International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Data Mining Techniques to Improve Efficiency of Analyzing Medical Data

Dr. Rafath Samrin, Shahana Tanveer

Professor, Department of CSE, Deccan College of Engineering and Technology, Hyderabad.
Associate Professor, CSE Department, Deccan College of Engineering &Technology, Hyderabad.

Abstract: As the Internet of medical Things emerge in the field of medicine, the volume of medical data is expanding rapidly and along with its variety. As such, clustering is an important procedure to mine the vast data. Many swarm intelligence clustering algorithms, such as the particle swarm optimization (PSO), firefly, cuckoo, and bat, have been designed, which can be parallelized to the benefit of mass data computation. However, few studies focus on the systematic analysis of the time complexities, the effect of instances (data size), attributes (dimensionality), number of clusters, and agents of these algorithms. In this paper, we performed comparative research for the PSO, firefly, cuckoo, and bat algorithms based on both synthetic and real medical data sets. Finally, we conclude which algorithms are effective for the medical data mining. In addition, we recommend the more suitable algorithms that have been developed recently for the different medical data to achieve the optimal clustering.

Keywords: Data mining, swarm intelligence, clustering algorithms, medical data analysis.

I.INTRODUCTION

As with time data mining becomes most valuable tool for extraction and manipulation of data and also helps to provide useful information based on which decisions are taken. Most of the system fails because they don't have the right tools which tackle most of the uncertainties. The industrial persons use these data mining tools to create more values from their system by optimizing the processes. Data mining becoming a powerful tool which evaluates and provides the best results or decisions based on the previous records. Data mining becomes the demand of the companies with strong consumer focus like finance, retail, communication and marketing. By using this retailer uses records related to customer purchases based on their history which helps in developing the products and promote them according to specific customers.

Data mining refers to as a field which deals with the search and research on the data. Mining is a term which means fetching or extraction of data from a large data set or we called as a huge data repository. Data mining basically categorized into two types Classification and Clustering. Both terms are different in nature from each other. In Classification there is a set of predefined classes and then find out the objects belongs to which class whereas in clustering firstly groups of objects can be prepared and then find out whether they relate with each other or not. Data mining is the step in the process of knowledge discovery in the databases which inputs the cleaned data, transform it, search the data by using some algorithms and produce the output patterns. It is also the relationship to the evaluation steps of the whole knowledge discovery in the databases process. Data mining is a new discipline lying at the interface of statistics, pattern recognition, database technology, machine learning and other areas. Figure 1 shows the whole process of data mining.

Clustering is a well-known problem in computer science. In recent years, scholars have applied swarm intelligence algorithms to solve the clustering problem. Some examples are the PSO clustering, Firefly

clustering, Bat clustering, etc. Swarm intelligence algorithms are popular in the optimization community. The core idea of swarm intelligence algorithms is imitating behaviors of creatures in nature, especially creatures that have a habit of swarming together, e.g. ants, reies, bees, etc. Researchers believed that there are some underlying reasons for their behavior, such as searching for food, being together with companions, evading obstacles, etc. It is found that swarm intelligence clustering approaches have more possibilities to deviate from the local optima, and therefore it is useful to apply swarm intelligence algorithms to solve clustering problems. Up-to-date, different kinds of swarm intelligence algorithms have been applied to clustering problems [1][5].

Clustering is the unsupervised classification of patterns into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. Clustering is a difficult problem combinatorial and differences in assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur. The paper here presents an overview of pattern clustering methods from a statistical pattern recognition perspective with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners. Different approaches to clustering data can be described. The other representations of clustering methodology are possible. At the top level, there is a distinction between hierarchical and partitioned approaches (hierarchical methods produce a nested series of partitions while partitioned methods produce only one) [8].

II.RELATED WORK

In literature, Tang et al. [4] have compared the performance of several swarm intelligence clustering approaches. However, there is no systematic experiment and analysis on how instances (data size), attributes (dimensionality), number of clusters, and number of agents can affect the performance of all those approaches. Therefore, this gives us the motivation to analyze the time complexities of four swarm intelligence clustering approaches (PSO, Firefy, Cuckoo and Bat) systematically in this paper. Then, by conducting experiments on synthetic and real data, we also confirmed that the assumption of their time complexity is correct. The experiments on synthetic data were conducted based on four aspects: data size, dimensionality, number of clusters and number of agents. In addition, we conducted experiments on real data to further confirm that our assumption is correct.

