IADIS International Journal on Computer Science and Information Systems
Vol. 10, No. 2, pp. 65-78
ISSN: 1646-3692

COMPARATIVE PERFORMANCE ANALYSIS OF CLUSTERING TECHNIQUES IN EDUCATIONAL DATA MINING

Kyle DeFreitas. *University of the West Indies, St Augustine, Trinidad and Tobago*

Margaret Bernard. *University of the West Indies, St Augustine, Trinidad and Tobago*

ABSTRACT

Clustering analysis provides a useful way to group objects without having previous knowledge about the data being analysed. In this paper, we first survey the research done on clustering analysis in education and identify the algorithms used. We then present a case-based experiment to show the relative performance of clustering algorithms with Learning Management System log data. We compare partition-based (K-Means), density-based (DBSCAN) and hierarchical (BIRCH) methods to determine which technique is the most appropriate for performing clustering analysis within the LMS. We conclude by showing that partition-based methods produce the highest Silhouette Coefficient values and the better distribution amongst the clusters. The BIRCH algorithm also performs fairly well and can act as a good starting point to find cluster groups in new datasets as the algorithm does not required that the number of clusters be specified a priori.

KEYWORDS

Clustering, Educational Data Mining, Learning Management Systems, Web Usage Mining, Moodle

1. INTRODUCTION

Learning Management Systems (LMS) provide educators with a platform to distribute information, to engage students and manage distance or online classes more effectively. As more educational content is delivered through web-based systems such as the Modular Objective Oriented Development Learning Environment (Moodle), the need to analyse the usage and impact of the content within the web-based environment is a high priority for many institutions. Moodle, like many LMS, will record every transaction and activity of the user within the system. These transactions are usually stored in relational databases, which can become relatively large. Due to the large numbers of users and interactions provided by the use of these web platforms, traditional tools of observations and analysis have given way to the reliance on data mining techniques to identify trends and useful patterns within the data. Some of the more useful techniques include statistical methods, data visualization, association rule mining, classification and clustering (Romero et al. 2008).

While data mining is traditionally applied to business and scientific problems, there is an increased interest in the application of data mining techniques for tackling problems within the educational domain (Baker & Yacef 2009). Over time a number of authors have proposed that the unique constraints of education require special attention and the term Educational Data Mining (EDM) is used to identify analysis that considers the additional requirements of the discipline. Some of these constraints include the domain of course content (Romero et al. 2004), the user model (Pahl & Donnellan 2002), the purpose of the reporting (Romero & Ventura 2007), the process of learning (Li et al. 2004) and the process for analysis within the LMS environment (DeFreitas & Bernard 2014).

A number of previous studies have been conducted that evaluated the performance of clustering algorithms. For example, Berkhin (2006) provides a detailed survey of the various techniques, while Maulik and Bandyopadhyay (2002) evaluate the general performance of the clustering algorithms. While the previously mentioned papers utilize artificial datasets or datasets supplied with the toolkits, this paper will consider the performance of the algorithms within the educational context utilizing real data generated from the Moodle installation of a tertiary institution offering a variety of courses in different disciplines. We consider datasets with large numbers of students and diverse courses.

This paper establishes which clustering algorithm is most appropriate for performing analysis on web log data for learning management systems. It will highlight the potential characteristics of the algorithms and the data within web-based educational systems, which can account for the observations made during the comparison. It will consider the impact of feature selection on the performance of the clustering analysis. A number of studies have utilised K-Means algorithm on the basis of the ease of use, simplicity and performance of the algorithm. The questions we asked were: Was the use of K-Means the best approach? Are there other categories of algorithms that may perform better or act as a better starting point for analysis within EDM?

This paper is structured as follows: Section 2 gives an overview of the families of clustering algorithms along with a description and justification of the algorithm selected to represent each family of algorithms. Section 3 provides a survey of the literature of the research on clustering analysis in education. Section 4 describes the methodology for testing the performance of the algorithms with the datasets selected. Section 5 provides the results and discussion of the performance of the respective algorithms. Finally, Section 6 concludes by summarizing findings and identifying possible future work and implications of the work.

2. BACKGROUND

Clustering analysis is one of the techniques used in data mining and involves the partitioning of a set of data objects into subsets. These subsets or clusters are used to organize objects in such a way that each object within in a cluster are similar to one another yet they are dissimilar to objects which belong to other clusters (Han et al. 2012). Clustering of usage logs can be used to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to display personalised web content to users (Srivastava et al. 2000). It can also be used to gain potential insights into the way successful learners utilize the web-environment and resources which can provide appropriate recommendations to other learners that are struggling with content (Zaiane & Luo 2001) and generating recommendations specific to the learning style of the student (Klašnja-Milićević et al. 2011). Clustering analysis can identify student groups with similar usage patterns for improving the performance of other types of analysis within educational environments (Bogarín et al. 2014). Perera et al. (2009) improves group feedback and engagement by utilizing clustering analysis to provide targeted information based on usage similarities extracted with clustering and sequential pattern mining in order to improve the level and quality of feedback.

