International Journal of Applied Engineering & Technology

A ROBUST CLUSTERING INFORMATION MINING MODEL FOR SELECTIVE DISSEMINATION OF INFORMATION

Titus K. Rotich Too[1], Cheruiyot Wilson[2] and Daniel Otanga[3]
[1, 2]School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and technology, Nairobi, Kenya
[3]School of Computing and Information Technology, Masinde Muliro University of Agriculture and technology, Nairobi, Kenya
[1] titusrotich14@gmail.com, [2] wilchery68@gmail.com and [3] otanga@mmust.ac.ke

ABSTRACT

Currently, the rate at which information is being generated is exponential. This abundance of information has created computational challenge. Addressing this challenge, clustering emerges as a pivotal solution, focusing on the organization of unlabeled feature vectors into clusters. The primary objective is to group samples within a cluster in such a way that their similarities are maximized, distinguishing them from samples in different clusters. Notably, clustering operates under the assumption that the number of clusters is predetermined, with no a priori information provided about the data. This research solves the overwhelming abundance of information by leveraging the K-means clustering algorithm, to enhance information retrieval and facilitate more efficient navigation through extensive datasets within library contexts.

Keywords: Algorithm, Clustering, K-means, Mining.

INTRODUCTION

The amount of information materials held by academic libraries is enormous and ever increasing at an astonishing rate as new pieces of information are now not only present in the physical form but also in digital sources. These large numbers of information resources are a challenge to librarians' on how they can effectively and efficiently avail such relevant information resources to users [1].

Librarians should increasingly co-operate and collaborate with computer and information systems professionals within the academic universities to meet perceived needs in this area. The collaboration will ensure that skilled staff can be shared or put to the most productive use [2].

With the development of big data mining technology, it plays a very important role in the university library management system [3]. Information mining involves sets of methods to discover patterns, associations or, in general, interesting expertise from large amounts of data. For the past ten to twenty years, volumes of saved digital data, the memory capabilities and the computing electricity have grown, also has the want to take gain of all that potential [4].

Clustering is one of the strategies used in Data Mining. Clustering analysis is the manner of identifying statistics that are comparable to every other. This aid to recognize the differences and similarities between the data. This is once in a while known as segmentation and helps the users to understand what is going on within the database [5**].**

Data mining finds useful information from data sets, which can either be structured or unstructured, It is used in the libraries through the process of Bibliomining Stanton,[6]. The Idea of abstraction ant –colony clustering algorithm for information mining, as characterized by [7] is portrayed as "the way toward removing substantial, already obscure understandable, and noteworthy data from expansive databases and utilizing it to settle on urgent business choices"..

A. How does Information Mining work?

Data mining refers to extracting or mining knowledge from large amounts of data. The term is actually a misnomer. Thus, data mining should have been more appropriately named as knowledge mining which emphasis on mining from large amounts of data. The key properties of data mining are

i. Automatic discovery of patterns

ii. Prediction of likely outcomes

iii. Creation of actionable information

iv. Focus on large datasets and databases

There are four varieties of information mining [8] as listed below:

i. **Classes:** Stored information is used to find out information in predestined get-togethers. For example, a restaurant gadget may want to mine patron purchase data to pick out when customers go to and what they conventionally demand. This information may want to be used to manufacture extra via having each day specials.

ii. **Clusters:** Data items are assembled by means of reasonable associations or consumer tendencies. For example, statistics can be mined to apprehend grandstand portions or client affinities, define clusters as segments or similar records that share a number of properties and are so considered homogenous. Clustering works by distinguishing gatherings of clients who seem to have comparable inclinations and isolating gatherings who have altogether different inclinations. In contrast to arrangement, the class mark of each gathering is obscure. This is the best approach to normally fragment information into indistinct gatherings, called grouping. Interestingly, arrangement is doling out information into characterized groupings, (Bandaru et al., 2017). In brief, a great grouping strategy delivers high quality clusters with high intra-class yet low.

iii. **Associations:** Data can be mined to identify associations. The aim of associations is to establish links between individual records. Associations could either be through: associations' discovery, sequential pattern discovery or similar time sequence discovery. One of the broadly utilized instances of information mining is the disclosure of affiliation rules, particularly showcase bushel investigation. Colossal measures of client buy information are gathered every day at the checkout counters of shopping centers and retailers are keen on buying the conduct of their clients. This procedure, affiliation rules, every now and again discovered co-buy things, whereby things that are regularly bought together are racked more like each other. Additionally, the revealed connections can be spoken to as affiliation rules. This has given retailers a chance to strategically pitching their items to clients.

iv. **Sequential patterns:** Information is mined to anticipate behavior patterns and trends. that has the following processes:

Assembling a collection of data analysis.

