ISSN 2277-3061
Volume 15 Number6
International journal of Computers and Technology

Data Mining Algorithms: An Overview

Sethunya R Joseph[1], Hlomani Hlomani[2] Keletso Letsholo[3]
[1] Computer Science Department, Botswana International University of Science and Technology, Palapye, Botswana
Sethunya.joseph@studentmail.biust.ac.bw
[2] Computer Science Department, Botswana International University of Science and Technology, Palapye, Botswana
hlomanihb@.biust.ac.bw
[3] Computer Science Department, Botswana International University of Science and Technology, Palapye, Botswana
letsholok@.biust.ac.bw

ABSTRACT

The research on data mining has successfully yielded numerous tools, algorithms, methods and approaches for handling large amounts of data for various purposeful use and problem solving. Data mining has become an integral part of many application domains such as data ware housing, predictive analytics, business intelligence, bio-informatics and decision support systems. Prime objective of data mining is to effectively handle large scale data, extract actionable patterns, and gain insightful knowledge. Data mining is part and parcel of knowledge discovery in databases (KDD) process. Success and improved decision making normally depends on how quickly one can discover insights from data. These insights could be used to drive better actions which can be used in operational processes and even predict future behaviour. This paper presents an overview of various algorithms necessary for handling large data sets. These algorithms define various structures and methods implemented to handle big data. The review also discusses the general strengths and limitations of these algorithms. This paper can quickly guide or be an eye opener to the data mining researchers on which algorithm(s) to select and apply in solving the problems they will be investigating.

Keywords

big data, data mining, knowledge discovery, data mining algorithms

Academic Discipline And Sub-Disciplines

Computer Science

SUBJECT CLASSIFICATION

Data Mining and Algorithms

TYPE (METHOD/APPROACH)

Survey/Review

1. INTRODUCTION

A number of definitions about data mining have been laid forth by various researchers. Some have defined data mining as a process of discovering useful or actionable knowledge in large scale data [1, 4]. According to Zaki and Meira [8] data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data [8]. Another definition of data mining as coined by Ozer [2] and Garcia et.al. [26], is the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Data mining also means knowledge discovery from data which describes the typical process of extracting useful information from raw data [1, 3]. As pointed out by Kamruzzaman, Haider and Hasan [11], many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases, or KDD. This observation is quiet true if one can closely look at the interpretations which have been made by several researchers such [1, 2, 3, 10] about data mining. However as pointed out by Kamruzzaman, Haider and Hasan [11], data mining is also treated simply as an essential step in the process of knowledge discovery in databases.

2. SCOPE AND OBJECTIVE

Based on document analysis, this paper reviews and summarizes the information on data mining concept and some of the data mining algorithms. It gives the general overiew or background about data mining and the algorithms which are usually used to mine data. It gives a background about the categories of these algorithms. It then gives a discussion on the strengths and limitations of these data mining algorithms. The research paper is intended to give an understating to researchers, scholarly peers , learners, data miners, companies and anyone who wish to stay abreast with the data mining and the algorithms which are commonly used in data mining.

3. DATA MINING ALGORITHMS

A data mining algorithm is a set of heuristics and calculations that creates a data mining model from data [26]. It can be a challenge to choose the appropriate or best suited algorithm to apply to solve a certain problem. Even though one can use different algorithms to perform the same tasks, each algorithm yield a different set of results, and some algorithms can even produce more than one type of results. Some algorithms can perform classification process, that is, they can predict one or more discrete variables, based on the other attributes in the data set. Some algorithms perform regression

purposes, they can predict more or continuous variables based on the other attributes in the data set. As pointed out by Microsoft [26], some algorithms can perform segmentation,they divide data into groups,or clusters of items that have similar properties.While some algorithms can be associative by finding correlations between different attributes in a set, some can be used for sequence analysis processes, that is they can be used to summarise sequence or episodes in data, such as a web path flow [26]. However, all of the aforementioned types of algorithms can be categorized into two large categories: Supervised learning and Unsupervised learning algorithms.

The following sub-sections briefly discuss the two categories: supervised and unsupervised learning. Several examples of some of the aforementioned algorithms in each of the said categories are also given as a summary in Table 1 and 2. Basically both Table 1 and 2 show a general discussion of some of the strengths and limitations of some of these algorithms.

