Development of a Corruption Detection Algorithm using K-means Clustering

Conference Paper · November 2018

DOI: 10.1109/ICAEEE.2018.8642985

2 authors, including:

Mohammad Abu Yousuf
Jahangirnagar University
106 PUBLICATIONS **1,133** CITATIONS

Development of a Corruption Detection Algorithm using K-means Clustering

Md. Tawheedul Islam
Institute of Information Technology
Jahangirnagar University
Dhaka, Bangladesh
tawheed.inbox@gmail.com

Mohammad Abu Yousuf
Institute of Information Technology
Jahangirnagar University
Dhaka, Bangladesh
yousuf@juniv.edu

***Abstract*— Corruption in Bangladesh has been a continuing problem. It is arduous to detect all types of corruption but the corruption which effects directly upon the general people from the government or other organizations can be detectable. It has proposed an algorithm combining data mining technique to find corruption. In this paper, an intelligent system has been developed by forming a user evaluation interface. When a person receives a service from an organization then he/she can submit his/her opinion against the specific individual or organization anonymously. After putting their viewpoints, the algorithm having a modified K-means clustering will be executed and output will be sorted for the person of the organization according to their corruption level.**

Keywords—corruption; data-mining; clustering; k-means

I. INTRODUCTION

Corruption means "the abuse of entrusted power for private gain". It is classified as grand, petty and political, depending on the amounts of money and the sector where it occurs [1]. There are many forms of corruption, like: Bribery, Embezzlement and Fraud, Extortion, Conflict of Interests, Favoritism, Nepotism, Cronyism, Political Corruption etc. [2]. From the above forms of corruption, we can detect corruption and the corrupted person as well in which set up a service provider directly deals with his service receiver. There are some sectors where corruption is immensely affected. Public Services, Land Administration, Tax Administration, Customs Administration, Public Procurement, Judicial System, Law Enforcement, Banking Sector, Health Service, Transportation are most corrupted sectors [3]. A model has been proposed in this paper including a modified K-means algorithm which will determine the corruption level of an organization and the corrupted person. The opinions of service receivers and employees of an organization will be stored in a database. The proposed algorithm including data mining technique will help to create some different group as well as cluster according to corruption level of the corrupted people. The clustering is performed by finding similarities among data according to characteristics found in raw data [4]. The term "clustering" is commonly used in several research communities as a technique for grouping of unlabeled data [5]. Actually, clustering process is a continuous and an iterative process of knowledge discovery from enormous quantities of raw and unorganized data [6]. Suggestion and complaint portal [7], ratings and reviews systems in e-commerce [8] or online services [9], teaching evaluation report system [10] are also can exploit this idea.

II. PROPOSED MODEL AND ALGORITHM

A. Proposed Model to Detect Corruption

After receiving a service, the service receiver will get a login credential through e-mail. Later the user requires to visit the evaluation portal to provide his/her opinion anonymously. Similarly, the opinion has been taken from employees of the service provider internally. All the opinion has been stored in a database. After running the proposed algorithm, the output will appear. This result has been sorted by "Static Centroid K-means Clustering" as a final result. The proposed model to detect corruption is shown in Fig. 1.

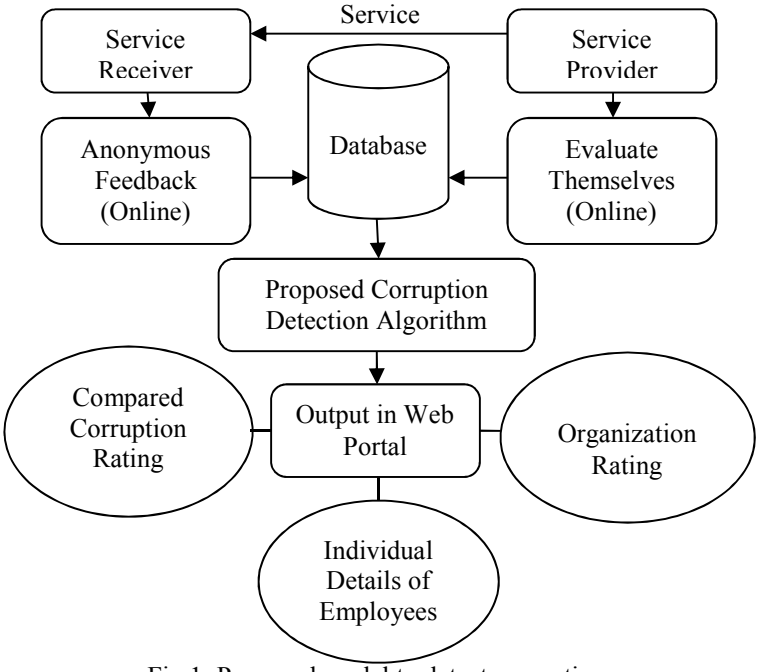

Fig 1: Proposed model to detect corruption

B. Proposed Algorithm

The proposed "Corruption Detection Algorithm" is a combination of a "Proposed Algorithm" and "Proposed Static Centroid K-means clustering" (Fig. 5) which is a modified version of traditional K-means clustering [11]. Actually, the "Proposed Algorithm" will use the "Proposed Static Centroid K-means clustering" to sort the corrupted employees of an organization. The conventional K-means clustering algorithm has been modified than Hierarchical Clustering or Density-

978-1-5386-8252-4/18/$31.00 ©2018 IEEE

based Clustering algorithm to make the centroid static. It is not possible to assume any specific number of clusters in Hierarchical Clustering and the Density-based Clustering has dynamic centroid. The following algorithm (Fig. 2) has been proposed to detect corruption. This Corruption Detection Algorithm will be used for both service receiver's opinions and evaluation of employees themselves.

