New York Science Journal

Websites:
http://www.sciencepub.net/newyork
http://www.sciencepub.net

Emails:
newyorksci@gmail.com
editor@sciencepub.net

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

# Testing the Effectiveness of Using a Clustering Algorithm to Label Datasets

Ashwag Maghraby, Fawaz AlBatati

College of Computers, Umm Al-Qura University, Makkah, Kingdom of Saudi Arabia
aomaghraby@uqu.edu.sa , s44280427@st.uqu.edu.sa

**Abstract:** Big Data is generated in huge quantities and at tremendous speeds, and the vast majority of this big data is unlabeled datasets, unlike labeled datasets, which are very few and limited. The labeled dataset is organized and formatted which makes it easier to search and analyze. In contrast, an unlabeled dataset has no explanation, label, tag, class, or name for the features which makes it more difficult to process and analyses. For many researchers performing analytics on such Unlabeled vast data become challenging. This paper studied the effectiveness of utilizing a Clustering Algorithm Technique such as K-Medoids, K-Means, and Hierarchical to convert unlabeled datasets into a labeled form. The experiment results showed that the K-Medoids algorithm's performance is the best, followed by the Hierarchical algorithm and the K-Means algorithm, which are close in performance with a relative preference for the Hierarchical algorithm.
[Ashwag Maghraby, Fawaz AlBatati. **Testing the Effectiveness of Using a Clustering Algorithm to Label Datasets.** *N Y Sci J* 2024;17(5):21-31]. ISSN 1554-0200 (print); ISSN 2375-723X (online). http://www.sciencepub.net/newyork. 04. doi:10.7537/marsnys170524.04.

**Keywords:** Unsupervised Learning; Clustering Algorithm; Unlabeled data; labeling datasets

## 1. Introduction

Supervised learning is the process of learning a prediction model using a set of labeled datasets. It is well known that prediction model accuracy often rises as more labeled datasets become available.

Labeled datasets are often difficult to collect since the labeling process is frequently done by experts in the field from which these data are extracted. On the contrary, unlabeled datasets are widely available because labeling is a time consuming and laborious operation.

Different researchers have used different methods to solve the issue of few and scarcity of labeled datasets, among these researchers were those who dealt directly with datasets by using methods for labeling unlabeled datasets.

The majority of previous studies used supervised learning in [1] [2], semi-supervised learning in [3] [4] [5] [6], and deep learning in [5] with a small quantity of labeled data and a large amount of unlabeled data to create and learn their models. In contrast, few studies used unsupervised learning in [7] [8] [9] [10] and deep learning in [10] with only unlabeled data in [7] [8] for creating and learning the models.

This study offers a completely human-free solution to this pressing problem (labeling datasets) using AI itself. In order to turn unlabeled datasets into labeled datasets, this research provides a novel method employing unsupervised learning (Clustering Algorithms). These algorithms were used in conjunction with preprocessing methods to produce models that demonstrated and validated the efficacy of utilizing clustering algorithms for labeling datasets.

The remaining sections in this paper are organized as follows: firstly, section 2 literature survey provides a theoretical background for studies that used different methods to solve the issue of the scarcity of labeled datasets. Secondly, in section 3 approaches, methodologies, algorithms, and methods were explained. Thirdly, in sections 4 and 5 experiment and results discussion explains the steps of the experiment for implementing the clustering algorithm and discusses the experiment's result. Finally, in section 6 conclusion and recommendations provides conclusions and suggestions that align with the study's major themes.

## 2. Literature survey

Through extrapolating and reviewing the studies related to the subject of this research, they can be divided into three paradigms as follows:

### A. Supervised Learning Paradigm

Blum, A., et al. [1] used substituted labels by utilizing the structure between the patterns' dispersion to the various sensory modalities. Where, De Sa, et al. [2], used two supervised learning algorithms: co-training and multi-modality. Co-training makes use of two

classifiers to learn from various views on the data**,** whereas multi-modality employs substituted labeling to assign labels to unlabeled data. Algorithms were evaluated on the tasks of classifying web pages and vowel recognition, and they were shown to be effective in achieving good performance even when there is a small amount of labeled data.

**B.Semi-Supervised Learning (with Deep Learning) Paradigm**

A variety of tasks, including information access, remote sensing image processing, natural language processing, and semi-supervised clustering, are covered in four publications[3][4][5][6]. Twenty distinct semi-supervised learning algorithms are proposed among the publications, using a range of methods including classification, clustering, and dimensionality reduction. In the publications, the suggested algorithms are tested on various real-world data sets, and it is demonstrated that the algorithms can achieve good performance even when there is a small amount of labeled data. Vittaut JN, et al. [3] used a limited quantity of labelled data with a larage number of unlabeled data. In [4] study, Forestier, G. et al. employed a small amount of labelled data together with unlabeled data in this study. Bouchachia, A. researcher in [5] used partially labelled data along with unlabeled data, while Pavithra, M., et al. [6] used high-dimensional sparse data and a small number of labelled examples as seeds while researchers in [5] employing semi-supervised learning and deep learning (neural networks).

