IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 2, February 2020

A Novel Approach for Constraint Selection using Genetic Algorithm

Mandana Gholamigazafrudy[1]

Tehran, Iran[1]

Abstract: K-means algorithm has simple implementation and high speed. However, this algorithm is not capable of using the side information. Since this deficiency is effective on the performance of the algorithm, some improvements like Cop-k-means and CVQE algorithms have been designed. However, they encounter with another shortage which is equalizing the importance of constraints. To resolve this problem, the present paper proposes a mechanism to select the constraints of a data set through the use of CVQE clustering algorithm and the Imperialist Competitive Algorithm. The improvement measure in this method is the minimum constraint violation, reduction in the inter-cluster distances, and increase the distance between separate clusters. Thus, Davies Bouldin index is utilized to compare the results of the proposed algorithm with those of Cop-k-means, and CVQE algorithms. After clustering four data sets using these algorithms, the results proved that the proposed algorithm performs successfully in improving the constrained clustering.

Keywords: Clustering, Constrained Clustering, Constraint Selection

I. INTRODUCTION

Machine learning is derived from sciences such as pattern recognition and computational learning theory. The main purpose of machine learning is for the system to be able to perform the modeling operation based on the input data and present its prediction as the output, based on the created model. The machine learning methods are divided into three categories, including supervised learning, unsupervised learning, and semi-supervised learning. In supervised learning, which consists of clustering methods, all of the data are labeled, and a set of input and output pairs are given to the system so that the system can make a model from input to output. Factors such as a large number of data, high costs of data labeling, unavailability of the labels, limited time for data labeling, and many other factors prevent the widespread use of supervised algorithms for solving many real-world problems [1].

Clustering is a method to recognize large patterns and unlabeled datasets and is able to widely categorize object collections. Clustering algorithms are very useful and have applications in fields such as machine learning, data mining [2], image processing, text processing, web mining, pattern recognition, economy, medical, banking, and countless issues in today's world [3]. Various methods have been proposed for clustering, including hierarchical, centroid-based, density-based, and grid-based clustering [4]. However, the unsupervised nature of the clustering and also the algorithm unawareness of the data types and clustering purposes have caused these algorithms to not perform appropriately in some cases. In fact, it is very hard to find a comprehensive clustering algorithm that is capable of handling all of the issues [2].

On the other hand, additional information about the clusters is available in many cases. This additional information could be the label of the classes for a subset of objects, or additional information about the similarity of object pairs, or the user settings for the manner of sample categorization. The semi-supervised field has grown to find a way of matching the available additional information [5]. The additional information about the datasets is given to the algorithm as constraints. The constraints are added to avoid empty clusters or clusters with low member numbers [6]. Constrained clustering is a semi-supervised method to increase the accuracy of the clustering algorithm [7].

Semi-supervised learning can be considered a combination of supervised learning and unsupervised learning algorithms. As labeled data is usually not available in sufficient quantities and is costly to provide, semi-supervised learning helps the system provide more accurate results by employing unlabeled data that is easier to access and combining it with the hidden knowledge in the limited-access labeled data.

In semi-supervised clustering, constraints are used to obtain a high-quality result [3], [4], [5] or to partition the dataset so that it reflects the user's perception of the data [6]. However, to the same extent that the existence of appropriate constraints can be useful in improving clustering results, inadequate constraints can disrupt the performance of clustering algorithms. One of the most important problems concerning the constrained clustering is the selection of useful constraints.

Nonetheless, since the constraint selection problem is a Non-deterministic Polynomial-time (NP) optimization problem [8], an exact solution has yet to be defined for constraint selection problems and meta-heuristic algorithms are used to

solve them. A proper approach for constraint selection is to utilize evolutionary algorithms because they are very strong and powerful in solving optimization problems.
The main objective of this paper is constraints selection using an evolutionary optimization algorithm named Imperialist Competitive Algorithm (ICA) to improve constrained clustering. To achieve this, the improved version of Cop-k-means constrained clustering algorithm, known as Constrained Vector Quantization Error (CVQE) that can use side information and data labels is used for dataset clustering.