1: Input: X= $\{s_1, s_2, \dots \dots \dots \dots \dots \dots, s_K\}$
2: Determine: *upper and lower limit of* X
3: Count: *number of user*= N
4: Average: $f(z) = \sum_{k=0}^{n} \frac{X^{(n)}}{N}$
5: if X *contain maximum or minimum value*
6: then *reject*
7: else *input the value to database*
8: Output: if 50 *or less data set*
9: then *execute static centroid k-means*
10: 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,

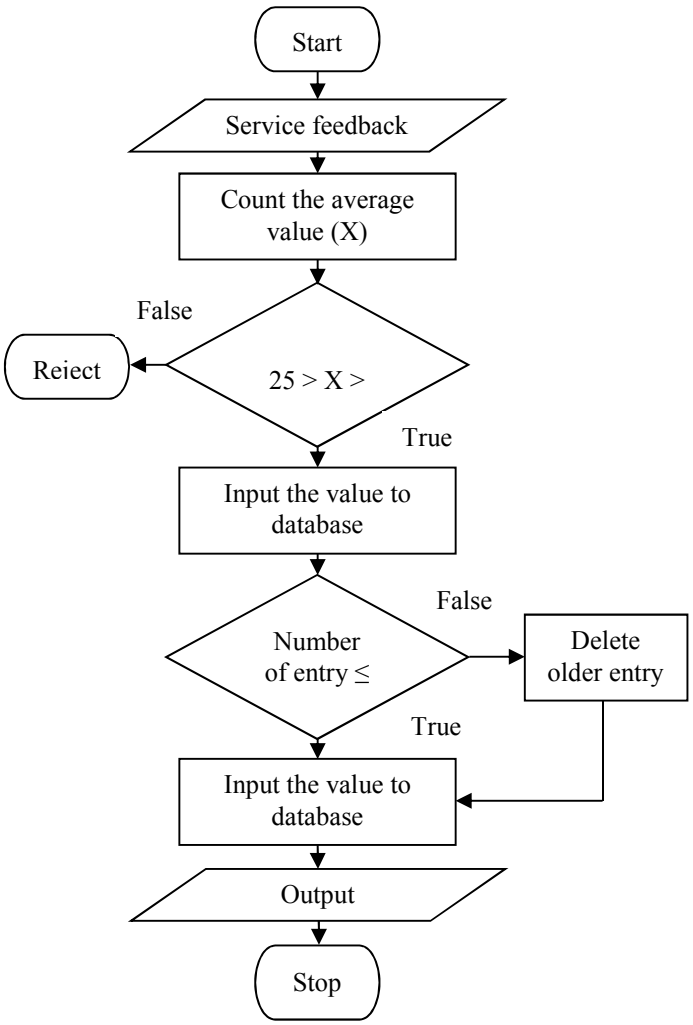

Fig 3: Flowchart of the proposed algorithm

When the service receiver will visit the evaluation portal to provide his/her opinion, the algorithm will start to execute through a web server. The service receiver will see only those employees names on the evaluation interface who are directly involved in the service. The service receiver might have less idea about those employees who served indirectly or who even didn't serve at all. So, there will have to a tracking server and the employees have to login with their individual credential. The server will trace the employee list who served a specific service receiver. So, the evaluation portal will show only those employees list who have directly provided the service to the service receiver. Then he/she will select the employee one by one and have a form to evaluate the employees. In this algorithm, there will have 5 psychological statements to evaluate the employee's honesty, professionalism, helpfulness etc. The service receiver will see 5 options (strongly disagree, disagree, neutral, agree, strongly agree) against every statement. The options have provided points from 1 to 5. So, the points range will be 5 to 25 then average points will be calculated. It is unusual for a person to get maximum or minimum points in all statements. So, to avoid the outlying values (maximum and minimum points) the algorithm will filter it.

