Applied Computer Science, vol. 15, no. 3, pp. 73–86
doi:10.23743/acs-2019-23

Submitted: 2019-07-16
Revised: 2019-08-08
Accepted: 2019-09-03

Data Mining, classification, clustering, association, regression, algorithms bottleneck

Workineh TESEMA[*]

INEFFICIENCY OF DATA MINING ALGORITHMS AND ITS ARCHITECTURE: WITH EMPHASIS TO THE SHORTCOMING OF DATA MINING ALGORITHMS ON THE OUTPUT OF THE RESEARCHES

Abstract

This review paper presents a shortcoming associated to data mining algorithm(s) classification, clustering, association and regression which are highly used as a tool in different research communities. Data mining researches has successfully handling large amounts of dataset to solve the problems. An increase in data sizes was brought a bottleneck on algorithms to retrieve hidden knowledge from a large volume of datasets. On the other hand, data mining algorithm(s) has been unable to analysis the same rate of growth. Data mining algorithm(s) must be efficient and visual architecture in order to effectively extract information from huge amounts of data in many data repositories or in dynamic data streams. Data visualization researchers believe in the importance of giving users an overview and insight into the data distributions. The combination of the graphical interface is permit to navigate through the complexity of statistical and data mining techniques to create powerful models. Therefore, there is an increasing need to understand the bottlenecks associated with the data mining algorithms in modern architectures and research community. This review paper basically to guide and help the researchers specifically to identify the shortcoming of data mining techniques with domain area in solving a certain problems they will explore. It also shows the research areas particularly a multimedia (where data can be sequential, audio signal, video signal, spatio-temporal, temporal, time series etc) in which data mining algorithms not yet used.

[*] Jimma University, Faculty of Computing, Department of Information Technology, Jimma, Ethiopia, workineh.tesema@ju.edu.et

1. INTRODUCTION

Data mining is a process of exploring huge data, typically business related data which is called big data (Rehman, 2017). This process is performed to find hidden patterns and relationship present in the data. The overall objective of the data mining process is to extract knowledge from a large dataset and transform it into a comprehensible structure for further use (Bavisi, Mehta & Lopes, 2014). Data mining is the process of automatically finding implicit, previously unknown, and potentially useful information from large volumes of data (Yadav, Wang & Kumar, 2013). Therefore, the role of data mining algorithms has become vital to researchers in science, medicine, business, and security domains. Recent advances in data extraction techniques have resulted in tremendous increase in the input data size of data mining applications (Kalyani, Bharathi & Rao, 2016). Data mining algorithm is the tool that involves retrospective analysis to extract diamonds of knowledge from historical data and predict outcome of the future (Talia, Trunfio & Marozzo, 2016). Data mining can automate the process of finding patterns and relationships in raw data and the results can be utilized for decision support. That is why data mining is used, especially in science, business, health, security, and informatics areas (Massaro et al., 2017). Data mining is a technology that uses various techniques to discover hidden knowledge from heterogeneous and distributed historical data stored in large databases, warehouses and other massive information repositories (Kotu & Deshpande, 2015).

In researches, data mining algorithm is the tool used for retrieving hidden knowledge from noisy data. Hence, the volume of data enhanced from time-to-time it is difficult to retrieve knowledge. So, the role of data mining was help to retrieving using different algorithms. There are a lot of data mining algorithms that were used to conduct researches. Many researchers and students were used the existing algorithms. Some algorithms need data to be feed into it to get certain knowledge (Kotu & Deshpande, 2015). Sometimes the quality of data has impact on the performance of the algorithms; hence the algorithms were measured by its performance and evaluations metrics on imported dataset (Massaro, Maritati & Galiano, 2018).

Therefore, data mining algorithm is the heart of the data-mining process. These algorithms determine how cases are processed and hence provide the decision-making capabilities needed to classify, segment, associate, and analyze data for processing.

