A Study on Clustering Facebook Multimedia Based on Metadata—A Comparative Analysis

Prashant Bhat and Pradnya Malaganve

1 Introduction

Knowledge discovery designates achieving useful or most important information among huge set of data which is gathered from several data warehouses [1] and other data sources. To attain this, data mining techniques are essentially used. In our research we are utilizing cluster techniques and algorithms to extract useful information by grouping the instances in different Clusters. Cluster is an unsupervised approach for grouping the instances of a data set [2]. The word unsupervised means there is no label for the instances where in supervised method contains label for the instances. In the present work we have explained the process of three clustering algorithms that are Expectation Maximization (EM), Simple K-Means, and Hierarchical Clusterer. These algorithms are applied on a data set of a cosmetic company's Facebook page. This data set contains 19 attributes such as total interactions, type, likes, and shares. These attributes are considered as Metadata of the dataset and 500 instances are present in the dataset. The attribute "type" has taken for the observation which further contains four kinds of instance that are Link, Status, Photo, and Video. We have used a method, i.e., Classes to Cluster Evaluation for all three Cluster algorithms and tested using WEKA data mining tool to get the essential results [3]. Based on the confusion matrix, time taken to build the model and number of incorrectly clustered instances, the comparison of all three algorithms is made and result is carried out to analyze in depth to prove the best suitable clustering algorithm for Facebook data set [4, 5].

P. Bhat · P. Malaganve (✉)
Garden City University, Bengaluru, India

R. Agrawal et al. (eds.), *Cyber Security in Intelligent Computing and Communications*,
Studies in Computational Intelligence 1007,
https://doi.org/10.1007/978-981-16-8012-0_7

1.1 Expectation Maximization

Expectation Maximization is a method of estimating max probable variables even when missing values present in the data set [6]. It is a repetitive process which generates the loop between two modes, namely, E-mode, i.e., estimation mode and M-mode, i.e., maximization mode [7]. In this approach E-mode strives to estimate the missing variables then the M-mode strives to develop the variables present in the data set to put the data into the model in a better way [8].

Expectation Maximization Algorithm

Step 1: Estimating latent or missing variables of the data set.
Step 2: Maximizing the variables that are present in the data set.

1.2 Simple K-Means Cluster

It is unsupervised learning algorithm that divides same number of instances [9, 10] to all the clusters as the algorithm shown below [11].

Simple K-Means Cluster Algorithm

Step 1: "n" number of instances are considered.
Step 2: All the instances are classified in "k" number of clusters.
Step 3: Mean value of the instances is calculated for "k" number of clusters.
Step 4: All the instances are compared with the mean value.
Step 5: The values which are near to mean value are exchanged to respective Clusters.
Step 5: Form new Cluster.
Step 6: Repeat Step 4 and Step 5 till instances are grouped correctly in each Clusters.

1.3 Hierarchical Clusterer

Rather than unstructured cluster, Hierarchical Clusterer is more informative and well-structured cluster. Below algorithm shows the process of Hierarchical Clusterer.

Hierarchical Clusterer Algorithm

Step 1: Form the Proximity or similarity matrix.
Step 2: Let each instance be a cluster.
Step 3: Combine two nearest clusters.
Step 4: Repeat Step 3 till single Cluster remains (Fig. 2).