This section highlights the various types of clustering and will justify the selection of algorithms used to represent each of the methods identified.

2.1 Clustering Methods

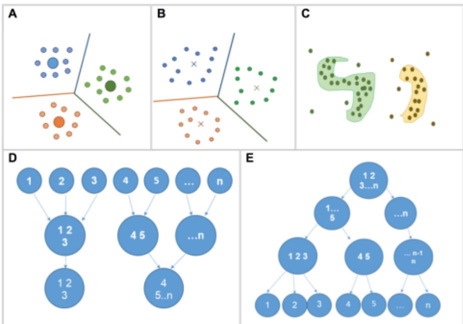

Figure 1. Family of clustering algorithms

Clustering algorithms can be classified by the data type analysed, similarity measure used to group data and the theory used to define the cluster (Sivogolovko & Novikov 2012). However, there is a great level of consensus on classifying algorithms on the basis of the techniques used to create the partitions (Berkhin 2006). In this paper we focus on three families of techniques: Partitioning, Hierarchical and Density-based methods as proposed by Kim & Han (2009) and Han et al. (2012). While other techniques such as fuzzy clustering (Pakhira et al. 2004), graph partitioning and grid-based methods (Berkhin 2006) exist, the selected clustering methods are some of the most commonly used and are available in various data mining tools such as WEKA (Mark Hall et al. 2014) and SciKit-Learn (Pedregosa et al. 2011). These techniques and data mining tools are also commonly used in educational data mining analysis (Romero et al. 2008).

A) *Partitioning methods*: Partitioning algorithms will attempt to group the **n** objects into a specified **k** partitions (k <= n) where each partition (sub-group) represents a cluster such that the objects within a cluster are “similar” to one another and “dissimilar” to objects in other clusters. Each cluster can be represented by a centroid, which can either be objects within the dataset (k-Medoids) illustrated in Figure 1A or values within the domain space (K-Means) as illustrated in Figure 1B.

B) *Hierarchical methods*: Hierarchical-clustering methods group data objects into a hierarchy or tree of clusters. Hierarchical clustering initializes a cluster system as a set of singleton clusters (agglomerative case) or a single cluster of all points (divisive case) and proceeds iteratively merging or splitting the most appropriate cluster(s) until the stopping criterion is achieved (Berkhin 2006). Agglomerative and divisive hierarchical clustering is illustrated in Figures 1D and 1E respectively.

C) *Density-based methods*: Density-based clustering utilizes the local densities to determine how objects are grouped to create clusters (Kriegel et al. 2011). These local densities, referred to as neighbourhoods, allow for the identification of clusters of arbitrary shapes as illustrated in Figure 1C compared to the elliptical and convex shapes of the partitioning methods that solely rely on the distance between objects within the data space.

2.2 Clustering Algorithms

2.2.1 K-Means

The K-Means algorithm will group **n** items into **k** user-defined clusters (k <= n). This grouping is done on the basis of minimizing the sum of squared distances between items and the respective centroids. This reorganization will continue for each iteration of the algorithm, until it no longer creates a change in organization of objects within the clusters or until some threshold is reached. The centroid for the K-Means algorithm is represented as the mean value in the data-space calculated from the attribute values of objects within the cluster.

While there are other partitioning algorithms such as the Partitioning Around Medoids (Kaufman & Rousseeuw 1990), which uses an actual object within the object space as the centroid and Clustering Large Applications based upon Randomized Search (Ng & Han 2002), which utilizes spatial features of objects which facilitate analysis of not only point objects but more complex polygon-based objects as well, the K-Means algorithm performs acceptably well with a wide variety of data sets (Maulik & Bandyopadhyay 2002) and can therefore be a representative algorithm for the partitioning methods.

2.2.2 BIRCH

The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is an agglomerative hierarchical clustering algorithm that performs well on large datasets with high dimensionality (Witten et al. 2005). BIRCH creates a height-balanced clustering feature tree of nodes that summarizes data by accumulating its zero, first, and second moments of the cluster. The resulting clustering feature (CF) is used to compute the centroids and measures the compactness and distance of the cluster. This storage of the statistical information such as the number of data points, linear sum of N points, and the sum of squares of N points in the CF helps to reduce the additional recalculations and makes the merging of sub-cluster incrementally more efficient.

