International Journal of Scientific Research and Modern Education (IJSRME)
Impact Factor: 6.225, ISSN (Online): 2455 – 5630
(www.rdmodernresearch.com) Volume 2, Issue 1, 2017

A SHORT-LIVED LEARNING ON DATAMINING CONCEPTS

D. Sudha* & R. Senthilkumar**
* Assistant Professor, Department of Computer Science, A.V.C College (Autonomous), Mayiladuthurai, Tamilnadu
** Assistant Professor, Department of Computer Science, Dharmapuram Adhinam Arts College, Mayiladuthurai, Tamilnadu

Cite This Article: D. Sudha & R. Senthilkumar, "A Short-Lived Learning on Datamining Concepts", International Journal of Scientific Research and Modern Education, Volume 2, Issue 1, Page Number 192-196, 2017.

Abstract:

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it. Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc To analyze, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields. This paper divulges more number of applications of the data mining and also focuses scope of the data mining which will helpful in the further research.

Key Words: Data Mining, KDD, Machine Learning, Prediction, Training Set & Prediction Set

1. Introduction:

. In the present situation there is enormous amount of data being collected and stored in databases everywhere across the world. It is not difficult to find the repositories with Terabytes of data in organizations and research fields. There is huge collection of data present and it is very difficult to extract important pieces of information out of it and without automatic extraction methods this information is practically impossible to mine. Year after year many algorithms were created to extract important information from large sets of data. There are different methodologies to approach this problem like classification rule, association rule, clustering, etc. The input for the classification is the training data set, whose class labels are already known. Classification analyse the training data set and constructs a model based on the class label, and aims to assign class label to the future unlabeled records. Since the class field is known, this type of classification is known as supervised learning.

2. Literature Review:

Recently many Data mining research were done in the various domains, such as Mobile commerce. Paper [1] proposed cluster-Based Temporal Mobile Sequential pattern mine (CTMSP-Mine) to discover the cluster-Based Temporal Mobile sequential pattern. They used techniques two techniques such as Co-smart-cast algorithm to cluster the mobile transaction sequences. In this algorithm, they proposed LBS-alignment to evaluate similarity of mobile transaction sequences and GA-based time segmentation algorithm to find the most suitable time intervals. After clustering and segmentation user cluster table and time interval table are generated. CTMSP-Mine algorithm to mine CTMSPs from mobile transaction database according to user cluster table and time table. In online, they predict subsequent behaviors according to user's previous mobile transaction sequences and current time mining an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data Mining is widely used in diverse areas. There are number of commercial data mining system available today yet there are many challenges in this field.

Paper [2] deals Students Mood recognition during online self-assessment test .They used exponential logic and its formulas for computation. Student's previous answers and slide bar status are considered as input. Total Number of questions for online self-assessment test, Student's goal, and slide bar value are used as variables for exponential logic. This system identifies student's current status of mood and gives appropriate feedback. Limitation of this system is student's manually selecting their mood using slide bar without any automation.

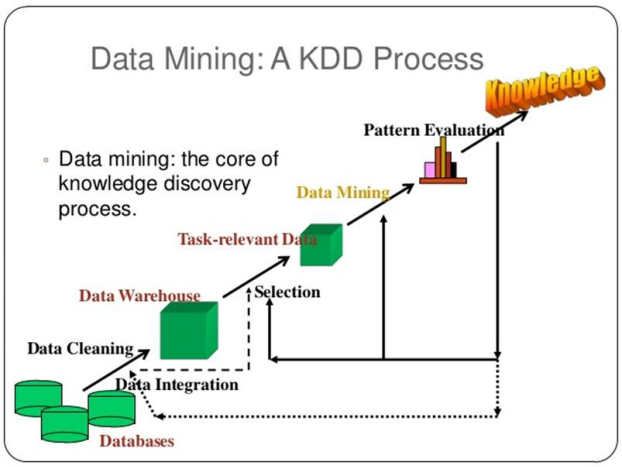
Paper [3] focused on how to improve aspect- level opinion mining for online customer reviews. They proposed the Joint Aspect/Sentiment model (JAS) to extract aspects and aspect dependent sentiment lexicons from online customer reviews in a unified framework. They used Gibbs Sampling algorithm.

