IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

A Comprehensive Survey of Dynamic Data Mining Process in Knowledge Discovery from Database

D. Ramana Kumar[1*], S. Krishna Mohan Rao[2]

[1]Dept. of Computer Science and Engineering, Jawaharlal Nehru Technological University Hyderabad, Hyderabad, Telengana State, India
[2]Principal, Gandhi Institute for Technology, Bhubaneswar, Oddisha, India

[*]*Corresponding Author: ramanad74@gmail.com, +91-9989633282*

Available online at: www.ijcseonline.org

Accepted: 16/Dec/2018, Published: 31/Dec/2018

Abstract : Data mining and knowledge discovery in databases have been considered as an important research area in education and industry. This survey presents an overview, a description and future direction which denotes a standard for knowledge discovery using dynamic data mining process model. The paper mentions particular real-world applications, data mining techniques, challenges incorporated in real-world application of knowledge discovery, current and future research concepts in the field. The applications to both academic and industrial concerns are discussed. The major target of the survey is the integration of the research in this particular area and thereby assisting in improving the existing model by using dynamic data mining. The bonding between the knowledge discovery and dynamic data mining in real world is reviewed with appropriate examples. The survey critically evaluates the area of knowledge discovery database to inform users about various models and to develop various models using dynamic data mining. The knowledge discovery database management standards will help in promoting the industry growth and pushing the industry beyond the edge.

Keywords—Knowledge Discovery Database, Data Mining, Dynamic Data Mining Real World Application

I. INTRODUCTION

Rapid growth in the huge database has lead to data mining in various sectors of business like industry, administration and technological applications. It is important for business decisions, information management and query processing tasks. Data mining help to discover patterns and links that can help spot key opportunities in the market, sales trend, developing smart marketing campaigns and trend prediction. Data mining methods involves the intersection of statistics, machine learning and database systems like MySQL, Microsoft access, SQL server etc. Among the tools that are used in the data mining process, R software is built with all the possible community support and libraries. Data mining tasks include summarization, classification, clustering and trend analysis. They are important data mining techniques helpful in the creation of business value. The term 'Knowledge discovery in databases', also referred as KDD .It is coined by Gregory Piatetsky-Shapiro in the year 1989 is a widely used process in search of knowledge from the existing data. KDD is referred to the high level application of certain data mining methods in terms of machine learning, statistics, machine learning, artificial intelligence and data visualization. There also available many theoretical and philosophical considerations for KDD techniques [1,3]

Purpose of KDD is a multi-step process of extraction useful information from huge database by using data mining algorithms. The process of KDD involves data preparation, identification of pattern and knowledge extraction. It also involves the process of understanding of application domain, purpose of the end-user, creation of target data set like data samples, data cleaning, formatting and pre-processing. Thus, KDD is also considered as one of the top research field for both machine learning and database researchers [1, 3]. However, huge challenges are also rising on the path of KDD. Problem in the KDD process involves with mapping low level data into other forms of data [2]. One of the prime issues faced by the researchers in the path of KDD is data security. Malicious activities are increasing on a simultaneous scale within the system. Some other form of challenges is found in the form traditional data algorithms that are designed to identify only well-defined objects. As result, when there are varieties of data that are often found in the undefined category like swirling bodies of water and ocean nutrients that appear as changes in the sea data. Another challenge in the KDD process is prediction issues in the presence of missing values of data. Near about 5-15% of missing data is not often easy to manage in the prediction scenario, whenever interpretation becomes very difficult. The

application challenges faced by working in KDD applications are summarized as:

1. Mining methodology: Certain issues are prevalent while carrying out the data mining tasks. They are in the form of data uncertainty, unwanted data or noise and incomplete data.

2. Feature Engineering: The most important task of data mining is to track a user's prediction criteria or background and visualization of data mining results. The mining environment should be highly user interactive and exploratory. With the help of interactive form of mining, users must be allowed to dynamically change the focus of the search in order to refine the mining query on the basis of final results.

