27/Proceed27.pdf*

[32] Cohen, J. J., Olivia, C., Rud, P., "Data Mining of Market Knowledge in The Pharmaceutical Industry". *Proceeding of 13th Annual Conference ofNorth-East SAS Users Group Inc., NESUG2000, Philadelphia Pennsilvania, September 24-26 2000.*

[33] Elovici, Y., Kandel, A., Last, M., Shapira, B., Zaafrany, O., "Using Data Mining Techniques for DetectingTerror-RelatedActivitiesonthe Web". Web Site: *www.ise.bgu.ac.il/faculty/mlast/papers/JIW_Paper.pdf*

[34] Solieman, O. K., "Data Mining in Sports: A Research Overview, A Technical Report, MIS MastersProject, August 2006". Web Site:*http://ai.arizona.edu/hchen/chencourse/Osama-DM_in_Sports.pdf*

[35] Maciag, T., Hepting, D. H., Slezak, D., Hilderman, R. J., "Mining Associations for Interface Design". *Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 4481, pp. 109-117, June 26, 2007.*

[36] Foster, D. P. and Stine, R. A., "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy".*JournaloftheAmericanStatisticalAssociation,Alexandria,VA,ETATS-UNIS,vol.99,* ***ISSN****0162-1459,pp.303-313January15,2004*

[37] Kraft, M. R., Desouza, K. C., Androwich, I., "Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population". *IEEE, Proceedings of the 36thHawaii International Conference on System Sciences, 0-7695-1874-5/03, 2002.*

[38] Kusiak, A.,Kernstine,K. H.,Kern,J.A.,McLaughlin, K.A.,and Tseng,T.L.,"DataMining:Medical and Engineering Case Studies". *Proceedings of the Industrial Engineering Research 2000 Conference, Cleveland, Ohio, pp. 1-7,May 21-23, 2000.*

[39] Ansari, S., Kohavi, R., Mason, L., and Zheng, Z., "Integrating E-Commerce and Data Mining: Architecture and Challenges". *Proceedings of IEEE International Conference on Data Mining, 2001*.

[40] Jadhav, S. R., and Kumbargoudar, P., "Multimedia Data Mining in Digital Libraries: Standards and Features".*Proceedings of conference Recent advances in Information Science and Technology READIT – 2007, pp 54-59, Organized by Madras Library Association - Kalpakkam Chapter & Scintific information Resource Division, Indira Gandhi Center for Atomic research, Department of Atomic Energy, Kalpakkam, Tamilnadu,India. 12-13 July 2007.*

[41] Anjewierden, A.,Koll¨offel,B.,and HulshofC.,"Towardseducationaldata mining: Usingdata mining

methods for automated chat analysis to understand and support inquiry learning processes". *International Workshop on Applying Data Mining in e-Learning, ADML'07, Vol-305, Page No 23-32, Sissi, Lassithi - Crete Greece, 18 September, 2007.*

[42] Romero, C., Ventura, S. and De-Bra, P. "Knowledge Discovery with Genetic Programming for Providing Feedback to Courseware Authors, Kluwer Academic Publishers, Printed in the Netherlands, 30/08/2004".

[43] Chen, H., Chung, W., Xu Jennifer, J., Wang, G., Qin, Y., Chau, M., "Crime Data Mining: A General Framework and Some Examples". *Technical Report, Published by the IEEE Computer Society, 0018-9162/04, pp 50-56, April 2004.*

[44] Chodavarapu Y., "Using data-mining for effective (optimal) sports squad selections". *Web Site: http://insightory.com/view/74//using_data-mining_for_effective_(optimal)_sports_squad_selections*

[45] Jensen, Christian, S., "Introduction to Temporal Database Research," Web site: http://www.cs.aau.dk/~csj/Thesis/pdf/chapter1.pdf

[46] Vajirkar, P., Singh, S., and Lee, Y., "Context-Aware Data Mining Framework for Wireless Medical Application*". Lecture Notes in Computer Science (LNCS), Volume 2736, Springer-Verlag. ISBN 3-540-40806-1, pp. 381 – 391.*

*****