Available online at www.worldscientificnews.com

World Scientific News

An International Scientific Journal

WSN 41 (2016) 267-274 EISSN 2392-2192

A Review on Data Mining Algorithms for Internet of Things

M. Bhuvaneswari[1], K. Merlin Jeba[2], V. Srividhya[3]

Avinashilingam University, Coimbatore, India

[1-3]E-mail address: bhuvana.msc.mphil@gmail.com , merlin.mphil@gmail.com , vidhyavasu@gmail.com

ABSTRACT

The fields of computer science and electronics have merged to result into one of the most notable technological advances in the form of realization of the Internet of Things (IoT). Internet of Things (IoT) is an innovative idea which will transform the real world objects into intelligent virtual objects in the near future. Nowadays internets of things are generating an enormous data, that data are applied in data mining to extract the knowledge. To many, the massive data generated or captured by IoT are considered having highly useful and valuable information. Some data mining algorithms are used to analyze the potential data in internet of things.This paper begins with a discussion of the IoT and then, it describes about three data mining algorithms such as clustering, classification, and association rules. At last the application areas of Internet of things are discussed in this paper.

***Keywords*:** Internet of things, IoT, Clustering, Association Rule, Classification, IoT applications

1. INTRODUCTION

The Internet of Things will produce large volumes of data. IoT and its relevant technologies are integrating classical networks with networked instruments and devices. It has been playing an essential role ever since it appeared which covers from traditional tools to

(Received 01 February 2016; Accepted 10 February 2016; Date of Publication 11 February 2016)

general household objects [1] and has been attracting the attention of researchers from academia, medical, industry, and government in recent years. There is a great vision that all things can be easily controlled and observed, and it can communicate with each other through internet, and can even make decisions among themselves [2]. In order to make IoT smarter, many of analysis technologies are introduced into IoT; one of the main technology is data mining.

Data mining involves discovering novel, andpotentially useful patterns from large data sets and applying algorithms to the extraction of hidden information. The aim of any data mining process is to build an efficient predictive or descriptive model of a large amount of data. Based on a broad view of data mining algorithms, data mining is the process of discovering potential knowledge from large amounts of data stored in either databases, data warehouses, or other information repositories.

2. DATA MINING ALGORITHMS

The classification, clustering, association analysis are the main data mining functionalities.

(i) **Classification** is the process of finding a set of models or functions that describe and distinguish data classes or concepts, for the purpose of predicting the class of objects whose target label is unknown.

(ii) **Clustering** analyzes data objects without consulting a known target model.

(iii) **Association** analysis is the discovery of association rules displaying attribute-value conditions that commonly occur together in a given set of data.

2. 1. Classification

Nowadays a major part of decision-making with the help classification algorithm is possible. Given an object, assigning it to one of predefined target categories or is called classification. The goal of classification is to accurately predictthe target class for each case in the data [3]. For example, a classification model could be used to select loan applicants as low, medium, or high credit risks [4]. There are a lot of methods to classify the data includes. The research structure of classification is shown in Figure 1.

(i) **Decision tree** is a flowchart type structure that has nodes and edges. The nodes are either root or internal nodes. The root is a special node and internal nodes represent conditions that are used to test the attribute values. The terminal nodes represent the classes.

(ii) Classification and Regression Trees (CART) is a nonparametric decision tree algorithm. It produces either classification or regression trees, based on whether the response variable is categorical or continuous. CHAID (chi-squared automatic interaction detector) and the improvement researcher [5] focus on dividing a data set into exclusive and exhaustive segments that differ with respect to the response variable.

(iii) The KNN (*K*-Nearest Neighbor) algorithm is introduced by the Nearest Neighbor algorithm which is designed to find the nearest point of the observed object. The main idea of the KNN algorithm is to find the *K*-nearest points [6].