This section briefly reviews several swarm intelligence algorithms, literature that involves application of swarm intelligence algorithms to solve clustering problem, as well as articles comparing swarm intelligence clustering algorithms. For the current optimization problems, it is difficult to search the optima when the search space is very large. Therefore, Kennedy and Eberhart [6] proposed Particle Swarm Optimization (PSO) to obtain an approximate optimum with partially searching the search space. In this way, it is highly efcient as it does not require searching the whole search space and its strategy ensures its accuracy is quite good. This was the first time that the strategy of a group of individuals was presented to the swarm intelligence community. Later
on, Yang [7] proposed the Firefly algorithm by imitating the behavior of this insect. The basic idea is that one Firefly will be attracted by another. The attractiveness is defined to be proportional to their brightness, which is mathematically represented by the fitness in clustering problems. Subsequently, Yang and Deb [8] also proposed another swarm intelligence algorithm called the Cuckoo algorithm, which imitates the behavior of cuckoos laying eggs. In particular, each Cuckoo (agent) will lay an egg in a random nest and that egg will

randomly be dumped or kept by the host of that nest in one generation. Furthermore, Yang [9] proposed the third swarm intelligence algorithm in 2010, called Bat algorithm, whereby the basic idea is imitating bats to sense distance by echolocation.

As various swarm intelligence algorithms were proposed, Van der Merwe and Engelbrecht [5] became the first to suggest clustering by PSO. To the best of our knowledge, his was the first paper proposed to adopt a swarm intelligence algorithm to solve the clustering problem. After that, Senthilnath et al. [3] proposed the Firefly clustering approach.

Algorithm 1 PSO Clustering Algorithm

Input: A set of points $P = \{p_1, p_2, \ldots, p_l\}$ and three parameters w, c_1 and c_2
Output: An agent A with best fitness

Initialize $\mathcal{A} = \{A_1, A_2, \ldots, A_k\}$;
Calculate $\mathcal{F} = \{F_1, F_2, \ldots, F_k\}$

Initialize $\mathcal{PA} = \{PA_1, PA_2, \ldots, PA_k\}$;
Calculate $\mathcal{PF} = \{PF_1, PF_2, \ldots, PF_k\}$;
Initialize GA;
Calculate GF;
Initialize $\mathcal{V} = \{V_1, V_2, \ldots, V_k\}$;
For before stop criterion meets **do**
  For each A_i **do**
    Update A_i by Equation (2);
    Calculate F_i
    If $F_i < PF_i$ **then**
      $PA_i = A_i$;
      $PF_i = F_i$
    End
    If $PF_i < GF$ **then**
      $GA = PA$;
      $GF = PF_i$;
    End
  End
End

PSO clustering algorithm

Recently, Ameryan et al. [1] and Saida et al. [2] also proposed new clustering algorithms based on the Cuckoo algorithm. Tang et al. [4] has compared the performance of several swarm intelligence clustering algorithms in 2012. However, none of the above papers have systematically compared the time complexities of all four swarm intelligence clustering algorithms (pertaining to PSO, Firey, Cuckoo, and Bat). Furthermore, none of the above papers have systematically analyzed the effect of data size, dimensionality, number of clusters and number of agents to all four swarm intelligence clustering algorithms.

III. EXPERIMENT RESULT

In this section, parameters for every clustering algorithms are introduced in the first place. Next, the experiments are conducted on synthetic data for comparing the efficiency and effectiveness of four approaches. The synthetic data are scaled from four aspects (data size l, dimensionality n, number of clusters m and number

of agents k) so as to compare different approaches from different perspectives. Afterwards, we also conduct the experiments based on real data sets to show our experiments on synthetic data are reasonable. Finally, six medical data sets are tested as case studies. In this paper, all our experiments were conducted on a computer with an Intel Xeon E5-1650 CPU at 3.5GHz, with 64 GB memory. The operating system was Windows 7 and programming language is Matlab with development environment of Matlab 2014a.

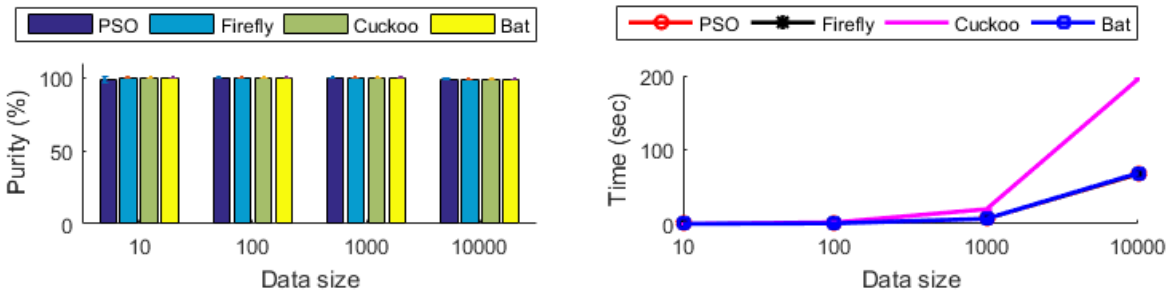

Fig 1. Results on the synthetic data when data size is different. (a) Purity. (b) Time.

