Outliers Detection Based on Partitioning Around Medoids

Ei Ei Htwe
Computer University (Mandalay)
eieihtwe007@gmail.com

Abstract

Clustering is the process of grouping a set of objects into classes or clusters so that objects within a cluster have similarity in comparison to one another, but are dissimilar to objects in other clusters. Clustering analysis is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes. This system is intended to cluster Iris Plants (Setosa, Versicolor and Virginica) by using Partitioning Around Medoids (PAM) clustering algorithm. Partitioning Around Medoids (PAM) algorithm processes step of methods for selecting initial medoids. It also calculates the distance matrix once and uses it for finding new medoids at every iterative step. On further analysis of Iris data, the researchers found that Setosa is a clearly separable cluster while the other two clusters, Versicolor and Virginica, have significant overlap with each other. All clustering methods were able to identify Setosa more or less correctly, but made mistakes on Versicolor and Virginica. So, there are to detect inconsistent data which are outliers in the clusters. By using this system, the user can know about Iris plants, at the same time, they can learn outlier detection by PAM clustering algorithm.

1. Introduction

Outliers are the set of objects that are considerably dissimilar from the remainder of the data. Outlier detection is an extremely important problem with a direct application in a wide variety of application domains, including fraud detection, identifying computer network intrusions and bottlenecks, criminal activities in e-commerce and detecting suspicious activities. Clustering is a popular technique used to group similar data points or objects in groups or clusters. Clustering is an important tool for outlier analysis. Several clustering-based outlier detection techniques have been developed. Most of these techniques rely on the key assumption that normal objects belong to large and dense clusters, while outliers form very small clusters [4]. This system uses Partitioning Around Medoids (PAM). PAM attempts to determine k partitions for n objects. The algorithm uses the most centrally located object in a cluster (called medoid) instead of the cluster mean. The system uses Iris plants dataset. There are fifty plants of each species with the four measurements on each plant: petal length, petal width, sepal length, and sepal width [3]. PAM is more robust than the k-means algorithm in the presence of noise and outliers. This is because the medoids produced by PAM are robust representations of the cluster centers and are less influenced by outliers and other extreme values than the means. Small clusters are then determined and considered as outlier clusters. To detect the outliers in the rest of clusters, computer the Absolute Distances between the Medoid, μ, of the current cluster and each one of **the Points, p_i,** in the same cluster (i.e., $|p_i - \mu|$). The produced value will be termed (ADMP) [4].

2. Related Work

Clustering has been studied extensively for more than 40 years and the researchers discovered many clustering algorithms due to its wide applications.

For partitioning methods, k-means algorithm is the simplest and most commonly used clustering algorithm employing a square error criterion. In k-means, each cluster is presented by the center of the cluster. It is relatively scalable and efficient in clustering large data sets. K-means algorithm is by far the most popular clustering tool used in scientific and industrial applications. K-means algorithm is computationally fast, and iteratively partitions a data set into k-disjoint clusters, where the value of k is and algorithmic input. The goal is to obtain the partition with the smallest square error. After that, the researchers developed some partitioning clustering algorithm by the extension of k-means algorithm. Because k-means algorithm has some drawbacks such as it is sensitive to noise and outliers. The k-medoids algorithm of PAM in which each cluster is represented by the most centrally located objects called medoids. PAM is more robust than k-means in the presence of noise and outliers

because a medoid is less influenced by outliers or other extreme values than a mean.

Clustering-based approaches consider clusters of small sizes as clustered outliers. In these approaches, small clusters (i.e., clusters containing significantly less points than other clusters) are considered outliers. The advantage of the clustering-based approaches is that they do not have to be supervised. Moreover, clustering-based techniques are capable of being used in an incremental mode (i.e., after learning the clusters, new points can be inserted into the system and tested for outliers) [4].

3. Theory Background

3.1 Outliers

Data mining, in general, deals with the discovery of non-trivial, hidden and interesting knowledge from different types of data. With the development of information technologies, the number of databases, as well as their dimension and complexity, grow rapidly. It is necessary what we need automated analysis of great amount of information. The analysis results are then used for making a decision by a human or program. One of the basic problems of data mining is the outlier detection [6].