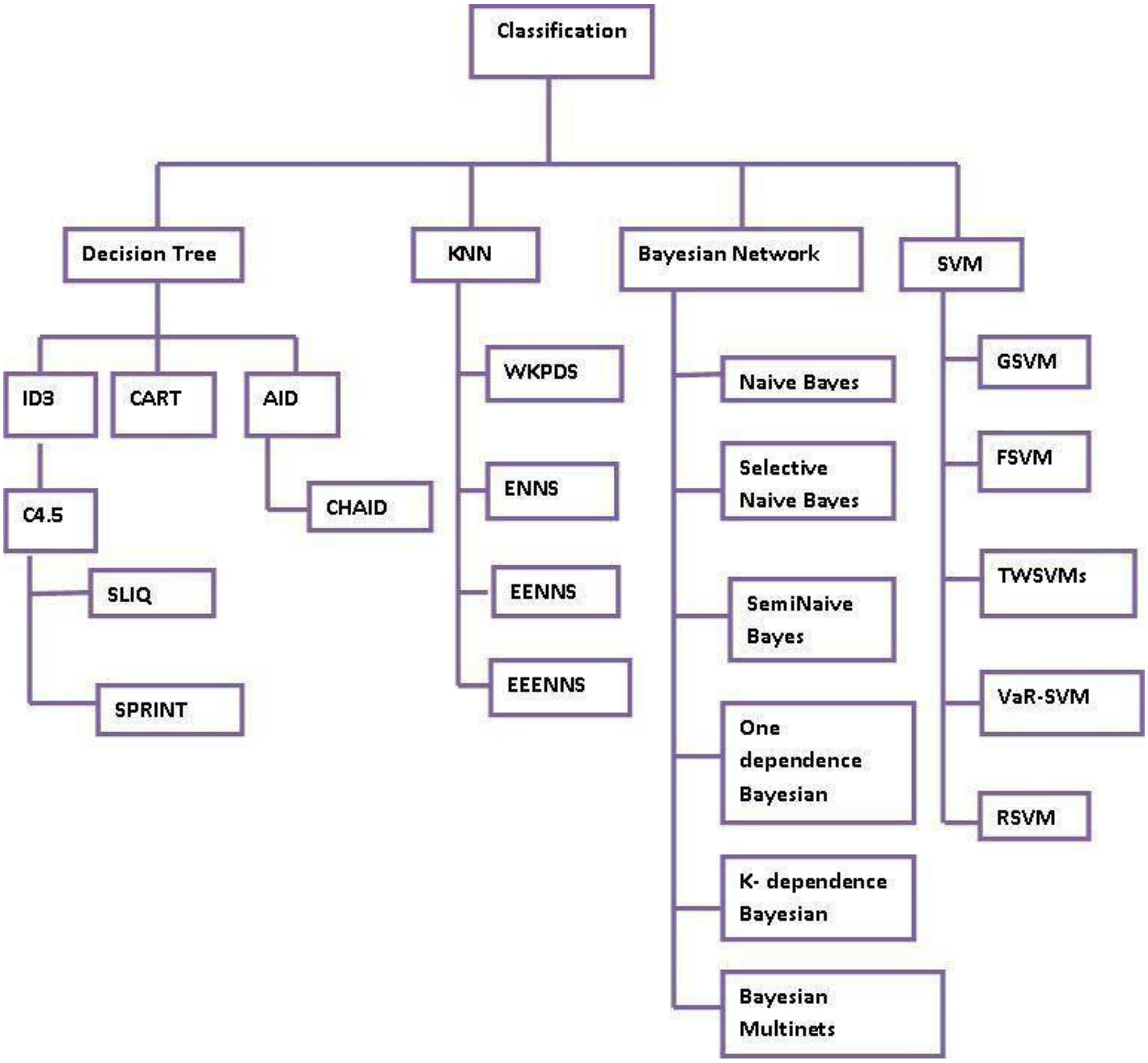

Figure 1. The research structure of classification

(iv) Bayesian networks are directed acyclic graphs whoseeach nodes represent random variables in the Bayesian sense. Edges represent conditional dependencies; nodes which are not connected represent variables which are conditionally independent of each other. Based on Bayesian networks, these classifiers have much strength, like model interpretability and accommodation to complex data and classification problem settings [7].

(v) Support Vector Machines algorithm is supervised learning model with associated learning algorithms that analyze data and recognize patterns, which is based on statistical learning theory. SVM produces a binary classifier, the so-called optimal separating hyper planes, through an extremely nonlinear mapping of the input vectors into the high-dimensional feature space. SVM is widely used in text classification [8], marketing, pattern recognition, and medical diagnosis [9].

