Advancement in Anti-Corruption Measures: Crafting a K-means Clustering-Based Algorithm for Enhanced Detection of Corruption Patterns

Nishat Tasnim
Department of Computer Science and Engineering
University of Asia Pacific
Dhaka, Bangladesh
Email: 21201149@uap-bd.edu

Nahida Afrin
Department of Computer Science and Engineering
University of Asia Pacific
Dhaka, Bangladesh
Email: 21201011@uap-bd.edu

Samira Afsana
Department of Computer Science and Engineering
University of Asia Pacific
Dhaka, Bangladesh
Email: 21201013@uap-bd.edu

***Abstract*—Addressing the persistent issue of corruption in Bangladesh, this paper endeavors to tackle the challenge of detecting corruption, particularly those instances that directly impact the general populace originating from governmental or organizational sources. We propose a novel algorithm that integrates data mining techniques to identify corrupt practices. The core of our approach involves the development of an intelligent system, incorporating a user evaluation interface. When individuals avail themselves of services from an organization, they have the opportunity to anonymously submit their opinions regarding specific individuals or organizations. Subsequently, a refined K-means clustering algorithm is employed to analyze the data, and the output is organized to rank individuals or organizations based on their corruption levels.**

***Index Terms*—corruption; data-mining; clustering; k-means**

I. Introduction

Corruption, defined as "the abuse of entrusted power for private gain," encompasses various forms such as grand, petty, and political, depending on the magnitude of money involved and the sector affected (1). Examples of corruption include bribery, embezzlement, fraud, extortion, conflict of interests, favoritism, nepotism, cronyism, and political corruption (2). Detection of corruption is particularly crucial in setups where a service provider directly interacts with their service receiver. Certain sectors bear a disproportionate impact of corruption, including public services, land administration, tax administration, customs administration, public procurement, the judicial system, law enforcement, banking, health services, and transportation (3). This paper proposes a model incorporating a modified K-means algorithm to determine the corruption level of an organization and the individuals involved. Opinions from service receivers and employees are stored in a database. The proposed algorithm, integrating data mining techniques, facilitates the creation of distinct groups and clusters based on the corruption levels of individuals. Clustering involves identifying similarities among data based on characteristics found in raw data (4). The term "clustering" is widely used in various research communities as a technique for grouping unlabeled data (5). The clustering process is continuous and iterative, contributing to knowledge discovery from vast quantities of raw and unorganized data (6). This concept can be applied in suggestion and complaint portals (7), ratings and reviews systems in e-commerce (8), online services (9), and teaching evaluation report systems (10).

II. Proposed Model and Algorithm

A. Proposed Model to Detect Corruption

After receiving a service, the service receiver will receive login credentials via email. Subsequently, users are required to visit the evaluation portal to provide their opinions anonymously. Similarly, opinions are solicited from internal employees of the service provider. All collected opinions are stored in a database. Upon executing the proposed algorithm, the output will be generated. This result is sorted using the "Static Centroid K-means Clustering" to obtain the final result. The proposed model for detecting corruption is illustrated in Fig. 1.

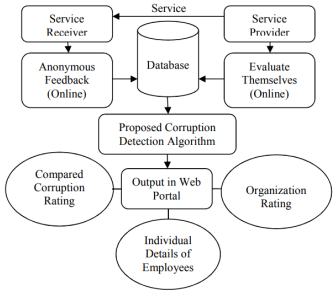

Fig. 1. Proposed Model to Detect Corruption.

B. Proposed Algorithm

The proposed *Corruption Detection Algorithm* is a combination of the *Proposed Algorithm* and *Proposed Static Centroid K-means clustering* (Fig. 5), which is a modified version of the traditional K-means clustering (11). The *Proposed Algorithm* utilizes the *Proposed Static Centroid K-means clustering* to

categorize the corrupted employees of an organization. The conventional K-means clustering algorithm has been modified compared to Hierarchical Clustering or Density-based Clustering algorithms to maintain the centroid as static. Hierarchical Clustering does not assume a specific number of clusters, and Density-based Clustering has a dynamic centroid. The proposed algorithm (Fig. 2) has been designed to detect corruption. This Corruption Detection Algorithm will be applied to both service receivers' opinions and the evaluation of employees themselves.

Input: $X = \{S_1, S_2, \ldots, S_K\}$;
Determine: *upper and lower limit of X*;
Count: *number of users* N;
Average: $f(z) = \sum_{n=0}^{k} \frac{X^n}{N}$;
if *X contains maximum or minimum value* **then**
    reject;
else
    input the value to the database;
Output:
if *50 or fewer data sets* **then**
    execute static centroid k-means;
else
    delete older date then *execute static centroid k-means*;

Fig. 2. Proposed Corruption Detection Algorithm

The flowchart of the proposed algorithm is depicted in Fig. 3.