An outlier is an observation of the data that deviates from other observations so much that it arouses suspicions that it was generated by a different mechanism from the most part of data. Outlier detection has many applications, such as data cleaning, fraud detection and network intrusion. The existence of outliers can indicate individuals or groups that have behavior very different from the most of the individuals of the dataset. Frequently, outliers are removed to improve accuracy of the estimators. But sometimes the presence of an outlier has a certain meaning, which explanations can be lost if the outlier is deleted.

3.2 Outliers in Clustering

The outlier detection problem in some cases is similar to the classification problem. The main concern of clustering-based outlier detection algorithms is to find clusters and outliers, which are often regarded as noise that should be removed in order to make more reliable clustering. Some noisy points may be far away from the data points, whereas the others may be close. The far away noisy points would affect the result more significantly because they are more different from the data points. It is desirable to identify and remove the outliers, which are far away from all the other points in cluster. So, to improve the clustering such algorithms use the same process and functionality to solve both clustering and outlier discovery [6].

3.3 Partitioning Around Medoids (PAM) Algorithm

Input: The number of clusters k and a database containing n objects.

Output: A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoid.

Method:

Use real object to represent the cluster

1. Select ***k*** representative objects arbitrarily
2. For each pair of non-selected object ***h*** and selected object ***I*** **(Go to process 1:)**, calculate the total swapping cost TC_{ih} **(Go to Process 2:)**
3. For each pair of ***i*** and ***h***,
   If $TC_{ih} < 0$, ***i*** is replaced by ***h***
   Then assign each non-selected object to the most similar representative object
4. repeat steps 2-3 until there is no change[2];

Process 1: To determine whether a non-medoid object, h, is a good replacement for a current medoid, i, the following four cases are exmained for each of the non-medoid objects, j.

Case 1: j currently belongs to medoid i. If i is replaced by h as a medoid and j is closest to one of t, i is not equal to t, then j is reassigned to t.

Case 2: j currently belongs to medoid i. If i is replaced by h as a medoid and j is closest to h, then j is reassigned to h.

Case 3: j currently belongs to medoid t, t is not equal to i. If i is replaced by h as a medoid and j is still closet to t, then the assignment does not change.

Case 4: j currently belongs to medoid t, t is not equal to i. If i is replaced by h as a medoid and j is closet to h, then j is reassigned to h.

Process 2: PAM Clustering: Total Swapping Cost

$$TC_{ih} = \sum_j C_{jih}$$

- i is a current medoid, h is a non-selected object
- Assume that i is replaced by h in the set of medoids
- $TC_{ih} = 0$;
- For each non-selected object $j \neq h$: TC_{ih}+=d(j,new_medj)-d(j,prev_medj):
- new_medj = the closest medoid to j after i is replaced by h
- prev_medj = the closest medoid to j before i is replaced by h

3.4 Euclidean Distance

The system uses Euclidean distance measure between two points:

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

Where $X=x_1, x_2, \ldots\ldots\ldots x_n$ and $y_1,y_2,\ldots\ldots\ldots y_n$ are two points. It calculates distance based on sepal length, sepal width, petal length and petal width.

3.5 The Purpose Of Studying Outliers

Outliers can provide useful information about the process. An outlier can be created by a shift in the location (mean) or in the scale (variability) of the process. Though an observation in a particular sample might be a candidate as an outlier, the process might have shifted. Sometimes, the spurious result is a gross recording error or a measurement error. Measurement systems should be shown to be capable for the process they measure. Outliers also come from incorrect specifications that are based on the wrong distributional assumptions at the time the specifications are generated [5].

4. Architecture of the System

First, perform PAM algorithm, producing a set of clusters and a set of medoids (cluster centers).

Small clusters are determined and considered as outlier clusters.

A small cluster is defined as a cluster with fewer points than half the average number of points in the k clusters.

To detect the outliers in the rest of clusters, computer the Absolute Distances between the Medoid, μ, of the current cluster and each one of **the Points, p_i,** in the same cluster (i.e., $|p_i - \mu|$). The produced value will be termed (ADMP). If the ADMP value is greater than a calculated threshold, T, then the point is considered an outlier: otherwise, it is not. The value of T is calculated as the average of all ADMP values of the same cluster multiplied by (1.5). The basic structure of the proposed method is as follows [4]:

Step 1. Perform PAM clustering algorithm to produce a set of *k clusters.*

Step 2. Determine small clusters and consider the points (objects) that belong to these clusters as outliers.