2. 2. Clustering

Clustering algorithms divide data into meaningful groups represented in Figure 2. The patterns in the same group are similar in some sense and patterns in different group are dissimilar in the same sense. Searching for clusters involves unsupervised learning [10]. In information retrieval, for example, the search engine clusters billions of web pages into different groups, such as news, reviews, videos, and audios. One straightforward example of clustering problem is to divide points into different groups [11]. The research structure of clustering is revealed in the following Figure 2

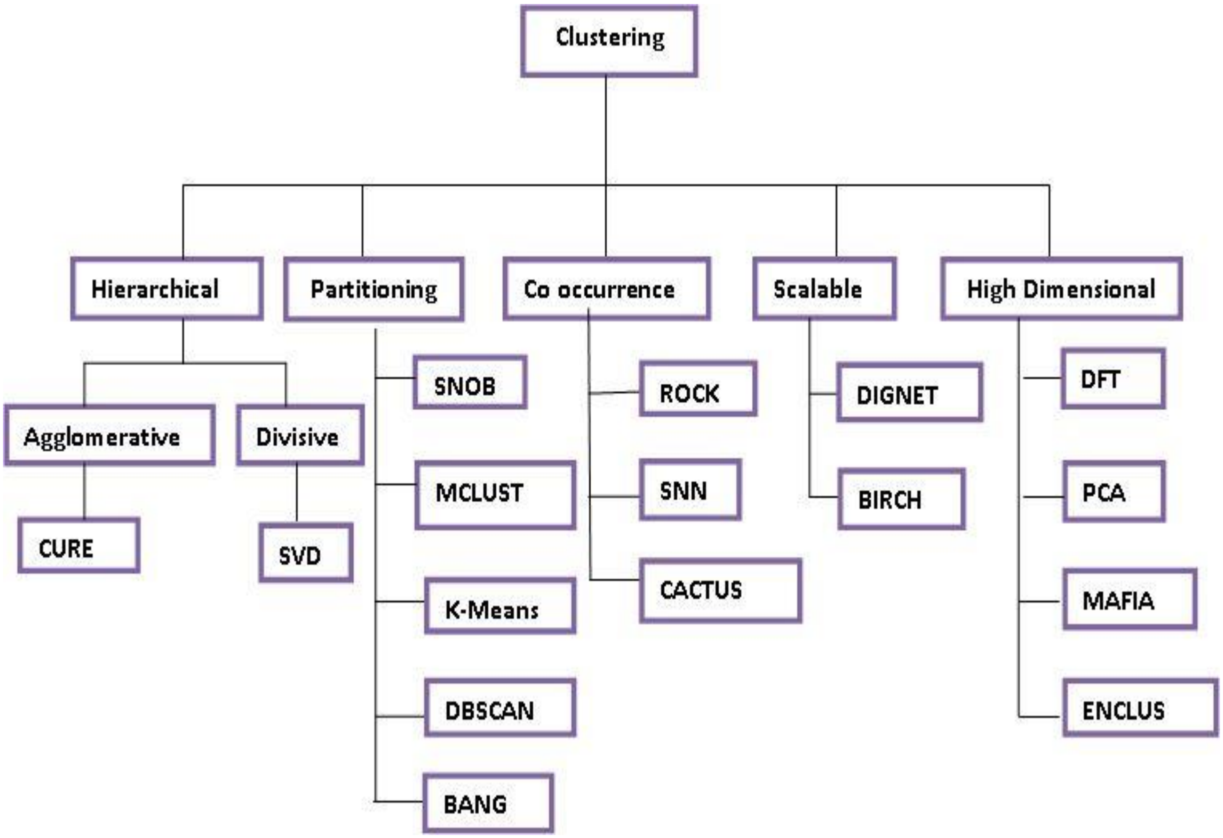

Figure 2. The research structure of clustering

(i) Hierarchical clustering method combines data objects into subgroups; those subgroups merge into larger and high level groups and so forth and form ahierarchy tree. Hierarchical clustering methods have two classifications, agglomerative (bottom-up) and divisive (top-down) approaches.

(ii) Partitioning algorithms discover clusters either byiteratively relocating points between subsets or by identifying areas heavily populated with data. Density-based partitioning methods attempt to discover low-dimensional data, which is dense connected, known as spatial data.