2.2.3 DBSCAN

The Density Based Spatial Clustering of Application with Noise (DBSCAN) is a density-based technique that separates data points into three parts: Core points (points that are at the interior of the cluster), Border points (points which fall within the neighbourhood of the core point) and Noise points (points that are not a core point or a border point). It utilizes a defined minimum radius (ε) and minimum cluster size (minPts) to determine which density areas within the space can be considered a cluster.

DBSCAN is sensitive to the choice of epsilon. Small values of epsilon will cause clusters to be categorised as noise while large values result in denser clusters to be merged. While DBSCAN does not require the number of clusters to be specified it may perform poorly when the number of dimensions for the objects is high as the computation required to calculate the relative density of objects in respect to each other is dependent on the dimensions.

Table 1. Summary of the comparison between algorithms utilized in the experiment

Algorithm	Type	Metrics used	Strengths	Weaknesses
K-Means	Partitioning	Sum of Squared Distance between points	1. Performs well on wide variety of datasets 2. Can work reasonably well on large and high dimensional datasets	1. Requires trying various values of k to determine the best cluster 2. Performance is dependent on the initialization of algorithms 3. Only works well on convex and elliptical data
BIRCH	Hierarchical	Euclidian Distance between points	1. Useful for analysing data where there may be an underlying hierarchical structure within the data 2. Efficiently analyses large datasets	1. Does not scale very well to data with high dimensionality
DBSCAN	Density Based	Euclidian Distances between nearest points	1. Can identify clusters of arbitrary shapes 2. Does not require number of clusters for initialization	1. Poor handling of outliers may affect accuracy of cluster members

2.3 Clustering Validation

Clustering validation techniques are used to evaluate the performance of clustering algorithms and can be grouped as internal and external measures. External measures utilize the ground of truths, i.e. an externally provided class labels to determine how well the created clusters represent the true groupings of the supplied dataset. Internal measures evaluate the properties and organization of the clusters created in order to determine how well the clustering algorithm performed. Internal validation methods evaluate how closely the objects in the clusters are and how distinct a cluster is from another. The Silhouette Coefficient (SC) is an internal validity measure that evaluates the clustering performance based on the pairwise-difference between and within cluster distances (Liu et al. 2010). The silhouette values closer to 1 represent clusters that are more accurately clustered while scores closer to -1 represents incorrect clustering. In a comparison between 11 internal validity measures, Liu et al. (2010) identified that the silhouette performs well on a variety of data types that vary in structure, noise and skewed distributions. A similar comparison by Sivogolovko and Novikov (2012) demonstrates that the silhouette coefficient measure works well between partitioning and density-based clustering methods across various data types. Though not ideal for non-convex clusters, there is still an acceptable performance, which makes the Silhouette Coefficient an acceptable validity measure to compare the performance between the categories of clustering algorithms.

3. CLUSTERING IN EDUCATION

This section provides an overview of clustering analysis within the educational context. We review several publications from 2009 to 2015, and organize the research according to purpose and algorithms used. The intent is that this would provide valuable insights into how clustering analysis is used within educational environment.

Table 2. Survey of Clustering analysis in education from 2009 to 2015

Authors	Purpose	Algorithm
(Permata Alfiani & Ayu Wulandari 2015)	Classification	K-Means
(Campagni et al. 2015)	Evaluation	K-Means
(Bogarín et al. 2014)	Evaluation	EM
(Bovo et al. 2013)	Evaluation	K-Means, X-Means, EM, Hierarchical
(Antonenko et al. 2012)	Evaluation	Ward, K-Means
(Tam et al. 2012)	Evaluation	K-Means
(López et al. 2012)	Classification	K-Means, X-Means, EM, Hierarchical, FarthestFirst, sIB
(Tair & El-halees 2012)	Evaluation	K-Means
(Mashat et al. 2012)	Classification	K-Means, SOM, Fuzzy Clustering
(Klašnja-Milićević et al. 2011)	Classification	Unknown
(Ayesha et al. 2010)	Classification	K-Means
(Amershi & Conati 2009)	Exploration	K-Means
(Perera et al. 2009)	Exploration	K-Means
(Chen & Chen 2009)	Exploration	K-Means, Fuzzy clustering

Clustering in education has been extensively explored by a number of authors as summarized in Table 2. In the table we highlight the various purposes for which clustering analysis was applied, and, where available, we highlight the algorithm that is utilized for the actual analysis. In this table we group the purpose of the analysis into three categories: Classification, Evaluation and Exploration. The Classification group refers to the use of clustering algorithm to formulate models which can be subsequently used for classification tasks such as prediction. The Evaluation group refers to use of clustering techniques to create groups in order to gain a better understanding of the behaviours of users and content within the educational environment. Exploration uses clustering techniques to provide the building block for further analysis of the data. This may be for subsequent clustering analysis or the input into other analysis techniques.