3.1 SUPERVISED LEARNING

The supervised learning algorithms are those for which the class attribute values for the dataset are known before running the algorithm. This data is called labelled data or training data [1]. Instances in this set are tuples in the format (x, y) where x is a vector and y is the class attribute, commonly a scalar. Supervised learning builds a model that maps x to y. The task is to find a mapping m(.) such that *m(x) = y.* The unlabelled dataset or test dataset is also provided, in which instances are in the form(x,?) and y values are unknown. Given *m*(.) learned from training data and x of an unlabelled instance, m(x) can be computed which results in the prediction of the label for the unlabelled instance [5].

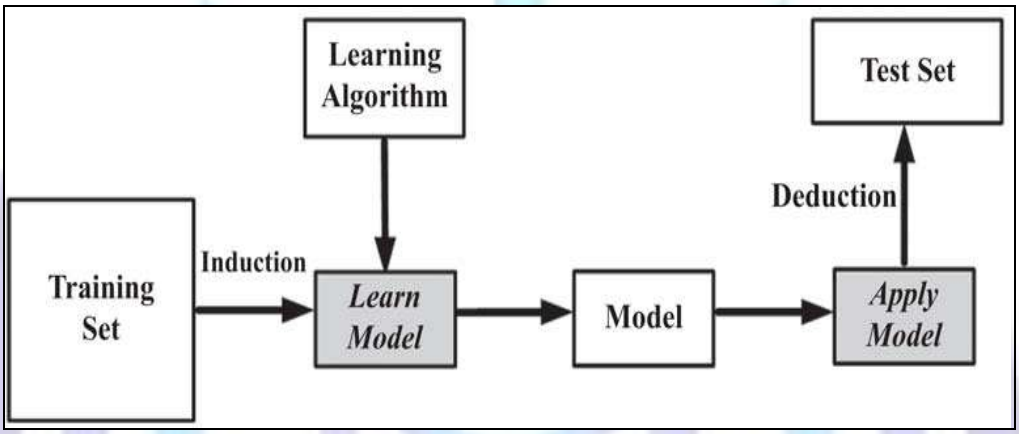

Figure 1: Supervised learning process [5].

Figure 1 shows the systematic process of a supervised learning algorithm. This process starts with a training set (i.e. labelled data) where both features and labels (class attribute values) are known. A supervised learning algorithm is run on the training set in a process known as induction. In the induction process, the model is generated. The model maps the features values to the class attribute values. The model is used on a test set to predict the unknown class attribute values (deduction process).

Supervised learning can be divided into i) classification and ii) regression. When the class attribute is discrete, it is called classification; when the class attribute is continuous, it is regression [1, 5, 6].

i) CLASSIFICATION

In classification, class attribute values are discrete. Given a set of data elements classification maps each data element to one of a set of pre-determined classes based on the difference among data elements belonging to different classes. The goal is to discover rules that define whether an item belongs to a particular subset or class of data [6]. For example, when trying to determine which households will respond to a direct mail campaign, look at rules that separate the “probables” from the not probables. Then use IF-THEN rules in a tree-like structure to represent the predictions and classify the set of data items.

Classification Process: According to Han, Kamber and Pei [7], classification is a two-step process model construction describing a set of predetermined classes. Each tuple is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction is training set. The model is represented as classification rules, decision trees, or mathematical formulae. Then the model will be used for classifying future or unknown objects [1].

Some of the examples of classification methods include decision tree learning, naive Bayes classifier, K-nearest neighbour classifier, and classification with network information. Also, regression methods examples include K-means, linear regression and logic regression. See Table 1, it shows a summary of examples some of the supervised learning classification algorithms. It shows a discussion of their strengths and limitations when handling data in general.

ii) REGRESSION

In regression, class attribute values are real numbers. For instance, if we wish to predict the stock market value (class attribute) of a company given information about the company (features). The stock market value is continuous; therefore, regression must be used to predict it. The input to the regression method is a dataset where attributes are represented using x_1, x_2,…, x_m (also known as regressors). The class attribute is represented using Y (also known as the dependent variable), where the class attribute is a real number and the relation between Y and the vector X = (x_1, x_2,…, x_m) [5]. Examples of regression methods include linear regression and logic regression. See Table 1, it shows a summary of examples some these supervised learning regression algorithms.