(iii) Co-occurrence in order to handle categorical data, researchers, change data clustering to pre clustering of items or categorical attribute values.

(iv) Scalable clustering research faces scalability problemsfor computing time and memory requirements.

(v) High dimensionality data clustering methods are designed to handle data with hundreds of attributes.

2. 3. Association Analysis

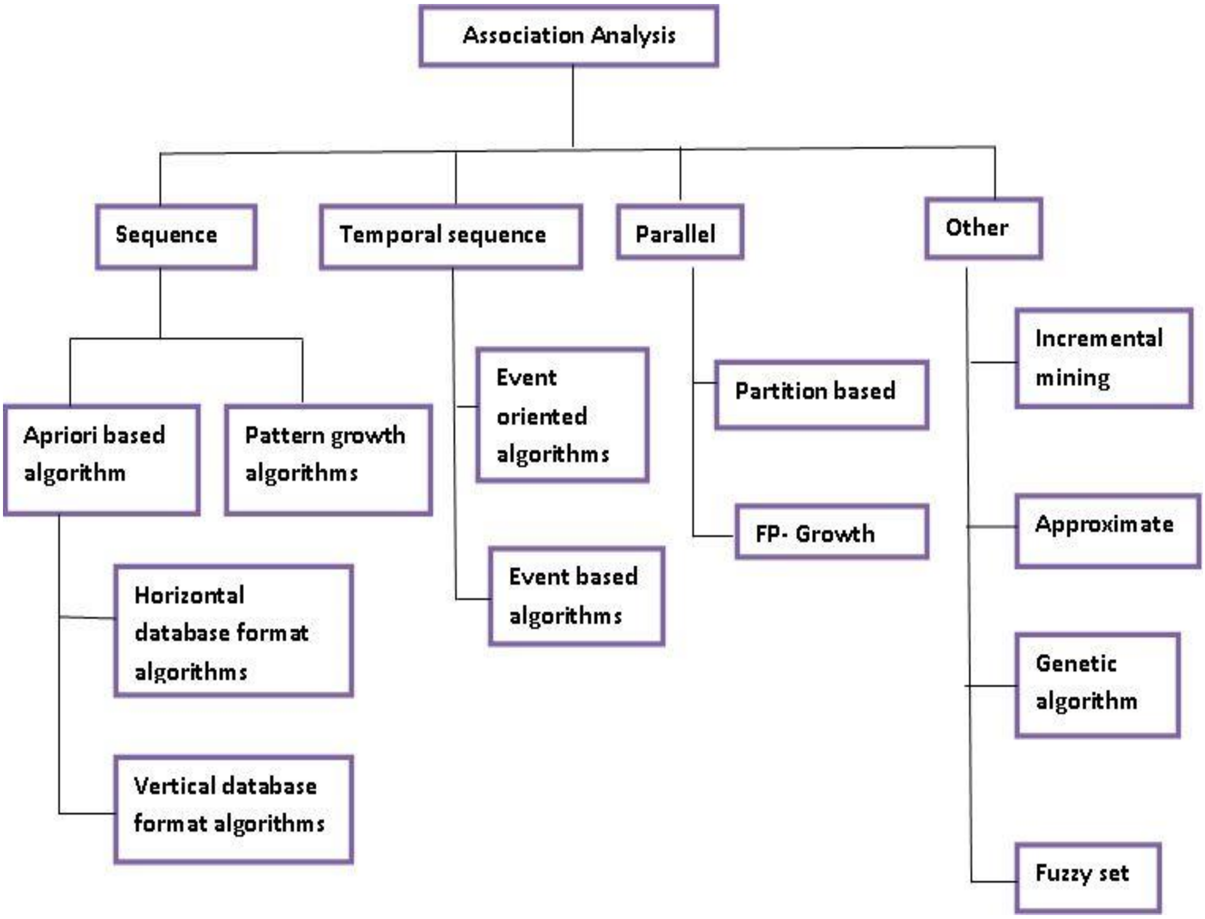

Figure 3. The research structure of association analysis

Association rule mining [12] focuses on the market basket analysis or transaction data analysis, and it targets discovery of rules showing attribute value associations that occur frequently and also help in the generation of more general and qualitative knowledge which in turn helps in decision making [13]. The research structure of association analysis is exposed in Figure 3.