For the rest of the clusters (not determined in Step 2).

Begin

Step 3. For each cluster j, compute the $ADMP_j$ and T_j values.

Step 4. For each point i in cluster j, if $ADMP_{ij}>T_j$ then classify point i as an outlier; otherwise not.

End.

The system explains the functionalities of the system in next section.

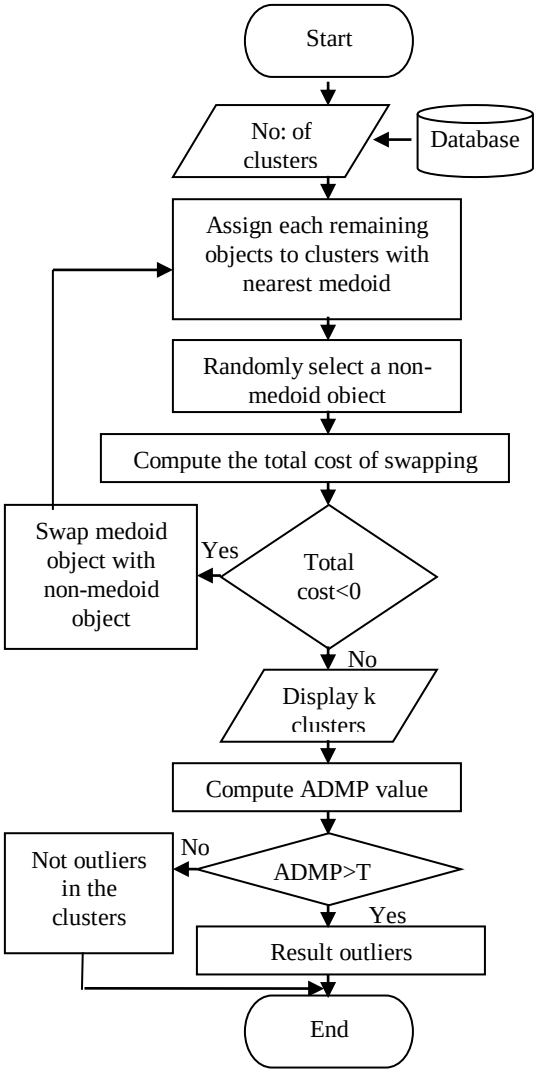


Figure 1: System Flow Diagram

4.1 Functionalities of the System

There are four processes of the system.

Process 1: Perform PAM clustering algorithm.

Process 2: Determine small clusters.

Process 3: Compute the ADMP values and T (for the rest of the clusters).

Process 4: If $ADMP_{ij} > T_j$ then classify point i as an outlier; otherwise not.

Process 1: In PAM, k partitions for n objects are formed. Initially randomly k medoids are chosen out

of set of objects. Medoid representing a cluster is most centrally located object in the cluster. Each remaining object is clustered with the medoid to which it is the most similar based on the distance between the object and medoid by using Euclidean distance. The strategy then replaces one of the medoids by one of the non-medoids as long as quality of resulting cluster is improving. This quality is estimated using a cost function that measures the average dissimilarity between an object and the medoid of its cluster.

In the iterative process a non-medoid object is randomly chosen for replacement with current medoids. Each replacement causes movement of some objects from one cluster to the other cluster. Each time a reassignment occurs a difference in square error E is contributed to the cost function. Therefore the cost function calculate the difference is square error value if a non-medoid object replaces current medoid. The total cost of swapping is the sum of costs incurred by all non-medoid objects. If the total cost is negative then replacement of medoid with non-medoid object is good since the actual square error would be reduced. The process is iterated until good replacements of medoids are found. In the end k-medoids are returned [1].

Process 2: A small cluster is defined as a cluster with fewer points than half the average number of points in the k clusters.

Process 3: Compute the Absolute Distances between the Medoid, μ, of the current cluster and each one of the points, p_i, in the same cluster (i.e, $|p_i - \mu|$). The value of T is calculated as the average of all ADMP values of the same cluster multiplied by (1.5).