II. RELATED WORKS

In 2007, DereK Greene et al. [9] attempted to find limitations in the dataset that would have the most significant effect on guiding the clustering algorithm to find the more accurate solutions. In their method, they eliminated the constraints that fail to improve clustering.
Besides, in 2007, Yi Hong et al. [10] proposed an algorithm called Uncertainty based Assignment order Learning Algorithm (UALA) to tackle the sensitivity problem of constrained clustering algorithms. The suggested algorithm ranks all dataset samples based on clustering uncertainty, which is calculated using multiple clustering algorithms.
In 2008, Kumar et al. [11] presented a new approach. The proposed method that is based on active learning basis, the Farthest First Query Selection (FFQS) algorithm was used for clustering. The algorithm consists of two steps: exploration and consolidation. In the exploration step, the main body of the clusters is formed by cannot-link constraints, and in the consolidation step, points that are not located in the main body are selected randomly, and the user is asked for each of the points within the clusters to achieve a must-link constraint. The major challenge with this approach is the direct involvement of human resources with the algorithm. Although this practice can create more accurate clusters due to the use of expert knowledge, it is very costly and increases the clustering time. Human error is added to the clustering process as well.
In 2009, Carlos Ruiz et al. [12] proposed a method for automatic constraint selection for a semi-supervised clustering algorithm that, unlike many previous methods, performs the constraint selection process automatically. This algorithm receives a set of labeled data as input and ranks the data based on their application. It then selects the constraints from the top-ranked samples.
Viet-Vu Vu et al. [13] proposed a method to improve the FFQS algorithm in 2010. They weighted the data points using the K- Nearest Neighbors Graph (K-NNG) and determined the candidate data points to query from the user based on the weight of each data point and the strong paths between that point and the other points. Furthermore, the authors succeeded in developing an algorithm that can create new constraints based on previous constraints with the use of mathematical equations.
In 2012, Viet-Vu Vu et al. [14] proposed active query selection to develop constrained clustering. The suggested algorithm initially forms a set of candidate constraints to obtain an appropriate constraint set using the K-NNG. In the second step of the proposed mechanism, a new constraint set is used to evaluate the ability of the selected constraint set in clusters separation.
In 2016, Sahoo et al [15] used semi-supervised clustering to process medical texts. They first used the unsupervised NSGA-II-clus clustering method to obtain different partitions. Then they employed the additional available information to select the best solution from a set of final non-solvable solutions. To this end, 10% of the side information in the form of 28 must-link and cannot-link constraints was used to rank in the irregular solutions set.

III. RESEARCH METHOD

To enhance the clustering quality in the CVQE algorithm in the proposed method, ICA algorithm is used to select useful Constraints.

***Step 1*: Generating the initial population for the constraint selection**
In this step, the initial countries are randomly generated. Each country has an N number of Pi, where N is the number of constraints of the input dataset. The variable values of Pi for the formation of each country randomly are 0 or 1. Then the equivalent pair of must-link and cannot-link constraints of any country equal to zero is eliminated from the constraints set. In the end, a subset of must-link and cannot-link constraints remains.

Step 2: Clustering and fitness calculation
Once the country related to the constraint selection is formed, the power of each country is calculated. To do this, the CVQE algorithm starts to cluster the data. The CVQE algorithm depends on the order of the input data and when each data is entered, the algorithm looks for the nearest centre of the cluster to it. In the case the placement of data within the considered cluster does not violate the constraints, the data is assigned to the cluster. Otherwise, the algorithm seeks another cluster near the data, assigning the data to which will not violate the limitation. If placing the data in all clusters violates the constraints, the algorithm calculates the value of CVQE based on Eq. (1) for the considered data with respect to each cluster.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 2, February 2020

$$\mathrm{CVQE}=\sum_{j=1}^{k}\left(\mathrm{CVQE}_j\right)$$
$$\mathrm{CVQE}_j=\frac{1}{2}\sum_{x_a\in c_j} D(x_a,\mu_j)^2 + \frac{1}{2}\sum_{x_a\in c_j,(x_a,x_b)\in c_{=},y_a\neq y_b} D(\mu_{y_a},\mu_{y_b})^2 + \frac{1}{2}\sum_{x_a\in c_j,(x_a,x_b)\in c_{\neq},y_a=y_b} D(\mu_{y_a},\mu h_{(y_b)})^2$$

Eq. (1). Calculation of the CVQE value.[16]

