Supported by tetfund TERTIARY EDUCATION TRUST FUND

PERFORMANCE EVALUATION OF SELECT DATA MINING SOFTWARE TOOLS FOR DATA CLUSTERING

A. O. Ameen*, A. O. Bajeh, B. A. Adesiji, A. O. Balogun and M. A. Mabayoje
Department of Computer Science, University of Ilorin, Kwara State, Nigeria
*Corresponding author: aminamed@unilorin.edu.ng

Received: October 12, 2017 **Accepted:** May 25, 2018

Abstract: Data mining is used to discover knowledge from information system. Clustering is one of the techniques used for data mining. It can be defined as a technique of grouping un-labelled data objects such that objects belonging to one cluster are not similar to the objects belonging to another cluster. Data mining tools refer to the software that are used for the process of efficiently analysing, summarizing and extracting useful information from different perspectives of data. This paper presents a comparative analysis of four open-source data mining software tools (WEKA, KNIME, Tanagra and Orange) in the context of data clustering, specifically K-Means and Hierarchical clustering methods. The results of the performance analysis based on the execution time and quality of clusters showed that WEKA tool outperforms the other tools with the lowest SSE of 199.7308 with an average execution time of 1.535 seconds. Knime has SSE of 222.217 but with an average execution time of 7.13 seconds, and then Tanagra with SSE of 269.3902 and average execution time of 2.01 seconds, Orange has the poorest performance with SSE of 388.78.

Keywords: Clustering, data mining, hierarchical, K-means, KNIME, tanagra, WEKA

Introduction

Data mining is the extraction of intriguing, relevant, constructive, previously unexplored and substantially valuable patterns or information from huge stack of data that can be used to make valid predictions (Prakash & Aarohi, 2015). It involves an integration of techniques from multiple disciplines such as database and data warehouse technology, machine learning, statistics, pattern recognition, neural networks, data visualization and information retrieval (Durairaj & Ranjani, 2013). In clustering, objects in the same cluster are as similar as possible and objects in different clusters are as dissimilar as possible. Although clustering is often used interchangeably with classification, they are two different concepts: clustering is a descriptive model that groups objects without prior knowledge of the classes while classification assigns objects to predefined set of classes based on trained model (Osama, 2008; Balogun *et al*., 2016). Clustering is sometimes referred to as unsupervised learning because there is no already-known result to guide the algorithm whereas, estimated values are compared with known results in supervised learning algorithms such as classification.

Data mining tools refer to the software that can be used for extracting and summarizing useful information from different perspectives of data. Notable among the available tools are Wakaito Environment for Knowledge Analysis (WEKA) (Hall *et al*., 2009); Orange, KNIME and R (Spector, 2004); Tanagra, Tavera and Rapid miner (King & Elder, 1998). In this study, four of these tools that support clustering are discussed and studied. This study will make it easier for a beginner to understand the concept of data clustering and to choose suitable tools for any of the clustering algorithms since not all the tools support all clustering algorithms and every tool has its own advantages and disadvantages (Mulik & Gulawani, 2012; Subathra *et al*., 2015)

This study is aimed at evaluating the performance of the four open-source tools in clustering breast cancer dataset, specifically the Wisconsin Breast Cancer Database Original (WBC) dataset collected from the University of California Irvine (UCI) repository. Feature selection was done to decrease the dimensionality of the datasets. The performance of the tools were measured based on time taken to build clusters, number of clusters built and the quality of the clusters measured as the sum of squared error function.

The other parts of this paper are organised as follows: a review of related works is presented in section 2; Section 3 describes the clustering analysis; section 4 presents the clustering concepts. The experimental setup for the study and tools description is presented in Section 5. Results are presented and discussed in Section 6 and the paper is concluded in Section 7.

Various contributions have been added to existing knowledge in the literature regarding data mining in the past few years, most importantly; clustering techniques. This section presents the review.

Tamije *et al.* (2011) presented a paper on the performance analysis of clustering algorithms in brain tumor detection of MR images. They analysed various clustering techniques which are K-means, Self-Organizing Map, Hierarchical Clustering and Fuzzy C-means on the basis of execution time taken and accuracy, to track tumor objects in Magnetic Resonance (MR) brain images. They combined the algorithms one by one and applied histogram clustering. K-means and Hierarchical clustering obtained better results. Sankar (2011) clustered customer data using IBM Intelligent Miner. An organization retail smart store dataset was used and clustered using demographic clustering technique. High, medium and low-value customers were identified at the end of the study. Maryam *et al.* (2012) implemented and compared four clustering algorithms namely K-means, Single-Linkage, DBSCAN and Self-Organizing map. F1 measure was used in order to measure accuracy and quality of clustering. They also compared between two methods, mapping to two dimensional space and statistical average and concluded that mapping method is better.