Process 4: If the ADMP value is greater than a calculated threshold, T, then the point is considered an outlier, otherwise it is not.

4.2 Result Set

Table 4.1 Sample Iris Dataset

No	Sepal Length	Sepal Width	Petal Length	Petal width
1	5.4	3.9	1.7	0.4
2	4.6	3.4	1.4	0.3
3	4.8	3.4	1.6	0.2
4	4.3	3.0	1.1	0.1
5	4.9	2.4	3.3	1.0
6	6.6	2.9	4.6	1.3
7	5.0	2.0	3.5	1.0
8	5.9	3.0	4.2	1.5
9	7.6	3.0	6.6	2.1
10	4.9	2.5	4.5	1.7
11	6.5	3.2	5.1	2.0
12	6.4	2.7	5.3	1.9

Iris data consists of three species of Iris: Iris Setosa, Iris Versicolor and Iris Virginica. Iris dataset is obtained from UCI Machine Learning Repository. There are fifty plants of each species with the following four measurements on each plant: sepal length, sepal width, petal length and petal width.

Process 1: Perform PAM clustering algorithm

Table 3.2 Centroids from Iris Dataset

No:	Sepal length	Sepal width	Petal length	Petal width
1	5.4	3.9	1.7	0.4
5	4.9	2.4	3.3	1.0
9	7.6	3.0	6.6	2.1