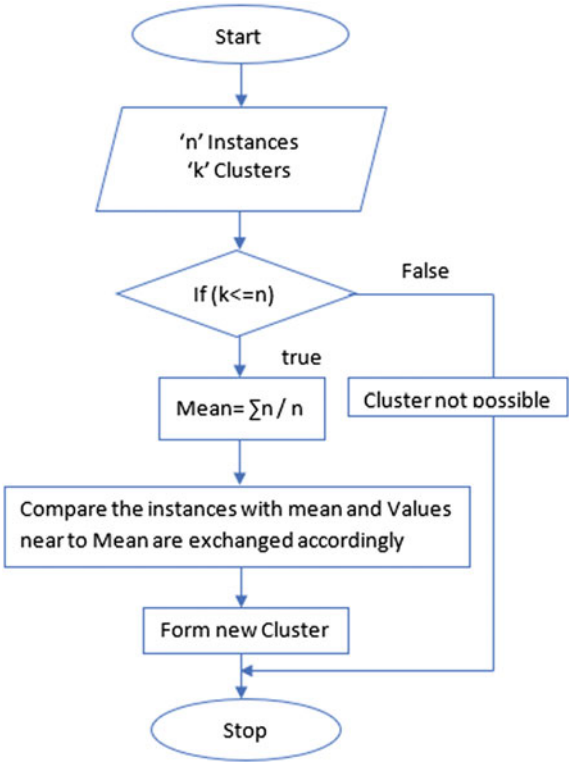

Fig. 1 Flowchart for simple K-means algorithm

1.4 Classes to Cluster Evaluation

In the present work we have used a single method, i.e., Classes to Cluster Evaluation for all three above explained algorithms. This method applies Brut Force approach to find minimum class label errors to Clusters followed by a constraint that one class label can be assigned to only one Cluster. If any Cluster returns "No Class" that indicates all the instances under that particular Cluster are considered as incorrectly Clustered instances. In WEKA, Classes to Cluster Evaluation method initially ignores the instances and directly generates the Clusters. Then at the time of testing, it assigns the instances to the Clusters based on majority values of instances within each Cluster. And related confusion matrix will be formed.

2 Proposed Model for Clustering Multimedia Based on Metadata

Figure 3 represents the methodology that carries out achieving the detailed compar-

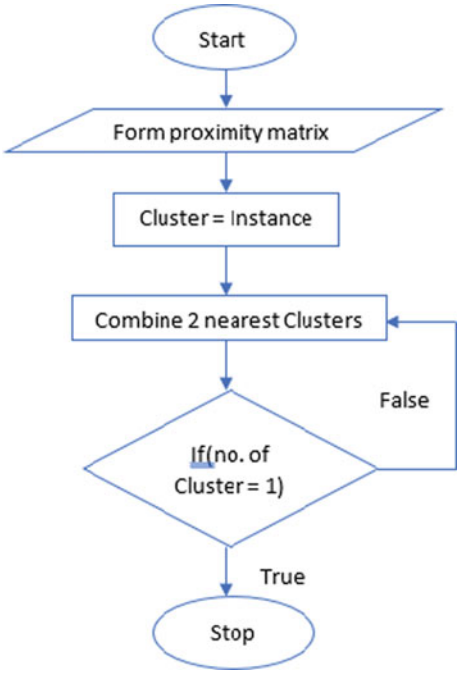

Fig. 2 Flowchart for hierarchical clusterer algorithm

ison analysis of Expectation Maximization, Simple K-Means, and Hierarchical Clusterer algorithms expecting for knowledge discovery and group data into respective clusters. We have shown five steps in the proposed model to achieve cluster algorithm results.

(1) Meta Data extraction process
(2) Pre-processing
(3) Cluster techniques
(4) Classes to Cluster Evaluation
(5) Result Analysis

Meta Data can be determined as data about data. As we have used cosmetic company's Facebook page data in the present work hence the Meta Data are URL of web page, number of likes, shares, comments type of the content uploaded, etc. These Meta Data need to be extracted from the web [12].

In this work, Info extractor tool is used for extracting the dataset that contains 19 attributes and 500 instances. Extracted data is stored in .CSV (Comma Separate Value) or .ARFF (Attribute Relation File Format) files for further findings. Initially the extracted data will be unrefined or raw.

Hence, we move to the next stage, i.e., Pre-processing. The term Unrefined means the dataset may contain huge amount of noise in it. For example, missing values in the dataset or the dataset may contain such values which cannot be understood and are meaningless. So, the unrefined data will be purified. In data mining several [13]

