I J C T A, 9(28) 2016, pp. 217-223

MULTI-OBJECTIVE CLUSTERING AND OPTIMIZATION

Sheik Faritha Begum* A. Rajesh and K. P. Kaliyamurthie*****

Abstract: This paper discusses the optimization of clustering over ill-structured datasets. The dataset used in this experiment has derived from the measures of sensors used in a urban waste water treatment plant. The objective of this experiment is to categorize the quality of water from the treatment plant. Water Quality is indicated by the following parameters: Suspended Solids(SS),Biological Oxygen Demand(DBO),Chemical Oxygen Demand(DQO), Sediment(SED),Conductivity(COND). In this paper K-means algorithm is used for clustering the data set. The optimization of clustering is discussed in terms of number of clusters which results in a natural categorization of the dataset based on the quality parameters of the water. K-Means was chosen for its simplicity and efficiency. Moreover K-Means algorithm is the best suitable algorithm for clustering numerical datasets.

Keywords: K-Means, Optimization, Clustering, Categorization, Water Quality Assessment, Water Quality parameters.

1. INTRODUCTION

Many Researches have been done so far to detect and diagnosis faults in simulated waste water treatment plant.Many statistical techniques have been developed to extract process information from large amount of data. But these type of plants are varying in the process load due to their nature causes like rain and temperature change. So theoretical modeling has become a challenging task over period of time. Also interpreting the measures obtained by these methods attracts major attention in research. This paper addresses the interpretation of results which in turn helps to analyze the working condition of waste water treatment plant. Processing waste-water treatment is completely nonlinear,hence to be treated in a nonlinear way.

2. RIVER POLLUTION INDEX (RPI):

Assessment of river quality currently by E.P.A is purely based on a index known as "River Pollution Index"(RPI). RPI is an indicator used for determining the level of pollution of a river. Usually this index value is calculated by identifying the concentration of four water quality parameters: Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD5), Suspended Solids (SS) and Ammonia Nitrogen (NH3-N). The comparison baselines and level of pollution using RPI are shown below:

Water Quality/Item	*Non(Slightly)- polluted*	*Lightly-polluted*	*Moderately-polluted*	*Severely-polluted*
Dissolved Oxygen(DO)mg/L	DO ≥ 6.5	6.5 > DO ≥ 4.6	4.5≥DO≥2.0	DO<2.0
Biochemical Oxygen Demand (BOD5)mg/L	BOD5 ≤ 3.0	3.0 < BOD5 ≤ 4.9	5.0≤BOD5≤15.0	BOD5>15.0
Suspended Solids(SS) mg/L	SS ≤ 20.0	20.0 < SS ≤ 49.9	50.0≤SS≤100	SS>100
Ammonia Nitrogen(NH3-N)mg/L	NH3-N ≤ 0.50	0.50<NH3-N≤0.99	1.00 ≤ NH3-N ≤ 3.00	NH3-N>3.00
Point Scores	1	3	6	10

* Research Scholar,Department of Computer Science and Engineering, Bharath University, Chennai, India. Email: sfaritha@gmail.com

** Professor, Department of Computer Science and Engineering, C. Abdul Hakeem College of Engineering and Technology, Vellore, India. **Email:** amrajesh73@gmail.com

*** Professor, Department of Computer Science and Engineering, Bharath University, Chennai, Tamil Nadu, India. Email: sfaritha@gmail.com

The most powerful task in data mining process is clustering which is used to discover interesting patterns in underlying data. The problem of clustering is defined as a partition of data into group of patterns called clusters where data points in each cluster is more similar to each other and the points in the different cluster are dissimilar to each other. The basic functionality of clustering process is to explore the patterns into meaningful groups so that we can derive some useful interpretations and can make some conclusions over them. Nowadays clustering is applied in many fields which includes medical sciences, engineering sciences and life sciences. Although the primary objective of clustering is grouping of patterns, optimizing the results to fix some criteria becomes major issue. So in this paper clustering has been posed as optimization problem. Generally the parameters considered for optimization during clustering is the cluster characteristics like connectivity,separation between clusters, and compactness within cluster. Also a general approach for posing clustering as an optimization problem is to use some internal and external cluster validity indices to optimize the clustering results,hence to prove how good the generated cluster was?