Using this table, we can note a number of interesting trends. Firstly, the K-Means algorithm remains a popular choice for performing clustering analysis among authors. As expected the majority of studies utilize clustering analysis for evaluation, however twenty percent (20%) utilize clustering for creating models for further exploration and thirty-six percent (36%) use it as the basis for subsequent classification. While some authors use clustering analysis by itself, a number of authors such as Campagni et al., (2015), utilize clustering techniques in conjunction with sequential patterns or other non-related patterns for extracting information.

While the use of the K-Means algorithm is particularly notable in literature, the justification for its use is often overlooked or not fully developed. For example (Amershi & Conati 2009), (Perera et al. 2009) uses the K-Means algorithm because of its simplicity and scale wells within a certain limit. The purpose of this study is to extend this knowledge by considering a case based analysis of the comparative performance across the categories of clustering technique.

One of the interesting observations of the survey of publications is that many authors suggest that there is a need to test the results across a wider cross section of courses in order to generalize the results obtained from their experiments. They also showed that the number of attributes has a strong relationship with the accuracy of the algorithms. Some algorithm performance decreased when removing attributes. In this case study, we analysed the performance of clustering algorithms over a large number of students and across a wide set of diverse courses; we considered also the impact of dimensionality to the generated clusters and performance of the analysis.

4. METHODOLOGY

The Moodle LMS database contained 4,894,199 log records generated by 12,400 students over 921 courses offered by the tertiary institution between 2008 and 2015. To gain an insight into how clustering analysis can be done within different courses, the performance of the algorithm was evaluated on different sizes of datasets. The total number of users within the course was used to select which courses were analysed. Five courses were selected for the analysis: C1496, the course with the maximum (1496) number of students; C333, the course with the number of students (333) in the upper quartile; C236, the course with the median (236) number of students; C162, the course with the number of students (162) in the lower quartile; and C136, the course with the minimum (136) number of students above 100, our stipulated threshold.

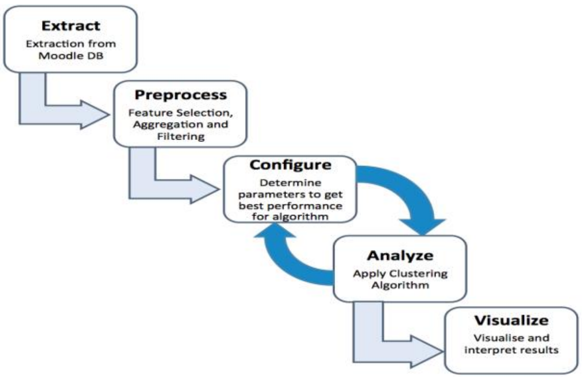

Figure 2. Steps for performing clustering evaluation

The pre-processing of the data was accomplished in three steps: *data summarization, feature selection and data filtering*.

The *data summarization* was used to generate the feature vector that represented each individual user within the course. The feature vector consisted of the number of times a user viewed resources, posted to a forum, and attempted quizzes within the LMS. These dimensions were selected as they represented the tasks performed most frequently within the LMS. The vector also contained four time dimensions: the total number of times the user accessed the system in the early morning, morning, afternoon and night. These were selected as they facilitated a decomposition of the time of day and frequency that a student would perform activities within the course. The seven features of the vector attempt to represent both the activities performed with a breakdown of the time users selected. From our initial investigation of the data, not all courses may have utilized forums and quizzes and this model will ensure that feature vector can represent these courses as well.

Table 3 provides a further explanation of the dimensions within the feature vector that represented information about a student within a single course.

Table 3. Feature vector generated from data summarization of Moodle log data

Name	Description	Extraction method from Moodle Logs
Views (d1)	Number of times the user viewed any activity or resource within the LMS	Aggregate all view related actions such as "view", "view form" and "view discussion" for each user within the course.
Posts (d2)	Number of times the user submitted a post within the forums of the course	Aggregate all actions related to contributing to the forums. It includes actions such as "add post", "update post" and "add discussion" for each user within the course.
Quizzes (d3)	Number of times the user attempted quizzes within the course	Aggregate all actions related to attempting quizzes. It includes actions such as "attempt" or "continue attempt" for each user within the course.
Early Morning / Late Night (d4)	Number of times the user performed the previously stated activities between 12:00 am and 6:00 am	Aggregate all actions for the three activities for the specified time period.
Morning (d5)	Number of times the user performed the previously stated activities between 6:00 am and 12:00 pm	
Afternoon (d6)	Number of times the user performed the previously stated activities between 12:00 pm and 6:00 pm	
Night (d7)	Number of times the user performed the previously stated activities between 6:00 pm and 12:00 am	

The *feature selection* chooses a combination of the dimensions produced in the data summarization stage for analysis. The experiment uses the seven dimensions (d1 … d7) identified in Table 3 and a reduced number of dimensions in an attempt to determine if the dimensionality affects the performance of the algorithms relative to each other. Two dimensions were selected for the reduced feature-vector analysis, the total number of views (d1) and the frequency of operations in the afternoon (d6). These features were selected because they had consistently significant values for each of the courses selected for analysis.

