ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

A Comprehensive Study on Distributed Data Mining and Learning Algorithms

Sanjay Kumar Sen, Dr. B.K. Ratha.

Asst. Professor, Computer Science & Engg, Orissa Engineering College, Bhubaneswar, Odisha, India

Head of the Department , PG Dept. of Computer Science and Application, Utkal University, Vani Vihar, Bhubaneswar. Odisha, India

ABSTRACT: In this paper, we provide a comprehensive overview of data mining, distributed data mining, techniques and learning algorithms used for data mining as well as potential areas of data mining application. Further, the paper presents an introduction to agent and multi-agent technology, its advantages and application areas.

KEYWORDS: data mining, distributed data mining technique, agent, multi-agent technology, learning algorithm.

I. INTRODUCTION

Data mining is an emerging technology that has made revolutionary change in the information world. The term "data mining" (often called as knowledge discovery) refers to the process of analysing data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of an organizational system.

Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Therefore, data mining consists of major functional elements that transform data onto data warehouse, manage data in a multidimensional database, facilitates data access to information professionals or analysts, analyse data using application tools and techniques, and meaningfully presents data to provide useful information.

According to the Gartner Group, data mining is the process of discovering meaningful new correlation patterns and trends by sifting through large amount of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques (Larose, 2005). Thus use of data mining technique has to be domain specific and depends on the area of application that requires a relevant as well as high quality data.

More precisely, data mining refers to the process of analysing data in order to determine patterns and their relationships. It automates and simplifies the overall statistical process, from data source(s) to model application. Practically analytical techniques used in data mining include statistical methods and mathematical modelling. However, data mining and knowledge discovery is a rapidly growing area of research and application that builds on techniques and theories from many fields, including statistics, databases, pattern recognition, data visualization, data warehousing and OLAP, optimization, and high performance computing[2] . Worthy to mention that online analytical processing (OLAP) is quite different from data mining, though it provides a very good view of what is happening but cannot predict what will happen in the future or why it is happening. In fact, blind applications of algorithms are not also data mining. In particular, data mining is a user centric interactive process that leverages analysis, technologies and computing power, or a group of techniques that find relationships that have not previously been discovered. So, data mining can be considered as a convergence of three technologies such as increased computing power, improved data collection and management tools, and enhanced statistical algorithms[3].

II. DATA MINING PROCESS

Data Mining is an iterative process (Figure -2.1) consists of the following list of stages:

i. Data cleaning
ii. Data integration

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

iii. Data selection
iv. Data transformation
v. Data mining
vi. Pattern evaluation
vii. Knowledge presentation