2. LITERATURE SURVEY:

In paper[6] the authors experimented K-means clustering over credit approval and soybean data sets.The experimental results proved that k-means clustering works efficiently over very large datasets. Additionally the same experiment to prove the above statement has also been done on two real world data sets with half a million objects applications. In paper[7] the authors proposed a model which incorporates multi-layered network architecture along with back propagation learning mechanism. This model consists of two phases. In phase1 in order to reduce the number of samples to be supplied to second phase, K-means algorithm is applied to the training set. In phase2 to the neural network runs automatically by selecting the optimal set of samples.In paper[8] the authrs used new initialization method is to select well separated initial points which have the potential of forming high-quality clusters. The experimental results show that improved K-means algorithm generates the clusters having high structural similarity which in turn produces the high percentage of sequence segments.In paper[9], the authors introduced a term called *Centroid Ratio* to compare results of two clusterings. In order to avoid local optima problem of k-means they have used. This centroid ratio is then used in prototype-based clustering by introducing a *Pairwise Random Swap* clustering algorithm The swap strategy in the algorithm alternates between simple perturbation to the solution and convergence toward the nearest optimum by *k*-means. The centroid ratio is shown to be highly correlated to the mean square error (MSE) and other external indices. Moreover, it is fast and simple to calculate In paper[9] the authors proposed KMSVM algorithm which combines K-means clustering method along with Support Vector Machine(SVM) by requiring additional input parameter to be determined named number of cluster. This paper discusses the strategy to identify the determination of input parameters.

In paper[10],the authors integrated K-means clustering along with nature inspired optimization algorithms to generate global optima. Enhancing the cluster quality by finding global optima, unprecented performance have been produced by anticipating new hybrids. Experimentation has been applied for image segmentation. Also the algorithm speeds up the clustering process by doing convergence of global optimal along with many search agents.. In paper[11],the aimed at presenting an algorithm which improves Kmeans clustering with partitioning accuracy and better time complexity. They have reduced the number of objects needed for examining similarity each during iteration. Spatial data structure called range tree is used in this approach. n this paper time complexity depends on number of objects examined, number of iterations and number of clusters..

In paper[12] In this paper the authors insist that there is no need for re-distribution of data elements for every iterstion since only few data elements are changing their cluster node during clustering process. The

authors have introduced optimized version of traditional Kmeans algorithm to optimize the running of an algorithm..Experimental results shows that the propoased algorithm reduces 70% of total running time.

3. CLUSTERING OF WASTE WATER TREATMENT PATTERNS USING K-MEANS:

In existing systems clustering has been performed on structured data sets. In this model we have performed clustering in ill-Structured datasets.

3.1 Dataset Description:

This dataset[13] derived from the daily measures of sensors in a waste water treatment plant(WWTP). Actually The objective of deriving this dataset is to classify the operational state of the plant to predict faults at each of the stages of the treatment process. This domain has been stated as an ill-structured domain.

- Number of instances: 527
- Number of Attributes: 07
- Attribute Information:
- All attributes are numeric and continuous
- N. Attrib.

1. PH-S (output pH)
2. DBO-S (output Biological demand of oxygen)
3. DQO-S (output chemical demand of oxygen)
4. SS-S (output suspended solids)
5. SSV-S (output volatile suspended solids)
6. SED-S (output sediments)
7. COND-S (output conductivity)

3.2 Architecture

The daily measures of sensors in a waste water treatment plant is retrieved and loaded as dataset. Data preprocessing is done to remove missing values. Also the attributes other than numerical are filtered. K-means clustering is done over the pre-processed dataset in order to obtain the centroid clustered model and clustered set by fixing the parameters which includes number of clusters, similarity measures and maximum runs.The cluster model possess each individual clusters along with its items.It also possess centroid value for each attributes for all clusters. In order to measure the performance, the ratio of intra cluster distance and inter cluster distance is measured and recorded as DBvalue.