(i) For the first catalog of association analysis algorithms, the data will be processed sequentially. The apriori based algorithms have been used to discover intra transaction associations and then discover associations; there are lots of extension algorithms.

(ii) In some area, the data would be a flow of events and therefore the problem would be to discover event patterns that occur frequently together. It divides into 2 parts: event-based algorithms and event-oriented algorithms

3. APPLICATION AREAS FOR THE INTERNET OF THINGS

From building and home automation to wearables, the IoT touches every facet of our lives. In this identified six key areas for the IoT with potential for exponential growth. The main application areas for the IoT are revealed in Figure 4.

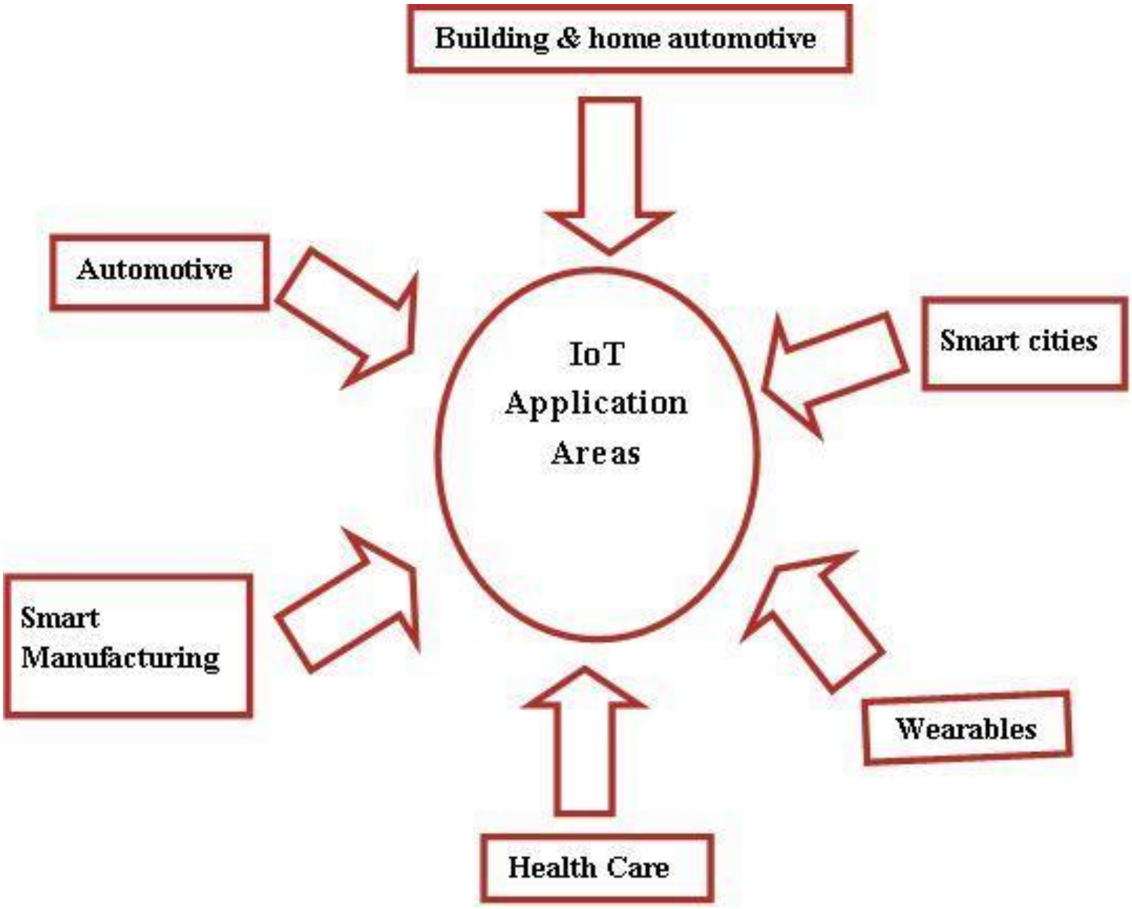

Figure 4. Application areas for IoT

(i) Building and Home automation

Home automation is an appealing context for the Internet of Things (IoT). This is where a home's electrical devices are connected to a central system that automates those devices based on user input. For example, push a button and the shades go up, or give a voice command and lights turn on. IoT is the magic dust that turns the automated home into the smart home.