1.1. Shortcoming of Data Mining Algorithms

On the other hand, data mining algorithms have been unable to maintain the same rate of growth. Consequently, there is an increasing need to understand the bottlenecks associated with the execution of these algorithms in modern architectures. According to Ozisikyilmaz B. (2009) to analysis of the data mining applications,

there is an architecture problem. The architecture variation in user expectations and satisfaction relative to the actual hardware performance to develop more efficient architectures that are customized to end-users.

Additionally, data mining algorithms were need preprocessing stages to retrieve hidden knowledge (Massaro, Barbuzzi, Vitti, Galiano, Aruci & Pirlo, 2016). This pre-processing stage is time and budget consuming, because commonly the input dataset owns features (noisy, null values, missed values, duplicate values, incomplete values), which require a transformation and cleaning step, in order to match the input format and assumptions of the data mining algorithms being considered (Wimmer & Powell, 2015). On the other hand, the data mining stage involves typically the use of one or more inductive learning algorithms, requiring that the user iterates over several steps, especially when the results are not good enough, either in terms of performance (accuracy) or understanding of the rules generated for the model (Al-Khoder & Harmouch, 2015). The other bottleneck of data mining algorithms is dynamic nature of data which refers to high voluminous and continuously changing information which is not stored earlier for analyzing and processing like static data. It is difficult to maintain dynamic data as it changes with time (Gulli & Pal, 2017). Many algorithms are used to analyze the data of interest. These data can be sequential, audio signal, video signal, spatio-temporal, temporal, time series etc (Nguyen, Woon, & Ng, 2015).

1.2. Recent Researches Towards Data Mining Algorithms

The finding of the research paper by Massaro et al. (2018) to improve the best performing predictive model k-Nearest Neighbor (k-NN) exhibited the best performance. The performance comparison has been performed between Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), gradient boosted trees, decision trees, and deep learning algorithms. The gradient boosted trees approach represents an alternative approach having the second best performance.

The other finding of the study conducted by S.R. Joseph et al. (2016) on data mining algorithms presents overview of various algorithms necessary for handling large datasets and the strengths and limitations of data mining algorithms. This work argued that data mining algorithms classification, regression, and clustering have limitations as a large of volume of dataset. These algorithms have a bottleneck to handle a large volume data, to identify associations, to identify patterns and to analyze a large size of datasets. The study clearly showed that choosing the appropriate algorithm for specific purpose is a big challenge. While different algorithms used to perform the same business task, each algorithm come up with a different result, and some algorithms yield more than one type of results. Therefore, data mining algorithms has a jam to identify pattern from a huge dataset and they produce different results.

Clustering is one of a data mining algorithm which is used to partition meaningful data into useful clusters which can be understandable and has analytical value. In the paper after giving a brief viewpoint of data mining and clustering techniques, a comparative study of various partitioning algorithms was done. The paper analyses four important partitioning algorithms known as K-means, K-Medoids, CLARA and CLARANS. The study presents a comparative table to understand merits and demerits of each of the algorithms. The analysis shows that CLARA and CLARANS are comparatively more efficient and scalable than other algorithms. However algorithms such as K-Means and K-Medoid can be further modified to make them equally efficient and scalable. The paper shows that not all partitions algorithms are efficient to handle large datasets. The exploration of partitioning algorithms opens new vistas for further development and research (Swarndeep Saket & Pandya, 2016). However, further research is required to study efficiency parameters of each of the partitioning algorithms.

Clustering is the basic composition of data mining analysis, plays a significant role in different modern science research. On one hand, many tools for cluster analysis have been created, along with the information increase and subject intersection. On the other hand, each clustering data mining algorithm has its own strengths and weaknesses, due to the complexity of information (Xu & Tian, 2015). The main purpose of the paper is to introduce the basic and core idea of each commonly used clustering algorithm, specify the source of each one, and analyze the advantages and disadvantages of each one. It is hard to present a complete list of all the clustering algorithms due to the diversity of information, the intersection of research fields and the development of modern computer technology. So 19 categories of the commonly used clustering algorithms, with high practical value and well-studied, are selected and one or several typical algorithm(s) of each category is(are) discussed in detail so as to give readers a systematical and clear view of the important data analysis method, clustering (Xu & Tian, 2015).