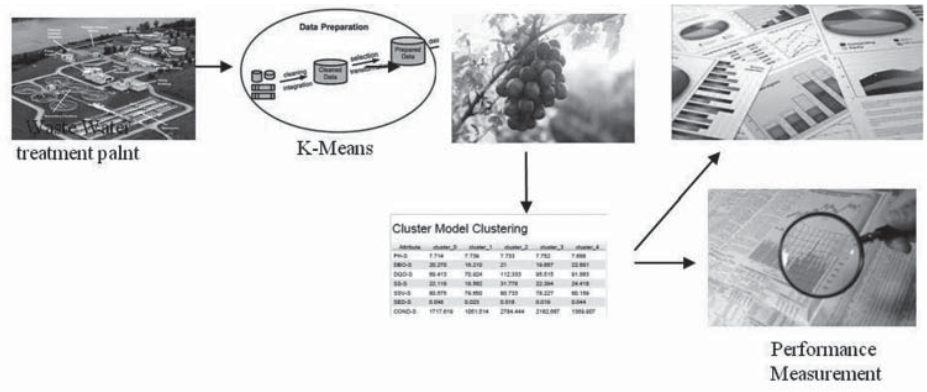

Figure 1: Architecture diagram for K-means clustering applied over waste water treatment patterns

3.3 K-Means

K-means algorithm assigns the n number of data objects into K clusters with the target of minimizing the sum of squared distance of cluster center.Initially k number of cluster centers are randomly picked and on each iteration each data objects are assigned to the cluster whose mean is nearer to it in terms of Euclidean distance. The mean of newly formed cluster is calculated and reassigning of data objects in accordance with the calculated mean has been done.The above process is repeated until none of the data object has to be reassigned.

By applying the above described methodology in different versions namy algorithms have been proposed in literature. But none of the algorithms produce the optimal result by running Kmeans at single time.The only way to produce the optimal result is to run Kmeans clustering with varying number of cluster centers and obtaining the best.

Simply K-means is a well known unsupervised clustering algorithms solving common clustering problem. The primary goal is to fix a centroid individually for K number of cluster. Fixing up of centroids receives more attention since different centroids yields different results. So, the better idea is to select centroids for k number of clusters which is far away from each other.Next assigning the data objects or points nearer to the centroid is taken place. After all the points assigned to its cluster and none of the points left,the initial phase stops.Now recalculation or fixing up of new centroid is done based on newly formed clusters. Again calling the previous step, reassignment of data objects closer to the newly formed centroid is done.This process is repeated until when it is not possible to fix up new centroids or all data objects are assigned to the same clusters in which resides previously.In oreder to obtain optimization, we need to optimize some objective function.In this paper we tried to minimize a objective function called squared error function. The objective function[15]

$$J = \sum_{j=1}^{k} \sum_{i=1}^{x} \left\| x_i^{(j)} - c_j \right\|^2 \tag{1}$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a distance measure between a data point $x_i^{(j)}$ and the cluster centroid.

The algorithm is composed of the following steps:

1. Fix K centroids representing from the data objects which are being clustered.
2. Assign each data object to the cluster whose centroid is closer to the object.
3. When no objects left to assign, fixing up of new K centroids is done based on data objects assign.
4. Repeat Step2 and Step3 until the newly formed centroid is same as previous one. After this achievement of optimization has done by using any objective function.