The CVQE index is a criterion to show the amount of constraints violation of the data at the time of joining a cluster. Since the ideal state in the constrained clustering is the lowest violation of the constraints, the data is assigned to a cluster that has the lowest amount of CVQE, or in other words, has the lowest amount of constraint violation.
Then the center of a cluster to which the data is assigned is updated such that if the must-link constraint is violated, centers of clusters with two pairs of must-links move toward each other, and in the next iteration, the chance of placing these two data within the same cluster increases. Moreover, under the circumstances the cannot-link constraint is violated, the center of the cluster with two data moves towards the nearest cluster so that in the next iterations the chance of placing the data in different clusters increases, refer to (2).

$$\mu_j=\frac{\sum_{x_i\in c_j}[x_i+\sum_{(x_i,x_a)\in C_{=},y_i\neq y_a}\mu_{y_a}+\sum_{(x_i,x_a)\in C_{\neq},y_i=y_a}\mu h_{(y_a)}]}{|\mu j|+\sum_{(x_i,x_a)\in C_{=},y_i\neq y_a}1+\sum_{(x_i,x_a)\in C_{\neq},y_i=y_a}1}$$

Eq. (2). Updating cluster centers with violates restrictions.[16]

As the main objective of clustering is to reduce the radius of clusters the average radius of clusters in each iteration is considered as the fitness function of constrains selector settlement. This value is obtained using (3), where x_i denotes the *i*th data at each cluster, μ_j is the center of the jth cluster, R_i shows the radius of the ith cluster, K represents the number of clusters, and $MaxR_i$ gives the maximum distance between the data within the Kth cluster (radius of the Kth cluster) and the center of that cluster.

$$R_i = \mathrm{Max}\left(D\left(x_i,\mu_j\right)\right)$$
$$\mathrm{AverageR} = \frac{1}{K}\sum_{i=1}^{K} R_i$$

Eq. (3). Calculation of the average radius of clusters.

Then, to the number of empires, the most powerful countries will be selected to play the role of imperialist, and other countries are considered as their colony randomly.

Step 3: Movement of the colonies toward the empire countries
In this stage, after the formation of the initial population, the colonies will move toward their empires as equation (4), and their power will be calculated [14].

$$\{x\}_{new} = \{x\}_{old} + U\left(0,\beta\times d\right)$$

Eq. (4). Colonies moving toward their empires

This equation indicates that the new location of a colony equals the sum of the initial (old) location and a random value, which is the uniform 0 and the random value of β multiplied by the distance between the colony and the empire.

Step 4: The revolution
The chances of the colonies for revolution are evaluated. If each colony is capable of revolution, a number of its elements will be selected randomly, and their value will be replaced by random numbers. Then, the power of the revolutionary colony will be calculated. In each empire, it will be checked to make sure that the power of the colony is not higher than the empire, and if a colony were more powerful than the empire, the empire would be replaced by the colony. The total power of the empires, which is derived from the power of imperialist and the colonies, will be calculated based on the equation (5).

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 2, February 2020

$$T.C._n = COST\left(imperialist_n\right) + \varsigma Mean\left\{COST\left(Colonies\ of\ \ empire_n\right)\right\}$$

Eq. (5). The total power of the empires

Step 5: Imperialist competitive

Due to the powers calculated for all of the empires, the weakest colony of the weakest empire, will be recognized (since the problem is minimal, the empire with the higher power number will be the weakest). This colony exits the intended empire colonies, and another empire will be selected by the roulette wheel so that this colony can join its colonies. This process will be continued until the final conditions are met. It means that whether the program reaches the maximum repeat or ideally, only one empire remains as the optimum point and the result of the issue.