Adhoc data mining: Similar to query languages like SQL, high-level data mining query languages must be provided with freedom to define adhoc data mining taks.Relevant sets of data for the purpose of analysis must be provided with specification, domain knowledge, types of knowledge that is to be mined, constraints to be specified on the patterns discovered. There must be flexibility in handling mining requests and interactive in terms of users.

3. Efficiency and scalability: Data mining algorithms must be developed in a easier manner enhancing efficiency, performance, scalability and optimization.

4. Database diversity: There are available wide variety of data sets like structured data, semi-structured-data and unstructured data, stable set of data, dynamic data, sensor data, spatial data, multimedia data, social network data, web data, software program code. Therefore the data mining tool must be diverse in application to handle all sorts of data according to their nature of data.

5. Data mining and society: There are certain social impacts of data mining. Misuse and personal data violation are strictly avoided in order to provide personal data security. Invisible data mining is one such area where users are able to get recommendation based on their pattern of search. But, these systems must improve their functionality so that users can easily perform data mining without prior knowledge of data mining algorithms [4,5].Real world data are collected over a certain range of period ranging from seconds to years. Thus the traditional data may not be able to get fit into the data change over the time, leading to degradation in the process of prediction. The old form of data may not often fulfil the present data analysis criteria or the requirement. There comes the role of dynamic data mining process. In this process, wide variety of data can be extracted from a set of large dynamic data by consideration of validity of date and real time [5].

The rest of the paper is organized as follows, Section I contains the introduction of knowledge discovery databases, Section II contain the related work elaborating on the need of KDD, Section III contain details pertaining to data mining and knowledge discovery in the real world, Section IV contain the process models of KDD, Section V explain the data mining procedures and models of KDD process, Section VI describes the comparative analysis of different techniques proposed in the literature, and Section VII concludes the research work with future directions.

II. NEED FOR KNOWLEDGE DISCOVERY DATABASE (KDD)

The conventional technique of transforming data into knowledge is concerned on manual evolution and the interpretation. For instance, the specialists assists in evaluating the current trends and modifications happening in the health care data providing the report detailing the analysis to the health organization in health-care field. The report attained becomes the fundamental for the decision making in future and planning for the management of the health care. For several utilization, this type of manual assessment of an informational index is moderate, costly, and emotional. In generally, as databases volumes implement rapidly changes, this sort of manual information development is getting to be entirely unrealistic in different areas. Databases are increasing in size in two ways such as (1) the number N of records in the database and (2) the number d of fields or parameters to an object. Databases containing on the request of N = 109 articles are ending up progressively normal, for example, in the galactic sciences. Therefore, the quantity of fields without. Much of a stretch be on the request of 102 or indeed, even 103, for example, in restorative symptomatic utilizations. We trust that this activity is absolutely not one for people; then, development work should be robotized. The need to scale up human assessment capacities to dealing with the huge number of bytes that we can gather is both monetary and logical. Organizations uses database to increase focused benefit, increment proficiency, and give high profitable administrations to clients. Information we catch about our condition are the important proof we use to manufacture hypotheses, models of the universe we live in. Since PCs have permitted people to store a huge number of databases than we can process, it is as it were characteristic to swing to computational systems to enable us to uncover significant instance structures from the Gigantic volumes of databases. Therefore, KDD is an attempt to solve an issue that the computerized data information made an unavoidable truth for every one of us.

III. DATA MINING AND KNOWLEDGE DISCOVERY IN THE REAL WORLD

KDD applications are widely used and are useful in the operational deployment of the system in solving real-world problems in the file of science and business. Field of astronomy is the primary user of KDD and its best example is SKICAT. This is one of the data systems in order to perform image analysis by the astronomers.This system has processed near about 3 terabytes of sky data that showed detection of 109 objects and performed their classification by outperforming humans and the existing techniques. KDD is an important process in business especially to handle investment, finance, marketing, Telecommunication, manufacturing and internet agents.