1.3. Types of Data Mining Algorithms

Data mining involves primarily the following four classes of tasks presented below.

1.3.1. Classification

a. Decision Tree

Decision trees arrange information in a tree-like structure, classifying the information along various branches. Each branch represents an alternative route, a question. This structure can be used to help you predict likely values of data attributes.

b. CART

Another type of algorithm is CART which is educated from trained dataset. Decision tree has a leaf node where non-leaf node represents a feature and each leaf represent a value that can take the feature. Decision tree instances are classified by path starts from the root and ends at a leaf node branches based on instance feature values (Zafarani, Abbasi & Liu, 2014). Construction of decision trees is based on heuristics.

c. J48

J48 is a Java implementation of C4.5 in Weka package is referred to as J48. This algorithm is used to handle both nominal and numeric values as well as handle missing values.

d. C4.5

C4.5 is used to continue data and avoids over fitting of data. Over fitting is a big problem at the time of displaying result of decision tree. C4.5 improves computational efficiency and handles training data with missing and numeric value.

1.3.2. Clustering

Clustering algorithms are data mining algorithms that work related data grouped together. Therefore, it groups like data together in various groups. There is no prearranged arrangement for grouping these data; alike data are assimilated, and the analysis of the significance of such groupings is left to the user. Clustering is useful to see patterns in data, such as trying to identify geographic regions that are likely to respond well to a certain sales campaign.

a. K-means Algorithm

K-means algorithm is the most largely used algorithm in many researches. This algorithm is a commonly used for grouping technique. The nature of k-means is starts with a collection of data and attempts to group them into ‘k’ number of groups based on certain specific distance measurements. K-means clustering algorithm generates a specific number of disjoint, flat (non-hierarchical) clusters.

b. DBSCAN

DBSCAN is also another algorithm works based on the density. This clustering algorithm has played a great role in finding the density of the non-linear shapes of the structure. DBSCAN uses the concept of density reach ability and density connectivity.

c. K-Nearest Neighbor

K-nearest neighbor (KNN) is an algorithm works based on the similarity measure. It stores available cases and classifies new cases based on a similarity measure, for instance distance functions. A case is categorized by a majority vote of its neighbors, with the case being assigned to the most common amongst its k nearest neighbors measured by a distance function. If k = 1, then the case is assigned to the class of its nearest neighbor. This algorithm is used in statistical estimation and pattern recognition as a nonparametric technique.

d. Expectation-Maximization

Expectation-Maximization (EM): EM-based algorithm is a soft clustering technique; it is robust to noise and able to handle missing data. In the initial step, the EM-based algorithm guesses the parameters of the model, such as the maximum number of clusters and their centers. Then, it iteratively performs two alternating steps, the Expectation (E) and the Maximization (M). In the Expectation step, for each data object, current parameters of the model used to calculate its membership for each cluster. In the Maximization step, re-estimate the model parameters; for instance, re-calculate the new centers to maximize the likelihood between the model and the assumed model (Xu & Tian, 2015).

e. Support Vector Machine

One of the initial shortcomings of the SVMs is its costly computational complexity in the training phase, which leads to inapplicable algorithms in the large datasets. However, this problem is being solved with great success. One approach is to break a large optimization problem into a series of smaller problems, where each problem only involves a couple of carefully chosen variables so that the optimization can be done efficiently. The process iterates until all the decomposed optimization problems are solved successfully.

A more recent approach is to consider the problem of learning SVMs as that of finding an approximate minimum enclosing ball of a set of instances (Štulec, Petljak & Kukor, 2016). These instances, when mapped to an N-dimensional space, represent a core set that can be used to construct an approximation to the minimum enclosing ball. Solving the SVMs learning problems on these core sets can produce good approximation solutions in very fast speed. For example, the core vector machine and the further ball vector machine (Otha & Higuci, 2013) can learn SVMs for millions of data in seconds.

f. PageRank

PageRank produces a static ranking of Web pages in the sense that a PageRank value is computed for each page off-line and it does not depend on search queries. The algorithm relies on the democratic nature of the Web by using its vast link structure as an indicator of an individual page's quality. In essence, PageRank interprets a hyperlink from page x to page y as a vote, by page x, for page y. However, PageRank looks at more than just the sheer number of votes, or links that a page receives. It also analyzes the page that casts the vote. Votes casted by pages that are themselves "important" weigh more heavily and help to make other pages more "important". This is exactly the idea of rank prestige in social networks (Xu & Tian, 2015).