Present data to a data mining software

Interpret the results

Apply the results to a new problem

METHODOLOGY

In the past, Data services like Selective Dissemination of Information (SDI) and Current Awareness (CA) services are the methods in which users are fed with updates from the library. Information mining model will offer a different approach, where users will be fed with updates from there areas of study [7].

The research methodology approach that was used was quantitative experimental research design. Quantitative research design is the technique and measurements that produces quantifiable/discrete values [9]. The collected data results from empirical observations and measures. These methods require a good amount of time and planning. They always tend to have closed ended responses.

Quantitative research is considered as an analytical approach towards research. Quantitative researchers, as [10] elaborate, regard the world as being outside of themselves and there is an objective reality which is independent of any observations.

A. Problem Statement

Currently, the rate at which information is being generated is exponential. This abundance of information has created computational challenge. Addressing this challenge, clustering emerges as a pivotal solution, focusing on the organization of unlabeled feature vectors into clusters. The primary objective is to group samples within a cluster in such a way that their similarities are maximized, distinguishing them from samples in different clusters. Notably, clustering operates under the assumption that the number of clusters is predetermined, with no a priori information provided about the data.

This research solves the overwhelming abundance of information by leveraging the K-means clustering algorithm, to enhance information retrieval and facilitate more efficient navigation through extensive datasets within library contexts.

B. Research Objective

The man objective of this research implemented a robust K-means clustering model for selective dissemination of information for library users. The model was based on K-means Clustering which improved the start of art by enhancing parralization for the computing problem regarding large datasets on the basis of k-Centroid through data aggregation which enhanced the data points index access of similar data clusters

C. Clustering Algorithms in Data Mining

Clustering is concerned with the grouping of unlabeled feature vectors into clusters, such that samples within a cluster are more similar to each other than samples belonging to different clusters. Usually, it is assumed that the number of clusters is known in advance, but otherwise no prior information is given about the data. Applications of clustering can be found in the areas of communication, data compression and storage, database searching, pattern matching, and object recognition [13].

Apriori algorithm produces the association rules which can be applied for large database. Collaborative filtering Algorithm is used to recommend the books of similar user profiles. For a recommendation system, data collection, processing data in addition with user data is required, where user ratings play a crucial role. Automatizing the support count estimation in Apriori algorithm can be done to improve the efficiency of the system [6].

Ant colony streamlining strategy for ideally grouping N objects into K bunches. The algorithm utilizes appropriated specialists which copy the manner in which proper ants discover a most restrained way from their domestic to sustenance supply and back. This algorithm has been actualized and tried on a few reenacted and true datasets. The presentation of this calculation is contrasted and other accepted stochastic/heuristic strategies viz. hereditary algorithm, reproduced toughening and tabu search. The computational recreations discover extremely promising consequences related to the nature of association found, the normal variety of ability assessments and the getting ready time required. The insect kingdom clustering algorithm is a swarm-based approach for taking care of grouping issues that is propelled by the conduct of subterranean insect settlements in clustering their bodies and arranging their hatchlings. For most datasets, the reflection subterranean insect colony clustering results are unmatched (Gao, 2016). It is a subterranean insect settlement clustering algorithm for ideally grouping N objects into K groups. The algorithm utilizes the international pheromone refreshing and the heuristic records to construct clustering arrangements and uniform hybrid administrator to additionally decorate preparations determined by ants [14].

Enhanced ant-colony clustering algorithm utilizing an information blend component is proposed to enhance the computational productivity and exactness of the subterranean insect colony grouping algorithm. The reflection subterranean insect province clustering algorithm is utilized to group benchmark issues, and its execution is contrasted and the subterranean insect settlement clustering algorithm and different strategies utilized as a part of existing writing. The execution of the algorithm is checked by genuine datasets and contrasted and those of the subterranean insect province grouping algorithm and different algorithms proposed in past investigations. This investigation with utilization of reflection insect settlement clustering algorithm has the ground works on data mining [8].

Artificial bee colony algorithm is a meta-heuristic algorithm which models the behavior of honey bee searching for a quality food supply. The widespread synthetic bee colony algorithm consists of three components: meals sources, employed bees and unemployed bees. A meals supply represents a viable solution of problem. Every employed bee will bind to a food source. Unemployed bees have two kinds: onlookers and scouts. Onlookers update food source by means of neighborhood search and scouts search for new food supply randomly. This model has two primary behaviors: update contemporary data sources and locate new data sources [8].