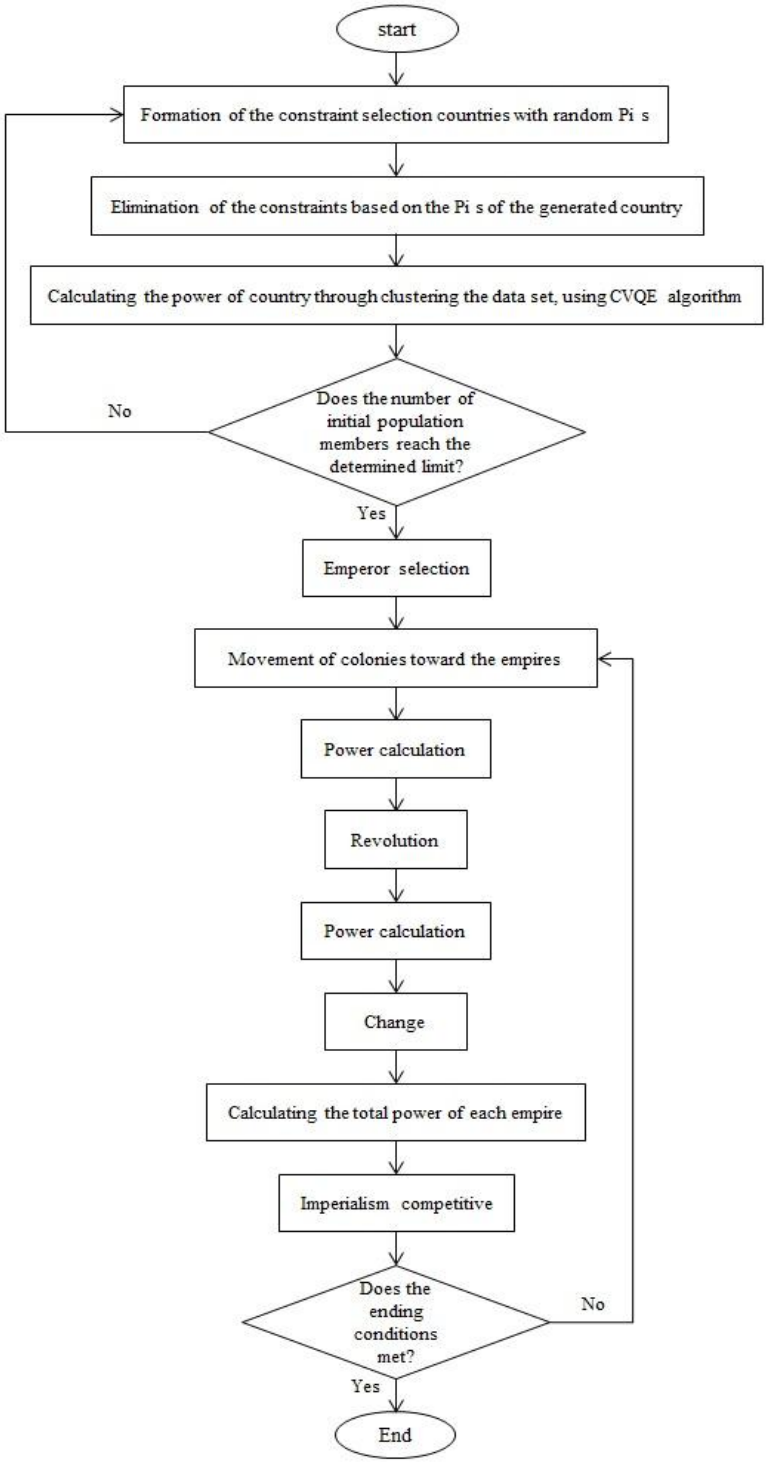

Fig. 1. Flowchart of the proposed mechanism.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 2, February 2020

IV. SIMULATION RESULTS

The Davies Bouldin criterion was used to evaluate the performance of each one of algorithms, including the proposed algorithm, the Cop-K-means clustering algorithm, and the CVQE clustering algorithm. This criterion is shown as DB for simplicity. In order to calculate the DB, first, the average value of the intra-cluster distance is calculated for each cluster, then, the inter-cluster distance is calculated. The combination of the outer cluster and inter-cluster distances is calculable through Eq. (6).

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$$

Eq. (6). Combining in-cluster and outer-cluster distance

where S_i and S_j are the average inter-cluster distances for the ith and jth clusters, and $M_{i,j}$ expresses the outer-cluster distance between the ith and jth clusters. Finally, the maximum values of $R_{i,j}$ are averaged and the result is DB. The lower the value of DB, the lower the intra-cluster distance and the higher the inter-cluster distance, which is one of the conditions of an ideal clustering. To evaluate the performance of algorithms, each algorithm is run 10 times in the clustering of the mentioned datasets and the time duration and the DB values for each run are measured. The average run time and DB values for 10 runs are taken into account as the final result.

A. *Assessment of the proposed algorithm based on the data of Iris, wine, heart, and glass*

In order to evaluate the performance of the algorithms, each algorithm was run 10 times in the clustering of the mentioned datasets, and the running time and DB value are measured each time. The average of running time and DB of 10 executions of each algorithm is considered as the final result.