1.3.3. Regression

On the other hand, regression and association were algorithms largely used for a research purpose.

a. Linear Regression

This method is works based on statistics for predicting the value of a dependent and independent variables where the relationship between the variables can be described with a linear model.

b. Logistic Regression

This algorithm is based on the classification analog of regression. Logistic regression is preferable when trees in the same situations and effects are small and predictors contribute additively (when there are no interactions).

1.3.4. Association

a. Apriori approach

The bottlenecks of the apriori approach is reduces the size of candidate frequent itemsets by using apriori property. However, it still requires two nontrivial computationally expensive processes. It requires as many database scans as the size of the largest frequent itemsets. In order to find frequent k-itemsets, the apriori algorithm needs to scan database k times. Breadth-first (i.e., level-wise) search a candidate generation and test the frequency of true appearance of the itemsets. It may generate a huge number of candidate sets that will be discarded later in the test stage.

One of the most popular data mining approaches is to find frequent itemsets from a transaction dataset and derive association rules. Given a set of transactions, the problem of mining association rules is to generate all association rules that have support and confidence no less than the user-specified minimum support (called minsup) and minimum confidence (called minconf), respectively. Finding frequent itemsets (itemsets with support no less than minsup) is not trivial because of the computational complexity due to combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence no less than minconf. Apriori and AprioriTid, proposed by R. Agrawal & R. Srikant (2015), are seminal algorithms that are designed to work for a large transaction dataset.

b. FP-Growth Algorithm

To break the two drawbacks of Apriori algorithm, FP-growth algorithm is used. FP-growth requires constructing FP-tree. For that, it requires two passes. FPgrowth uses divide and conquer strategy. It requires two scans on the database. It first computes a list of frequent items sorted by frequency in descending order (F-List) and during its first database scan. In the second scan, the database is compressed into a FP-tree. This algorithm performs mining on FP-tree recursively. There is a problem of finding frequent itemsets which is converted to searching and constructing trees recursively. The frequent itemsets are generated with only two passes over the database and without any candidate generation process. There are two sub processes of frequent patterns generation process which includes: construction of the FP-tree, and generation of the frequent patterns from the FP-tree (Kumbhare & Chobe, 2014).

1.5. Data Mining Architecture

Data visualization techniques are becoming very useful methods to discover patterns in datasets, because they impact directly the human visual system, currently the most powerful pattern recognizer and discoverer (Shneiderman, 2003). There is a wide variety of techniques, which may be used in the several stages of the KDD process: in the pre-processing stage, to get a rough feeling of the features of the dataset; in the data mining stage, to discover patterns, such as clusters of items, correlations or dependencies among attributes, or to visualize the model produced by the data mining algorithm, in order to have a better understanding how the responses are generated by the model. Among the most common visualization techniques to mine knowledge from data are 2D, 3D scatter-plots, and scatter-plot matrix.

1.6. Limitations of Data Mining Algorithms

As this review paper shows that the bottleneck of data mining algorithms especially in researches are inefficiency of the algorithms on a large size of datasets and multimedia data. Data mining algorithms were not handling a large volume, heterogeneous data including multimedia data and spatial data (Wu et al., 2007). The performance of data mining algorithms are low on multimedia data (video signal, audio signal, sequential data, spatio-temporal, temporal, time series) and spatial data. Additionally, the presentation of data mining results to easily view and understand the output of the data mining algorithms there is a need to use knowledge representation (decision tree, rules, equations, semantic networks) and visualization techniques (such as graphs, bar charts, etc. (Huang & Hou, 2017).