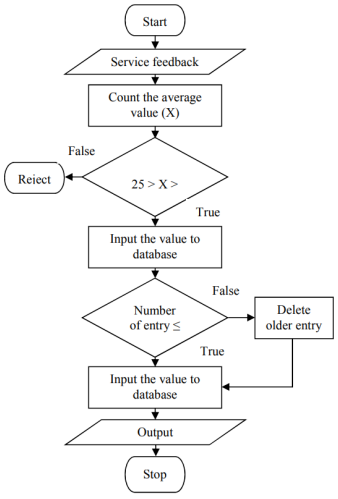

Fig. 3. Flowchart of the proposed algorithm.

When the service receiver visits the evaluation portal to provide their opinion, the algorithm starts executing through a web server. The service receiver sees only those employees' names on the evaluation interface who are directly involved in the service. To address potential knowledge gaps about indirectly involved employees, a tracking server is employed, and employees must log in with their individual credentials. The server traces the list of employees who served a specific service receiver. Consequently, the evaluation portal displays only those employees who directly provided the service. The service receiver then selects employees one by one and evaluates them using a form that includes 5 psychological statements related to honesty, professionalism, helpfulness, etc. Each statement offers 5 options (strongly disagree, disagree, neutral, agree, strongly agree), each assigned points from 1 to 5. The points range from 5 to 25, and the average points are calculated. To avoid extreme values (maximum and minimum points), the algorithm filters them.

C. Proposed Static Centroid K-means Clustering Algorithm

Static Centroid K-means clustering is almost similar to the traditional *K-means Clustering*. The traditional K-means algorithm is based on a simple idea: Given a set of initial clusters, assign each point to one of them, then each cluster center is replaced by the mean point on the respective cluster (12). For the proposed fixed centroid K-means clustering, the difference is that the centroid value is defined manually, and it will be static for all the centers. The proposed static centroid K-means clustering is shown in Fig. 4.

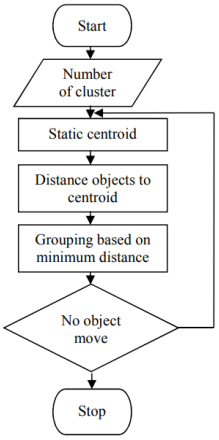

Fig. 4. Flowchart of the proposed static centroid K-means algorithm.

Here is the algorithm for the proposed static centroid K-means clustering in Fig. 5.

Select static K cluster centroids
    $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}$;
repeat
    $c^{(i)}$ = index (from 1 to K) of cluster centroid closest to $X^{(i)}$;
    for $k = 1$ *to* K **do**
        μ_K = average (mean) of points assigned to cluster k;
until *for* $i = 1$ *to* m;

Fig. 5. Proposed static centroid K-means algorithm.

III. Datasets and Experiments

A. Experiment with Test Dataset

In this section, we present an experiment conducted with a 2-dimensional dataset. The x-axis represents the sum of average points from internal employees, while the y-axis represents the sum of average points from service receivers. The objective is to detect corruption for an employee named "A." The x-values are taken from Table II, and the y-values are taken from Table I. Other employees' scores are assumed as a test dataset from Table III.

TABLE I. POINTS FROM SERVICE RECEIVER

Service Receiver	Q. 1	Q. 2	Q. 3	Q. 4	Q. 5
Person 1	4	5	3	4	2
Person 2	5	4	3	5	3
Person 3	4	5	5	5	4
Average Points	4.3	4.7	3.3	4.7	3

TABLE I

TABLE II. POINTS FROM EMPLOYEES

Employees	Q. 1	Q. 2	Q. 3	Q. 4	Q. 5
Person 1	5	5	4	4	5
Person 2	5	4	5	5	3
Person 3	4	5	4	5	4
Average Points	4.7	4.7	4.3	4.7	4

TABLE II

TABLE III. SUM OF AVERAGES

Employee Name	Sum of Average Points (From Employees)	Sum of Average Points (From Service Receiver)
A	22.3	20
B	24.3	23.4
C	18.5	17.5
D	17.3	18.2
E	10.8	7.6
F	18.6	14.3
G	20.4	19.5
H	17.2	22.8
I	20.7	21.6
J	12.7	14.3
K	15.5	16
L	16.7	17.5

TABLE III

Corruption score of employees is shown in Fig. 6 which is drawn from Table III.