Table 1. The data of the outputs of three algorithms in clustering of the iris data set

Proposed algorithm	**CVQE**	**Cop-K-means**	
DB	DB	DB	Failed run
4.54139	4.683	4.328	7.1

Table 2. The data of the outputs of three algorithms in clustering the wine data set

Proposed algorithm	**CVQE**	**Cop-K-means**	
DB	DB	DB	Failed run
6.473	4.7.492	7.168	17.1

Table 3. The data of the outputs of three algorithms in clustering the heart data set

Proposed algorithm	**CVQE**	**Cop-K-means**	
DB	DB	DB	Failed run
4.258	12.384	12.347	23

Table 4. The data of the outputs of three algorithms in clustering the glass Data set

Proposed algorithm	**CVQE**	**Cop-K-means**	
DB	DB	DB	Failed run
7.325	7.988	8.982	36.9

V. CONCLUSION

In this article, an improved constrained clustering algorithm was presented. The proposed algorithm performs constraint selection using the Imperialist Competitive Algorithm. Then, it clusters the dataset, and when facing mandatory conditions of a constraint violation, it continues the clustering of the data by finding the best cluster that has the lowest amount of constraint violation. The performance of the proposed algorithm was tested with three datasets, including the Iris, the Wine, and the Heart datasets, which are extracted from the UCI database, and the average values of the results of 10 executions of the algorithm were calculated by the Davies-Bouldin criterion. Then, the results were compared with the results of the execution of the two constrained clustering algorithms, including the Cop-K-means and the CVQE algorithms. Finally, it was observed that:

1) The clustering results of the proposed algorithm is better than the two selected clustering algorithms.
2) The running time of the proposed algorithm is longer than the two Cop-K-means and CVQE algorithms.

REFERENCES

[1]. A. Aeen, Active Constraint Clustering with the Possibility of Ranking Constraints at the Sample Level, 2015.
[2]. Green, P.E., J. Kim, and F.J. Carmone, A preliminary study of optimal variable weighting in k-means clustering. Journal of Classification, 7th, Springer, Sept., 1990.
[3]. K. Wagstaff and C. Cardie, Clustering with instance-level constraints, AAAI/IAAI, 2000. 1097: p. 577-584.
[4]. I. Davidson and S. Ravi, Clustering with constraints: Feasibility issues and the k-means algorithm, SIAM international conference on data mining, 2005.
[5]. S. Basu, A. Banerjee & R. Mooney, Semi-supervised clustering by seeding, 19th International Conference on Machine Learning(ICML), 2002.
[6]. M. Halkidi, D. Gunopulos, N. Kumar, M. Vazirgiannis and C. Domeniconi, A framework for semi-supervised learning based on subjective and objective clustering criteria, 5th IEEE International Conference, 2005.
[7]. J. Wang, B. Xue, X. Gao and M. Zhang, A differential evolution approach to feature selection and instance selection, Pacific Rim International Conference on Artificial Intelligence, Sringer, 2016.
[8]. J.w. Shavlik and T.G. Dietterich, Readings in machine learning, Morgan Kaufmann, 1990.
[9]. D. Greene and P. Cunningham, Constraint selection by committee: An ensemble approach to identifying informative constraints for semi-supervised clustering, European Conference on Machine Learning, Springer, 2007.
[10]. Y. Hongand and S. Kwong, Learning assignment order of instances for the constrained K-means clustering algorithm, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009.
[11]. P.K. Mallapragada, R. Jin, and A.K. Jain. Active query selection for semi-supervised clustering, 19th International Conference on Pattern Recognition (ICPR), IEEE, 2008.
[12]. C. Ruiz, C. G. Vallejo, M. Spiliopoulou and E. Menasalvas, Automated constraint selection for semi-supervised clustering algorithm, Conference of the Spanish Association for Artificial Intelligence, Springer, 2009.
[13]. V.-V. Vu, N. Labroche, and B. Bouchon-Meunier. An efficient active constraint selection algorithm for clustering, 19th International Conference on Pattern Recognition (ICPR), IEEE, 2010.
[14]. V.-V. Vu, N. Labroche and B. Bouchon-Meunier, Improving constrained clustering with active query selection, Pattern Recognition, 2012.
[15]. P.Sahoo, A. Ekbal, S. Saha, D. Molla and Kaushik Nandan, Semi-supervised Clustering of Medical Text, The Clinical Natural Language Processing Workshop (ClinicalNLP), 2016.
[16]. D.T. Truong and R. Battiti, A Survey of Semi-Supervised Clustering Algorithms: from a priori scheme to interactive scheme and open issues, University of Trento, 2013.