However, many data mining algorithms work with static datasets. This requires that the algorithm be completely rerun any time the database changes. Since data mining problems are often not precisely stated, interfaces may be needed with both domain and technical experts. Although some techniques may work well, they may not be accepted by users if they are difficult to use or understand. Determining the intended use for the information obtained from the data mining tool is a challenge. Indeed, how business executives can effectively use the output is sometimes considered the most difficult part. Because the result are type that have not previously been known. Business practices may have to be modified to determine how to effectively use the information uncovered (Massaro, Maritati & Galiano 2018).

Some data mining algorithms, like k-NN, are easy to build but quite slow in predicting the target variables. Algorithms such as the decision tree take time to build but can be reduced to simple rules that can be coded into almost any application. The trade-offs between production responsiveness and build time need to be considered and if needed, the modeling phase needs to be revisited if the response time is not acceptable by business application. The quality of prediction, accessibility of input data and the response time of the prediction remain the most important quality factors in the business application. Therefore, this review work presents an inefficiency of data mining algorithms specifically in research work. The role of this review is also to give a clue idea for researchers and to what extent the efficiency of data mining algorithms helpful on the quality of retrieve hidden pattern from datasets. The summary of these algorithms has been shown as a table 1 below.

Tab. 1. Summary on Inefficiency of Data Mining Algorithms

No.	Algorithm	The study	Explored Bottleneck of the Algorithm
1.	Apriori	S. R. Joseph *et al.*, (2016) Data mining Algorithms: an overview, & G. Negandhi (2007) Apriori Algorithm Review for Finals	– It is reduces the size of candidate frequent itemsets by using Apriori property. – It requires many database scans as the size of the largest frequent itemsets. – It needs more search space and the input-output cost increases. – It increases the number of database scan hence it increases in computational cost. – It discovers a huge quantity of rules, when some being irrelevant.
2.	C4.5	N. Rehman (2017) Data Mining Techniques Methods Algorithms and Tools	– It requires target attribute that will have only discrete values. – Small change in data can cause different decision trees to be built. – In small training set, the C4.5 algorithm does not work very well (less accurate and/or efficient).
3.	CART	R. Zafarani, M. Abbasi and H. Liu (2014)	– It has unstable decision tree. – It splits variable only by one.
4.	Decision tree	R. Zafarani, M. Abbasi and H. Liu (2014), Social Media Mining, Cambridge University Press	– It cannot predict the value of a continuous attribute. – It provides error prone results on too many classes. – Irrelevant attribute affect construction of decision tree in a bad manner. – Small change in data can change the decision tree completely. – Performs poorly with many class and small data. – It does not function well with categorical variables having multiple levels. – It brought Over fitting problem.
5.	EM	N. Rehman (2017) Data Mining Techniques Methods Algorithms and Tools	– It is slow convergence. – Inability to provide estimation to the asymptotic variance-covariance matrix of the maximum likelihood estimator. – The EM algorithm does not require the gradient. – It can be used in cases where some data values are missing, although this is less relevant in the 1d case.

Tab. 1. Summary on Inefficiency… – cont.

6.	K-Means	J. Swarndeep Saket , & S. Pandya (2016) An Overview of Partitioning Algorithms in Clustering Techniques	– Every time starting with a random set of initial clusters and repeat a number of times to obtain an optimal. – It is difficult in comparing quality of the clusters produced. – It is difficult to predict what K should be when fixed number of clusters can make. – It doesn't work well with non-globular cluster.
7.	K-Mediod	"	– Due to its time complexity k-mediods is more costly than k-means algorithm. – It doesn't scale well for a large datasets. – The results produced and total run time depends upon initial partitions.
8.	KNN	S. Bavisi, *et al.*, (2014) A Comparative Study of Different Data Mining Algorithms	– It has poor run time performance. – It requires high calculation complexity – It considers no weight difference between samples. – It is sensitive to irrelevant and redundant feature. – Sensitiveness to noisy or irrelevant attributes. – It is a weak as a classifier for an IDS is its large storage requirements. – It is highly susceptible to the curse of dimensionality and slow in classifying test tuples.
9.	Naive Bayes	"	– It provides less accuracy. – The precision of algorithm decreases if the amount of data is less. – It will consider the probability of attribute to be zero (zero problem). – Independence assumption is often violated in the real world.
10.	Neural Networks	H.-C Huang, &, C.-I. Hou (2017)	– It has poor interpretability. – It takes time for long training. – Inability to interpret the learned model. – It has high complexity.
11.	PageRank	D. Xu & Y. Tian (2015) A Comprehensive Survey of Clustering Algorithms	– It is computed value for each page off-line and it does not depend on search queries. – Older pages may have higher rank – so even if a new page has some very good contents but it may not have many links in the early state. – PageRank can be easily increased.

