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 Database

D. Ramana Kumar[1*], S. Krishna Mohan Rao[2], K. Rajeshwar Rao[3]

[1]Department of Computer Science and Engineering, JNTUH, Hyderabad, India
[2]GIET, Bhubaneswar, Odisha, India
[3]Department of Computer Science and Engineering, SIET, Hyderabad, India

[*]*Corresponding Author: ramanad74@gmail.com, Tel.: +91 -93983-21758*

Available online at: www.ijcseonline.org

Accepted: 13/Dec/2018, Published: 31/Dec/2018

Abstract— Data mining and knowledge discovery in databases have been considered as a significant research area in industry. This survey presents an overview, description and future directions which depict a standard for knowledge discovery and data mining process model. The paper mentions particular real-world applications, specific data mining techniques, challenges involved in real-world application of knowledge discovery, current and future research ideas in the field. The applications to both academic and industrial problems are discussed. The main target of the review is the consolidation of the research in this particular area and thereby helping in enhancing the existing model by embedding other current standards.

Keywords— Knowledge discovery database, data mining, real world application

I. Introduction

The rapid growth in the number of the available databases in industrial, administrative and of the applications has shown to extract the knowledge from the huge amounts of data. Through the extraction in the databases the large databases will help in serving the rich and reliable source of knowledge generation and verification and the knowledge discovered can be applied to the information management, query processing, decision making and many other applications. Thus, the knowledge discovery databases (KDD) are considered as one of the best research topics in both machine learning and database researchers (Hemanth et al., 2011) [25]. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD) (Fayyad et al., 1996) [3]. Numerous philosophical considerations are present for the knowledge discovery databases and helps in the development of the different KDD techniques (Prasad et al., 2017) [26]. Knowledge discovery in databases is defined as the algorithmic and statistical data analysis to extract the knowledge and some useful patterns from vast amount of data. The KDD field is always looking for the development of the method and techniques for making sense of the data. The problem with the process of KDD is mapping the low-level data into other forms which is more compact. KDDM concerns the entire knowledge extraction process, including how the data is stored and accessed, how to develop efficient and scalable algorithms that can be used to analyse massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine [3]. Most classification methods assume that the data conforms to a stationary distribution. However, the real-world data is usually collected over certain periods of time, ranging from seconds to years, and ignoring possible changes in the underlying concept, also known as concept drift, may degrade the predictive performance of a classification model. Moreover, the computation time, the amount of required memory, and the model complexity may grow indefinitely with the continuous arrival of new training instances. Data mining is a part of a process called knowledge discovery in databases. The application challenges faced by KDD are as follows,

- High dimensionality: There are not only large number of records but also large number of fields so that the problem of dimensionality is high. The type of the knowledge database create problem in terms of the size increasing in the search. It also gives an chance where the data mining techniques finds the different types of patterns which are not valid.
- Overfitting: The algorithm searches for the good parameters for one model using a limited set of data and it does not model general patterns but also noise specific to the particular dataset which results in poor performance.

- Larger databases: The databases which contains hundreds of fields and tables with n number of records with large data volumes with efficient algorithms.
- Assessing statistical significance: The problem occurs when the system is searching for many possible models. Considering an example if a system tests N models with a significance level of 0.01 and on an average with pure random data few models will be considered as significant.
- User interaction: Many of the present KDD methods are not interactive and cannot easily incorporate prior knowledge about a problem.
- Integration with other systems: The stand-alone discovery system is not that useful and the issues of integration are DBMS integration, spreadsheets and visualization tools. All of steps that are performed before carrying out data mining, such as data selection, data cleaning, pre-processing, and data transformation (Han et al., 2006)[27].

Association rule techniques are used for data mining if the goal is to detect relationships or associations between specific values of categorical variables in large data sets. There may be thousands or millions of records that have to be read and to extract the rules for, but the question is what will happen if there is new data, or there is a need to modify or delete some or all the existing set of data during the process of data mining. In the past user would repeat the whole procedure, which is time-consuming in addition to its lack of efficiency. There are various data streaming challenges which are affecting the real time applications of the data generated. From this, the importance of dynamic data mining process appears and for this reason this problem is going to be the main topic of the proposed research. The basic data mining tasks consists of a number of processes such as Time series analysis, association analysis, classification, regression cluster analysis and summarization.