A. Marketing

In order to analyze the customer database and categories the customer groups in terms of their behavioural search, the database marketing systems are used as the primary application. In the business week it was revealed that majority of the retailers are maximizing their outcomes with the help of database marketing systems [6]. One more interesting system that is available for the retailers is the market-based analysis systems to track the customer behavioural pattern and increase their performance in the market [7].

B. Investment

In several companies, the data mining can be utilized for investment, but do not define the systems. One instance assumed is the LBS capital management. This utilizes the systems experts, neural networks and few genetic algorithms to organize the portfolios. This system has functioned better in the broad stock market [8].

C. Manufacturing

The model called as CASSIOPEE which was implemented as a section of the combine task between the general electric and SNECMA was utilized by the major European airlines to diagnose and estimate the problems for the Boeing 737.Clustering techniques were used to eliminate the faults.

D. Fault detection

The FAIS model [9] in the treasury of US financial crimes network implementation is utilized to predict the financial transactions which represent money laundering performance

IV. KDDM PROCESS MODELS

A KDDM model consists of a set of process which needs to be followed by practitioners while implementing the KDMM projects. These models are defined in secured tasks. It ranges from the errand of understanding the venture space and information, through information arrangement and investigation, to assessment, understanding, and use of the created outcome. All proposed models likewise accentuate the iterative concept of the model as far as number of feedback loops and repetitions, which are activated by a modification procedure. The development of the standard KDDM model was started several years ago and consists of nine steps and research has aids on developing the new models rather than improving the design of the single model. In the nine-step model, the first model includes the academic research features which also has several important business concerns. The various features present in all of the KDDM models consist complex and time-consuming data preparation tasks [10]. The nine steps involves the implementing the application area, developing the target data, data cleaning and pre-processing, data reduction and projection, selecting the data manipulation task and interpreting the patterns with the discovery knowledge.

A. Applications and impact of KDDM models

The applications used by various algorithms differ and are simpler than the industrial applications. The nine-step model is presented by an industrial DM software system called as Mind set and has been employed in number of KDDM projects. The various models have several applications and also involve assessment the data from the retailer. The industrial project concerning customer cross sales and a research project concerning evaluation of marketing internet data have been analysed [11]. The CRISP-DM model has determined few research projects such as performance assessment of heating, ventilation and HVAC systems and evaluation of the thrombosis data [12], analysis of retail store data [13], implementation of a new techniques for collaborative of the KD projects through offering the support for the distributed teams [14]. The six step model incorporates the development of a computerized system for assessment of SPECT bull eye images [15], developing and mining a database of cardiac SPECT images [16].

V. THE DATA-MINING STEP OF THE KDD PROCESS

In various applications, the data mining algorithms are presented in the KDD data-mining process. The primary objectives of data-mining are investigated together with the algorithms employed and the details of the algorithms that include various techniques. There are two major steps such as verification and discovery are estimated in the knowledge discovery process. The users queries are verified by limiting the system in verification step and the new patterns are determined by the system in discovery step. The KDD is further divided into two steps such as prediction and description the systems that finds patterns to predict the future behaviour. Therefore, the data-mining algorithms are developed in the form of discovery related which contributes the fitting models and predicting the patterns. In this, the model fitting consists two steps which are the statistical and

logical techniques. The statistical techniques permits the nondeterministic process in the model and the logical techniques permits the deterministic process. In general, the data-mining techniques are based on the prediction algorithms includes machine learning and pattern recognition like classification, clustering, etc. [17].