**C.Unsupervised Learning (with Deep Learning) Paradigm**

In [7], AlBatati and Alarabi used only unlabeled data to train a K-means clustering algorithm. They then used the clusters to label the data. This approach was effective for labeling spatial data. In [8], Pius Owoh et al. used a combination of unsupervised and supervised learning to label unlabeled sensor data. They first used K-means clustering to cluster the data. Then, they used a support vector machine (SVM), K-nearest neighbor (KNN), and naive Bayes (NB) algorithm to classify the data. This approach achieved high accuracy in labeling the data. In [9], Tashfin Ansari et al. used only labeled data to train a K-means clustering algorithm. They then used the clusters to predict the labels of new data. This approach was effective for predicting the clusters of the IRIS dataset. In [10], Dara et al. used a self-organizing map to cluster unlabeled data and then inferred potential labels from the clusters. They also used a multi-layer perceptron (MLP) algorithm to train the model on both labeled and inferred data. This approach improved the performance of the MLP algorithm on a variety of benchmark tasks.

The issues raised in the study [1] [2] [3] [6] are the use of labeled datasets, even if their use is limited and small.It also brings attention to the absence of clustering algorithms in the papers analyzed. While in [6], the researchers discovered that clustering accuracy and performance significantly improved when co-training, multi-modality, and semi-supervised clustering algorithms were combined with labeled data.

In [4], the result was due to the predictive model for the labeled data and not due to the clustering algorithm used. In [5], a different clustering algorithm than the one used in this paper was used, and the performance of the model was improved by using the MLP pre-labeling method with only labeled data and not due to the used of clustering algorithms. In [7], a method for labeling unlabeled spatial data using the K-means clustering algorithm is demonstrated. In [8], unlabeled data was clustered using the K-means algorithm and then the outputs were used as labeled data to train three Supervised algorithms and the performance improvement was due to training the Supervised algorithms not due to the K-means algorithm. In [9], the K-means clustering algorithm was used on only one labeled dataset to demonstrate how to calculate the optimal K-value of clustering. In [10], a different clustering algorithm than the one used in this paper was used, and the result was thanks to the use of labeled data and the MLP algorithm.

By using clustering-based unsupervised learning algorithms, this research aimed to address the limitations of working with unlabeled datasets. The ultimate goal was to supply labeled datasets to artificial intelligence researchers so they could use them for productive research. It employed three Clustering-based unsupervised learning algorithms: K-Medoids, K-Means, and Hierarchical clustering algorithms on unlabeled data to identify patterns and group similar data points together, and then the clustering output was compared with the actual class label to obtain the best algorithm that achieved the highest accuracy. Table 1 (a) and (b) summarize the key differences between the methodology used in this paper and earlier research, and it serves as the primary justification for this scientific.

**3. Clustering algorithms approaches**

This section gives an overview of Clustering Algorithms. It describes their uses, benefits, advantages, and disadvantages and how to implement and activate them.

Clustering is a type of unsupervised learning technique that seeks to group together instances that are strongly connected to one another in order to make instances from one cluster more comparable to those from other clusters [11].

This paper presents three clustering algorithms for unsupervised learning to convert unlabeled datasets into categorized datasets: K-means, K-Medoids, and Hierarchical clustering.