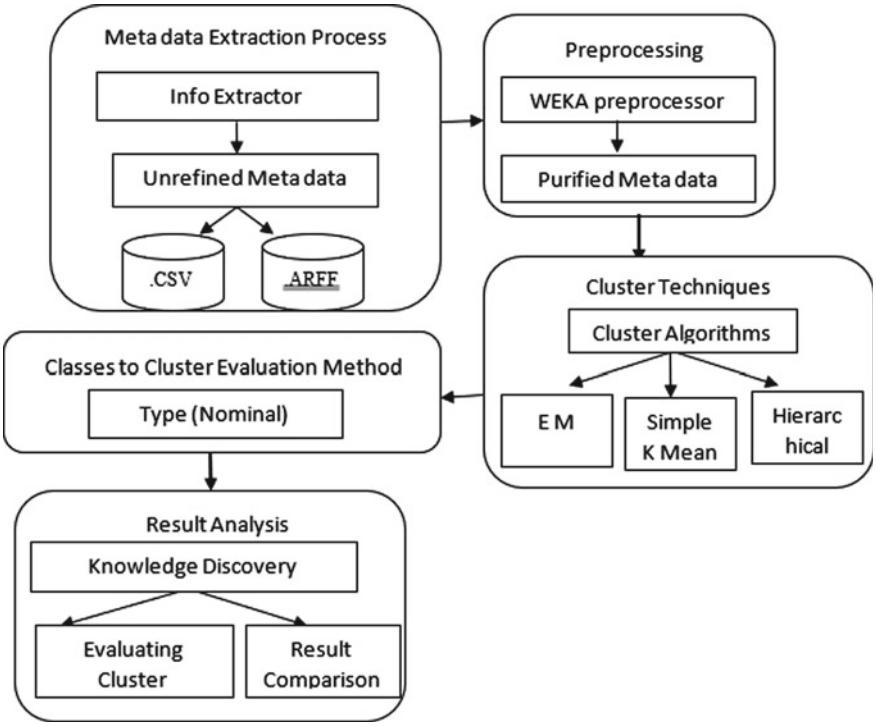

Fig. 3 Proposed model for clustering multimedia based on metadata

techniques are available to fill the missing values. For example, we can use most frequently appeared value of the respective column to fill the gap, by calculating the mean value of remaining instances the missing value can be filled, one global value can be declared such as "null" to fill the missing and so on. Using these techniques manually we can fill the gap in the dataset which is termed as pre-processing the noisy data. In present work we have used WEKA tool for pre-processing as the dataset is large in size.

The very next step carries three cluster algorithms that are Expectation Maximization, Simple K-Means, and Hierarchical Clusterer algorithm for the experiment. All three respective algorithms and flowcharts are defined in the introduction section algorithms are to be applied on the dataset for knowledge discovery. The resultant parameters are compared and analyzed in detail.

In the introduction section we have explained the process of the method, Classes to cluster evaluation. To group four instances that are Photo, Status, Video, and Link into four different clusters this method is used. And these four instances belong to "type" Metadata. "Type" contains nominal values, i.e., non-numeric.

The final step is to determine the relationship between variables and to compare the findings of all three algorithms considered in this research work. Evaluating cluster results and analyzing the result in depth leads to knowledge discovery.

2.1 Attributes Descriptions of Table 1

Page total likes: It indicates the total number of users those who have liked the cosmetic company's Facebook page.

Type: This attribute indicates the content type, whether the content is link, video, photo, or status.

Category: It indicates the characterization of the manual content.

Post month: This attribute indicates in which month the post is published.

Post week: This attribute indicates in which week the post is published.

Post hour: This attribute indicates at what time the post is published.

Table 1 Attributes descriptions

Name of attribute	Data type	Description
Page total likes	Numeric	Number of users who liked the company's page
Type	String	Type of content (Link, Photo, Video, Status)
Category	Numeric	Manual content characterization
Post month	Numeric	Post published month
Post week	Numeric	Post published week
Post hour	Numeric	Post published hour
Paid	Numeric	Company paid to the Facebook for advertising
Lifetime post total reach	Numeric	Number of unique users who saw the page post
Lifetime post total impression	Numeric	Number of times a post from a page is displayed
Lifetime engaged users	Numeric	Number of unique users clicked anywhere in the post
Lifetime post consumers	Numeric	Number of users clicked anywhere in the post
Lifetime post consumptions	Numeric	Numbers of clicks anywhere in the post
Lifetime Post Impressions by people who have liked your Page	Numeric	Number of impressions by users who have liked the page
Lifetime Post reach by people who like your Page	Numeric	Number of unique users saw a page post because they have liked it
Lifetime People who have liked your Page and engaged with your post	Numeric	Number of unique users liked and clicked anywhere in the post
Comment	Numeric	Number of comments of the post
Like	Numeric	Number of likes of the post
Share	Numeric	Number of shares of the post
Total interactions	Numeric	Sum of comments, likes and shares of the post

Paid: This attribute shows whether the cosmetic company has paid to the Facebook for its advertisement. Attribute values will be in the form of yes/no.

Lifetime post total reach: It shows the number of unique users who have seen the page post.

Lifetime post total impression: It indicates the number of times the post from company's page has appeared whether it is clicked or not. For example, first time when it is updated, second time, if a friend put any comment on it or if a friend shares it.

Lifetime engaged users: It shows the number of unique users who have clicked anywhere in a post.

Lifetime post consumers: This attribute indicates the total number of users who have clicked on the page.

Lifetime post consumptions: It shows the total number of clicks anywhere in a post.

Lifetime Post Impressions by people who have liked your Page: It shows the number of impressions only from the users who have liked a page.

Lifetime Post reach by people who like your Page: It is the total number of unique users who saw a page post only because they have liked it.

Lifetime People who have liked your Page and engaged with your post: It shows the number of unique users who have liked a page and also clicked anywhere in a post.

Comment: Total number of comments present on a post.