In Paper [4] a novel weakly supervised cybercriminal network mining method which can uncover both explicit and implicit relationships among cybercriminals based on their conversational messages posted on online social media. Mined two types of semantics such as transactional and collaborative relationships among cybercriminals using context-sensitive Gibbs sampling algorithm. They used probabilistic generative model to extract multi-word expressions describing two types of cyber-criminal relationships in unlabeled messages. They used concept level approaches to better grasp the implicit semantics associated with text.

Research [5] focused on classifying internet users based on their internet user behavior to offer him/her enhanced services. To do this, they have collected data such as Timestamp, IP address, URL and 10 keywords from proxy Cache and origin web server and categorized user behavior. To cluster users, two kinds of categorization algorithms are used such as "hard clustering" producing a partition, and an algorithm of "soft clustering" discovering overlapping clusters. The first algorithm is a method of hierarchical agglomerative clustering (HAC).

3. KDD Process:

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1) shows data mining as a step in an iterative knowledge discovery process.



4. Data Mining Applications:

Data mining is highly useful in the following domains –

- ✓ Market Analysis and Management
- ✓ Corporate Analysis & Risk Management
- ✓ Fraud Detection

Apart from these, data mining can also be used in the areas of production control[6], customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid.

4.1 Market Analysis and Management:

Listed below are the various fields of market where data mining is used −

Customer Profiling − Data mining helps determine what kind of people buy what kind of products.

Identifying Customer Requirements − Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.

Cross Market Analysis − Data mining performs association/correlations between product sales.

Target Marketing − Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.

Determining Customer purchasing pattern − Data mining helps in determining customer purchasing pattern.

Providing Summary Information − Data mining provides us various multidimensional summary reports.

4.2 Corporate Analysis and Risk Management:

Data mining is used in the following fields of the Corporate Sector –

Finance Planning and Asset Evaluation − It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.

Resource Planning − It involves summarizing and comparing the resources and spending.

Competition − It involves monitoring competitors and market directions.

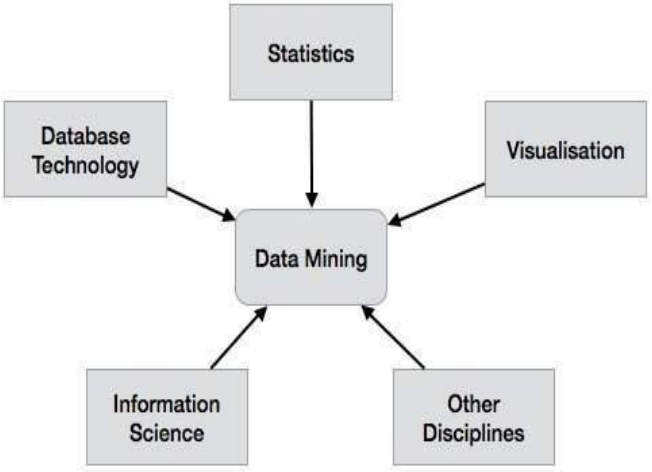
4.3 Fraud Detection:

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms.

5. Data Mining System Classification:

A data mining system can be classified according to the following criteria [7] –

- ✓ Database Technology
- ✓ Statistics
- ✓ Machine Learning
- ✓ Information Science
- ✓ Visualization
- ✓ Other Disciplines



Apart from these, a data mining system can also be classified based on the kind of (a) databases mined, (b) knowledge mined, (c) techniques utilized, and (d) applications adapted.

5.1 Classification Based on the Databases Mined:

Database system can be classified according to different criteria such as data models, types of data, etc. And the data mining system can be classified accordingly. For example, if we classify a database according to the data model, then we may have a relational, transactional, object-relational, or data warehouse mining system.

5.2 Classification Based on the kind of Knowledge Mined:

We can classify a data mining system according to the kind of knowledge mined. It means the data mining system is classified on the basis of functionalities such as –

- ✓ Characterization
- ✓ Discrimination
- ✓ Association and Correlation Analysis
- ✓ Classification
- ✓ Prediction
- ✓ Outlier Analysis
- ✓ Evolution Analysis

5.3 Classification Based on the Techniques Utilized:

We can classify a data mining system according to the kind of techniques used. We can describe these techniques according to the degree of user interaction involved or the methods of analysis employed.