Table 1 (a). summary of previous researches and paper approach

| Previous Studies | Datasets selection | AI / ML methods | Algorithms | Difference from this paper |
|---|---|---|---|---|
| Blum, A., et al. [1] | Large set of unlabeled data + small set of labeled data | Supervised Learning | Co-Training (Naive Bayes) | Use labeled data + Supervised Learning |
| De Sa, et al. [2] | substituted label by utilizing the structure in between dispersion of patterns to the various sensory modalities | Supervised Learning | MULTI-MODALITY (NN labeling + Self-Supervised piecewise-linear) | Use labeled data + Supervised Learning |
| Vittaut JN, et al. [3] | small number of labeled data + large number of unlabeled data | Semi-Supervised Learning | CML + CEM | Use labeled data + Semi-Supervised Learning |
| Forestier, G. et al. [4] | limited labeled data + unlabeled data | Semi-Supervised Learning | Nine different algorithms: SL + DL + CLM + SRIDHCR + SCEC + RC + SK + CK + SLEMC | Use labeled data + Semi-Supervised Learning |
| Bouchachia, A. [5] | partly labeled data + unlabeled data | Semi-Supervised Learning & Deep Learning (Neural Networks) | 1- radial basis function network (RBFN) 2- multi-layer perceptron (MLP) 3- Pre-labeling based on nearest neighborhood for MLP 4- clustering 5- Gaussian mixture models based on the expectation-maximization (EM) 6- Seeded clustering | Use labeled data + Semi-Supervised Learning & Deep Learning |
| Pavithra, M., et al. [6] | A small number of labeled examples + unlabeled examples | Semi-Supervised clustering | K-means clustering & Hierarchical clustering | Use labeled data + Semi-Supervised Learning |
| AlBatati, F. & Alarabi, L. [7] | only unlabeled data | Unsupervised Learning | K-means Clustering | Didn't use K-Medoids, and Hierarchical Clustering + Use K-means Clustering with only three preprocessing techniques + Use only one dataset |
| Pius Owoh, N., et al. [8] | Only unlabeled data | Unsupervised Learning and Supervised Learning | K-means clustering and Support Vector Machine (SVM) + K-Nearest Neighbor (KNN) + Naive Bayes (NB) | Didn't use K-Medoids, and Hierarchical Clustering |

Table 1 (b). summary of previous researches and paper approach

| Previous Studies | $SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$ | I / ML methods | Algorithms | Difference from this paper |
|---|---|---|---|---|
| Tashfin Ansari et al. [9] | Only labeled data | Unsupervised Learning | K-Means Clustering | Use only labeled data with one dataset + Didn't use K-Medoids, and Hierarchical Clustering |
| Dara, R. et al. [10] | unlabeled data + potential labeling inferred labeling from clusters | Unsupervised Learning & Deep Learning (Neural Networks) | A multi-layer perceptron (MLP) + Clustering | Use labeled data + Unsupervised Learning & Deep Learning |

**Input:**

- $k$: the number of clusters,
- $D$: a data set containing $n$ objects.

**Output:** A set of $k$ clusters.

**Method:**

(1) arbitrarily choose $k$ objects from $D$ as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

Figure 1. The steps of K-means Clustering algorithm [12].

**A. K-means Clustering algorithm**

The K-means clustering method represents every cluster by its associated cluster centroid. The technique would repeatedly use the following phases to split the input data into K-distinct clusters, see Figure 1 [12]:

Phase1: Assign each instance to its nearest centroid to create K-clusters.

Phase2: Recalculate each cluster's centroid.

**Measures of K-means algorithm Validity**

The Internal Validation Criteria (Unsupervised) measure is used to evaluate a clustering structure's quality without taking into account outside data. To begin, determine the number of the ideal clusters (K-value). K-class labels will then be created as a result of the K-means algorithm's division of the datasets into (K-clusters). This study employs the elbow technique to locate the elbow point by Sum of Square Error (SSE) measurement to identify the best cluster number (K-value) in the K-means algorithm. The SSE measurement's potential uses are depicted in (1) [13].

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x) \quad (1)$$

In cluster $C_i$, $x$: data point, $m_i$: its representative point

**B.K-medoids Clustering algorithm**

The K-Medoids algorithm, developed to address outliers in the k-means algorithm, uses actual objects to represent clusters, minimizing dissimilarities between objects, instead of using the mean value of objects as a reference point in a cluster. This makes it more robust to noise and outliers but is more computationally expensive. The Phases of K-Medoids algorithm are shown in Figure 2 [14]:

Phase1: Pick (k) randomly from the dataset; (k = clusters number).

Phase2: Each point of data is assigned to the cluster that contains its nearest medoid.

```
begin
  Initialize tree T to root containing D;
  repeat
    Select a leaf node L in T based on pre-defined criterion;
    Use algorithm A to split L into L_1 ... L_k;
    Add L_1 ... L_k as children of L in T;
  until termination criterion;
end
```

Figure 2. The algorithm of K-Medoids clustering [14].

Phase 3: For each data point in cluster i, the distinction from every other data point is calculated and then added. A point in an ith cluster is designated as the medoid when the calculated sum of its distances from other locations is at its smallest value.

Phase4: Repeat steps 2 and 3 as necessary to achieve convergence, which is when the medoids stop moving.

**Input:**

- $k$: the number of clusters,
- $D$: a data set containing $n$ objects.