Like: Total number of likes present on a post.

Share: Total number of shares on a post.

Total Interactions: This attribute is the total number of comments, number of likes, and number of shares on a post.

3 Results and Analysis

3.1 Confusion Matrix

Confusion Matrix is to identify all the clustered instances of a dataset. By this matrix formation, we can identify correctly clustered and incorrectly clustered instances. The confusion matrix of all three algorithms is shown in Tables 2, 3, and 4.

Table 2 is formed using WEKA and the first row of the confusion matrix is assigned to classes, i.e., Cluster 0, Cluster 1, Cluster 2, and Cluster 3. And the remaining values of the matrix indicate all the instances of the dataset which are grouped as different clusters. Table 2 delivers that Expectation Maximization algorithm has formed four clusters and it has divided the instances in respective clusters as below.

Cluster 0 is holding 34 "status" instances that are clustered correctly.

Cluster 1 is holding 183 correctly clustered instances which are "photo".

Cluster 2 is holding 4 correctly clustered instances which are "video".

Table 2 Confusion matrix expectation maximization

0	1	2	3	← Assigned to Class
101	183	47	95	| Photo
34	1	9	1	| Status
8	7	2	5	| Link
3	0	4	0	| Video

Cluster 0 ← Status
Cluster 1 ← Photo
Cluster 2 ← Video
Cluster 3 ← Link

Table 3 Confusion matrix simple K-means

0	1	2	3	← Assigned to Class
40	51	307	28	| Photo
8	0	35	2	| Status
4	0	16	2	| Link
4	0	3	0	| Video

Cluster 0 ← Status
Cluster 1 ← No Class
Cluster 2 ← Photo
Cluster 3 ← Link

Table 4 Confusion matrix hierarchical clusterer

0	1	2	3	← Assigned to Class
40	51	307	28	| Photo
8	0	35	2	| Status
4	0	16	2	| Link
4	0	3	0	| Video

Cluster 0 ← Status
Cluster 1 ← No Class
Cluster 2 ← Photo
Cluster 3 ← Link

Cluster 3 is holding 5 “link” instances that are clustered correctly.

Table 3 represents the confusion matrix of Simple K-Means algorithm and the number of correctly clustered instances belong to four different categories are shown below which is determined using WEKA Data Mining tool.

Cluster 0 is holding 8 “status” instances that are clustered correctly.

Cluster 1 is holding 0 instances.

Cluster 2 is holding 307 correctly clustered instances which are “photo”.

Cluster 3 is holding 2 “link” instances that are clustered correctly.

Simple K-Means is containing Cluster 1 as null or no class that indicates the algorithm has not clustered video instances from the dataset.

Table 4 represents the confusion matrix of Hierarchical Clusterer algorithm and below are the number of correctly clustered instances which are formed with the help of WEKA.

Cluster 0 is holding 423 "photo" instances that are clustered correctly.

Cluster 1 is holding 0 instances.

Cluster 2 is holding 0 instances.

Cluster 3 is holding 0 instances.

In Hierarchical Clusterer, Cluster 1, Cluster 2, and Cluster 3 are having 0 instances. This indicates that the algorithm has not clustered video, status, and link instances correctly in particular group.

3.2 Analysis of Table 5

Table 5 gives the distinct result of all three algorithms. As we observe all the readings of Table 5, Expectation Maximization has incorrectly clustered 274 instances out of 500 instances, i.e., 54.8% instances are clustered wrong. Hence correctly clustered instances are 226. In Simple K-Means algorithm findings, incorrect clustered instances are 183, i.e., 36.6% instances are not clustered correctly. Hence correctly clustered instances are 317. Compared to Expectation Maximization algorithm, Simple K-Means algorithm has better numbers while clustering the instances in particular group. The algorithm Hierarchical Clusterer has incorrectly clustered 77 instances and 15.4% instances of the whole dataset is incorrect. Hence correctly clustered instances are 423 and by observing incorrectly clustered values of all three algorithms, we can say that Hierarchical Clusterer is having very less instances which are clustered incorrectly. Hence Hierarchical clusterer can produce the best accuracy in clustering the Facebook dataset in a better way.

The time taken to build all three models is minimum as all the algorithms have taken less than 1 s to get executed. And according to the observations, it clearly indicates that the Simple K-Means cluster takes the least time to generate the model.