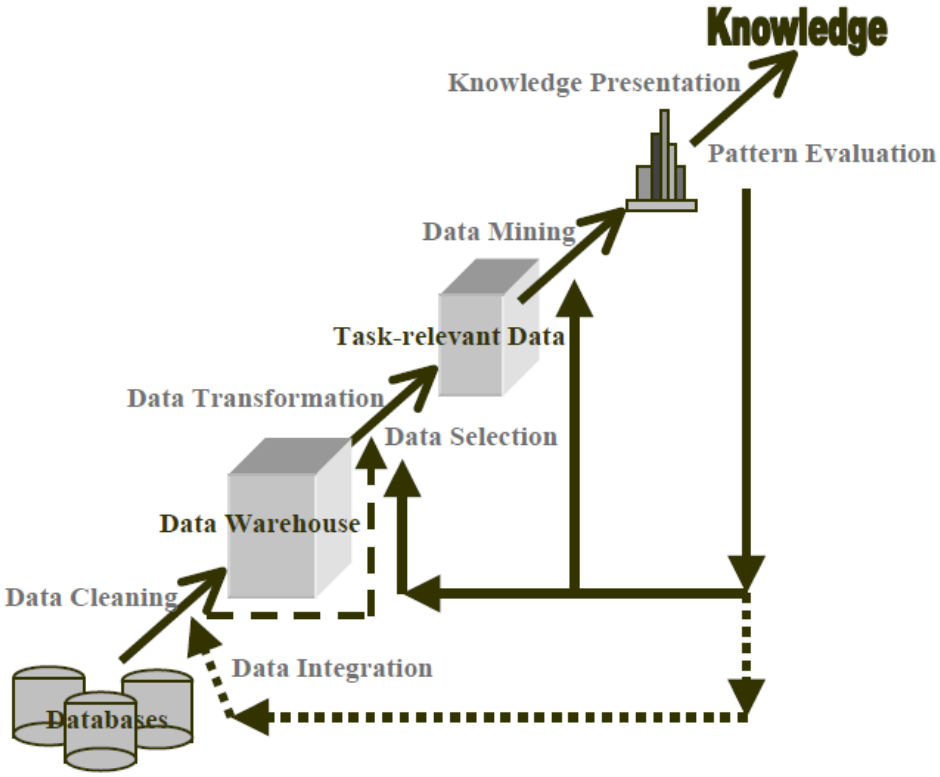

Figure-2.1: Data Mining or Knowledge Discovery Process

Data cleaning: This task handles missing and redundant data in the source file. The real world data can be incomplete, inconsistent and corrupted. In this process, missing values can be filled or removed, noise values are smoothed, outliers are identified and each of these deficiencies are handled by different techniques.
Data Integration: Data integration process combines data from various sources. The source data can be multiple distinct databases having different data definitions. In this case, data integration process inserts data into a single coherent data store from these multiple data sources.
Data Selection: In the data selection process, the relevant data from data source are retrieved for data mining purposes.
Data Transformation: This process converts source data into proper format for data mining. Data transformation includes basic data management tasks such as smoothing, aggregation, generalization, normalization and attributes construction.
Data Mining: In Data mining process, intelligent methods are applied in order to extract data patterns. Pattern evaluation is the task of discovering interesting patterns among extracted pattern set. Knowledge representation includes visualization techniques, which are used to interpret discovered knowledge to the user.
Pattern Evaluation: During data mining, a large number of patterns may be discovered. However, all those patterns may not be useful in a particular context. It is highly required to assess the usefulness of the discovered patterns based on some criteria, so that truly useful and interesting patterns representing knowledge can be identified.
Knowledge Presentation: Finally, the mined knowledge has to be presented to the decision-maker using suitable techniques of knowledge representation and visualization.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

III. METHODS AND LEARNING ALGORITHMS

Researchers identify two fundamental goals of data mining: prediction and description. Prediction makes use of existing variables in the database in order to predict unknown or future values of interest, while description focuses on finding patterns describing the data the subsequent presentation for user interpretation. The relative emphasis of both prediction and description differs with respect to the underlying application and technique. There are several data mining techniques fulfilling these objectives. Some of these are classification, clustering, association and pattern discovery.

- **Classification:** Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Some well-known classification models are:
    a) Decision Tree
    b) Bayesian Network
    c) Neural Networks
    d) Support Vector Machines (SVM)
- **Clustering:** Clustering is a technique for identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as pre-processing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. Some commonly used clustering methods are:
    a) Partitioning Methods
    b) Hierarchical Agglomerative (divisive) methods
    c) Density based methods
    d) Grid-based methods
    e) Model-based methods
- **Association rule:** Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value. Some association rule types are:
    a) Multilevel association rule
    b) Multidimensional association rule
    c) Quantitative association rule

IV. DATA MINING APPLICATIONS

Data mining is a way to discover new meaning in data, performs data processing using sophisticated data search capabilities and statistical algorithms, which can be utilized in any organization or system that needs to determine the patterns or relationships implicit in a large data warehouse for better strategies. It can be reasonably beneficial to any corporate industries, financial institutions, retailers, pharmaceutical firms, security agencies, government departments,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

online service providers, libraries, and individual researchers too. It can be used for a variety of applications in both public and private sectors. Corporate industries and financial institutions often use data mining to increase sales, reduce costs, improve market performance, enhance customer base by means of developing models for credit scoring, risk assessment, fraud detection, etc.

Recently data mining have been increasingly used in public sectors for many purposes such as detecting fraud and waste, measuring and improving program performance, identifying fraudulent claims for payment, adjusting resource allotments, predicting crime patterns and locations, identifying terrorist activities, tracking individual terrorists, controlling aviation traffics, etc. In no doubt data mining becomes an essential tool for homeland safety and security, marketing, process control, manufacturing, network detection, and many others. Information analysts can provide a reasonable level of assurance to their results for commercially and otherwise viable through data mining efforts.

V. META-LEARNING

There are the two major limitations to traditional view of learning (i.e. base-learning approach):

- Data patterns are usually embedded in the predictive model itself, successive training of the same learner over the same data fails to acquire any additional knowledge.
- There is no easy way of extracting, sharing and reusing acquired knowledge among different domains or analysts.

Meta-learning overcomes the problem by intelligent data mining processes with the ability to learn and adapt based on previously acquired experience. This limits the amount of user input necessary to perform informed data analysis task, which may be good either for running multiple tasks at once without overwhelming the analyst, or for automatic decision making without any need for user intervention when the user himself may lack the expertise. Moreover, such system can learn from every new task, thus being more experienced and informed over time, providing new levels of adaptation to newly introduced obstacles[4]

The primary goal of meta-learning is the understanding of the interaction between the mechanism of learning and the concrete contexts in which that mechanism is applicable. Learning at the meta-level is concerned with accumulating experience on the performance of multiple applications of a learning system. The main aim of current research is to develop meta-learning assistant, which are able to deal with the increasing number of models and techniques, and give advice dynamically on such issues as model selection and method combination.