**Output:** A set of $k$ clusters.

**Method:**

(1) arbitrarily choose $k$ objects in $D$ as the initial representative objects or seeds;
(2) **repeat**
(3) assign each remaining object to the cluster with the nearest representative object;
(4) randomly select a nonrepresentative object, $o_{random}$;
(5) compute the total cost, $S$, of swapping representative object, $o_j$, with $o_{random}$;
(6) **if** $S < 0$ **then** swap $o_j$ with $o_{random}$ to form the new set of $k$ representative objects;
(7) **until** no change;

Figure 3. Bottom-Up Agglomerative Hierarchical algorithm [15].

```
begin
  Initialize n × n distance matrix M using D;
  repeat
    Pick closest pair of clusters i and j using M;
    Merge clusters i and j;
    Delete rows/columns i and j from M and create
      a new row and column for newly merged cluster;
    Update the entries of new row and column of M;
  until termination criterion;
  return current merged cluster set;
end
```

Figure 4. Top-Down Divisive Hierarchical algorithm [15].

**Measures of K-medoids algorithm Validity (Internal Criteria)**

The Sum of the Absolute Error (SAE) measure is used to evaluate the quality of a clustering structure without taking into account external data. The elbow method can be used to find the optimal number of clusters (K-value) in the k-medoids algorithm by finding the point where the SAE measure starts to decrease rapidly. Equation (2) shows how the SAE measurement may be used [14].

$$E = \sum_{i=1}^{k} \sum_{\boldsymbol{p} \in C_i} dist(\boldsymbol{p}, \boldsymbol{o}_i) \quad (2)$$

E: SAE to all dataset's objects p, oi: object's representative of Ci

**A.Hierarchical Clustering algorithm**

Hierarchical Clustering algorithm is a widely used method for understanding taxonomies, utilizing a dendrogram as a visual representation of nested clusters, which can be cut to any desired number of clusters [15]. Hierarchical clustering has two primary forms:

1- Agglomerative (Bottom-Up) Methods: Beginning with the points as independent clusters, the clusters' closest pairs are merged at each stage until only one cluster (or k-clusters) are left, see Figure 3 [15].

2- Divisive (Top-Down) Methods: Split one cluster at a time, starting with an all-inclusive cluster, until only one or k-clusters remain that each containing a single point, see Figure 4 [15].

**Measures of Hierarchical Algorithm Validity**

The External Validation Criteria (Supervised) measure is used to evaluate a clustering structure's quality by taking into account outside data.

**1) Entropy (Measurement of disorder):**

The degree to that every cluster contains of an instances from a single class. The possible smallest value for entropy is 0.0, which occurs when all objects in a cluster are from the same. In other words, there's no disorder in the cluster. The more considerable value of entropy, doing more disorder there is in the associated cluster [16].

**2) Purity:**

Another indicator of how many objects of a particular class are included in the same cluster. Purity levels near 0 indicate a bad cluster, whereas purity values around 1 indicate ideal clustering [16].

**D. Algorithms Metrics**

**1) Accuracy:**

The accuracy of a model is a measure of how well the model can classify new data. It is calculated by dividing the number of correctly classified test set tuples by the total number [16].

**2) Error rate:**

The error rate of a model is the proportion of test set data points incorrectly classified by the model [16].

**3) *F*-measure (F-score):**

Instead of using precision and recall separately, we can combine them into a single measure called the F1 score. The F1 score is calculated using the harmonic mean of precision and recall. This means it gives more weight to cases with high precision and recall [11].

Precision: calculated by dividing the number of correctly classified positive tuples by the total number of tuples classified as positive [11].

Recall: what percentage of positive tuples are correctly labeled [11].

**4) Homogeneity:**

A perfectly homogeneous clustering is one where all data points in a cluster have the same class label. Homogeneity measures how close a clustering algorithm is to this ideal [16].

**5) Completeness:**

A complete clustering is perfect if one in which all data points from the same class are clustered together. Completeness indicates how close to perfection the clustering algorithm is [16].

**6) V-measure:**

It is a measure of clustering quality based on homogeneity and completeness. It is a harmonic mean of these two measures, which gives more weight to homogeneity when they are close. V-measure is independent of the absolute values of the labels, which means that it does not change if the labels are permuted [16].

**4. Methodologies**

This section presents a novel approach to label unlabeled datasets by evaluating clustering algorithms on labeled datasets and providing experimental evidence of these algorithms' efficacy.

**A. Experiment Steps**

Figure 5 shows the experiment steps:

1) First, Downloading the dataset: This step involves the following actions: describe the dataset, display information about the dataset, and display the dataset attributes information.