Tab. 1. Summary on Inefficiency… – cont.

12.	SVM	S. R. Joseph *et al.*, (2016) Data Mining Algorithms: An Overview	– It is hard to interpret and memory intensive. – It is a classifier high algorithmic complexity and extensive memory requirements. – Depend on the choice of the kernel. – High complexity. – Difficult to design multi-class classifiers.
13.	DBSCAN	D. Xu, & Y. Tian (2015) A Comprehensive Survey of Clustering Algorithms	– It has trouble when the clusters have widely varying densities. – It also has trouble with high-dimensional data. – It is expensive when the computation of nearest neighbors requires computing all pairwise proximities. – To set an input parameters it is high sensitivity. – It is poor cluster descriptors.
14.	Linear regression	S. R. Joseph *et al.*, (2016) Data Mining Algorithms: An Overview	– It is limited to predicting numeric output. – It has a lack of explanation about what has been learned can be a problem. – It doesn't work well for data with continuous or binary outcomes.
15.	Logistic Regression	"	– It is a classic problem on text classification. – It is not stable when one predictor explain the response variable. – It requires more assumptions and sensitive to outliers.
16.	Density Based	"	– It fails in various density clusters. – It fails in neck type of dataset.

3. CONCLUSION

This review article attempts to answer which data mining algorithm(s) is efficient in the research work, the bottleneck associated with the algorithms and research area in data mining. Data mining algorithms are a tool used to retrieve hidden knowledge from datasets. It is a KDD process to identify the pattern and its relationship in huge dataset. Recent advances in data extraction techniques have resulted in tremendous increase in the input data size of data mining applications. Different algorithms provide different perspectives on the complete nature of the pattern. However, the size of dataset was increased from time-to-time in different fields for research purposes. This review paper basically to show

a directions and used as a guide and help the researchers specifically to identify the shortcoming of data mining techniques with domain area in solving a certain problems they will explore. It also shows the research areas not yet started particularly in multimedia where data can be sequential, audio, video, spatio-temporal and time series. Therefore, the reviewed paper shows that it becomes difficult to handle a large datasets in order to identify associations and patterns. And it indicates an overview of architecture and algorithms used in large datasets.

REFERENCES

Agrawal, R., & Srikant, R. (2015). Fast algorithms for mining association rules. In *Proc. of the 20th International Conference on Very Large Data Bases (VLDB)* (pp. 487–499). Santiago, Chile.

Al-Khoder, A., & Harmouch, H. (2015). Evaluating Four Of The most Popular Open Source and Free Data Mining Tools. *International Journal of Academic Scientific Research*, *3*(10), 13–23.

Bavisi, S., Mehta, J., & Lopes, L. (2014). A Comparative Study of Different Data Mining Algorithms. *International Journal of Current Engineering and Technology*, *4*(5), 3248–3252.

Gulli, A., & Pal, S. (2017). *Deep Learning with Keras-Implement neural networks with Keras on Theano and Tensor Flow*. Birmingham, UK: Packt Publishing.

Huang, H. C., & Hou, C. I. (2017). Tourism Demand Forecasting Model Using Neural Network. *International Journal of Computer Science & Information Technology (IJCSIT)*, *9*(2), 19–29.

Joseph, S. R., Hlomani, H., & Letsholo, K. (2016). Data Mining Algorithms: An Overview. *International journal of Computers and Technology*, *15*(6), 6806–6813.