5.4 Classification Based on the Applications Adapted:

We can classify a data mining system according to the applications adapted. These applications are as follows –

- ✓ Finance
- ✓ Telecommunications
- ✓ DNA
- ✓ Stock Markets
- ✓ E-mail

6. Data Mining - Cluster Analysis:

Clustering [8] is the process of making a group of abstract objects into classes of similar objects.

6.1 Applications of Cluster Analysis:

- ✓ Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.

- ✓ Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- ✓ In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- ✓ Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- ✓ Clustering also helps in classifying documents on the web for information discovery.
- ✓ Clustering is also used in outlier detection applications such as detection of credit card fraud.
- ✓ As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

6.2 Requirements of Clustering in Data Mining:

The following points throw light on why clustering is required in data mining –

- ✓ Scalability − We need highly scalable clustering algorithms to deal with large databases.
- ✓ Ability to deal with different kinds of attributes − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- ✓ Discovery of clusters with attribute shape − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- ✓ High dimensionality − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- ✓ Ability to deal with noisy data − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- ✓ Interpretability − The clustering results should be interpretable, comprehensible, and usable.

6.3 Clustering Methods:

Clustering methods can be classified into the following categories –

- ✓ Partitioning Method
- ✓ Hierarchical Method
- ✓ Density-based Method
- ✓ Grid-Based Method
- ✓ Model-Based Method
- ✓ Constraint-based Method

6.3.1 Partitioning Method:

Suppose we are given a database of ‘n’ objects and the partitioning method constructs ‘k’ partition of data. Each partition will represent a cluster and $k \leq n$. It means that it will classify the data into k groups, which satisfy the following requirements −

- ✓ Each group contains at least one object.
- ✓ Each object must belong to exactly one group.

6.3.2 Hierarchical Method:

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here –

- ✓ Agglomerative Approach
- ✓ Divisive Approach

6.3.3 Density-Based Method:

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

6.3.4 Grid-Based Method:

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantage:

- ✓ The major advantage of this method is fast processing time.
- ✓ It is dependent only on the number of cells in each dimension in the quantized space.

6.3.5 Model-Based Method:

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points. This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

6.3.6 Constraint-Based Method:

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

7. Conclusion:

This paper gives a general introduction of data mining, the process of discovering interesting knowledge from large amounts of data stored in information repositories. It is also shown that data mining technology can be used in many areas in real life including biomedical and DNA data analysis, financial data analysis, the retail industry and also in the telecommunication industry. One of the biggest challenges for data mining technology is managing the uncertain data which may be caused by outdated resources, sampling errors, or imprecise calculation. From this study, we could conclude that data mining is an ever growing research field with interdisciplinary applications increasing.

8. References:

1. Eric Hsueh-Chan Lu, Vincent S. Tseng, Member, IEEE, and Philip S. Yu, Fellow, IEEE "Mining Cluster-Based Temporal Mobile Sequential Patterns in Location- Based Service Environments" IEEE Transac Tions On Knowledge And Data Engine Ering, Vol. 23, NO. 6, JUNE 2011.
2. Ranjeeta Rana, Mrs. Vaishali Kolhe, "Analysis of Students Emotion for Twitter Data using Naïve Bayes and Non Linear Support Vector Machine Approaches" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 5 3211 – 3217, May 2015,
3. Pratima More, Archana Ghotkar," A Study of Different Approaches to Aspect-based Opinion Mining" International Journal of Computer Applications (0975 – 8887) Volume 145 – No.6, July 2016
4. Parvathy G., Bindhu J.S, "A Probabilistic Generative Model for Mining Cybercriminal Network from Online Social Media: A Review", International Journal of Computer Applications (0975 - 8887) Volume 134 - No.14, January 2016.
5. Simpa Jindal, "A Data Mining Approach for Evaluating Usage of Internet amongst Various Universities", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 6, June 2016.
6. M. J. Shaw et al/Decision Support Systems 31 2001 127–137 "Knowledge management and data mining for marketing"
7. Megha Gupta, Naveen Aggarwal, "Classification techniques analysis", NCCI 2010 -National Conference on Computational Instrumentation CSIO Chandigarh, INDIA, 19-20 March 2010.
8. L.V. Bijuraj, "Clustering and its Applications", Proceedings of National Conference on New Horizons in IT - NCNHIT 2013.