Table 1: Examples of supervised learning algorithms commonly used in data mining

SUPERVISED LEARNING ALGORITHMS			
i) CLASSIFICATION			
ALGORITHM	**HOW IT WORKS**	**STRENGTH**	**LIMITATIONS**
1.Decision Tree Learning(C4.5,ID3,CART)	-Decision is learnt from training the data set -Each non-leaf node in a tree represent a feature and each branch represent a value that the feature can take -Instances are classified by following a path that starts at the root node and ends at a leaf by following branches based on instance feature values [5] -Construction of decision trees is based on heuristics	-Multiple Decision trees can be learned from the same data set -They can both correctly predict the class attribute values for all instances in the dataset -Implicitly perform variables screening or feature selection -Requires relatively little effort from users for data preparation -Non-linear relationships between parameters do not affect tree performance -Easy to interpret and explain -Allow the addition of new possible scenarios.[5]	-Decision trees are constructed recursively from training data using a top-down greedy approach in which features are sequentially selected.[5] -Without proper pruning or limiting tree growth, they tend to over-fit the training data, making them somewhat poor predictors. -Low-Performance since you need to 'redraw the tree' every time you wish to update your CART model and poor resolution on data with complex relationships among the variables. -They are only two possibilities (left-right) at each node, hence there are some variable relationships that decision trees just cannot learn. -Practically limited to classification.
Naive Bayes Classifiers (NB)	-Bayesian network is a model that encodes probabilistic relationships among variables of interest [13]. -In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features [12]. -Naïve Bayesian technique is generally used for intrusion detection in combination with statistical schemes, a procedure that yields several	-Easy to use. -Effective if the training set is large enough. -Ability to learn; as the training set gets larger, the results get more and more accurate (intelligence) [5]. -Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. -If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than	-Does not consider the sequence of words(Non-relevant word feature) -It cannot learn interactions between features. -Is based on the so Bayesian theorem and is particularly suited when the dimensionality of the inputs is high -Considerably higher computational effort is required [14]. -In Naïve Bayes approach it is assumed

	advantages, including the capability of encoding interdependencies between variables and of predicting events, as well as the ability to incorporate both prior knowledge and data [17].	discriminative models like logistic regression, so you need less training data. - Naïve Bayesian classifiers simplify the computations and exhibit high accuracy and speed when applied to large databases. -Bayesian classifiers give satisfactory results because focus is on identifying the classes for the instances, not the exact probabilities) [12].	that the data attributes are conditionally independent [13] which is not always so
K-Nearest Neighbour	-K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition as a non-parametric technique. -A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbour [17].	- The k-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours -It is analytically tractable -KNN is simple in implementation and it uses local information, which can yield highly adaptive behaviour. -A major strength of the KNN algorithm is that it lends itself very easily to parallel implementations [16]	- It is a type of lazy learning where the function is only approximated locally and all computations are deferred until classification. -One of the weaknesses of the K-Nearest Neighbours Algorithm as a classifier for an IDS is its large storage requirements -KNNs are also known to be highly susceptible to the curse of dimensionality and slow in classifying test tuples
Support Vector Machine	-An SVM maps input (real valued) feature vectors into a higher dimensional feature space through some nonlinear mapping. SVMs are developed on the principle of structural risk minimization [16]. -Structural risk minimization seeks to find a hypothesis (h) for which one can find lowest probability of error whereas the traditional learning techniques for pattern recognition are based on the minimization of the empirical risk, which attempt to optimize the performance of the learning set	-High accuracy -SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space [18] -have the ability to update the training patterns dynamically whenever there is a new pattern during classification -Nice theoretical guarantees regarding over-fitting(Support Vector Machines are able to model complex nonlinear decision boundaries and are	-Memory-intensive, hard to interpret, and kind of annoying to run and tune -You do not have a nice probabilistic interpretation, -You cannot easily update your model to take in new data (using an online gradient descent method) - A disadvantage of SVMs as a classifier is its high algorithmic complexity and extensive memory requirements [19]]. This consequently makes the speed both in training and testing slow.

		less prone to over fitting than other methods) -Used in text classification problems where very high-dimensional spaces are the norm [12].	
ii) REGRESSION			
Linear Regression	Linear Regression is a statistical procedure for predicting the value of a dependent variable from an independent variable when the relationship between the variables can be described with a linear model.	-Linear regression implements a statistical model that, when relationships between the independent variables and the dependent variable are almost linear, shows optimal results - Linear regression is often used to model non-linear relationships [20]. -Linear regression is useful for data with linear relations or applications for which a first-order approximation is adequate [21].	-Linear regression is limited to predicting numeric output. - A lack of explanation about what has been learned can be a problem -Does not work well for data with continuous or binary outcomes
Logistic Regression	Logistic regression is classification analog of linear regression. It is preferable to trees in the same situations that linear regression is preferable to regression tress-when effects are small and when predictors contribute additively (no interactions).	-Lots of ways to regularize the model, and one do not have to worry about the features being correlated, - One can easily update the model to take in new data (using an online gradient descent method), - Use it if one wants a probabilistic framework (e.g., to easily adjust classification thresholds, to say when you're unsure, or to get confidence intervals) - Or if you expect to receive more training data in the future that you want to be able to quickly incorporate into your model [5]. -Logistic regression is intrinsically simple, it has low variance and so is less prone to over-fitting. -logistic regression is faster	-Text classification is a classic problem -It is unstable when one predictor could almost explain the response variable, because the coefficient of this variable will be boosted to as high as possible, here is a case when people turn to discriminant analysis -Requires more assumptions and is sensitive to outliers [25].