A.Data mining methods

In practical, the main objectives of data-mining algorithms are prediction and description processes. The prediction process includes the number of variables and fields in the datasets for determining the future pattern. The description aims on predicting the human interpretable patterns defining the database. Classification is a techniques which learning a function and maps the datasets into predefined classes [18].In the knowledge discovery, the classification techniques are classifying the trends in financial markets [19] and objects identification in the huge image databases [20].The purpose of classification is that estimating the probability that a patient will survive by giving the results of a set of diagnostic tests. Clustering is a task where it determines the finite set of groups to define the database [21]. For instances, clustering application involves discovering homogeneous subpopulations for customers in marketing database. Summarization includes the techniques for determining a compact definition for the subset of data. Then, evaluation is attained by tabulating the mean and standard deviations for all the fields. Dependency modeling consists determining a model that defines particular dependencies between the parameters. There are two types of dependency models such as structural and quantitative. One of the widely used learning techniques is the decision tree algorithm which is used to interpret and represented as if then-else conditions. This algorithm functions well in noisy data and does not required any prior knowledge of data to classify the medical database based on the disease, loan applicant by feasibility of payment and networks malfunction by attacks [22]. Support Vector machines (SVM) algorithms is based on structural risk reduction process which is related to regularization theory. SVM uses training methods for learning classification and regression rules from the data.SVM has been described as an effective tool for several aspects of data-mining involves classification, regression [23]. In Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) is one of the important networks that were designed to model temporal sequences and its long range dependencies with better accuracy than other neural networks.[24,34] LSTM are very complex to train since it needs memory-bandwidth-bound estimation. It is difficult for hardware designer and mainly restricts the usages of neural networks. In general, LSTM need four MLP layers per cell to execute and for each sequence time-step [17, 24,]. SVM have been presented as a efficient tool for many aspects of data mining algorithms such as classification, regression and detection. SVM are a type of machine learning model which are the most popularly used in statistical learning applications [24]. Fuzzy logic offers a powerful way to classify a concept in an abstract based on the fuzzy set theory. In data mining algorithms, K-means clustering are capable of extracting patterns from huge amount of database. The combination of fuzzy logic with data mining algorithms will produce more abstract patterns at greater level than at the data level [25]. Evaluation of the data points into clusters is based upon the distance between cluster cancroids and data point in K-Means clustering [26]. Accuracy of k-means clustering depends on the value of k and finding the suitable number of clusters is challenging task. The Naïve Bayes classifier offers a prediction techniques based on the inferences of probabilistic graphic models that represents the probabilistic dependencies underlying a specific model using a graph structure [27]. A probabilistic graphical method is a graph where nodes denote the random variables, and the arcs denote the conditional dependence considerations. Therefore, it offers a dense representation of joint probability distributions. An undirected graphical model is known as a Markov network and a directed graphical model is called as a Bayesian network [28]. Random forest classifier is a prediction technique that integrates the bagging and the random selection of features to create a set of decision trees [29]. The random subset selection of features is a type of random subspace technique, which implements stochastic bias presented by Eugene Kleinberg. The performance evaluation of Random Forest classifiers are used to estimate the Classification accuracy using decision tables. Therefore, Random Forest classifier works better than other data-mining classification techniques. It functions a nonlinear, smooth mapping of high-dimensional data onto the elements of a regular and low-dimensional array. The technique transforms to geometrical relationships between points in a two-dimensional map from a non-linear statistical relation between data points in a high dimensional space. Maximum likelihood Gaussian classifiers consider inputs are un correlated distributions for various classes and varies only in mean values. Gaussian classifier is using the Bayes decision theorem [30].

Comparative Analysis

Table 1. Comparative Analysis of Techniques

Title	Technique used	Advantages	Disadvantages
A Review on Support Vector Machine for Data Classification [24].	SVM	This algorithm solves the all pattern classification issues effectively.	It consumes more time for execution. Automatic Feature detection NA.
Knowledge Discovering Data mining Using Soft Computing.[25].	Fuzzy logic	Simple to implement and understand.	Fuzzy logic required more tuning
Application of data	naïve	This classifier	It has more error

mining to network intrusion detection: classifier selection model. [27]	Bayes	is simple to develop.	rate in practical applications. Automatic Feature detection is Not Applicable.
A Clustering Algorithm for Intrusion Detection.[31]	K-Means	This Technique effective to noisy training data	This techniques works slower than other clustering approaches
KNN Model-Based Approach in Classification. [32,33].	The K-Nearest Neighbour (KNN) algorithm	The K-Nearest Neighbour (KNN) Classifier is a classifier that works well on recognition problems.	It is slow for real-time prediction if there are a large number of training examples and is not robust to noisy data.