(ii) Smart cities

Smart City is the product of accelerated development of the new generation information. The main features of a smart city include a high degree of information technology integration and a comprehensive application of information resources. The essential components of urban development for a smart city should include smart technology, smart industry, smart services, smart management and smart life.

(iii) Smart manufacturing

Manufacturing worldwide is on the cusp of a revolution. Technologies based on the IoT have the potential to radically improve visibility in manufacturing to the point where each unit of production can be "seen" at each step in the production process. Batch-level visibility is being replaced by unit-level visibility. This is the dawn of smart manufacturing.

(iv) Wearables

IoT is interconnecting uniquely identifiable devices through Internet. It is the next evolution where devices can interact with other devices. In IoT paradigm, devices can represent itself digitally, no longer the device will just relate but will be connected to surrounding devices and database data.With the broadest portfolio in the industry provides highly efficient ultra-low power solutions for the wearable's market.

(v) Health care

The Internet of Things could be a game changer for the healthcare industry. It is transforming healthcare industry by increasing efficiency, lowering costs and put the focus back on better patient care.

4. CONCLUSION

The Internet of Things concept arises from the need to manage, automate, and explore all devices, instruments, and sensors in the world. In order to make wise decisions both for people and for the things in IoT, data mining technologies are integrated with IoT technologies for decision making support and system optimization. Data mining involves discovering novel, interesting, and potentially useful patterns from data and applying algorithms to the extraction of hidden information. In this paper, reviewed the three data mining algorithms are classification, clustering and association. The application areas for the IoT are also discussed. The major application area are Building and Home automation, Smart cities, Smart manufacturing, Wearable's, Health care.

References

[1] Q. Jing, A. V. Vasilakos, J. Wan, J. Lu, and D. Qiu, Security of the internet of things: perspectives and challenges. *Wireless Networks*, vol. 20, no. 8, pp. 2481-2501, 2014

[2] C.-W. Tsai, C.-F. Lai, and A. V. Vasilakos, Future internet of things: open issues and challenges. *Wireless Networks*, Vol. 20, no. 8, pp. 2201-2217, 2014

[3] G. Kesavaraj and S. Sukumaran. A study on classification techniques in data mining. in *Proceedings of the 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT '13)*, pp. 1-7, July 2013

[4] S. Song, Analysis and acceleration of data mining algorithms on high performance reconfigurable computing platforms [Ph.D. thesis], Iowa State University, 2011.

[5] M. van Diepen and P. H. Franses, Evaluating chi-squared automatic interaction detection. *Information Systems*, Vol. 31, no. 8, pp. 814-831, 2006

[6] D. T. Larose. k-nearest neighbor algorithm, in *DiscoveringKnowledge in Data: An Introduction to DataMining*, pp. 90-106, JohnWiley & Sons, 2005

[7] C. Bielza and P. Larranaga, Discrete bayesian network classifiers: a survey. *ACM Computing Surveys*, Vol. 47, no. 1, article 5, 2014

[8] T. Joachims, Text categorizationwith support vector machines: learning with many relevant features, in *Machine Learning: ECML-98*, Vol. 1398, pp. 137-142, Springer, Berlin, Germany, 1998.

[9] L. Yingxin and R. Xiaogang, Feature selection for cancer classification based on support vector machine. *Journal of Computer Research and Development*, Vol. 42, no. 10, pp. 1796-1801, 2005

[10] S. Ansari, S. Chetlur, S. Prabhu, G. N. Kini, G. Hegde, and Y. Hyder. An overview of clustering analysis techniques used in data mining. *International Journal of Emerging Technology and Advanced Engineering*, Vol. 3, no. 12, pp. 284-286, 2013

[11] R. Agrawal, T. Imieli ´nski, and A. Swami, Mining association rules between sets of items in large databases, in *Proceedings of the ACMSIGMOD International Conference on Management of Data (SIGMOD '93)*, pp. 207-216, 1993

[12] A. Gosain andM. Bhugra, A comprehensive survey of association rules on quantitative data in data mining, in *Proceedings of the IEEE Conference on Information & Communication Technologies (ICT '13)*, pp. 1003-1008, JeJu Island, Republic of Korea, April 2013