Meta-learning differs from *base-learning* in the scope of the level of adaptation; whereas learning at the base-level is based on accumulating experience on a specific learning task (e.g., credit rating, medical diagnosis, mine-rock discrimination, fraud detection, etc.), learning at the meta-level is based on accumulating experience on the performance of multiple applications of a learning system. If a base-learner fails to perform efficiently, one would expect the learning mechanism itself to adapt in case the same task is presented again. Meta-learning is then important in understanding the interaction between the mechanism of learning and the concrete contexts in which that mechanism is applicable. Briefly stated, the field of meta-learning is focused on the relation between tasks or domains and learning strategies. In that sense, by learning or explaining what causes a learning system to be successful or not on a particular task or domain, we go beyond the goal of producing more accurate learners to the additional goal of understanding the conditions (e.g., types of example distributions) under which a learning strategy is most appropriate. According to[5], there are several basic applications of meta-learning:

- Selecting and recommending machine learning algorithms,
- Employing meta-learning in KDD,
- Employing meta-learning to combine base-level machine learning systems,
- Control of the learning process and bias management,
- Transfer of meta-knowledge across domains

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

VI. DISTRIBUTED DATA MINING

With increasing globalization, organizations operate on multiple locations and collaborate with each other that lead to sharing of data at distributed data sources. Further, intense competition in business warrants a pressing need for information-rich decision-making that demands cohesive and integrated knowledge from massive distributed data which is often high dimensional in nature. Consequently, data mining systems tend to shift from centralized and stand-alone ones to complex distributed systems. However, data distribution over a network with limited bandwidth coupled with data security and privacy make centralized data mining process infeasible. Thus distributed data mining (DDM) emerges as a natural requirement. It is a framework to mine mission-critical information paying careful attention to distributed data and computing resources available at the distributed data sources [7]. DDM deals with the problem of extracting hidden patterns from the distributed data by performing local data analysis at individual data sources for generating partial data models, and combining these partial data models in order to develop a global model at a location where decision-making takes place. Meta-learning is used for this purpose[8]. Thus distributed data mining process is complex and involves many techniques and iterative steps. This motivates parallel and distributed computation in the distributed data mining system [9]. A general approach to DDM is depicted in Figure-2.2.

Recently, multi-agent system has been acknowledged as a powerful technology to develop distributed systems in general and distributed data mining systems in particular [10]. Agents being capable of flexible autonomous actions in a dynamic and open environment are very attractive for distributed data mining. Additionally, an agent's ability to move from one data location to another elegantly handles local data analysis task and avoids the need to transfer huge quantity of data over the network that causes high communication overhead and bandwidth issues[1].

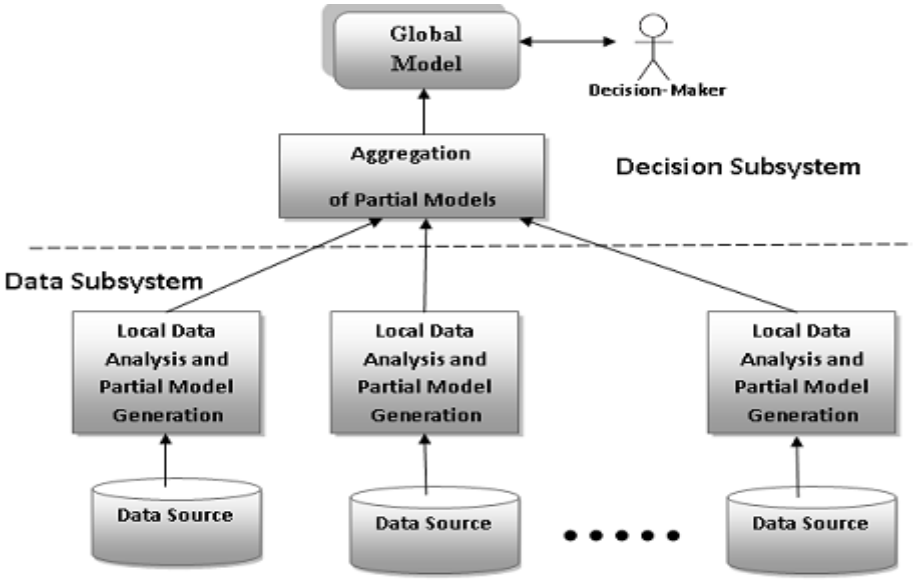

Figure-2.2: A Generic Framework of DDM

VII. AGENTS

Agents are usually defined as computer entities that are suited to some environment, and that are able to carry out their tasks in an autonomous manner. The environments in which agents operate may be dynamic, unpredictable, and uncertain; thus it is desirable for agents to exhibit some form of computer intelligence in order to interact with their environment. There are a number of capabilities that intelligent agents are expected to display[12]. Reactivity: Intelligent agents should be able to anticipate tasks, according to their environment, so that they can respond in a timely fashion.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

- Proactiveness: Intelligent agents should be able to explore alternatives to achieve their design objectives.
- Social ability: Intelligent agents should be capable of interacting with other agents (and possible humans) in order to satisfy their design objectives.