		and more reliable when the dimension gets large. -Logistic regression is less complex and easier to inspect	

3.2 UNSUPERVISED LEARNING

This is the unsupervised division of instances into groups of similar objects. Normally when discussing the unsupervised learning, most researchers focus on clustering [1, 5]. In clustering, the data is often unlabelled. Thus, the label for each instance is not known to the clustering algorithm. This is the main difference between supervised and unsupervised learning. Any clustering algorithm requires a distance measure. Instances are put into different clusters based on their distance to other instances [9]. The most popular distance measure for continuous features is the Euclidean distance:

d (X,Y) = $\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_n-y_n)^2}$. Table 2 show some of the examples of the unsupervised learning algorithms. It shows a discussion of their strengths and limitations when handling data in general.

Table 2: Examples of supervised learning algorithms commonly used in data mining

UNSUPERVISED LEARNING ALGORITHMS			
CLUSTERING			
ALGORITHM	**HOW IT WORKS**	**STRENGTH**	**LIMITATIONS**
K-Means	- K-Means methodology is a commonly used clustering technique. In this analysis the user starts with a collection of samples and attempts to group them into 'k' Number of Clusters based on certain specific distance measurements. -K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. [20].	-With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small). -K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular [23, 26]. -It is well suited to generating globular clusters. The K-Means method is numerical, non-deterministic and iterative	-The K-Means algorithm is repeated a number of times to obtain an optimal clustering solution, every time starting with a random set of initial clusters [20]. -Difficulty in comparing quality of the clusters produced (e.g. for different initial partitions or values of K affect outcome). -Fixed number of clusters can make it difficult to predict what K should be. -Does not work well with non-globular cluster [23].
Density Based	Density based clustering algorithm has played a vital role in finding non-linear shapes structure based on the density. -It uses the concept of density reachability and density connectivity [24].	-Does not require a prior specification of number of clusters. -Able to identify noise data while clustering. - Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is able to find arbitrarily size and arbitrarily shaped clusters [24].	-DBSCAN algorithm fails in case of varying density clusters. -Fails in case of neck type of dataset [24].
Apriori	-The Apriori Algorithm is an influential algorithm for mining frequent item sets for boolean association rules [25]. -Apriori uses a "bottom up" approach, where frequent subsets are extended	-Allow the pruning of many associations -Uses large item set property -Easily parallelized -Easy to implement [25].	-More search space is needed and I/O cost will increase [25]. -Number of database scan is increased thus candidate generation will increase hence increase in computational cost - Normally discover a huge quantity of rules, some

			being irrelevant
K-Means	- K-Means methodology is a commonly used clustering technique. In this analysis the user starts with a collection of samples and attempts to group them into 'k' Number of Clusters based on certain specific distance measurements. -K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. [20].	-With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small). -K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular [23, 26]. -It is well suited to generating globular clusters. The K-Means method is numerical, non-deterministic and iterative	-The K-Means algorithm is repeated a number of times to obtain an optimal clustering solution, every time starting with a random set of initial clusters [20]. -Difficulty in comparing quality of the clusters produced (e.g. for different initial partitions or values of K affect outcome). -Fixed number of clusters can make it difficult to predict what K should be. -Does not work well with non-globular cluster [23].

4. CONCLUSION

Due to the increase in the amount of data coming from everywhere(online: blogging,social media,databases,etc.), it has become difficult to handle the data, to find associations, patterns and to analyse the large data sets. Consequently, large numbers of technologies are being developed for the extraction of meaningful data from huge collections of textual data using different text mining techniques. Different tools, algorithms and methods which are being used to mine and analyse the data, perform differently on the data collections as has been indicated in the review which has been made in this paper. Choosing the best algorithm to use for a specific analytical task can be a challenge. While you can use different algorithms to perform the same business task, each algorithm produces a different result, and some algorithms can produce more than one type of result.

5. ACKNOWLEDGEMENTS

Special thanks are passed to BIUST for funding this research

REFERENCES

[1] P. Gundecha and H. Liu, Mining Social Media: A Brief Introduction. Tutorials in operations research, 2012.