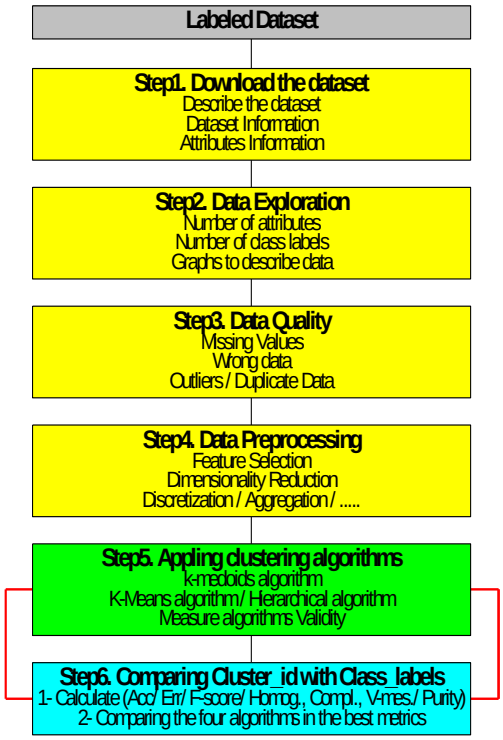



Figure 5. The steps of the experiments to prove and confirm the effectiveness of using a clustering algorithm for labeling datasets.

2) Second, Data Exploration: Exploratory data analysis (EDA) is the initial step in data analysis, involving statistical and visualization methods to identify patterns and problems in the dataset, aiding in the selection of appropriate models or algorithms. This step involves the following actions: display the number of dataset instances, display the number of dataset attributes, and display a statistical summary of numerical attributes.
3) Third, Data Quality: This research identifies data quality issues in a dataset and suggests strategies to address them. Key strategies include: 1- removing missing values, assessing missing data values, and disregarding missing data; 2- removing Outliers which are data objects that interfere with data analysis; and 3- dealing with duplicate data, often from heterogeneous sources, which can be addressed without necessarily removing duplicate data.
4) Fourth, Data Preprocessing: Preprocessing techniques are applied to a dataset to improve its cost, quality, and time. These include feature subset selection, dimensionality reduction, discretization and binarization, aggregation, sampling, and feature creation. The goal is to reduce dimensionality, eliminate irrelevant features, convert data to ordinal attributes, and reduce data variability by incorporating multiple attributes into one new attribute.
5) Fifth, Applying Clustering Algorithms:The following operations are carried out on the dataset in this step, if required: build clustering models using the K-Medoids, the K-Means, and the Hierarchical algorithms, then measure the validity of these algorithms to find the best number of clusters.
6) Sixth, Comparing Cluster_id with Class_labels: This step compares the found Cluster_id with actual Class_labels by calculating the following metrics: Accuracy, Error rate, *F*-measure (F-score), Purity, and (Homogeneity, Completeness, V-measure), then comparing the four algorithms in the best metrics.

**B. Datasets**

This section descried Iris Dataset.

**1) Value of Data**

Fisher's Iris dataset, a well-known pattern recognition database, transformed the field by offering an extensive compilation of measurements for iris flowers. Three distinct species of iris flowers: setosa,

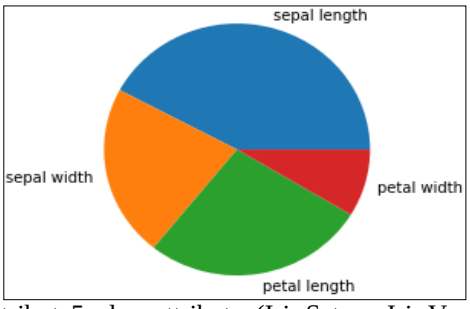
versicolor, and virginica are the subject of this dataset. Nevertheless, mistakes in the dataset for example, the second and third features of the 38th sample, "4.9,3.6,1.4,0.1," and the fourth feature measurement in the sample, "4.9, 3.1, 1.5, 0.2" have been fixed in subsequent studie, enabling researchers to carry out additional analysis and create pattern recognition algorithms.

**2) Data Description**

Figure 6 shows a sample from the. The dataset contains three classes (setosa, versicolor, and virginica) of fifty instances each; each class refers to a certain kind of iris plant. The attribute "class of iris plant" is the prediction attribute [17].

**3) Attribute Information**

This dataset consists of 5 attributes: Attribute 1: sepal length in cm; Attribute 2: sepal width in cm; Attribute 3: petal length in cm; Attribute 4: petal width in cm;




Attribute5: class attribute: (Iris Setosa, Iris Versicolour, Iris Virginica) [17].

A graph representing the "iris" dataset, which displays the distribution of instances based on attributes, and the attribute with more instances of beings than the others are shown in Figure 7.

| | sepal length | sepal width | petal length | petal width | class |
|---|---|---|---|---|---|
| 0 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 1 | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| 4 | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| ... | ... | ... | ... | ... | ... |
| 145 | 6.7 | 3.0 | 5.2 | 2.3 | Iris-virginica |
| 146 | 6.3 | 2.5 | 5.0 | 1.9 | Iris-virginica |

Figure 6. A sample from "iris" dataset.