II. RELATED WORK

1. Need for Knowledge discovery database (KDD)

The existing methods of turning data into knowledge is always concerned on manual analysis and the interpretation made. Considering the example in health-care field the specialists helps in analysing the current trends and changes happening in the health care data and helps in providing the report detailing the analysis to the health organization. The report made or obtained becomes the basis for the decision making in future and planning for the management of the health care. For numerous other applications, this type of manual examining of an informational index is moderate, costly, and exceptionally emotional. In actuality, as databases volumes develop drastically, this sort of manual information investigation is getting to be totally unrealistic in numerous areas. Databases are expanding in size in two different ways such as (1) the number N of records in the database and (2) the number d of fields or characteristics to an object. Databases containing on the request of N = 10^9 articles are ending up progressively normal, for instance, in the galactic sciences. Thus, the quantity of fields d can without much of a stretch be on the request of 10^2 or indeed, even 10^3, for instance, in restorative symptomatic applications. We trust that this activity is absolutely not one for people; subsequently, investigation work should be robotized, at any rate somewhat. The need to scale up human examination capacities to dealing with the vast number of bytes that we can gather is both monetary and logical. Organizations utilize information to increase focused advantage, increment proficiency, and give more profitable administrations to clients. Information we catch about our condition are the essential proof we use to manufacture hypotheses also, models of the universe we live in. Since PCs have empowered people to accumulate a larger number of information than we can process, it is as it were characteristic to swing to computational systems to enable us to uncover significant examples structures from the gigantic volumes of databases. Thus, KDD is an endeavour to address an issue that the computerized data information made an unavoidable truth for every one of us. [3]

2. Data Mining and Knowledge Discovery in the Real World

The successful KDD application is obtained by the large amount of interest in KDD. The well documented examples of successful systems can be considered as KDD applications and can also be deployed in the operational use of the real-world problems such as in business and science. The primary application in science is astronomy and a biggest achievement was by SKICAT which is one of the systems used by the astronomers to perform the image analysis, classification and separating of the sky objects [3]. The processing of 3 terabytes of image data resulting from the survey of sky images shows that the order of 10^9 objects are detectable and SKICAT can outperform humans and existing techniques in classification of the objects in sky. In business the major applications of the KDD include marketing, finance, manufacturing, telecommunications and fraud detection.

2.1 Marketing: The primary application is database marketing systems which helps in analysing the customer databases to find different customer groups and display the behaviour. The business week (Berry 1994) reveals that the over half of the retailers are planning to use the database marketing and can get good outcomes. The one more noticeable marketing application is market-basket analysis (Agarwal et al., 1996) [5] systems which finds the patterns available for the retailers.

2.2 Investment: The data mining can be used for investment in many of the companies but do not describe the systems. One example considered is the LBS capital management.

This uses the systems experts, neural nets and few genetic algorithms to manage the portfolios. This system has outperformed the broad stock market (Patil et al., 2016).[28]

2.3 Manufacturing: The system called s CASSIOPEE which was developed as a part of the joint project between the general electric and SNECMA was used by the major European airlines to diagnose and predict the problems for the Boeing 737.Clustering methods were used to remove the faults.

2.4 Fault detection: The FAIS system (Senator et al., 1995) from the treasury of US financial crimes enforcement network is used to identify the financial transactions which indicates money laundering activity.

3. KDDM process models:

A KDDM process model consists of a set of processing steps which needs to be followed by practitioners while conducting the KDMM projects. The procedures are explained in secured exercises. 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 outcomes. All proposed models likewise accentuate the iterative idea of the model as far as numerous 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 focused on developing the new models rather than improving the design of the single model. The nine-step model which was the first model to be proposed includes the academic research features which also has several important business issues. The CRISP-DM model is industry oriented. The several features present in all of the KDDM models include complex and time-consuming data preparation tasks (Brachman, 1996)[8]. The nine steps include the developing the application domain, creating the target dataset, data cleaning and pre-processing, data reduction and projection, choosing the data manipulation task and interpreting the patterns with the discovery knowledge.