[2] P. Ozer and I.G Sprinkhuizen-Kuyper, Data algorithms for classification, B**Sc Thesis Artificial** Intelligence,Radbound University Nijmegen, January 2008.

[3] J. Han, M. Kamber and J. Pei. Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, 2011.

[4] P. N. Tan, M. Steinbach and V, Kumar, Introduction to Data Mining, Pearson Addison Wesley, Boston, 2006.

[5] R. Zafarani, M. A. Abbasi and H. Liu, Social Media Mining, Cambridge University Press, 2014.

[6] J. Han and M. Kamber and Data Mining Concepts and Techniques, Morgan Kaufmann, pp. 60-101. (2000) Retrieved 05 Feb 2016

[7] J. Han, M. Kamber and J. Pei, J Data Mining Concepts and Techniques (3rd ed.) Chapter 8, Pp. 99-117. University of Illinois. Illinois, 2010.

[8] M. J, Zaki and W. Meira Jr, Data Mining and Analysis-Fundamental Concepts and Algorithms, Cambridge University Press. New York, USA.

[9] J .Han and M. Kamber, Data Mining Concepts and Techniques. Presentation Slides of Prof Anita Wasilewska, 2010.

[10] E. Garcia, C. Romeo, S. Ventura and T. Calders, Drawbacks and solutions of applying association rule mining in learning management systems, *Proceedings of the International Workshop on Applying Data Mining in e-Learning* 2007

[11] S. M. Kamruzzaman, F. Haider and A. R. Hasan, Text Classification Using Data Mining, ICTM, 2005.

[12] A. Adebowale, S.A. Idowu and A. Amarachi, Comparative Study of selected data Mining Algorithms Used for Intrusion Detection, *IJSCE,* , vol. 3, no.3, ISSN: 2231-2307, July 2013

[13] D. Barbara, N. Wu and S. Jajodia, "Detecting Novel Network Intrusions Using Bayes Estimators", Proceedings Of the First SIAM Int. Conference on Data Mining, (SDM 2001), Chicago, IL2001.

[14] E. Kesavulu, V. N. Reddy, and P. G. Rajulu, “A Study of Intrusion Detection in Data Mining”. Proceedings of the World Congress on Engineering 2011 Vol IIIWCE 2011, July 6 - 8, 2011, London, U.K, 2011.

[15] T. Lappas, and K. Pelechrinis, Data Mining Techniques for (Network) Intrusion Detection Systems, Department of Computer Science and Engineering Riverside, Riverside CA, 2006.

[16] W. Lee, S. J. Stolfo and K. W. Mok, “A data mining framework for building intrusion detection models,” In Proc. of the 1999 IEEE Symp. On Security and Privacy (pp. 120-132), Oakland, CA: IEEE Computer Society Press, 1999.

[17] http://www.saedsayad.com/k_nearest_neighbors.htm, [online], Accessed 24 February 2016.

[18] W. Lee, S.J. Stolfo and K.W. Mok, “Mining in a data-flow environment: Experience in network intrusion detection,” (Chaudhuri, S. & Madigan, D. Eds.). Proc. of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99), pp. 114-124, San Diego, CA: ACM, 1999.

[19] T. D. Lane, “Machine Learning Techniques for the computer security domain of anomaly detection”, Ph.D. Thesis, Purdue Univ., West Lafayette, IN, 2000.

[20] http://www.camo.com/resources/clustering.html, [online], accessed 24 February 2016.

[21]P. Komarek, Logistic Regression for Data Mining and High-Dimensional Classification, [Online] http://repository.cmu.edu/robotics 2004

[22]Data Mining hand-out, pp36-350, Nov 2013, [online], accessed 26th February 2016.

[23]http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/KMeans_Clustering_Clustering_Overview.htm, [online], accessed 26February 2016.

[24]https://sites.google.com/site/dataclusteringalgorithms/density-basedclusteringaalgorithm,[online]accessed 27th February 2016

[25] G. Negandhi, Apriori Algorithm Review for Finals, SE 157B, *Spring Semester*, 2007.

[26]. Microsoft. Data Mining Algorithms .*Analysi Services-Data Mining*. *2016*.

Authors’ Biography

Sethunya Rosie Joseph is PhD student for Computer Science at BIUST.She is employed as the Research Assistant at BIUST. Previously She has worked as a Computer Science lecturer. Joseph has several publications on Computer Science and Information Systems areas .

Dr Hlomani Hlomani is a Computer Science lecturer at BIUST.He has several publications on the area of Computer Science (Ontologies).

Dr Keletso Letsholo is a Computer Science lecturer at BIUST. He has several publications on the area of Computer Science(Natural language Processing and Software development).