VI. CONCLUSION

In this paper, review of the knowledge discovery database and data mining techniques are presented. In addition, this paper helps in defining various types of modelling and techniques in dynamic data mining. The goal of this survey is to combine the research in the area of KDDM with various machine learning algorithms and to develop different dynamic models. The KDDM standards will offers in developing the industry growth and push the industry beyond the edge. Training a deep neural network dynamically on new data is an improved approach for building a model which helps in detecting combination of new features dynamically and is a powerful way of improving accuracy there by increasing the correct number of predictions which will help industry in achieving better results on real world data.

REFERENCES

[1] K. S., Hemanth, C. M., Vastrad, S. Nagaraju, "Data Mining Technique for Knowledge Discovery from Engineering Materials Data Sets", In International Conference on Computer Science and Information Technology,Springer, Berlin, Heidelberg, pp. 512-522,2011

[2] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery in databases", AI magazine, Vol.17, Issue.3, pp.37, 1996

[3] M. Panda, M. R. Patra, "Evaluating machine learning algorithms for detecting network intrusions", International journal of recent trends in engineering, Vol.1, Issue.1, pp. 472, 2009.

[4] D. Y. Yeung, C. Chow, "Parzen-window network intrusion detectors" In Object recognition supported by user interaction for service robots, IEEE, Vol. 4, pp. 385-388, 2002.

[5] J. Han, M. Kamber, "Data Mining: Concepts and Techniques", 2/e San Francisco: CA. Morgan Kaufmann Publishers, an imprint of Elsevier. pp-5-38, 2006

[6] L. L. Berry, "Relationship marketing of services—growing interest, emerging perspectives", Journal of the Academy of marketing science, Vol. 23, Issue. 4, pp. 236-245, 1995.

[7] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, "Fast discovery of association rules", Advances in knowledge discovery and data mining, Vol. 12, Issue.1, pp. 307-328, 1996.

[8] C. Priyadharsini, A. S. Thanamani, "An Overview of Knowledge Discovery Database and Data mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue.1, 2014.

[9] T. E. Senator, H. G. Goldberg, J. Wooton, M. A. Cottini, A. U. Khan, C. D. Klinger, R. W. Wong, "Financial Crimes Enforcement Network AI System (FAIS) Identifying Potential Money Laundering from Reports of Large Cash Transactions." AI magazine, Vol.16, Issue.4, pp. 21, 1995.

[10] R. J. Brachman, T. Anand, "The process of knowledge discovery in databases". In Advances in knowledge discovery and data mining (1996, February), American Association for Artificial Intelligence, pp. 37-57.

[11] S. S. Anand, A. R. Patrick, J G. Hughes, D. A. Bell, "A data mining methodology for cross-sales", Knowledge-based systems, Vol.10, Issue.7, pp. 449-461, 1998.

[12] A. K. Jain, R. C. Dubes, "Algorithms for clustering data", 1988.

[13] S. Butler, "An investigation into the relative abilities of three alternative data mining methods to derive information of business value from retail store-based transaction data", Doctoral dissertation, BSc thesis, School of Computing and Mathematics, Deakin University, Australia, 2002.

[14] S. Moyle, M. Bohanec, E. Osrowski, "Large and tall buildings: a case study in the application of decision support and data mining", Kluwer International Series In Engineering And Computer Science, pp. 191-202, 2003.

[15] K. J. Cios, G. W. Moore, "Medical data mining and knowledge discovery: Overview of key issues", Studies in Fuzziness and Soft Computing, Vol.60, pp. 1-20, 2001.

[16] J. P. Sacha, K. J. Cios, L. S. Goodenday, "Issues in automating cardiac SPECT diagnosis", IEEE Engineering in Medicine and Biology Magazine, Vol.19, Issue.4, pp. 78-88, 2000.

[17] H. Sak, A. Senior, F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling". In Fifteenth annual conference of the international speech communication association ,2014.