Figure 7. A graph describes the "iris" dataset.

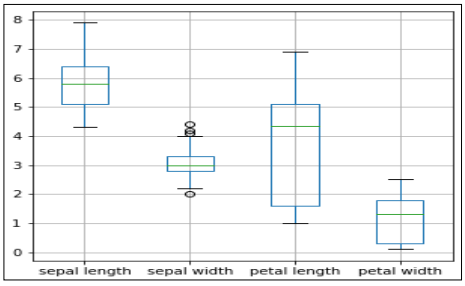



Figure 8. A graph shows there are outliers in the dataset.

**5. Experiment and results discussion**

The procedures used in the experiments to demonstrate and validate the efficacy of labeling datasets using a clustering algorithm described in Figure 5, have been applied in this section.

**A. Experiment**

**1) First, Downloading the dataset**

It can download the "iris" dataset from the link in [17].

**2) Second, Data Exploration**

The number of instances: 150 instances; The number of attributes: 5 attributes; The number of class labels: 3; and a statistical summary for numerical attributes is displayed.

**3) Third, Data Quality**

Check whether the selected dataset has any data quality issues and choose suitable strategies to deal with any issue.

There are no missing values, but there are outliers in attribute 2, "sepal width" as seen in Figure 8; processed by dropping all instances with outliers, the number of instances after processing outliers becomes 149, and there are duplicate data, number of duplicate rows: 3, processed by dropping all duplicated data, the number of instances after processing repeated data becomes 146.

**4) Fourth, Data Preprocessing**

Feature Selection: no need to apply the feature selection technique, due to the small size of the data in terms of the number of attributes.

**5) Fifth, Appling clustering algorithms**

**a)Using K-Medoids algorithm**

The approach validated using the kneed python package and elbow method, determining the

best K value which is three clusters, as illustrated in Figure 9.

**b) Using K-Means algorithm**

The approach validated using the kneed python package and elbow method, determining the best K value which is three clusters, as illustrated in Figure 10.

**c) Using Hierarchical algorithm**

The approach validated using purity and entropy metrics (lowest-value of entropy with highest-value of purity), determining the best K value which is four clusters , as illustrated in Figure 11.

**6) Sixth, Comparing Cluster_id with Class_labels**

**a) For K-Medoids algorithm**

Comparing the identified Cluster_Id with the real Class_Labels (Class attribute) is the next step. Table 2 shows the following values:

Accuracy= 93%; Error rate=7%; *F*-measure (F-score)= 93%; Purity= 0.93; Homogeneity= 0.80; Completeness= 0.80; and V-measure= 0.80.

**b) For K-Means algorithm**

Comparing the identified Cluster_Id with the real Class_Labels (Class attribute) is the next step. Table 2 shows the following values:

Accuracy= 89%; Error rate= 11%; *F*-measure (F-score)= 89%; Purity= 0.89; Homogeneity= 0.75; Completeness= 0.77; and V-measure:= 0.76.

**c) For Hierarchicalalgorithm**

Comparing the identified Cluster_Id with the real Class_Labels (Class attribute) is the next step. Table 2 shows the following values:

Accuracy= 89%; Error rate= 11%; *F*-measure (F-score)= 89%; Purity= 0.89; Homogeneity= 0.76; Completeness= 0.78; and V-measure= 0.77.

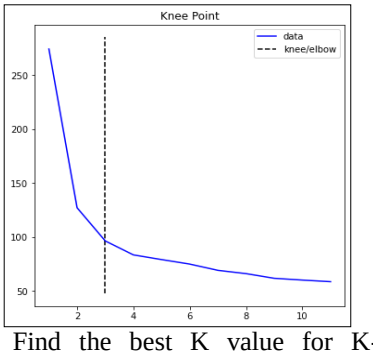



Figure 9. Find the best K value for K-Medoids algorithm using elbow method.