Pseudocode

The following pseudocode is defined in [16]

Algorithm 3.1 (k-means–)

Input: set of points X = {*x*1,.....*xn*}

A distance function *d* : X *x* X → □

Numbers *k* and l

Output: A set of *k* clusters centers C

C0 ← {*k* random points of X}

i ← 1

while (no convergence achieved) do

Compute *d*(*x* | Ci-1), for all *x* € X

Re-order the points in X such that

d(*x*1 | C*i* – 1) ≥.....≥ *d*(*xn* | C*i* – 1)

Li ← {*x*1,.....*x*l)

X*i* ← X \ L*i* = { *x*l + 1,.....,*xn*}

for (*j* €{ 1,...*k*}) do

P*j* ← {*x* € X*i* | *c*(*x* | C*i* – 1) = *ci* – 1, *j*}

ci,*j* ← mean(P*j*)

C*i* ← {*ci*, 1 ,....., *ci*, *k*}

i ← *i* + 1

Aiming to minimize the objective function, it is not guarantee that the algorithm stops everytime only after it achieves the optimization, since the cluster centroids are selected randomly.So in order to achieve optimization the K-means clustering algorithm has to be run multiple times by selecting different initial centroids.

3.4 Measures considered to fix optimal number of clusters:

Davies-Bouldin index: This index, DB, is defined in [17] as:

$$DB = \frac{1}{n}\sum_{i=1,i\neq j}^{n} \max\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \tag{2}$$

where the average of distance between the all data objects reside in cluster i to its cluster center is denoted as $\sigma_{i,}$ average of distance between the all data objects reside in cluster j to its cluster center is denoted as σ_j and number of cluster is denoted as *n*. $d(c_i, c_j)$ is calculated as distance between the two cluster centers. Minimum value of DB implies that the cluster is more compact. Also the number of clusters which gives the minimum DB is considered as the optimal number of clusters.

Average within cluster distance: At first the performance for all clusters are calculated . The average distance between points in a cluster is multiplied by the number of points is calculated by the cluster density performance operator. Euclidean distance is used as the distance measure. So a cluster containing 1 point would have an average distance of 0 because there no other points. Density should be equal to the distance between the points in a cluster which contains 2 points. The negative value is imposed by RapidMiner. Smallest absolute value is termed as best density and by negating this will lead to stop the process of optimization. The performance for all clusters is calculated by summing each cluster performance weighted by the number of points in each and dividing by the number of examples.

4. RESULTS:

4.1 K-Means: Results and Discussion:

K-means clustering is an exclusive clustering algorithm. In this type of clustering each object will reside in only one cluster and the similarity between the objects reside in a single cluster is maximum. The above

said similarity is calculated based on distance between the objects. Here is a simple explanation of how the *k*-means algorithm works. First, a term named "centroid" is introduced which is also called as the center of the cluster.Using Euclidean distance as a measure one can define the centroid of a cluster is a point whose attribute values are the average of the values of the corresponding attribute for all the data points in the cluster. By deciding how many clusters we would like to form, the algorithm starts. This value is *k*. Generally the value of *k* is a small integer, such as 2, 3, 4 or 5, but may be larger.

Using *k* as the initial set of centroids, assign the objects to the cluster which is nearer to the corresponding centroids. After assigning all objects, recalculate the centroid value and reassign the objects to the nearest centroid. Iterate this process until changes needed in assigning the objects.

Electrical Conductivity (COND) µmhos/cm:

Distilled water	: 0.5-3
River	: 500-1500
Inland Fresh water & Fisheries	: 150-1500
Irrigation purpose	: > 2000

When K = 5 :

In cluster 0: COND = 1717.619(Irrigation)

In cluster 1: COND = 1051.514(River)

In cluster 2: COND = 2784.444(Irrigation)

In cluster 3: COND = 2182.667(Irrigation)

In cluster 4: COND = 1369.907(River)

When K=3:

In cluster 0: COND = 1619.590(Irrigation)

In cluster 1: COND = 2338.974(Irrigation)

In cluster 2: COND = 1174.301(River)

4.2 Validation measures

Table 2.
DB index and Average within cluster distance for varying number of clusters

Indices	K = 3	K = 4	K = 5	K = 6	K = 7
DB-Index	– 0.589	– 0.564	– 0.599	– 0.651	– 0.519
Average within cluster distance	– 37441.949	– 18793.664	– 17636.211	– 11161.585	– 11094.728

- According to DB index value, the three optimal number of clusters are 6,5,3
- According to Average within cluster distance, the three optimal number of clusters are 3,4,5

By comparing DB Value and Average within cluster distance from above table, the optimal number of cluster for the given data set is concluded as 3. The clusters has been formed based on electrical conductivity of waste water from the treatment plant and the water in clusters are categorized into river water and irrigation water.