D. The Model

The main purpose of this research implemented a clustering information mining model to enhance selective dissemination of information for library users. The following figure 3.0 shows the implemented model.

Direct optimization of the likelihood function in this case was not a simple task, due to necessary constraints on the parameters and the complicated nature of the likelihood KNN function, which in general had a great number of local maxima and saddle-points.

The following fig. I shows the implemented model

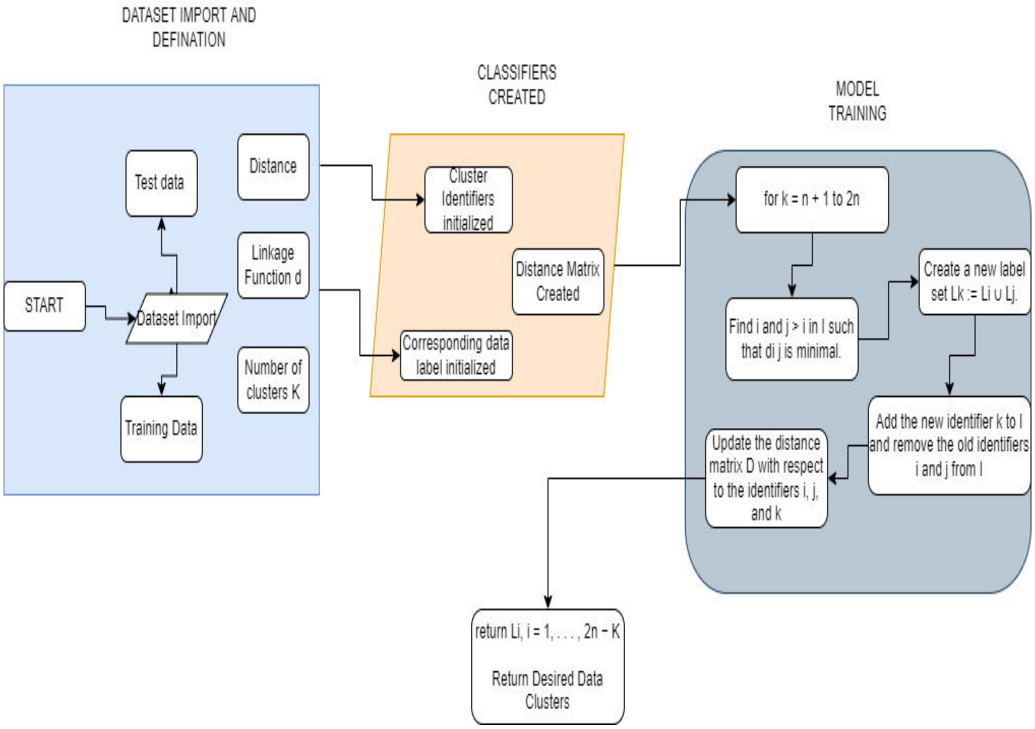
DATASET IMPORT AND DEFINATION
CLASSIFIERS CREATED
MODEL TRAINING
Test data
Distance
Linkage Function d
START
Dataset Import
Number of clusters K
Training Data
Cluster Identifiers initialized
Distance Matrix Created
Corresponding data label initialized
for k = n + 1 to 2n
Find i and j > i in I such that di j is minimal.
Create a new label set Lk := Li ∪ Lj.
Add the new identifier k to I and remove the old identifiers i and j from I
Update the distance matrix D with respect to the identifiers i, j, and k
return Li, i = 1, . . . , 2n − K
Return Desired Data Clusters

Figure 1. The clustering information mining model to enhance selective dissemination of information for library users

The dataset was divided into two-dimensional with P = 12 data points naturally clustered into K = 3 clusters. Points that are geometrically close to one another belong to the same cluster, and each cluster boundary is roughly marked using a uniquely colored solid curve. Each cluster center, also called a centroid is marked by a star symbol colored to match its cluster boundary**.**

The center of each cluster was drawn using a star symbol that matched the unique boundary color of each cluster. These cluster centers are often referred to in the jargon of machine learning as cluster centroids. The centroids allowed to understand about the dataset in the big picture sense – instead of P = 12 points we can think of our dataset grossly in terms of these K = 3 cluster centroids, as each represents a chunk of the data.

The dataset was imported online from a CSV file using python import library function NumPy. The dataset contained both training and test data that provided an enhanced validation of the model. NumPy python library works well with arrays (vectors and matrices), common mathematical functions like cos and sqrt and generating random numbers.