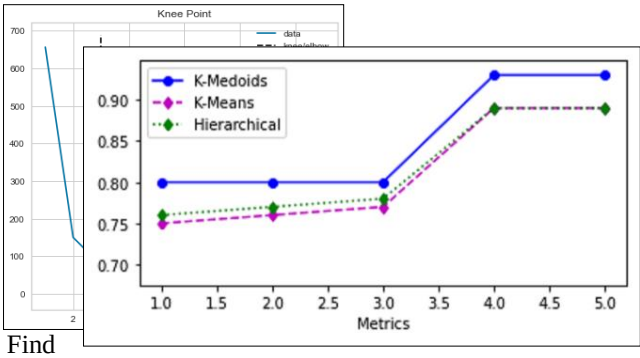



Figure 10. Find the best K value for K-Means algorithm using elbow method.

```
number of clusters = 1   Purity = 0.33   Entropy = nan
number of clusters = 2   Purity = 0.67   Entropy = 5.01
number of clusters = 3   Purity = 0.89   Entropy = 4.53
number of clusters = 4   Purity = 0.89   Entropy = 1.88
number of clusters = 5   Purity = 0.89   Entropy = 2.90
number of clusters = 6   Purity = 0.89   Entropy = 4.29
number of clusters = 7   Purity = 0.94   Entropy = 4.09
number of clusters = 8   Purity = 0.94   Entropy = 3.82
number of clusters = 9   Purity = 0.94   Entropy = 3.20
number of clusters = 10          Purity = 0.94   Entropy = 2.10
number of clusters = 11          Purity = 0.94   Entropy = 2.47
```

Figure 11. Find the best number of clusters for Hierarchical algorithm by using purity and entropy metrics.

**B. Results Discussion**

**1) Results Discussion of K-Medoids algorithm**

Comparing the metrics used in previous section for the K-Medoids algorithm, it is clear that Accuracy reaches 93% compared to the Error rate of 7%, and these are excellent rates so far. The metric of Purity (0.93) is close to 1, which is an excellent ratio, and the metrics of Homogeneity (0.80), Completeness (0.80), and V-measure (0.80) are close to 1, which is a very good ratio. Likewise, the metric of *F*-measure (F-score = 93%) points to an excellent ratio.

**2) Results Discussion of K- Means algorithm**

Comparing the metrics used in previous section for the K-Means algorithm, it is clear that

Accuracy reaches 89% compared to the Error rate of 11%, which is a very good ratio so far. The metric of
Table 2. A comparison between the three algorithms based on the measures deduced from Experiment

| | Acc. | Err. | F1 | Purity | Homog. | Compl. | V-mes. |
|---|---|---|---|---|---|---|---|
| **K-Medoids** | 93% | 7% | 93% | 0.93 | 0.80 | 0.80 | 0.80 |
| **K-Means** | 89% | 11% | 89% | 0.89 | 0.75 | 0.77 | 0.76 |
| **Hierarchical** | 89% | 11% | 89% | 0.89 | 0.76 | 0.78 | 0.77 |

Figure 12. Diagram of the three algorithms based on the metrics deduced from Experiment.

Purity (0.89) is close to 1, which is a very good ratio, and the metrics of Homogeneity (0.75), Completeness (0.77), and V-measure (0.76) are close to 1, which is a good ratio, relatively. Likewise, the metric of *F*-measure (F-score = 89%) points to a very good ratio.

**3) Results Discussion of Hierarchical algorithm**

Comparing the metrics used in previous section for the Hierarchical algorithm, it is clear that Accuracy reaches 89% compared to the Error rate of 11%, which is a very good ratio so far. The metric of Purity (0.89) is close to 1, which is a very good ratio, and the metrics of Homogeneity (0.76), Completeness (0.78), and V-measure (0.77) are close to 1, which is a good ratio, relatively. Likewise, the metric of *F*-measure (F-score = 89%) points to a very good ratio.

**C. Comparing algorithms in the best metrics**

Table 2 shows a comparison between the three algorithms based on the metrics deduced from Experiment and Figure 12 shows a diagram of the three algorithms based on the measures deduced from the experiment.

The Hierarchical algorithm and the K-Means algorithm are closely connected in performance with a relative preference for the Hierarchical algorithm, while the K-Medoids strategy performs the best overall, as shown in Table 2 and Figure 12. So clear that these techniques can be employed to convert unlabeled datasets into labeled datasets.

Finally, Figure 13 shows an image from the "iris" dataset after completing the experiment. It is clear from Figure 13 that a new attribute (Cluster ID) was added to the dataset after the completion of the experiment, which represents the cluster numbers based on which the dataset was divided using the clustering algorithms, which were compared with the actual class of the dataset by metrics of the clustering algorithm.

**6. Conclusion and future work**

This paper proposes a different solution and approach from related work by using unsupervised learning clustering algorithms for labeling datasets due to the strength of these algorithms, proven through multiple studies conducted on them [18] [19], in addition to the suitability of using them with unlabeled data because they not require a class label.

When comparing the result reached in this paper (thanks to the use of clustering algorithms) with the highest results in related studies, this research find the following: In [4], a higher result was obtained than that achieved in this paper, but it did not use all the clustering algorithms used in this paper, and the result was thanks to the predictive model for the labeled data and not because of the clustering algorithm used. In [5], an excellent result was obtained, and a different clustering algorithm than the one used in this paper was used, and the result was due to the MLP pre-labeling method using labeled data and not due to the clustering algorithms used. In [10], a higher result was obtained than that achieved in this paper, but used a different clustering algorithm, and the result was thanks to the use of labeled data and the use of the MLP algorithm. Even in the related works whose results were lower than those of this paper, the credit for obtaining those results was the use of algorithms other than clustering algorithms, such as co-training, Multi-Modality, Semi-supervised Clustering, Supervised algorithms, and Deep learning algorithms, as well as the use of other approaches different from the approach used in this paper, such as predictive model and MLP pre-labeling.