The *data filtering* applies the normalization (N) or scaling (S) filters on the data set before analysis. The filtering reduces the extreme differences in the attributes, to make calculation of distances and therefore clusters easier for the respective algorithms.

Following the pre-processing of the data, the data was *analysed* using the respective algorithms and each algorithm was *configured* until the best value for the silhouette coefficient was determined. For the K-Means, the value of k was adjusted ($2 \leq k \leq 10$); for the BRICH algorithm, the threshold ($0.01 \leq thd \leq 0.09$) and branching factors ($10 \leq bf \leq 50$) were adjusted and for the DBSCAN algorithm, epsilon (ε) ($0.01 \leq \varepsilon \leq 0.09$) and minimum points ($2 \leq$ min_pts ≤ 10) were adjusted.

The results were documented and tabulated in Table 4 which is highlighted in the next section. Further investigation into the comparison was done using histograms in Figure 3.

5. RESULTS

Table 4 below shows the relative clustering performance of the three selected algorithms, K-Means (with normalized and scaling filters), DBSCAN, and BIRCH applied to the five courses (C1496, C333, C236, C162, and C136). The table shows the results when all seven dimensions were selected as the feature vector as well as results when only two of the seven dimensions were selected.

Table 4. Results of clustering performance testing between algorithms used in the experiment

			C1496			C333			C236			C162			C136		
Algorithm	Dimension	Filter	Clusters	Silhouette Coefficient	Time (s)	Clusters	Silhouette Coefficient	Time (s)	Clusters	Silhouette Coefficient	Time (s)	Clusters	Silhouette Coefficient	Time (s)	Clusters	Silhouette Coefficient	Time (s)
K-Means	7	N	7	0.457	0.146	7	0.485	0.105	3	0.421	0.044	7	0.442	0.097	8	0.388	0.073
K-Means	7	S	2	0.668	0.056	3	0.760	0.020	3	0.641	0.077	2	0.855	0.047	2	0.905	0.011
DBSCAN	7	N	2	0.087	0.044	33	0.388	0.003	2	0.195	0.003	23	0.264	0.001	24	0.164	0.001
BIRCH	7	N	4	0.466	0.630	5	0.475	0.023	3	0.398	0.052	5	0.410	0.023	3	0.370	0.019
K-Means	2	N	2	0.704	0.020	7	0.902	0.031	2	0.926	0.011	2	0.873	0.037	2	0.928	0.009
K-Means	2	S	2	0.765	0.043	2	0.793	0.042	2	0.767	0.052	8	0.674	0.079	2	0.929	0.033
DBSCAN	2	N	2	0.780	0.039	16	0.906	0.009	2	0.922	0.003	2	0.859	0.002	1	-1.000	0.010
BIRCH	2	N	2	0.801	0.105	7	0.902	0.013	2	0.934	0.015	2	0.873	0.011	2	0.928	0.008

Figure 3 displays the Silhouette Coefficient given in Table 4 for each of these courses for each algorithm (with normalized pre-processing). Figure 3a shows the analysis on 7 dimensions and Fig 3b shows the analysis on 2 dimensions.

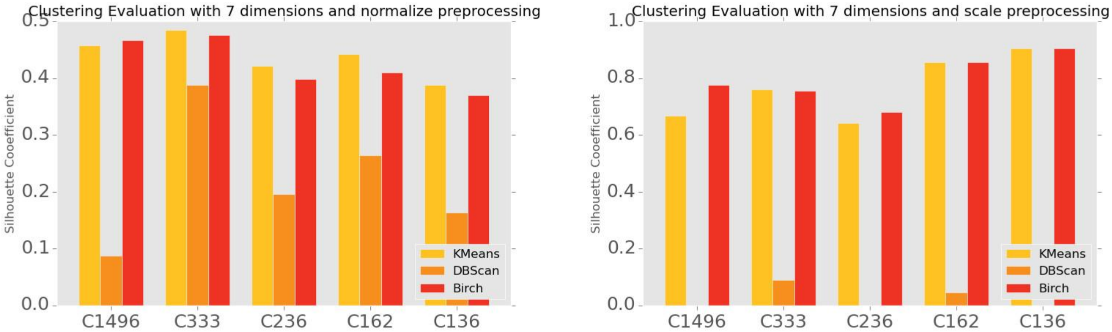

Figure 3. Clustering results for 7 and 2 dimensions and normalized preprocessing

Table 5. Results of clustering performance testing between algorithms used in the experiment