C. Proposed Static Centroid K-means Clustering Algorithm

Static Centroid K-means clustering almost similar with 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, 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,

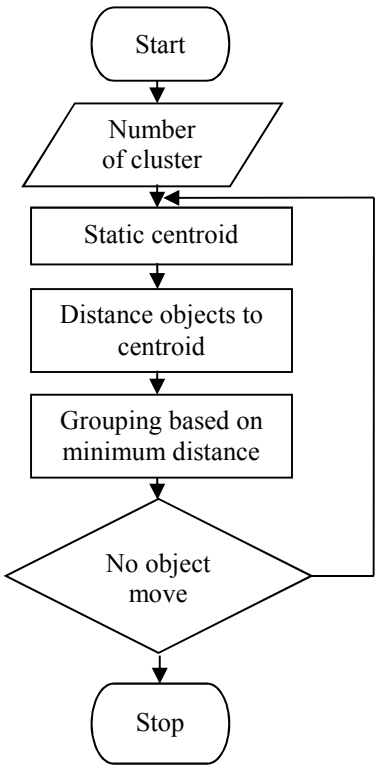

Fig 4: Flowchart of the proposed static centroid K-means algorithm

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

1: Select static K cluster centroids
$\mu_1, \mu_2, \ldots \ldots \ldots \ldots \ldots \ldots, \mu_K \in R$
2: Repeat {for i = 1 to m
4: $c^{(i)}$ = index (from 1 to K) of cluster centroid
5: closest to $x^{(i)}$
6: for k = 1 to K
7: μ_K = average (mean) of points assigned to cluster k}

Fig 5: Proposed static centroid K-means algorithm

III. Datasets and Experiments

A. Experiment with Test Dataset

Here, we have plotted 2-dimensional dataset. Sum of the average points from internal employees are on the x-axis and the sum of the average points from service receivers are on the y-axis. To detect corruption for an employee named "A", we supposed x value from Table II and y value from Table I. Assumed the other employee's scores as a test data set 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 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 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

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

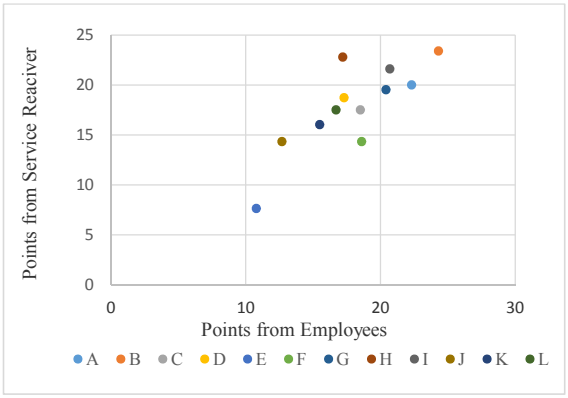

Fig 6: Two dimensional corruption score of employees

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,

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

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

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

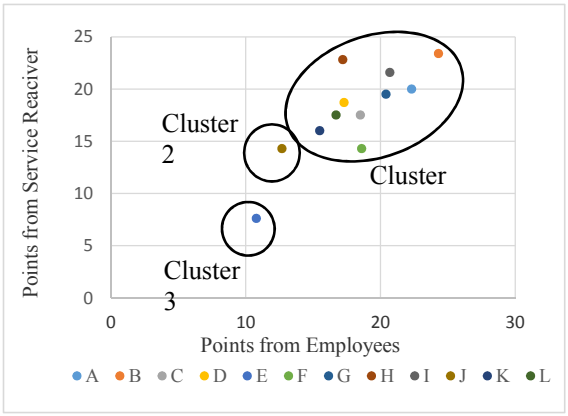

Fig 7: Three clusters by using traditional K-means algorithm

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,

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

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 are 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

Cluster 1 is for the group of honest people. Cluster 2 is for average (less honest) person and the Cluster 3 is for the corrupted employee's group. So, we get different three clusters from the proposed static centroid K-means algorithm which is shown in Fig. 8.

Fig 8: Three clusters by using static centroid K-means

B. *Result Analysis*

We can see from fig 7, for traditional K-means clustering where the density of data is high, a single large cluster (Cluster1) 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, 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 lowers are in Cluster 3 as well.

IV. CONCLUSION AND CHALLENGES

There is a lot of future scope of this model. The proposed model can be enhanced by adding comment & multimedia attachment option, public opinions field in output page etc. Auto mail sending option to concern organizations for 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 to data synchronization for the transfer issue and synchronized with an SMS based evaluation system. This algorithm can be applied as a digital complain and suggestion box. Nevertheless, there is some limitation to this online based idea. The main challenge of this system is, 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 family can be more effective to increase the morality.

REFERENCES

[1] e.V., T. (2018). *Transparency International - What is Corruption?* Transparency.org. Available at: https:/www.transparency.org/ what-is-corruption.

[2] Assignment Point. (2018). *Corruption in Bangladesh.* Available at: www.assignmentpoint.com/arts/social-science/corruption-bangladesh

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

[4] I.S. Dhillon and D.M. Modha, *"Concept decompositions for largesparse 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, "Aframework for projected clustering of high dimensional datastreams" in *Proceedings of the Thirtieth internationalconference 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*. Available at: 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*. Available at: 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*. Insightly. Available at: 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.* Available at: https://www.asd.k12.pa.us/apps/pages/ index.jsp?uREC_ID=417575&type=d&pREC_ID=1016663

[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. Available https://www.onmyphd.com/?p=k-means.clustering