3.1 Applications and impact of KDDM models

The research applications used by different models differ and are easier than the industrial applications. The nine-step model proposed by Fayyad et al., 1996[3] is incorporated into an industrial DM software system called as Mindset and has been applied in number of KDDM projects. The different models have different applications and also include analysing the data from the retailer. The industrial project concerning customer cross sales and a research project concerning analysis of marketing internet data (Anand et al., 1998)[29].The CRISP-DM model has found few research projects such as performance evaluation of heating, ventilation and HVAC systems and analysis of the thrombosis data (Jensen, 2001), analysis of retail store data (Butler, 2002), development of a new methodology for collaborative of the KD projects through providing the support for the distributed teams (Moyle et al., 2012).The six step model includes the development of a computerized system for diagnosis of SPECT bull eye images (Cios et al., 2000), creating and mining a database of cardiac SPECT images (Sacha et al., 2000).

4. The Data-Mining Step of the KDD Process

The data mining methods present in the data mining component of the KDD process has numerous applications. The major goals of data mining are presented along with the methods used and the description of the algorithms which incorporates different methods. The knowledge discovery aims at two important process and they are described as verification and discovery. In verification process the user's hypotheses is verified by limiting the system. In discovery process, the new patterns are found by the system autonomously. To predict the future behaviour of the entities the discovery goal is further subdivided into prediction and in the description the system finds the patterns for presentation to a user in an understandable form. So, the data mining presented is in the form of discovery oriented. These data mining involves the fitting models and determining the patterns. The model fitting involves two process and they are statistical and logical. The statistical approach allows for nondeterministic effects in the model whereas the logical model is deterministic. The most of data mining techniques are based on the tested techniques from machine learning, pattern recognition and statistics such as classification, clustering etc. (Smyth et al., 1996)[3]

4.1 Data mining methods

The significant aim of the data mining which is in practice are prediction and description. The prediction involves using some variables and fields in the database to predict the future values whereas description focuses on finding the human interpretable patterns describing the data. Classification is a method which helps in learning a function which maps a data item into predefined classes (Hand et al., 1981)[34]. The classification methods which are present in the knowledge discovery includes classifying the trends in financial markets (Hong et al., 1996)[6] and identification of the objects in large image databases (Weir et al., 1996). Regression is a type of learning a function that maps a data item to a real valued prediction variable. The application of regressions is many and helps in estimating the probability that a patient will survive by giving the results of a set of diagnostic tests. Clustering is a descriptive task where one finds the finite set of categories to describe the data (Jain et al., 1988)[9] The examples of clustering applications include discovering homogeneous subpopulations for consumers in marketing databases. Summarization involves the methods for finding a compact description for the subset of data. The minor example can be by tabulating the mean and standard deviations for all the fields. Dependency modelling involves finding a model that describes significant dependencies between the variables. The two levels where dependency model exists are at the structural level and quantitative level. Decision trees are widely used learning method which is easy to interpret and can be re-represented as if-then-else rules. It does not require any prior knowledge of data distribution, works well on noisy data and has been applied to classify medical patients based on the disease, equipment malfunction by cause, loan applicant by likelihood of payment (Goebel, 2014). Support vector machines are based on structural risk minimization principle which is related to regularization theory. This is used as a training algorithm for learning classification and regression rules from the data. The support vector machine has been introduced as a robust tool for many aspects of data mining including classification, regression and outlier detection (Buxton et al., 2001)[14]. Long Short-Term Memory (LSTM) is a specific recurrent neural network (RNN) architecture that was designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs. The support vector machines have been introduced as a robust tool for many aspects of data mining including classification, regression and outlier detection. Support vector machines are a specific type of machine learning algorithm that are among the most widely used for statistical learning problems (Bhavsar et al., 2012)[18]. Based on the fuzzy set theory fuzzy logic provides a powerful way to categorize a concept in an abstract way by introducing vagueness. The K means clustering data mining method are capable of extracting patterns automatically from large amount of data. The integration of fuzzy logic with data mining methods will create more abstract patterns at higher level than at the data level (Suresh et al., 2012). In K-Means clustering, (Kusum K. B.,2010) assignment of the data points to clusters is depend upon the distance between cluster centroid and data point. Accuracy of k-means clustering depends upon the value of k. determining the appropriate number of clusters is challenging area for researchers. Th Naïve Bayes classifier provides a simple approach based on the inferences of probabilistic graphic models which specify the probabilistic dependencies underlying a particular model using a graph structure (Huy A. N.,2008). In its simplest form, a probabilistic graphical model is a graph in which nodes represent random variables, and the arcs represent conditional dependence assumptions. Hence it provides a compact representation of joint probability distributions. An undirected graphical model is called as a Markov network, while a directed graphical model is called as a Bayesian network or a Belief network (Mrutyunjaya P et al., 2009). Random forest classifier method combines bagging and the random selection of features to construct a group of decision trees with controlled variation (Yeung et al., 2002)[19]. The selection of a random subset of features is a method of random subspace method, which is a way to implement stochastic bias proposed by Eugene Kleinberg. The performance of Decision Table and Random Forest classifiers are used to predict the classification accuracy. Based on this, Random Forest outperforms on the most techniques. It performs a nonlinear, ordered, smooth mapping of high-dimensional input data manifolds onto the elements of a regular, low-dimensional array (Manikandan R.,2003)[38]. The algorithm converts non-linear statistical relationships between data points in a high dimensional space into geometrical relationships between points in a two-dimensional map. Maximum likelihood Gaussian classifiers assume inputs are uncorrelated and distributions for different classes differ only in mean values. Gaussian classifier is based on the Bayes decision theorem (Richard O. D.,1973)[41].