Table 5 Comparison analysis based on correctly clustered instances and time taken to build the model

	Expectation maximization	Simple K-means	Hierarchical clusterer
Incorrectly clustered instances	274.0	183.0	77.0
	54.8%	36.6%	15.4%
Time taken to build the model	0.42 s	0.02 s	0.83 s

4 Conclusion

After analyzing all the variables and readings of all three algorithms, it's proven that Hierarchical Clusterer is the best suitable algorithm for clustering Facebook pages dataset in a better way, as the algorithm has correctly clustered 423 instances out of 500 instances which is highest compared to Simple K-Means and Expectation Maximization algorithms. Hence, we would like to conclude that, because of the structure and formation of Hierarchical Clusterer, the algorithm is capable of clustering the instances in a better way as it considers every instance as a cluster and go on combining nearest clusters until formation of a single cluster.

References

1. C. Maionea, D.R. Nelsonb, R.M. Barbosa, Research on social data by means of Cluster analysis. 2210–8327/Ó2018 Production and hosting by Elsevier B.V.
2. S. Ding, F. Wu, J. Qian, Research on data stream clustering algorithms. Artif. Intell. Rev. **43**, 593–600 (2015)
3. P. Bhat, P. Malaganve, P. Hegade, A new framework for social media content mining and knowledge discovery. IJCA **182**(36), 17–20 (2019)
4. A. Shensa, J.E. Sidani,· M.A. Dew,· C.G. Escobar-Viera, B.A Primack, Social media use and depression and anxiety symptoms: a cluster analysis. **42**(2), 116–128 (2018)
5. C.C. Aggarwal, C. Zhai, A survey of text clustering algorithms, in *Mining Text Data,* ed. by C. Aggarwal, C. Zhai (Springer, Boston, MA). https://doi.org/10.1007/978-1-4614-3223-4_4
6. https://machinelearningmastery.com/expectation-maximization-em-algorithm/
7. C.K. Reddy, H. Chiang, B. Rajaratnam, TRUST-TECH-based expectation maximization for learning finite mixture models. IEEE **30**(7), 1146–1157 (2008). https://doi.org/10.1109/TPAMI.2007.70775.
8. S.P. Algur, P. Bhat, Web video object mining: expectation maximization and density based clustering of web video metadata objects. I. J. Inf. Eng. Electron. Bus. **1**, 69–77 (2016). https://doi.org/10.5815/ijieeb.2016.01.08
9. Y.G. Jung, M.S. Kang, J. Heo, Clustering performance comparison using K means and expectation maximization algorithm. Biotechnol. Biotechnol. Equip. **28**(sup1), S44–S48. https://doi.org/10.1080/13102818.2014.949045
10. N. Dhanachandra, K. Manglem, Y.J. Chanu, Image segmentation using K means clustering algorithm and subtractive clustering algorithm, in *IMCIP* (2015)
11. https://www.tutorialride.com/data-mining/KMeans-Clustering-in-data-mining.htm
12. M. Othman, S.A. Mohamed, M.H.A. Abdullah, M.M. Yusof, R. Mohamed, A framework to cluster temporal data using personalised modelling approach, in *Ghazali, SCDM 2018. Advances in Intelligent Systems and Computing*, vol. 700 (Springer, Cham, 2018)
13. S. Harifi, E. Byagowi, M. Khalilian, *Comparative Study of Apache Spark MLlib Clustering Algorithm: DMBD 2017* (Springer International Publishing AG, 2017). LNCS **10387**, 61–73 (2017). https://doi.org/10.1007/978-3-319-61845-6_7
14. H. Jia, S. Ding, X. Xu, The latest research progress on spectral clustering. Neural Comput. Appl. **24**, 1477–1486 (2014)
15. R. Vaarandi, M. Pihelgas, LogCluster—a data clustering and pattern mining algorithm for event logs, in *CNSM*, Barcelona (2015), pp. 1–7
16. **81**, 1 March 2015. https://doi.org/10.1016/j.energy.2014.12.054
17. S. Ajani, M. Wanjari, An efficient approach for clustering uncertain data mining based on hash indexing and voronoi clustering, in *5th International Conference and Computational Intelligence and Communication Networks,* Mathura (2013), pp. 486–490

18. H. Nguyen, Y. Woon, W.A. Ng, Survey on data stream clustering and classification. Knowl. Inf. Syst. **45**, 535–569 (2015). https://doi.org/10.1007/s10115-014-0808-1
19. F.T. Giuntini et al., How do i feel? Identifying emotional expressions on facebook reactions using clustering mechanism. IEEE Access **7**, 53909–53921 (2019). https://doi.org/10.1109/ACCESS.2019.2913136
20. S. Moro, P. Rita, B. Val, Predicting social media performance metrics and evaluation of the impact on brand building: a data mining approach. J. Bus. Res. (Elsevier) (2016)