Experiments in this paper with K-Medoids, Hierarchical, and KMeans algorithms show that K-Medoids perform best, making it a viable solution for converting unlabeled data into labeled data.

Future work should explore alternative algorithms for K-Medoids, Hierarchical, and K-Means algorithms like DBSCAN (Density-based clustering) and C-means (fuzzy clustering) for multi-dimensional datasets, overcoming K-value determination issues and processing fuzzy multi-dimensional datasets. In addition to testing the best performance algorithm obtained in this paper on unlabeled datasets, with a detailed explanation of the steps to do this leading to converting the unlabeled datasets into labeled form.

**Acknowledgements:**

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Corresponding Author:**
Dr. Ashwag Maghraby
College of Computers, Umm Al-Qura University Makkah, Kingdom of Saudi Arabia
E-mail: aomaghraby@uqu.edu.sa

**References**

[1]. Blum, A., & Mitchell, T. (1998, July). Combining labeled and unlabeled data with cotraining. In Proceedings of the eleventh annual conference on Computational learning theory (pp. 92-100).
[2]. De Sa, V. R. (1994). Learning classification with unlabeled data. In Advances in neural information processing systems (pp. 112-119).
[3]. Vittaut JN., Amini MR., Gallinari P. (2002) Learning Classification with Both Labeled and Unlabeled Data. In: Elomaa T., Mannila H., Toivonen H. (eds) Machine Learning: ECML 2002. ECML 2002. Lecture Notes in Computer Science, vol 2430. Springer, Berlin, Heidelberg.
[4]. Forestier, G. and Wemmert, C. (2016) 'Semi-supervised learning using multiple clusterings with limited labeled data', Information Sciences, 361–362, pp. 48–65. doi: 10.1016/j.ins.2016.04.040.
[5]. Bouchachia, A. (2007). Learning with partly labeled data. Neural Computing and Applications, 16(3), 267-293.
[6]. Pavithra, M., & Parvathi, R. M. S. (2019). A Review article on Semi-Supervised Clustering Framework for High Dimensional Data. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 4, pp. 102-108.
[7]. AlBatati, F., & Alarabi, L. (2021). Labeling Big Spatial Data: A Case Study of New York Taxi Limousine Dataset. International Journal of Computer Science & Network Security, 21(6), 207-212.
[8]. Pius Owoh, N., Mahinderjit Singh, M., & Zaaba, Z. F. (2018). Automatic annotation of unlabeled data from smartphone-based motion and location sensors. Sensors, 18(7), 2134.
[9]. Tashfin Ansari, Dr. Almas Siddiqui, Awasthi G. K. (2021). Clustering Analysis using an Unsupervised Machine Learning Method. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN: 2456-3307, Volume 7 Issue 3, pp. 602-609.
[10]. Dara, R., Kremer, S. C. and Stacey, D. A. (2002) 'Clustering unlabeled data with SOMs improves classification of labeled real-world data', Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290), Neural Networks, 2002.IJCNN '02. Proceedings of the 2002 International Joint Conference on, Neural networks, IJCNN'02, 3, p. 2237. doi: 10.1109/IJCNN.2002.1007489.
[11]. Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
[12]. Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier. (Chapter 10 Cluster Analysis, Section 10.2.1 K-Means).
[13]. Rakesh, Verma (2009). Data Mining the HypertextBook. (Chapter 4 Cluster Analysis, Section 2 Sum of Squared Error Equation). [CrossRef]
[14]. Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier. (Chapter 10 Cluster Analysis, Section 10.2.2: K-Medoids & Sum of Absolute Error Equation).
[15]. Aggarwal, C. C. (2015). Data mining: the textbook. Springer (Chapter 6 Cluster Analysis, Section 6.4 Hierarchical).
[16]. Rosenberg, A., & Hirschberg, J. (2007, June). V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL) (pp. 410-420).
[17]. Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2), 179-188. Iris Dataset. [CrossRef]
[18]. Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001, July). Clustering algorithms and validity measures. In Proceedings Thirteenth International Conference on Scientific and Statistical Database Management. SSDBM 2001 (pp. 3-22). IEEE
[19]. Charu C. Aggarwal, C. C. (2015). Data mining: the textbook. Springer. ISBN 978-3-319-14142-8 (eBook). DOI 10.1007/978-3-319-14142-8.

4/22/2024