III. COMPARATIVE ANALYSIS

Title	**Technique used**	**Advantages**	**Disadvantages**
A Clustering Algorithm for Intrusion Detection	K-Means (Qiang W. V.,2004)	This classifier is robust to noisy training data	This performs slower than other clustering approaches
Application of data mining to network intrusion detection: classifier selection model.	naïve Bayes (Huy A. N.,2008)	This classifier is easy to implement and computation is simple	This algorithm has more error rate in practical considerations
Application of data mining to network intrusion detection: classifier selection model.	SOM (Huy A. N.,2008)	This works good with nonlinear dataset.	The computation time is high
A Review on Support Vector Machine for Data Classification	SVM (Bhavsar et al., 2012)[18]	All the pattern classification problems will be solved**.**	limitation is speed and size
Knowledge Discovery in Datamining Using Soft Computing	Fuzzy logic (Suresh et al., 2008)	Ease of implementation and easy to understand	Requires more tuning and simulation

IV. CONCLUSION AND FUTURE SCOPE

This paper presents a review of the knowledge discovery database and data mining methods involved in it. Also, this paper helps in describing different types of modelling and methods in data mining process. The bonding between the knowledge discovery and data mining in real world is reviewed with appropriate examples. The KDMM process models is surveyed and the description of different main models is provided. The gaol of the survey is to consolidate research in the area of KDDM to inform users about different models and to develop different models depending on the previous experiences. The KDDM standards will helps in promoting the industry growth and pushes the industry beyond the edge.

REFERENCES

[1] Kurgan, L. A., &Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review, 21(1), 1-24.

[2] Han, J., & Fu, Y. (1994, July). Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases. In KDD Workshop (pp. 157-168).

[3] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[4] Jing, Y., Li, T., Fujita, H., Wang, B., & Cheng, N. (2018). An incremental attribute reduction method for dynamic data mining. Information Sciences, 465, 202-218.

[5] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., &Verkamo, A. I. (1996). Fast discovery of association rules. Advances in knowledge discovery and data mining, 12(1), 307-328.

[6] Apte, C., & Hong, S. J. (1994, July). Predicting Equity Returns from Securities Data with Minimal Rule Generation. In KDD Workshop (pp. 407-418).

[7] Berry, L. L. (1995). Relationship marketing of services—growing interest, emerging perspectives. Journal of the Academy of marketing science, 23(4), 236-245.

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

[9] Jain, A. K., &Dubes, R. C. (1988). Algorithms for clustering data.