It is self-evident that the above capabilities can increase the degree of autonomy with which agents can perform tasks.

VIII. MULTI-AGENT SYSTEMS (MASs)

A Multi-Agent System (MAS) is a software system that employs a number of interactive agents to solve a problem in open and decentralized uncertain environments. A central feature of MAS is that there is no centralized control mechanism; agents are required to collaborate to achieve the design objective of a given MAS. A MAS has collective capabilities that an individual agent does not have. Thus, as listed in [6]:

- In a MAS, computational resources and capabilities are distributed across a network of interconnected agents to solve problems that are too large for an individual agent. A centralized system may be plagued by resource limitations, performance bottlenecks, or critical failures.
- A MAS allows for the interconnection and interaction of multiple existing legacy systems; by building an agent wrapper around such systems, so that they can be incorporated into an agent society.
- In a MAS, problems are modeled in terms of autonomous interacting component agents that allows these agents to operate in self directed manner.
- In a MAS information from sources that are spatially distributed is efficiently retrieved, filtered, and globally coordinated.
- Use of a MAS provides solutions in situations where expertise is spatially and temporally distributed. With respect to the social ability of agents, expertise and resources can be shared.
- Use of MAS enhances overall system performance, especially along the dimensions of: computational efficiency, reliability, extensibility, robustness, maintainability, responsiveness, flexibility and reuse due to its distributed nature.

IX. CONCLUSION

The decentralized control and autonomy properties of MAS can be argued to offer the most significant advantages with respect to the data mining process in that they allow individual agents to collectively process large amounts of data. Individual agents can typically process sub-sets of the data and then combine the local results to produce the desired global result, thus achieving computational efficiency advantages. This local processing also offers the advantage of privacy preservation in data mining situations where this is a requirement. The advantages offered by MAS with respect to expertise and resource sharing are clearly of significance. For many data mining activities it is well established that there is no single best algorithm suited to all types of data. A data mining MAS where individual agents are equipped with different data mining algorithms which produce separate results from which the best can be selected, is therefore clearly another benefit.

REFERENCES

1.Dasilva J, Giannella C, Bhargava R, Kargupta H, and Klusch M (2005), "Distributed data mining and agents", *Engineering Applications of Artificial Intelligence*, 18(7), pp.791–807.
2.Klosgen W and Zytkow J M (eds.) (2002), Handbook of data mining and knowledge discovery, OUP, Oxford.
3.Kantardzic M (2003), Data mining: concepts, models, methods, and algorithms, John Wiley, New Jersey.
4.Ricardo Vilalta, Christophe Giraud-Carrier, Pavel Brazdil, Carlos Soares (2004), Using Meta-Learning to Support Data Mining, International Journal of Computer Science & Applications, Vol. I, No. 1, pp. 31-45.
5.Soares, C., Giraud-Carrier, C., Brazdil, P., Vilalta, R. (2009), Meta-learning: Applications to Data Mining, Springer
6.Sycara K (1998), Multiagent systems, AI Magazine, 19(2), pp.79-92.
7.Park B and Kargupta H (2003), "Distributed Data Mining: Algorithms, Systems, and Applications," in *The Handbook of Data Mining* (N. Ye edited), Lawrence Erlbaum Associates, pp. 341–361.
8. Vilalta R. and Drissi Y. (2002), "A Perspective View and Survey of Meta-Learning", *Journal of Artificial Intelligence Review*, 18 (2), pp.77-95.
9. Zaki M, and Pan Y (2002), "Introduction: Recent developments in parallel and distributed data mining", *Journal of Distributed Parallel Databases*, 11 (2), pp.123–127.
10.Wooldridge M (2002), *An Introduction to Multi Agent Systems*, John Wiley & Sons Ltd

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

11. Zeng L., Li L., Duan L., Lu K., Shi Z., Wang M., Wu W., and Luo, P.(2012), Distributed Data Mining: A Survey, *Information Technology and Management*, 13(4), pp.403–409.
12. Wooldridge M and Jennings N. (1995), Intelligent agents: Theory and practice, Knowledge Engineering Review, 10(2), pp.115-152.

BIOGRAPHY

Sanjay Kumar Sen is a Research Scholar in the Computer Science & Engg., Department, of PG Department of Computer Science & Application, Utkal University, Vani Vihar, Bhubaneswar, Odisha, India. He received M.Tech degree in 2005 from School of Computer Science & Application , Utkal Univesity, Vanivihar, Bhubaneswar, Odisha, India. His Interest is Data mining,