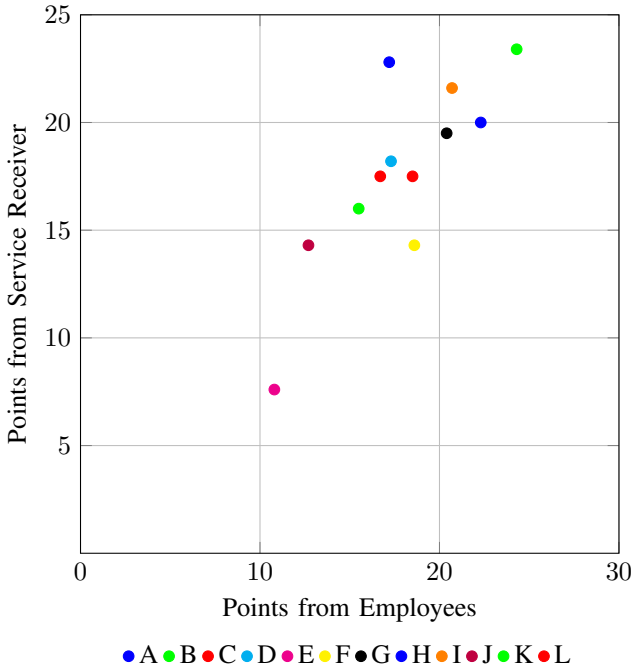

Fig. 6. Two-dimensional corruption score of employees.

Before calculating static centroid K-means clustering, we will execute traditional K-means clustering. For traditional K-means clustering, we can use the Sum of Squared Error (SSE). It calculates the Euclidean distance to the nearest centroid and computes the total sum of the squared errors. The different sets of clusters which are produced on different runs of K-means, we consider the one with the smallest squared errors (13). The SSE is defined as follows:

$$\underset{c}{argmin} \sum_{i=1}^{k} \sum_{x \in c_i} d(x, \mu_i) = \underset{c}{argmin} \sum_{i=1}^{k} \sum_{x \in c_i} \|x - \mu_i\|_2^2 \quad (1)$$

Where, c_i is the set of points that belong to cluster i. The K-means clustering uses the square of the Euclidean distance (9). After calculating, we get three clusters and the final centroids value which is shown in Table IV.

TABLE IV. INITIAL AND FINAL CENTERS

Cluster	Employee	Initial Centroid	Final Centroid
1	A,B,C,D,F,G,H,I,K,L	21,21	19.9,19.6
2	J	14.5,14.5	12.7,14.3
3	E	8.5,8.5	10.8,7.6

TABLE IV

So, from the traditional K-means clustering, we get the following three clusters (Fig. 7).

Fig. 7. Three clusters by using traditional K-means algorithm.

Now, we will execute the proposed static centroid K-means algorithm. For this algorithm, we will calculate the distance from all the values (employee's corruption score) to three static centroids. Then we will select the minimum distance for the appropriate cluster. Where, distance, D is defined as follows,

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{2}$$

According to our proposed algorithm, the corruption score (X) will be greater than 5 and less than 25, so we have considered three centroids as $C1(21, 21)$, $C2(14.5, 14.5)$, $C3(8.5, 8.5)$ to split the employees into three clusters. In case of the same distance, we have chosen the distance from the greater centroid (i.e. Score of employee D). The distance from all the employee's corruption score to three static centroids is shown in Table V.

TABLE V. DISTANCE FROM THREE CENTROID

Score of Employees	Distance From C1	Distance From C2	Distance From C3	Cluster
A(22.3, 20)	1.64	9.54	17.96	1
B(24.3, 23.4)	4.08	13.23	21.71	1
C(18.5, 17.5)	4.30	5.00	13.45	1
D(17.3, 18.2)	4.64	4.64	13.09	1
E(10.8, 7.6)	16.84	7.82	2.46	3
F(18.6, 14.3)	7.11	4.10	11.64	2
G(20.4, 19.5)	1.61	7.73	16.20	1
H(17.2, 22.8)	4.20	8.72	16.73	1
I(20.7, 21.6)	0.67	9.42	17.90	1
J(12.7, 14.3)	10.66	1.81	7.16	2
K(15.5, 16)	7.43	1.80	10.25	2
L(16.7, 17.5)	5.54	3.72	12.17	2

TABLE V

Cluster 1 is for the group of honest people. Cluster 2 is for average (less honest) persons, and Cluster 3 is for the corrupted employee's group. So, we get different three clusters from the proposed static centroid K-means algorithm, as shown in Fig. 7.