Algorithm	Courses	D	F	Clusters							
				1	2	3	4	5	6	7	8
K-Means	C1496	7	N	313	234	222	251	63	184	229	
DBSCAN	C1496	7	N	1292	16						
BIRCH	C1496	7	N	560	504	364	68				
K-Means	C333	7	N	15	77	77	46	32	73	13	
DBSCAN	C333	7	N	4	46	41	10	4	4	29	2
BIRCH	C333	7	N	89	41	101	92	10			
K-Means	C236	7	N	95	61	80					
DBSCAN	C236	7	N	186	8						
BIRCH	C236	7	N	70	94	72					
K-Means	C162	7	N	45	22	26	17	21	25	6	
DBSCAN	C162	7	N	17	7	6	3	11	8	2	2
BIRCH	C162	7	N	43	53	40	21	5			
K-Means	C136	7	N	24	6	26	9	11	18	16	26
DBSCAN	C136	7	N	11	2	6	5	4	3	5	8
BIRCH	C136	7	N	85	41	10					

The results demonstrate that for seven dimensions (Fig 3a), the K-Means and BIRCH algorithms gave comparable Silhouette Coefficients and they performed consistently better than the DBSCAN. This pattern is seen across all five courses of different class sizes. The K-Means generally produced a larger number of clusters (7 or 8) than the BIRCH (3 to 5 clusters). Also, the K-Means clusters had a more even distribution of students across clusters, making it the preferred algorithm. It is interesting that the BIRCH performs so well, and in many cases produces clusters comparable to the K-Means algorithm, as in our survey of algorithms used for clustering analysis, the BIRCH algorithm was hardly ever used. This is significant because, in the BIRCH algorithm, we are not required to specify the number of clusters a priori, and this can act as a good starting point to find cluster groups in new datasets. DBSCAN did not fare well in this experiment. For the 7-dimension analysis, it had a consistently lower Silhouette Coefficient as well as a much larger number of clusters, many of which had only a small number of students. In fact, above what is reported in Table 4, there were also a large number of isolated clusters with just a single student as outliers.

The results of Fig 3.b on two dimensions give a strong indication that the dimensionality of the dataset provided has an effect on the performance of the DBSCAN and the Hierarchical method (BIRCH). While the Silhouette Coefficients of the 2 dimensional sets may be larger, the number and spread between cluster groupings are much better when using the higher number of dimensions. DBSCAN produces better Silhouette Coefficients when using the smaller number of dimensions. The K-Means give the overall better result for the Silhouette Coefficients, number of clusters and the distribution of members across the clusters.

All algorithms were executed on the same machine. Comparative analysis of the execution times of the algorithms show little significant differences for the datasets provided. Even for the largest dataset C1496, the differences in execution times amongst the algorithms are small.