[10] Brodley, C. E., & Smyth, P. (1997). Applying classification algorithms in practice. Statistics and Computing, 7(1), 45-56.

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

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

[13] Priyadharsini, C., &Thanamani, A. S. (2014). An Overview of Knowledge Discovery Database and Data mining Techniques. International Journal of Innovative Research in Computer and Communication Engineering, 2(1).

[14] Burbidge, R., & Buxton, B. (2001). An introduction to support vector machines for data mining. Keynote papers, young OR12, 3-15.

[15] Du, J., Zhou, J., Li, C., & Yang, L. (2016, August). An overview of dynamic data mining. In Informative and Cybernetics for Computational Social Systems (ICCSS), 2016 3rd International Conference on (pp. 331-335). IEEE.

[16] Suresh, K., Jnaneswari, C., Kranthi, G. L., & Bindu, K. Knowledge Discovery in Datamining Using Soft Computing. vol, 3, 3952-3957.

[17] Al-mamory, S. O., &Jassim, F. S. (2013). Evaluation of different data mining algorithms with KDD CUP 99 Data Set. Journal of University of Babylon, 21(8), 2663-2681.

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

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

[20] Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis. A Wiley-Interscience Publication, New York: Wiley, 1973.

[21] Panda, M., & Patra, M. R. (2009). Evaluating machine learning algorithms for detecting network intrusions. International journal of recent trends in engineering, 1(1), 472.

[22] Ramadas, M., Ostermann, S., &Tjaden, B. (2003, September). Detecting anomalous network traffic with self-organizing maps. In International Workshop on Recent Advances in Intrusion Detection (pp. 36-54). Springer, Berlin, Heidelberg.

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

[24] Bharti, K. K., Shukla, S., & Jain, S. (2010). Intrusion detection using clustering. Proceeding of the Association of Counseling Center Training Agencies (ACCTA), 1.

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

[26] Adhikari, A., Jain, L. C., & Prasad, B. (2017). A State-of-the-Art Review of Knowledge Discovery in Multiple Databases. Journal of Intelligent Systems, 26(1), 23-34.

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

[28] Hiremath, R., & Patil, P. (2016). Astudy-Knowledge Discovery Approachesand Its Impact with Reference to Cognitive Internet of Things (Ciot). International Journal of Information, 6(1/2).

[29] Anand, S. S., Patrick, A. R., Hughes, J. G., & Bell, D. A. (1998). A data mining methodology for cross-sales. Knowledge-based systems, 10(7), 449-461.

[30] Butler, S. (2002). 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).

[31] Cios, K. J., & Moore, G. W. (2001). Medical data mining and knowledge discovery: Overview of key issues. Studies in Fuzziness and Soft Computing, 60, 1-20.

[32] Moyle, S., Bohanec, M., &Osrowski, E. (2003). Large and tall buildings: a case study in the application of decision support and data mining. KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE, 191-202.

[33] Sacha, J. P., Cios, K. J., &Goodenday, L. S. (2000). Issues in automating cardiac SPECT diagnosis. IEEE Engineering in Medicine and Biology Magazine, 19(4), 78-88.

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

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

[36] Suresh, K., Jnaneswari, C., Kranthi, G. L., & Bindu, K. Knowledge Discovery in Datamining Using Soft Computing. vol, 3, 3952-3957.

[37] Kusum K. Bharti, S. Shukla and S. Jain, “Intrusion detection using clustering”, vol.1, issue 2, 3, 4, pp.6, 2010

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

[39] Huy A. N., D. Choi,” Application of Data Mining to Network Intrusion Detection: Classifier Selection Model”, pp:1, 2008.

[40] Mrutyunjaya P. and M. Ranjan Patra,” Evaluating Machine Learning Algorithms for Detecting Network Intrusions”, International Journal of Recent Trends in Engineering, vol. 1, no.1, May 2009.

[41] Richard O. Duda and P. E. Hart, “Pattern Classification and Scene Analysis”, New York: Wiley, pp: 78, 1973.

[42] Qiang W., V. Megalooikonomou, “A Clustering Algorithm for Intrusion Detection”, pp:3, 2004