Ahmed *et al.* (2016) presented the Performance Analysis of Various Open Source Tools on Four Breast Cancer Datasets using Ensemble Classifiers Techniques. They performed an experimental study to investigate the quality of fusion methods for combining classifiers in an ensemble. They applied confusion matrix accuracy and 10-fold cross validation method to get the best subset of classifiers for each data set separately and concluded that Orange is the most superior of the four tools. Jyoti & Neha (2012) developed Document Similarity Soft Clustering (DSSC), a method of summarization via soft-clustering algorithm based on similarity function using hierarchical clustering technique. They used five datasets and analysed them with rapid miner and Tanagra data mining tools.

Prakash & Aarohi (2015) studied the performance analysis of clustering algorithms in data mining on WEKA tool. They analysed K-means, hierarchical, expectation maximization, MTree, Farthest first, Canopy, LVQ, Cascading k-means and DBScan algorithms and thereafter compared the results obtained on the basis of number of cluster, cluster instances, square error, time taken to build model and log likelihood. K-means was found to have performed better.

Raj *et al.* (2014) presented a study which compared clustering algorithms on the bank dataset in both normalized and un-normalized formats using WEKA data mining tool. They performed a comparative analysis of four clustering algorithms. They are k-means algorithm, hierarchical algorithm, expectation maximization algorithm and density based algorithm. K-means produced better results in terms of accuracy and efficiency compared to other algorithms.

Materials and Methods

Data clustering and tools

Data clustering

Clustering basically is the process of dividing a set of data into different groups/classes to find groups that are different from each other, and whose members are very similar to each other. There are broadly two types of clustering: Partition based clustering and hierarchical clustering (Shima, 2007).

i. Partition Based Clustering: it is based on the concept of relocating data objects between the clusters iteratively (Alan, 2000). The quality of the cluster is measured by the clustering criterion; sum of squared error (Hayaska *et al.*, 1996). K-means algorithm is one of the partition based algorithms. The principal concept of this clustering technique is to assign N clusters to each k data objects, where N is user-defined. For good results, the centroids/mean should be placed as far as possible from each other. Each data object of the dataset is associated with the nearest centroid until there is no data object pending. After the initial phase of grouping, the new centroid of each N clusters is determined. Once there are N new centroids, a new process of binding between the original data objects and the new centroids is started. Hence a loop is formed. As a result of the loop formation, the position of N centroids keep on changing until no more change in the position occurs. A good cluster must have low inter cluster similarity and high intra cluster similarity (Raj *et al.*, 2014).

ii. Hierarchical Clustering: Hierarchical algorithms combine or divide existing groups, creating a hierarchical structure, otherwise known as dendogram, which reflects the order in which groups are merged or divided. Hierarchical Clustering method forms clusters progressively and it is of two forms namely: Agglomerative and Divisive hierarchical algorithm. Agglomerative hierarchical algorithm works as follows (Andrew, 2014).
   a. Assign each of the data objects to a cluster each. That is, if there are N data instances, N clusters are formed with 1 data object each.
   b. Find the nearest pair of clusters and merge them together to form a pair, so that there will be N-1 clusters left
   c. Calculate the distance between the new cluster and each of the old ones.
   d. Repeat steps ii and iii until all the data objects have been clustered into cluster of size N.

Step iii can be performed using the metric technique and the linkage technique (Prakash & Aarohi, 2015). Metric technique is used to measure the distances between two data objects while linkage technique specifies the measurement between two clusters. Manhattan and Euclidean distance are the most used techniques in implementing Metric technique. Manhattan distance considers the sum of differences of the corresponding data objects while Euclidean distance is the shortest distance between the two data objects. The formulae for both metrics are given in Equations (1) and (2).