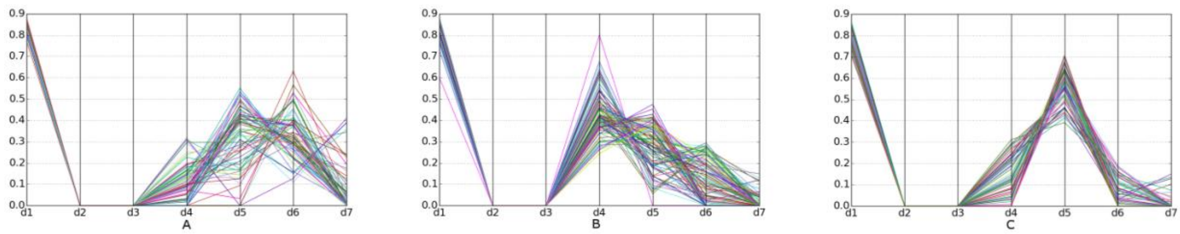

Figure 4. BIRCH Cluster composition for C236

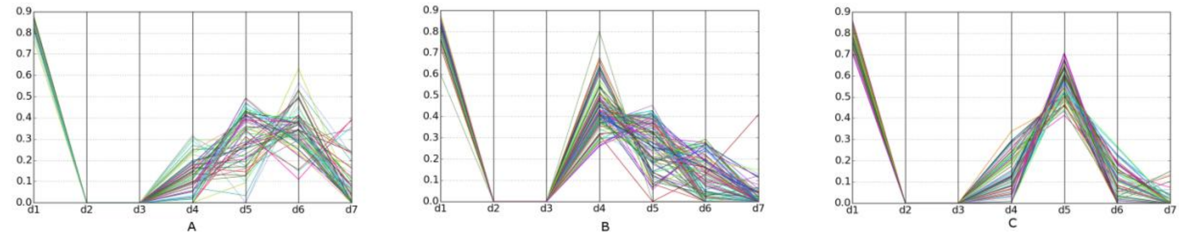

Figure 5. K-Means Cluster composition for C236

Figure 4 and 5 illustrates the composition of the clusters generated by the BIRCH and the K-Means algorithms within the C236 course respectively. What is immediately of interest is that the composition, as indicated by the grouping of lines, is identical between the result sets. Figure 4a and Figure 5a, show that the students within this group have an even participation within the course across the day. Figure 4b and 5b groups students based on a high level of participation during the morning (d4), a reasonable level of participation during the afternoon (d5) but very low participation during night (d6) and late night/early morning (d7). The final grouping in Figure 4c and 5c, represents groups of students that have a high participate level in the afternoon (d5) but generally very low levels of participation for the remaining hours of the day. The course C236 did not utilize the forums and quizzes activity within the Moodle

LMS. This resulted in no values for d2 and d3 which measured the total activity in forums, and activity with quizzes respectively.

The important implication of this comparison, is that the clusters generated between these two groups highlight very similar behaviours among the students within the course. Also when considered in light of a similar distribution between the groups indicated in Table 5, this re-emphasizes the potentials for the hierarchical methods such as the BIRCH to be used for clustering analysis within education.

6. CONCLUSION

Clustering analysis is particularly important during evaluation and exploratory data analysis, where researchers attempt to discover underlying trends that exist without any previous knowledge about the data that is generated within the LMS. However, the choice of which clustering technique and algorithm is determined by knowledge of the structure of the data, types of analysis to be drawn and the size of the dataset evaluated. Our survey of clustering analysis in education in section 3.1 shows that the majority of research papers used clustering for evaluation and exploration and in many cases only the more popular K-Means algorithm was used. The purpose of this study is to extend the knowledge about the performance of clustering algorithms by considering a case based analysis of the comparative performance across the categories of clustering technique.

To gain an insight into how clustering analysis can be done within different courses, the performance of three representative algorithms was evaluated on different sizes of datasets. This experiment demonstrates that there are common structures within independent courses, which can be used to determine the most appropriate technique to be used for analyses within the educational system. The experimental results confirmed that the Partition-based K-Means algorithm performed better than the Hierarchical BIRCH algorithm and the Density-based DBSCAN algorithm. It produces clusters with higher coefficients than the other two algorithms for the 2-dimension and 7-dimension datasets. Therefore, Partition-based methods can be regarded as the most appropriate technique to perform clustering analysis of educational log data. The good performance of the Partition-based method and relatively low performance of the Density-based method indicates that the data generated by LMS logs, tend to form groups around some centre (point or object) and analysis using this assumption will produce some of the more useful results for exploratory analysis of groupings.

The interesting performance of the BIRCH algorithm alludes to the need to investigate the possibility that there may be hierarchical structures within the log data of the LMS. An understanding of this hierarchy may provide interesting insights for understanding relationships that exist between the different dimensions, and user interaction within the LMS.

A future extension of this research will be the evaluation of the clusters to determine their potential impact within the educational context and apply a comparison of usefulness of the information generated by the techniques.

REFERENCES

Amershi, S. & Conati, C.C., 2009. Combining Unsupervised and Supervised Classification to Build User Models for Exploratory. *JEDM-Journal of Educational Data Mining*, Vol. No. 1, pp.1–54.

Antonenko, P.D., Toy, S. & Niederhauser, D.S., 2012. Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, Vol. 60, No. 3, pp.383–398.

Ayesha, S. et al., 2010. Data Mining Model for Higher Education System. *European Journal of Scientific Research*, Vol. 43, No.1, pp.24 – 29.

Baker, R.S. & Yacef, K., 2009. The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining*, Vol. No. 1, pp.3–17.

Berkhin, P.P., 2006. *A Survey of Clustering Data Mining Techniques*, Heidelberger, Berlin, Germany: Springer Berlin Heidelberg.

Bogarín, A. et al., 2014. Clustering for improving educational process mining. *Proceedins of the Fourth International Conference on Learning Analytics And Knowledge - LAK '14*, pp.11–15.

Bovo, A. et al., 2013. Clustering moodle data as a tool for profiling students. In *2013 Second International Conference on E-Learning and E-Technologies in Education (ICEEE).* pp. 121–126..

Campagni, R. et al., 2015. Data mining models for student careers. *Expert Systems with Applications*, Vol. 42, No. 13, pp.5508–5521.

Chen, C.-M. & Chen, M.-C., 2009. Mobile formative assessment tool based on data mining techniques for supporting web-based learning. *Computers & Education*, Vol. 52, Vol. 1, pp.256–273.