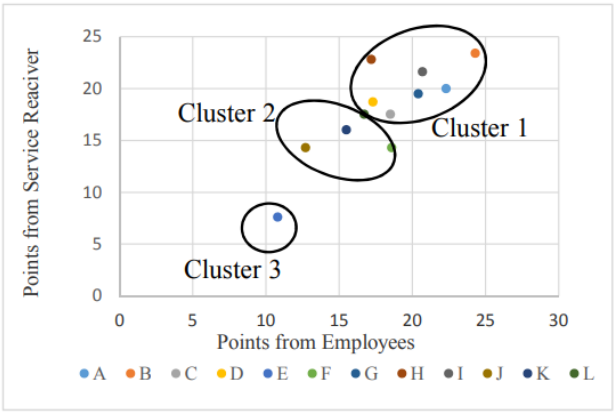

Fig. 8. Three clusters by using static centroid K-means algorithm.

B. Result Analysis

We can see from Fig. 7 that, for traditional K-means clustering where the density of data is high, a single large cluster (Cluster 1) has been created. In this situation, some lower values (i.e., F, K, and L) are in Cluster 1 as well as in the group of honest people, though they are average in corruption rating. We can solve this problem by using the proposed static centroid K-means clustering algorithm. According to this algorithm, centroids will be static. So, the lower value will not be an element of the upper cluster, and the upper value will not be an element of the lower cluster as well. We can see from Fig. 8 that the average corruption ratings are in Cluster 1, higher corruption ratings are in Cluster 2, and the lower ones are in Cluster 3 as well.

IV. Conclusion and Challenges

There is a lot of future scope for this model. The proposed model can be enhanced by adding a comment & multimedia attachment option, a public opinions field in the output page, etc. An auto mail sending option to concern organizations for a specific time duration can be developed. It is possible to generate a history graph to find the state of the corruption. The model will be capable of data synchronization for the transfer issue and synchronized with an SMS-based evaluation system. This algorithm can be applied as a digital complaint and suggestion box. Nevertheless, there are some limitations to this online-based idea. The main challenge of this system is that people need the Internet everywhere in the country. Finally, this model will be useful in society if the corrupted people are being faced with punishment. It is also needed to make awareness against corruption and its awful effect. Education from the family can be more effective to increase morality.

REFERENCES

[1] Transparency International. (2018). What is Corruption? Retrieved from https://www.transparency.org/what-is-corruption

[2] Assignment Point. (2018). Corruption in Bangladesh. Retrieved from www.assignmentpoint.com/arts/social-science/corruption-bangladesh

[3] Bliss, B. (2018). Bangladesh Corruption Report. Business Anti-Corruption Portal. Retrieved from https://www.business-anti-corruption.com/country-profiles/bangladesh/

[4] I.S. Dhillon and D.M. Modha, "Concept decompositions for large sparse text data using clustering," Machine Learning, vol. 42, issue 1, pp. 143-175, 2001.

[5] C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu, "A framework for projected clustering of high dimensional data streams" in Proceedings of the Thirtieth international conference on Very large databases-Vol 30, 2004, p. 863.

[6] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.

[7] Just.edu.jo. (2018). Complaints and suggestion system. Retrieved from http://www.just.edu.jo/Centers/HealthCenter/Pages/Complaints-and-suggestion-system.aspx

[8] Online Shipping Blog — Endicia. (2018). Customer Feedback and Why Your E-commerce Store Needs It - Online Shipping Blog — Endicia. Retrieved from https://online-shipping-blog.endicia.com/customer-feedback-and-why-your-e-commerce-store-needs-it/

[9] Team, L., Directors, B., News, P., Us, C., Programs, S. and Videos, A. (2018). How You Can Use Customer Feedback to Improve Your Business - Insightly. Retrieved from https://www.insightly.com/blog/how-you-can-use-customer-feedback-to-improve-your-business/

[10] Asd.k12.pa.us. (2018). Teacher Evaluation – Staff – Armstrong School District. Retrieved from https://www.asd.k12.pa.us/apps/pages/index.jsp?uREC_ID=417164&type=d&pREC_ID=912373

[11] P. Tan, M. Steinbach, A. Karpatne and V. Kumar, Introduction to data mining. New York, NY: Pearson Education, 2018.

[12] Shi Na, Liu Xumin, "Research on K-means Clustering Algorithm", IEEE Third International Conference on Intelligent Information Technology and Security Informatics, 2010.

[13] H. Gonçalves, "K-means clustering - algorithm and examples", Onmyphd.com, 2018. Retrieved from https://www.onmyphd.com/?p=k-means.clustering