Manhattan formula: $d = \sum_{i=1}^{n} |x_i - y_i|$ ……… (1)

Euclidean formula: $d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$ …… (2)

The process of data clustering involves identifying the dataset to be used for training the machine learning algorithm. Some of the datasets sometimes contain attributes that are irrelevant for clustering. Therefore, datasets must be pre-processed either manually or using pre-processing algorithms to remove the irrelevant attributes. The basic steps involved in clustering data are:

i. Selection of data mining tool: Selection of data mining tools depends on the type of machine learning algorithm supported by a particular tool and the algorithm to be used to solve the problem at hand as well as the nature/format and type of dataset to be used.

ii. Identification of required dataset: knowledge of the type of dataset supported by a particular clustering algorithm is very important because some algorithms do not perform well on datasets including outliers. An example of such algorithms is k-means algorithm.

iii. Data pre-processing: data pre-processing involves the detection of relevant attributes or features and removal of attributes that are irrelevant or redundant in building a model from the dataset before clustering. It helps to improve the speed of the clustering algorithms and increase comprehensibility (kumar & Minz, 2014).

iv. Selection of the clustering algorithm: choice of clustering algorithms to be used is very important and it depends on the problem to be solved.

v. Cluster validation and interpretation of results: clustering validation is one of the important issues essential to the success of clustering applications. Clusters can be validated using either external criteria or internal criteria. External validation uses external information that are not present in the data to evaluate the extent to which the clusters discovered by a clustering algorithm matches the one specified by the given class labels. Examples of external criteria measure are precision and recall, F-Measure, Purity and Entropy. Internal validation is often based on compactness and separation. Compactness is also known as cohesion, it measures how closely related the objects in a cluster are while separation measures how well-separated a cluster is from other clusters. Examples of internal criteria measure are Root-Mean-Square Standard Deviation (RMSSTD), R-Squared, Silhouette Index and so on.

Data mining tools

Over the years, several software tools for carrying out the basic functions required for data mining tasks have been developed. This section presents the most popular data mining tools that can perform K-Means and hierarchical clustering methods. These tools and clustering methods are the focus of this study.

i. Wakaito Environment for Knowledge Analysis (WEKA): WEKA toolkit is a cross-platform generally used toolkit for data mining that was originally developed at the University of Waikato in New Zealand (Hall, Frank, Holmes, Pfalminger, & Rontennam, 2009). It supports the following file formats: *.arff*, *.csv*, *.data*, *.bin*, *.dat* and *.xrff*.

ii. Konstanz Information Miner (KNIME): KNIME is an open source tool developed at the University of Konstanz in January 2014. It is also a cross-platform tool that works on Linux, OSX and Windows platforms, and has more than 100 nodes for analysing data. It can incorporate WEKA analysis module and R scripts through additional plug-ins. It supports the following file formats: *.json*, *.xml*, *.arff*, *.csv* and *.atom*. KNIME works on.

iii. Tanagra: Tanagra is a free data mining tool developed at Lumiere University Lyon 2, France in 2004 for research and academic purposes. It supports expectation maximization, hierarchical, k-means, Self-Organizing Map and neighbourhood graph clustering algorithms. It supports the following file formats: *.txt*, *.arff* and *.xls*.

iv. Orange: Orange was invented at University of Ljubljana, Slovenia. It is implemented in C++ and Python. It also supports files in *.csv*, and *.tsv* formats. Orange works on different versions of Linux, Apple's Mac OS X and Microsoft Windows.

Method

In other to test the performance of the data mining tools using clustering approach, each of the tools must pass through the stages depicted in Fig. 1 (Jyoti. & Neha, 2012):

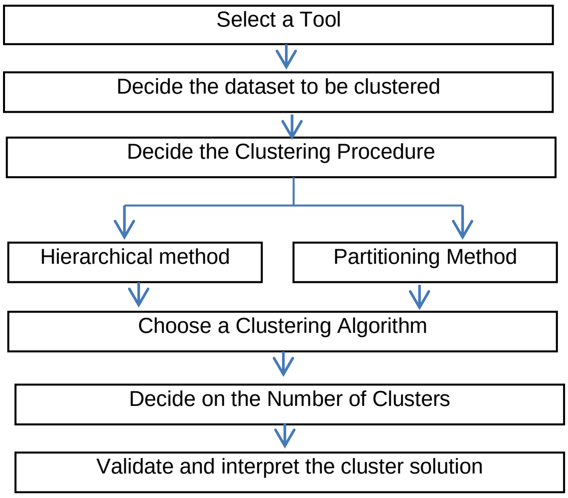
Select a Tool
Decide the dataset to be clustered
Decide the Clustering Procedure
Hierarchical method
Partitioning Method
Choose a Clustering Algorithm
Decide on the Number of Clusters
Validate and interpret the cluster solution

Fig. 1: Tool evaluation process

In this work, WEKA, Orange, KNIME and Tanagra tools were used to analyse Wisconsin Breast Cancer Original (WBC) dataset obtained from UCI machine learning repository. WBC has 699 instances and 12 attributes which are of type integer. The tools were evaluated based on their capability in performing clustering using two popular algorithms: K-means and hierarchical algorithms. The tools performances are measured on the basis of time-taken to build model, number of clusters formed, cluster instances and Sum of Squared Error (SSE) function also known as Within Sum of Squares. The computation of SSE is done using Equation (3):

$$SSE = \sum_{i=1}^{n}(x_i - \bar{x})^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Where: n is the number of observations, x_i represents the value of the ith observation and $\bar{x}$ is the mean of all the observations. SSE measures the cluster quality. A lower SSE value indicates that the instances in a cluster are very similar (homogeneous) while a high SSE indicates that the instances in the same cluster have a reasonable degree of differences between them and thus, they are not of good quality (Sayontan, 2016). It (SSE) can be computed for k-means algorithm alone because mean/centroid is required.