DeFreitas, K. & Bernard, M., 2014. A Framework for Flexible Educational Data Mining. In *The 2014 International Conference on Data Mining*. Las Vegas, USA, pp. 176–180.

Han, J., Kamber, M. & Pei, J., 2012. *Data Mining: Concepts and Techniques* 3rd ed., Massachusetts, USA: Morgan Kaufmann.

Kaufman, L. & Rousseeuw, P.J., 1990. Partitioning Around Medoids (Program PAM). In *Finding Groups in Data*. New Jersey, USA: John Wiley & Sons, Inc., pp. 68–125.

Kim, M.-S. & Han, J., 2009. A Particle-and-density Based Evolutionary Clustering Method for Dynamic Networks. In *Proceedings of VLDB Endowment*. VLDB Endowment, pp. 622–633.

Klašnja-Milićević, A. et al., 2011. E-Learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & Education*, Vol. 56, No. 3, pp.885–899.

Kriegel, H.-P. et al., 2011. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. No. 3, pp.231–240.

Li, J., Za, O.R. & Zaïane, O., 2004. Combining Usage, Content, and Structure Data to Improve Web Site Recommendation. In K. Bauknecht, M. Bichler, & B. Pröll, eds. *Proceedings in the International conference on e-commerce and web technologies*. Lecture Notes in Computer Science. Zaragoza, Spain: Springer Berlin Heidelberg, pp. 305–315.

Liu, Y. et al., 2013. Understanding of Internal Clustering Validation Measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. Sydney, New South Wales, pp. 1–6.

López, M. et al., 2012. Classification via clustering for predicting final marks based on student participation in forums. *Proceedings of the 5th International Conference on Educational Data Mining*, pp.4–7.

Mark Hall et al, 2014. The WEKA Data Mining Software: An Update. *SIGKDD Explorations Newsletter*, Vol. 11, No. 1, pp.10–18.

Mashat, A.F. et al., 2012. Efficient Clustering Technique for University Admission Data. *International Journal of Computer Applications*, 45(23), pp.39–42.

Maulik, U. & Bandyopadhyay, S., 2002. Performance evaluation of some clustering algorithms and validity indices. In *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. pp. 1650–1654.

Ng, R.T. & Han, J., 2002. CLARANS: a method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 14(5), pp.1003–1016.

Pahl, C. & Donnellan, D., 2002. Data Mining Technology for the Evaluation of Web-based Teaching and Learning Systems. In *Conference on E-Learning in Business, Government and Higher Education*. Montreal, Canada, pp. 1–6.

Pakhira, M.K., Bandyopadhyay, S. & Maulik, U., 2004. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, Vol. 37, No. 3, pp.487–501.

Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, pp.2825–2830.

Perera, D. et al., 2009. Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 6, pp.759–772.

Permata Alfiani, A. & Ayu Wulandari, F., 2015. Mapping Student's Performance Based on Data Mining Approach (A Case Study). *Agriculture and Agricultural Science Procedia*, 3, pp.173–177.

Romero, C. & Ventura, S., 2007. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1), pp.135–146.

Romero, C., Ventura, S. & De Bra, P., 2004. Knowledge discovery with genetic programming for providing feedback to courseware authors. *User Modelling and User-Adapted Interaction*, 14(5), pp.425–464.

Romero, C., Ventura, S. & García, E., 2008. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1), pp.368–384.

Sivogolovko, E. & Novikov, B., 2012. Validating cluster structures in data mining tasks. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops on - EDBT-ICDT '12*. EDBT-ICDT '12. New York, USA: ACM, p. 245.

Srivastava, J. et al., 2000. Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data. *ACM SIGKDD Explorations Newsletter*, 1(2), pp.12–23.

Tair, M.M.A. & El-halees, A.M., 2012. Mining Educational Data to Improve Students ' Performance : A Case Study. *International Journal of Information and Communication Technology Research*, 2(2), pp.140–146.

Tam, V., Lam, E.Y. & Fung, S.T., 2012. Toward a complete e-learning system framework for semantic analysis, concept clustering and learning path optimization. *Proceedings of the 12th IEEE International Conference on Advanced Learning Technologies, ICALT 2012*, pp.592–596.

Witten, I.H., Frank, E. & Hall, M.A., 2005. *Data Mining: Practical Machine Learning Tools and Techniques* 3rd ed., Massachusetts, USA: Morgan Kaufmann.

Zaiane, O.R. & Luo, J., 2001. Towards evaluating learners' behaviour in a web-based distance learning environment. In *Learning Technologies, 2001. Proceedings*. Madison, Wisconsin, USA, pp. 357–360.