Euclidean distance, which is defined as

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2}$$

$$d(2,1)=\sqrt{|5.4-4.6|^2+|3.9-3.4|^2+|1.7-1.4|^2+|0.4-0.3|^2}$$

$$=0.995$$

$$d(2,5)=\sqrt{|4.9-4.6|^2+|2.4-3.4|^2+|3.3-1.4|^2+|1.0-0.3|^2}$$

$$=2.3$$

$$d(2,9)=\sqrt{|7.6-4.6|^2+|3.0-3.4|^2+|6.6-1.4|^2+|2.1-0.3|^2}$$

$$=6.3$$

$$d(3,1)=\sqrt{|5.4-4.8|^2+|3.9-3.4|^2+|1.7-1.6|^2+|0.4-0.2|^2}$$

$$=0.8$$

$$d(3,5)=\sqrt{|4.9-4.8|^2+|2.4-3.4|^2+|3.3-1.6|^2+|1.0-0.2|^2}$$

$$=2.1$$

$$d(3,9)=\sqrt{|7.6-4.8|^2+|3.0-3.4|^2+|6.6-1.6|^2+|2.1-0.2|^2}$$

$$=6.1$$

$$d(4,1)=\sqrt{|5.4-4.3|^2+|3.9-3.0|^2+|1.7-1.1|^2+|0.4-0.1|^2}$$

$$=1.6$$

$$d(4,5)=\sqrt{|4.9-4.3|^2+|2.4-3.0|^2+|3.3-1.1|^2+|1.0-0.1|^2}$$

$$=2.5$$

$$d(4,9)=\sqrt{|7.6-4.3|^2+|3.0-3.0|^2+|6.6-1.1|^2+|2.1-0.1|^2}$$

$$=6.72$$

$$d(6,1) = \sqrt{|5.4-6.6|^2 + |3.9-2.9|^2 + |1.7-4.6|^2 + |0.4-1.3|^2}$$

$$= 3.4$$

$$d(6,5) = \sqrt{|4.9-6.6|^2 + |2.4-2.9|^2 + |3.3-4.6|^2 + |1.0-1.3|^2}$$

$$= 2.22$$

$$d(6,9) = \sqrt{|7.6-6.6|^2 + |3.0-2.9|^2 + |6.6-4.6|^2 + |2.1-1.3|^2}$$

$$= 2.38$$

$$d(7,1) = \sqrt{|5.4-5.0|^2 + |3.9-2.0|^2 + |1.7-3.5|^2 + |0.4-1.0|^2}$$

$$= 2.7$$

$$d(7,5) = \sqrt{|4.9-5.0|^2 + |2.4-2.0|^2 + |3.3-3.5|^2 + |1.0-1.0|^2}$$

$$= 0.5$$

$$d(7,9) = \sqrt{|7.6-5.0|^2 + |3.0-2.0|^2 + |6.6-3.5|^2 + |2.1-1.0|^2}$$

$$= 4.3$$

$$d(8,1) = \sqrt{|5.4-5.9|^2 + |3.9-3.0|^2 + |1.7-4.2|^2 + |0.4-1.5|^2}$$

$$= 2.92$$

$$d(8,5) = \sqrt{|4.9-5.9|^2 + |2.4-3.0|^2 + |3.3-4.2|^2 + |1.0-1.5|^2}$$

$$= 1.6$$

$$d(8,9) = \sqrt{|7.6-5.9|^2 + |3.0-3.0|^2 + |6.6-4.2|^2 + |2.1-1.5|^2}$$

$$= 3$$

$$d(10,1) = \sqrt{|5.4-4.9|^2 + |3.9-2.5|^2 + |1.7-4.5|^2 + |0.4-1.7|^2}$$

$$= 3.4$$

$$d(10,5) = \sqrt{|4.9-4.9|^2 + |2.4-2.5|^2 + |3.3-4.5|^2 + |1.0-1.7|^2}$$

$$= 1.4$$

$$d(10,9) = \sqrt{|7.6-4.9|^2 + |3.0-2.5|^2 + |6.6-4.5|^2 + |2.1-1.7|^2}$$

$$= 3.5$$

$$d(11,1) = \sqrt{|5.4-6.5|^2 + |3.9-3.2|^2 + |1.7-5.1|^2 + |0.4-2.0|^2}$$

$$= 3.98$$

$$d(11,5) = \sqrt{|4.9-6.5|^2 + |2.4-3.2|^2 + |3.3-5.1|^2 + |1.0-2.0|^2}$$

$$= 2.73$$

$$d(11,9) = \sqrt{|7.6-6.5|^2 + |3.0-3.2|^2 + |6.6-5.1|^2 + |2.1-2.0|^2}$$

$$= 1.9$$

$$d(12,1) = \sqrt{|5.4-6.4|^2 + |3.9-2.7|^2 + |1.7-5.3|^2 + |0.4-1.9|^2}$$

$$= 4.2$$

$$d(12,5) = \sqrt{|4.9-6.4|^2 + |2.4-2.7|^2 + |3.3-5.3|^2 + |1.0-1.9|^2}$$

$$= 2.7$$

$$d(12,9) = \sqrt{|7.6-6.4|^2 + |3.0-2.7|^2 + |6.6-5.3|^2 + |2.1-1.9|^2}$$

$$= 1.81$$

1, 5 and 9 are centroids in these clusters.

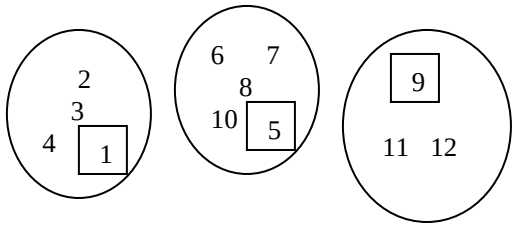


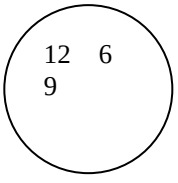
Figure 3.1 Clusters based on Euclidean distance of Iris Plants

In the iterative process a non-medoid object is randomly chosen for replacement with current medoids. Each replacement causes movement of some objects from one cluster to the other cluster. Each time a reassignment occurs a difference in square error E is contributed to the cost function. Therefore the cost function calculate the difference is square error value if a non-medoid object replaces current medoid. The total cost of swapping is the sum of costs incurred by all non-medoid objects. If the total cost is negative then replacement of medoid with non-medoid object is good since the actual square error would be reduced. The process is iterated until good replacements of medoids are found.

Table 3.3 Final Centroids from Iris Dataset

No	Sepal length	Sepal width	Petal length	Petal width
3	4.8	3.4	1.6	0.2
7	5.0	2.0	3.5	1.0
11	6.5	3.2	5.1	2.0

3, 7 and 11 are centroids in these clusters.