Analysis of the datasets on all the four data mining tools under study was performed on a computer system with the following configuration:

i. Processor: Intel(R) Core(TM) i5-3210M CPU at 2.50GHz
ii. RAM: 6GB
iii. Hard disk: 700GB
iv. Operating System: Windows 8.1 (64-bit)
v. Weka 3.8.1
vi. Orange 3
vii. Knime 3.3.1
viii. Tanagra 1.4.50

Results and Discussion

Table 1 presents the experimental results of the performance of each of the four tools used in the clustering of the collected dataset. Number of clusters (k) is user-specified for k-means algorithm for all the tools.

A lower SSE suggests that the instances in a cluster are similar (homogeneous) while a high SSE suggests that the instances in the same cluster have a reasonable degree of differences between them (Sayontan, 2016). From Table 1 above, Weka has the lowest SSE of 199.7308, followed by Knime that has SSE of 222.217, and then Tanagra with SSE of 269.3902, Orange has the highest SSE of 388.78. It can therefore be said that Weka produced the best quality clusters, followed by Knime, followed by Tanagra while the clusters produced by Orange is of less quality.

Table 1: Tool performance results

Tools	Algorithm	No. of Clusters (k)	Cluster Size	SSE	Execution Time (sec.)
WEKA	K-means	3	134 (19%) 449 (64%) 116 (17%)	199.7308	0.03
	Hierarchical	3	474 (68%) 220 (31%) 5 (1%)	N/A	3.04.
Orange	K-means	3	109 (16%) 127 (18%) 463 (66%)	388.780	N/A
	Hierarchical	3	464 (66%) 221 (32%) 14 (2%)	N/A	N/A
KNIME	K-means	3	220 (31%) 158 (23%) 321 (46%)	222.217	0.149
	Hierarchical	3	1 (0%) 480 (69%) 218 (31%)	N/A	14.11
Tanagra	K-means	3	208 (30%) 34 (5%) 457 (65%)	269.3902	0.047
	Hierarchical	3	220 (32%) 36 (5%) 443 (63%)	N/A	3.969

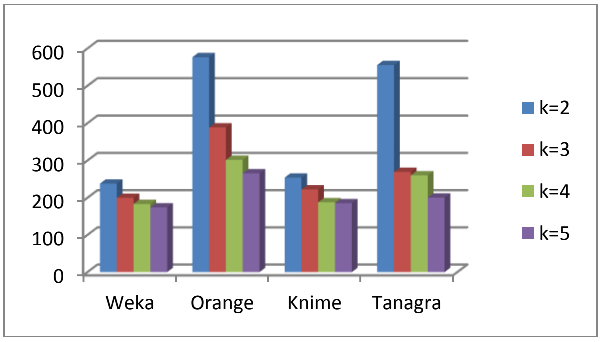

Fig. 2: Cluster quality representation using sum of squared error function

From Fig. 2, it can be observed that the SSE value decreases with an increase in the value of *k*. This means that as the value of k increases the quality of clusters produced by all of the tools increases too. It can also be observed that the SSE values for all values of k are the lowest in Weka tool while the SSE values are highest in Orange tool. It can therefore be said that Weka of all the analysed tools produced clusters of the best quality while the clusters produced by Orange are of least quality. In addition, for all the tools, hierarchical clusterer took more time to form clusters. Furthermore, Weka took the least execution time to form clusters for both k-means and hierarchical algorithms, followed by Tanagra tool, while Knime has the highest execution time. That of Orange tool could not be measured because it does not have the inbuilt execution time measurement functionality. However, for all the tools, k-means algorithm formed clusters faster than hierarchical algorithm.