Once the training dataset was created, it was time to choose the model to train. AutoML Google Colab was used to train the model that got an RMSE of 0. 198. AutoML used a number of sophisticated models such as neural architecture search, which build a learning networks one layer at a time, and TabNet, which is based on a sophisticated approach called sequential attention.

The key reason to using AutoML was its ability to write out evaluation data to BigQuery. That capability came in very handy for sliced evaluation—for example, it could easily analyze whether our model performs better on clustering the library data as could add the necessary code (to write out evaluation data) to our custom model which saved a lot of work within short time.

The dataset consisted of 9 data points that were independently generated from three bivariate normal distributions, whose parameters were assigned. For each of these three distributions, exactly 3 points were generated.

The datasets were clustered into three clusters that correspond to the three cases. To cluster the data into three groups, a possible model for the data was to assume that the points are drawn from an (unknown) mixture of three 2-dimensional Gaussian distributions.

This was a sensible approach, although in reality the data were not simulated in this way. It is instructive to understand the difference between the two models. In the mixture model, each cluster label Z takes the value {1, 2, 3} with equal probability, and be Bin (9, 1/3) distributed. However, in the actual simulation, the number of points in each cluster is exactly 3. Nevertheless, the mixture model would be an accurate (al- though not exact) model for these data.

The corresponding covariance matrices are initially chosen as identity matrices, which is appropriate given the observed spread of the data. Finally, the initial weights are 1/3, 1/3, 1/3. For simplicity, the algorithm stops after 3 iterations, which in this case was more than enough to guarantee convergence.

E. Dataset Population and Sample

The dataset population was obtained from a repository https://www.kaggle.com/datasets/notkrishna/goodreads-top-100-classical-books-of-all-time containing datasets that was acquired by scrapping goodreads.com. Goodreads is an American social cataloging website and a subsidiary of Amazon that allows individuals to search its database of books, annotations, quotes, and reviews. That dataset was 48 MB in size. The research was based on purposive sampling technique, a form of non-probability sampling in which decisions concerning the individuals to be included in the sample were based upon a variety of criteria which included specialist knowledge on the survey research issue and capacity and willingness to participate in the research.

The final sample size from the population of the dataset was 9 Megabytes in size with ten attributes. The final sample size included: book title, book language, book series, author, number of pages, average rating, number of ratings, book description and awards of the book. The data cluster linkages in the datasets included single linkage, complete linkage and group linkage.

The required software tools for the simulation environment were gathered differently. The required hardware computing environment required specifications were a minimum of 8 GB RAM, a, 500-gigabyte internal hard disk space and a GPU of 4.2 gigahertz clock speed. Google Colaboratory (a.k.a. Colab) is a cloud service based on Jupyter Notebooks for disseminating machine learning education and research [15]. It provided a runtime fully configured for deep learning and free-of-charge access to a robust GPU for faster model deployment.

RESULTS AND DISCUSSION

After implementing a clustering information mining model to enhance selective dissemination of information for library users, the immediate task was to check value and actualization of the model during the experiments with datasets. The experiments used data variables described in chapter three.

The purpose of modelling, analysis and discussion was to confirm that the model enhanced selective dissemination of information for library users. This chapter discusses in detail the modeling design, analysis and results of the model to an acceptable version that is of value.

A. Data Table Results

After the model was implemented, it was deployed into google Colab using AutoML where 3 clusters were realized on the datasets The data table results in the following Table 1 below

Table I: Data Table Results

The Clusters	Mean Vector	Covariance Matrix
Cluster 1	0.56	65
Cluster 2	0.51	67
Cluster 3	0.34	41

The data table results show that when there is increase in the value of one feature by one unit, the model equation produces two odds: one is the base and the other is an incremental value of a feature. The objective here was to look at the ratio of odds with every increase or decrease in the value of a feature. A change in a feature by one unit leads to changes in the odds ratio by a factor of exponential corresponding beta coefficients.

B. Results Comparison

The model results were compared with other approaches that have been used to clustering information mining model to enhance selective dissemination of information. The other models approach compared with investigated with the same objectives and the results were displayed in following table II below.

Table II: Comparison Results

Method	Accuracy percentage	Mean Vector	Covariance Matrix
Decision Trees	66%	0.35	44
organizing Maps	58%	0.45	55
Random Forest	65%	0.37	47
Proposed Approach	71%	0.29	29

The implemented model proved to be dynamic and overall best than the other machine learning approaches compared against to. Once the explainer model object was generated, individual clustering predictions and global predictions for generating explanations was inferred. In a clustering where you have two classes or multi-classes, you can generate separate feature importance for each class with respect to the features column. In this instance, you consider two records: mean vector where the model predicts the outcome correctly and covariance vector where the model incorrectly makes prediction.