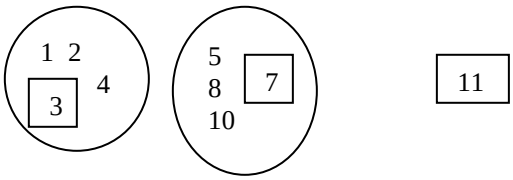


Figure 3.2 Clusters Using PAM algorithm

Process 2: There are not small clusters in this sample dataset.

For process 3 and process 4:

ADMP values in cluster with centroid 3.

$$|1-3| = |5.4-4.8| + |3.9-3.4| + |1.7-1.6| + |0.4-0.2| = 1.4$$

$$|2-3| = |4.6-4.8| + |3.4-3.4| + |1.4-1.6| + |0.3-0.2| = 0.5$$

$$|4-3| = |4.3-4.8| + |3.0-3.4| + |1.1-1.6| + |0.1-0.2| = 1.5$$

The total of ADMP values is 3.4.

Threshold $T_3 = \frac{3.4*1.5}{4} = 1.275$

$ADMP_1 > T_3$, $ADMP_4 > T_3$

So, 1 and 4 are outliers in cluster with centroid 3.

ADMP values in cluster with centroid 7.

$$|5-7| = |4.9-5.0| + |2.4-2.0| + |3.3-3.5| + |1.0-1.0| = 0.7$$

$$|8-7| = |5.9-5.0| + |3.0-2.0| + |4.2-3.5| + |1.5-1.0| = 3.1$$

$$|10-7| = |4.9-5.0| + |2.5-2.0| + |4.5-3.5| + |1.7-1.0| = 2.3$$

The total of ADMP values is 6.1.

Threshold $T_7 = \frac{6.1*1.5}{4} = 2.2875$

$ADMP_8 > T_7$, $ADMP_{10} > T_7$

So, 8 and 10 are outliers in cluster with centroid 7.

ADMP values in cluster with centroid 11.

$$|6-11| = |6.6-6.5| + |2.9-3.2| + |4.6-5.1| + |1.3-2.0| = 1.6$$

$$|9-11| = |7.6-6.5| + |3.0-3.2| + |6.6-5.1| + |2.1-2.0| = 2.9$$

$$|12-11| = |6.4-6.5| + |2.7-3.2| + |5.3-5.1| + |1.9-2.0| = 0.9$$

The total of ADMP values is 5.4.

Threshold $T_{11} = \frac{5.4*1.5}{4} = 2.025$

$ADMP_9 > T_{11}$

9 is outlier in cluster with centroid 11.

5. Conclusion

This system describes Partitioning Around Medoids (PAM) clustering algorithm which belongs to partitional clustering algorithms. Partitioning Around Medoids (PAM)_based on clustering algorithms for outlier detection is presented. We first perform the PAM clustering algorithm. Small clusters are then determined and considered as outlier clusters. A small cluster is defined as a cluster with fewer points than half the average number of points in the k clusters. The rest of outliers are then found (if any) in the remaining clusters based on calculating the absolute distances between the medoid of the current cluster and each of the points in the same cluster. PAM works effectively for small data sets, but does not scale well for large data sets. To deal with large data sets, a sampling-based method, called CLARA (Clustering LARge Applications) can be used. CLARANS has been experimentally shown to be more effective than both PAM and CLARA. The performance of CLARANS can be further improved by exploring spatial data structures, such as R*-trees, and some focusing techniques.

6. References

[1] D.K. Swami and R.C. Jain, Department of Computer Applications, "Partitioning Around Medoids for Classification".

[2] Kaufman and Rosseeuw, "Cluster Analysis Part 1", 1987.

[3] Mahesh Kumar and James B. Orlin b,"Scale-invariant Clustering with Minimum Volume Ellipsoids", 28, September, 2005.

[4] Moh'd Belal Al-Zoubi, "An Effective Clustering-Based Approach for Outlier Detection, Computer Information Systems Department, University of Jordan.

[5] Steven Walfish, "A Review of Statistical Outlier Methods", 2.Nov.2006.

[6] Svetlana Cherednichenko, "Outlier Detection In Clustering", University of Joensuu, Department of Computer Science, 24.January.2005.

[7] Blake, C. L. & C. J. Merz, 1998. UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/mlearn/MLRepository.html, University of California, Irvine, Department of Information and Computer Sciences.