Conclusions

This study has evaluated the performance of four data mining tools in performing K-means and hierarchical clustering. The tools are WEKA, Orange, Tanagra and KNIMEand their performance was based on their execution time and quality of clusters formed which is measured as Sum of Squared Error (SSE). The Wisconsin Breast Cancer (WBC) dataset obtained from UCI machine learning repository was used for the study. WEKA outperformed the other tools in terms of execution time; it formed the clusters faster when using both the K-means and Hierarchical clustering methods. WEKA is followed by Tanagra and KNIME in that order while execution time is not measured for the Orange tool because it does not have the inbuilt execution time measurement functionality. Thus, WEKA is the fastest in performing K-means and hierarchical clustering compared to other tools. This implies that WEKA is the choice to make when it is important to build clusters fast. Also, WEKA produced the best quality clusters, followed by KNIME, followed by Tanagra while Orange produced the least quality clusters. From the results obtained, it can be concluded that WEKA obtained the best result for clustering breast cancer dataset in terms of cluster quality and execution time taken.

Clustering is a vivid method. The solution is not exclusive and it firmly depends upon the analysts' choices (Reza & Shai, 2013). The validation of clustering structures is the most difficult and frustrating part of cluster analysis and it is often said that clusters are in the eyes of the beholder, therefore, the outcome of clustering should never be generalized (Enza, 2014).

As a future work, comparison between these four tools or more can be studied but using more or different evaluation metrics other than those considered in this study. Some other important factors include size of dataset and normalization can also be considered. Comparing between the results of these tools using huge and small datasets will definitely affect the cluster quality; also, using normalized or non-normalized data could also give different results. The tools can also be evaluated in the context of other machine learning tasks such as classification and association using datasets and algorithm designed for such tasks.

References

Ahmed AI, Attalah IH & Negm EM 2016. Performance Analysis of Open Source Tools on Four Breast Cancer Datasets Using Ensemble Classifiers Techniques. *Int. J. Engr. Res. & Techn. (IJERT),* 5(03), 610; available at http://www.ijert.org.

Alan DG 2000. An iterative relocation algorithm for classifying symbolic data. In: *Studies in Classification, Data Analysis and Knowledge Organization, Springer Link*, pp. 17 – 23.

Andrew M 2014. *Tutorial Slides.* Retrieved May 05, 2017, from http://www2.cs.smu.edu/-awm/tutorials/kmeans.html.

Balogun AO, Mabayoje MA, Salihu S & Arinze SA 2015. Enhanced classification via clustering using decision tree for feature selection. *Int. J. Appl. Information Sys. (IJAIS)*, 9(6): 11-16.

Enza M 2014. *Cluster Analysis.*

Hall M, Frank E, Holmes G, Pfalminger BW & Rontennam P 2009. The WEKA data mining software: An update. *ACM SIGKDD Exploration Newsletter, 11*.

Hayaska T, Toda N, Usui S & Hagaiwara K 1996. Least square error of function representation with discrete variable basis. *Proceedings of the 1996 IEEE Signal Processing Society Workshop.*

Jyoti K & Neha K 2012. Document delivery using clustering techniques and finding best cluster. *Int. J. Sci., Engr & Techn. Res., (IJSETR)*.

Maryam B, Mohammed F & Elnaz Z 2012. Review and comparison between clustering algorithms with duplicate entities detection purpose. *Int. J. Comp. Sci. & Emerging Techn.*, 108-114.

Osama AA 2008. Comparisons between data clustering algorithms. *The Int. Arab J. Information Techn.,* 5(3): 320-322.

Prakash S & Aarohi S 2015. Performance analysis of clustering algorithms in data mining in WEKA. *Int. J. Advances in Engr. & Techn.*, 1866 – 1868.

Raj B, Sunil S & Juhi S 2014. A comparative analysis of clustering algorithms. *Int. J. Comp. Applic.*, (0975-8887), 100: 15.

Reza BZ & Shai BD 2013. *A Uniqueness Theorem for Clustering.* Stanford edu.

Sankar R 2011. Customer data clustering using data mining technique. *Int. J. Database Mgt. Sys., (IJDMS),* 3(4): 1 – 11.

Sayontan S 2016. *Cluster Analysis for Marketing.* Retrieved April 5, 2017 from Understanding the Sum of Squared Error: www.clusteranalysis4marketing.com/interpretation/sum-of-squared-error-sse/

Shima J 2007. Measurement and comparison of clustering algorithms. *School of Mathematics and Systems Engineering*, 7.

Subathra P, Deepika R, Yamini K, Arunprasad P & Shriram K 2015. A study of open source data mining tools and its applications. *Res. J. Appl. Sci., Engr. & Techn.*, 1102.

Tamije P, Palanisamy V & Purusothaman T 2011. Performance analysis of clustering algorithms inbrain tumor detection of MR images. *European J. Scientific Res.* 62: 321 – 330.