CONCLUSION

The study implemented and evaluated a clustering mining model for selective dissemination of information at an academic library based on K-means clustering algorithm. After the model was implemented, it was deployed into google Colab using AutoML where 3 clusters were realized on the datasets. The model results were compared with other approaches that have been used to clustering information mining model to enhance selective dissemination of information. The other models approach compared with investigated with the same objectives and the results. The implemented model proved to be dynamic and overall best than the other machine learning approaches compared against with an accuracy percentage of 71%.

AREAS OF FUTURE RESEARCH

The comparison results from the experiments prove that the implemented recommendation k-mean clustering model approach for clustering information mining model to enhance selective dissemination of information for library users has the least percentage of mean vector and covariance matrix which resulted in a higher accuracy of 71%. When matrix vector and covariance vector are low, it brings an impression of an efficient model approach to clustering information mining model.

In future, work may be extended by adding suitable pre-processing approaches to improve the datasets as well as features selection approach to improve the classification accuracy. Future work should also extend on time series dynamic data that are in real time, thereby developing new technique against improved hybrid approaches.

REFERENCES

[1] L. Klain Gabbay and S. Shoham, “The role of academic libraries in research and teaching,” *J. Librariansh. Inf. Sci.*, vol. 51, no. 3, pp. 721–736, 2019, doi: 10.1177/0961000617742462.

[2] C. Stăiculescu and R. N. Elena Ramona, “University dropout. Causes and solution,” *Ment. Heal. Glob. Challenges J.*, vol. 1, no. 1, pp. 71–75, 2019, doi: 10.32437/mhgcj.v1i1.29.

[3] J. Apostolakis, “An introduction to data mining,” *Struct. Bond.*, vol. 134, no. August, pp. 1–35, 2010, doi: 10.1007/430_2009_1.

[4] R. Kruse and C. Borgelt, “Information mining,” *Int. J. Approx. Reason.*, vol. 32, no. 2–3, pp. 63–65, 2003, doi: 10.1016/S0888-613X(02)00088-9.

[5] K. L. Wagstaff, “Data Clustering,” *Adv. Mach. Learn. Data Min. Astron.*, no. September, pp. 543–561, 2012, doi: 10.1201/b11822-19.

[6] B. K. Nyaupane, “Machine Learning-- Clustering Algorithms,” no. September, 2022.

[7] S. S. Mesakar and M. S. Chaudhari, “A Review of Clustering Algorithms 1 1,2,” vol. 8491, no. October, pp. 4–6, 2013, doi: 10.5281/zenodo.7243829.

[8] G. Feng, J. Lin, and K. Wang, “Researches Advanced in Clustering Algorithms,” vol. 16, pp. 168–177, 2022.

[9] Dr. Elizabeth M. Minei, “How to Write an Introduction for a Research Paper - EduBirdie.com,” no. November, 2022, [Online]. Available: https://edubirdie.com/blog/research-paper-introduction.

[10] N. Pawar, “6. Type of Research and Type Research Design,” *Soc. Res. Methodol.*, vol. 8, no. 1, pp. 46–57, 2020, [Online]. Available: https://www.kdpublications.in.

[11] A. Sharma, “Review on Major Cyber security Issues in Educational Sector,” *Int. J. Comput. Sci. Eng.*, vol. 9, no. 12, pp. 26–29, 2021, doi: 10.26438/ijcse/v9i12.2629.

[12] Y. Zhao and X. Zhou, “K-means Clustering Algorithm and Its Improvement Research,” *J. Phys. Conf. Ser.*, vol. 1873, no. 1, 2021, doi: 10.1088/1742-6596/1873/1/012074.

[13] T. Ruzgas and M. Lukauskas, “Data clustering and its applications in medicine,” *New Trends Math. Sci.*, vol. 10, no. ISAME2022-Proceedings, pp. 067–070, 2022, doi: 10.20852/ntmsci.2022.465.

[14] N. Dey, A. S. Ashour, and G. N. Nguyen, “Deep learning for multimedia content analysis,” *Min. Multimed. Doc.*, vol. 1, no. 4, pp. 193–203, 2017, doi: 10.1201/b21638.

[15] T. Carneiro, R. V. M. Da Nobrega, T. Nepomuceno, G. Bin Bian, V. H. C. De Albuquerque, and P. P. R. Filho, “Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications,” *IEEE Access*, vol. 6, pp. 61677–61685, 2018, doi: 10.1109/ACCESS.2018.2874767.