Kalyani, J., Bharathi, H. N., & Rao, J. (2016) Stock Trend Prediction Using News Sentiment Analysis, *International Journal of Computer Science & Information Technology (IJCSIT)*, *8*(3), 67–76.

Kotu, V., & Deshpande, B. (2015). *Predictive Analytics and Data Mining – Concepts and Practice with RapidMiner*. Elsevier.

Kumbhare, T. A., & Chobe, S. V. (2014) An Overview of Association Rule Mining Algorithms. *International Journal of Computer Science and Information Technologies*, *5*(1), 927–930.

Massaro, A., Barbuzzi, D., Vitti, V., Galiano, A., Aruci, M., & Pirlo, G. (2016), Predictive Sales Analysis According to the Effect of Weather. In *Proceeding of the 2nd International Conference on Recent Trends and Applications in Computer Science and Information Technology* (pp. 53–55). Tirana, Albania.

Massaro, A., Galiano, A., Barbuzzi, D., Pellicani, L., Birardi, G., Romagno, D. D., & Frulli, L., (2017). Joint Activities of Market Basket Analysis and Product Facing for Business Intelligence oriented on Global Distribution Market: examples of data mining applications. *International Journal of Computer Science and Information Technologies*, *8*(2), 178–183.

Massaro, A., Maritati, V., & Galiano, A. (2018). Data Mining Model Performance of Sales Predictive Algorithms Based On Rapidminer Workflows. *International Journal of Computer Science & Information Technology (IJCSIT)*, 10 (3) 39–56. doi:10.5121/ijcsit.2018.10303

Negandhi, G. (2007). *Apriori Algorithm Review for Finals* (SE 157B). Spring Semester.

Nguyen, H.-L., Woon, Y. K., & Ng, W. K. (2015). A Survey on Data Stream Clustering and Classification. *Knowledge and Information Systems*, *45*(3), 535–569. doi:10.1007/s10115-014-0808-1

Otha, M., & Higuci, Y. (2013). Study on Design of Supermarket Store Layouts: the Principle of Sales Magnet, World Academy of Science. *Engineering and Technology*, *7*(1), 209–212.

Ozisikyilmaz, B. (2009). *Analysis, Characterization and Design of Data Mining Applications and Applications to Computer Architecture* (Unpublished doctoral dissertation). Northwestern University, Evanston, Illinois.

Rehman, N. (2017). Data Mining Techniques Methods Algorithms and Tools. *International of Computer Science and Mobile Computing*, *6*(7), 227–231.

Shneiderman, B. (2003). Inventing discovery tools: Combining information visualization with data mining. In *The Craft of Information Visualization Readings and Reflections Interactive Technologies* (pp.378-385). Morgan Kaufmann. doi:10.1016/B978-155860915-0/50048-2

Štulec, I., Petljak, K., & Kukor, A. (2016). The Role of Store Layout and Visual Merchandising in Food Retailing. *European Journal of Economics and Business Studies*, *4*(1), 139–152.

Swarndeep Saket, J., & Pandya, S. (2016). An Overview of Partitioning Algorithms in Clustering Techniques. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *5*(6), 1943–1946.

Talia, D., Trunfio, P., & Marozzo, F. (2016). *Data Analysis in the Cloud: Models and Techniques for Cloud-Based Data Analysis.* Elsevier Science.

Wimmer, H., & Powell, L. M. (2015) A Comparison of Open Source Tools for Data Science, In *Proceedings of the Conference on Information Systems Applied Research* (v8 n3651). Wilmington, North Carolina USA.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2007). *Top 10 algorithms in data mining*. London, UK: Springer-Verlag London Limited.

Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, *2*(2), 165–193. doi:10.1007/s40745-015-0040-1

Yadav, Ch., Wang, S., & Kumar, M. (2013) Algorithm and approaches to handle large Data-A Survey, *International Journal of Computer Science and Network*, *2*(3), 1307.5437.

Zafarani, R., Abbasi, M., & Liu, H. (2014). Social Media Mining. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139088510