In this paper we have used K-means clustering to cluster the waste water treatment datasets. DB index and Average within cluster distance are used as fitness value to fix the optimal number of clusters and electrical conductivity of the waste water is considered to categorize the clusters. In future other clustering algorithms can be used along with different validity measures. Also different parameters of the waste water will be considered for categorization.

References

1. U. Maulik and S. Bandyopadhyay, "Genetic algorithm based clustering technique," *Pattern Recognition.*, vol. 33, pp. 1455–1465, 2000
2. Chen, J., Liao, C. M., "Dynamic process fault monitoring based on neural network and PCA".Jour. of Process Control, vol. 12, pp. 277- 289, 2002
3. Johnson, R. A., Wichern, D. W., Applied Multivariate Statistical Analysis, Prentice-Hall, New Jersey, 1992
4. Daszykowski, M., Walczak, B., Massart, D. L., "Projection methods in chemistry". Chemometrics and Intelligent Laboratory Systems, vol. 65, pp. 97-112, 2003.
5. Vreˇcko, D., Hvala, N., Kocijan, J., Zec, M. "System analysis for optimal control of a wastewater treatment benchmark". Water sci. technol., vol. 43, pp. 199-206, 2001.
6. Zhexue Huang,"Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", ACSys CRC, CSIRO Mathematical and Information Sciences, GPO Box 664, Canberra, ACT 2601, Australia.
7. K. M. Faraoun and A. Boukelif,"Neural Networks Learning Improvement using the K-Means Clustering Algorithm to Detect Network Intrusions", International Journal of Computational Intelligence Volume 3.
8. Wei Zhong, Gulsah Altun, Robert Harrison, Phang C. Tai, and Yi Pan*, Senior Member," Improved K-Means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property",IEEE Transactions On Nanobioscience, Vol. 4, No. 3, September 2005.
9. Jiaqi Wang,Xindong Wu, Chengqi Zhang," Support vector machines based on K-means clustering for real- time business intelligence systems",Int. J. Business Intelligence and Data Mining, Vol. 1, No. 1, *2005*
10. Simon Fong,1 Suash Deb,2 Xin-She Yang,3 and Yan Zhuang1," Towards Enhancement of Performance of K- Means Clustering Using Nature-Inspired Optimization Algorithms", Hindawi Publishing Corporation, Scientific World Journal, Volume 2014, Article ID 564829, 16 pages.http://dx.doi.org/10.1155/2014/564829
11. M. N. Vrahatisand B. Boutsinas, P. Alevizos," The New *k*-Windows Algorithm for Improving the *k*-Means Clustering Algorithm", journal of complexity **18**, 375–391 (2002) doi:10.1006/jcom.2001.0633, available online at http://www.idealibrary.com.
12. Cosmin Marian Poteras, Marian Cristian Mihaescu, Mihai Mocanu Ahmed E. Hassan," An Optimized Version of the K-Means Clustering Algorithm, Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 695–699
13. Bache K,Lichman M.UCI Machine Learning Repository, University of California, School of Information and Computer Science: Irvine, CA, Available from:http://archive.ics.uci.edu/ml.Date Accessed:11.12.2013)
14. Alignment Hiroshi, Sawada Shoko Araki, Shoji Makino,"Underdetermined Convolutive Blind Source Separation via Frequency Bin-wise Clustering and Permutation ", Senior Member, IEEE, IEEE International Symposium on Circuits and Systems (ISCAS 2007)
15. Barileé, Barisi Baridam,"More Work on *K* -Means Clustering Algorithm: The Dimension-ality Problem ",International Journal of Computer Applications (0975 – 8887) Volume 44– No.2, April 2012.
16. Sanjay Chawla, Aristides, Gionisy," k-means: A unified approach to clustering and outlier detection",,http://www.siam.org/journals/ojsa.php
17. D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE Transaction on Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979