[18] D. J. Hand, "Deconstructing statistical questions. Journal of the Royal Statistical Society", Series A (Statistics in Society), pp. 317-356, 1994.

[19] J. Hall, G. Mani, D. Barr, "Applying computational intelligence to the investment process", Proceedings of CIFER-96: Computational Intelligence in Financial Engineering. Washington, DC: IEEE Computer Society ,1996.

[20] R. S. Manikantan, Ostermann, B. Tjaden, "Detecting Anomalous Network Traffic with Self-organizing Maps", Ohio University, pp.37, 2003.

[21] Y. Jing, T. Li, H. Fujita, B. Wang, N. Cheng, "An incremental attribute reduction method for dynamic data mining", Information Sciences, Vol.465, pp. 202-218, 2018

[22] H. A. Nguyen, D. Choi, "Application of data mining to network intrusion detection: classifier selection model", In Asia-Pacific Network Operations and Management Symposium, Springer, Berlin, Heidelberg, pp. 399-408, 2008

[23] R. Burbidge, B. Buxton, "An introduction to support vector machines for data mining", Keynote papers, young OR, Vol.12, pp. 3-15, 2001

[24] H. Bhavsar, M. H. Panchal, "A review on support vector machine for data classification", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol.1, Issue.10, pp. 185, 2012.

[25] K. Suresh, C. Jnaneswari, G. L. Kranthi, K. Bindu, "Knowledge Discovery in Datamining Using Soft Computing", Vol.3, pp. 3952-3957, 2012.
[26] K. Kusum, S. Bharti, Shukla, S. Jain, "Intrusion detection using clustering", Vol.1, Issue. 2, 3, 4, pp.6, 2010.
[27] A. N. Huy, D. Choi, "Application of Data Mining to Network Intrusion Detection: Classifier Selection Model", pp.1, 2008.
[28] P. Mrutyunjaya, M. R. Patra, "Evaluating Machine Learning Algorithms for Detecting Network Intrusions", International Journal of Recent Trends in Engineering, Vol.1, Issue.1, May 2009.
[29] M. Ramadas, S. Ostermann, B. Tjaden, "Detecting anomalous network traffic with self-organizing maps", In International Workshop on Recent Advances in Intrusion Detection September, Springer, Berlin, Heidelberg, pp. 36-54, 2003
[30] R. O. Duda, P. E. Hart, "Pattern Classification and Scene Analysis", New York: Wiley, pp: 78, 1973.
[31] W. V. Qiang, Megalooikonomou, "A Clustering Algorithm for Intrusion Detection", pp. 3, 2004.
[32] KNN Model-Based Approach in Classification, Gongde Guo1, Hui Wang , David Bell , Kieran Greer School of Computer Science, Queen's University Belfast, BT7 1NN, UK. partly European Commission project ICONS, project no. IST-2001-32429,2001.
[33] Priyanka, Sana Khan, Tulsi Kour, "Investigation on Smart Health Care Using Data Mining Methods", International Journal of Scientific Research in Computer Science and Engineering, Vol.4, Issue.2, pp.31-36, 2016.
[34] Prakash Singh , "Efficient Deep Learning for Big Data: A Review", International Journal of Scientific Research in Computer Science and Engineering, Vol.4, Issue.6, pp.36-41, 2016.

Authors Profile

Mr. D.Ramana kumar ,pursed Master of Computer Science from Dr.Baba Saheb Ambedkar Martwada University,Maharastra,India in 2000 and Master of Technology from Bharath University,Chennai,India in year 2009. He is currently pursuing Ph.D. from JNTU Hyderbad and currently working as Assosiate Professor in Department of Computer Science & Engineering,.
Email: ramanad74@gmail.com

Dr.S.KRISHNAMOHAN RAO
Principal, Gandhi Institute of Technology, Bhubaneshwar,Khurda,India. His areas of specializations are Mobile Computing, MANETS, Adhoc Sensor Networks and Computer Networks.
Email: krishnamohan6@yahoo.com

.