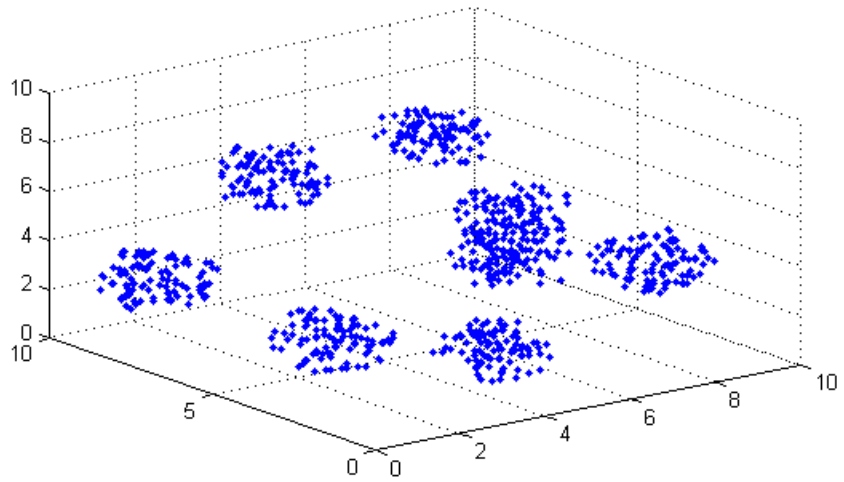

Fig 2. Example of the generated synthetic data.

IV.TEST ON REAL DATA

We also compared the four clustering approaches on three real data sets for further confirming of our conclusions. They are the Iris data set, Image Segmentation (IS) data set and Character Trajectories (CT) data set, respectively. Their descriptions are given below.

1) IRIS DATA SET

The Iris data set was first created by Fisher [11], and is widely used in the classification and clustering community as it is simple, clear, and proposed long ago. It contains 150 instances (data size) and 4 attributes (dimensionality) with 3 classes. The attributes represent sepal length, sepal

width, petal length and petal width. This data set is adopted as a simple tester for four approaches. The Iris data set can be downloaded from [3].

2) IS DATA SET

The Image Segmentation (IS) data set was created by the vision group at the University of Massachusetts. This data set contains 2,310 instances, 19 attributes and 7 classes. The attributes are 19 features extracted from the image, e.g. the column of the center pixel of the region, the number of the center pixel of the region, etc. The IS data set can be downloaded from [2].

3) CT DATA SET

The Character Trajectories (CT) data set was created by Williams et al. [14]. It has 2,858 instances, 15 attributes and 20 clusters. The CT data set originally contained only one attribute, which is a 3 by 205 matrix. Each column of the matrix represents a feature (they are x-axis value, y-axis value
and force of the pen).We vectorize this matrix to a vector with length 615 so that it can be conveniently transformed into an instance for clustering. The CT data set can be downloaded from [1].

The results of the four clustering approaches on the three real data sets are shown in Table2 and Figure 3. The results in Table 8 are as expected, being similar to the results. For purity, the four approaches are similar on all three data sets, except that Firefly appears to be slightly weaker (92.9%) compared to other three approaches (98.2% on average). For execution time, Cuckoo is still the slowest while the other three approaches are similar on all three data sets. As a systematical discussion of the performance (effectiveness and efficiency) of four algorithms was given above, the objective of our experiments on real data is validating the assumption and the detailed discussion is therefore omitted.

V.CASE STUDY ON MEDICAL DATA SETS

In this section, we analyzed 6 medical databases as case studies using the sEMG for Basic Hand Movements (sEMG) data set [1], Arrhythmia data set [7], Mice Protein Expression (MPE) data set [20], Heart Disease (HD) data set [5], Arcene data set [10] and Dorothea data set [10]. The description of each data set is given below:

sEMG Data Set: The sEMG for Basic Hand Movements (sEMG) data set [1] contains 900 instances, 6000 attributes and 6 classes from 5 healthy subjects (based on three females and two males). The 6 classes refer to six kinds of hand grasps data, which are holding spherical tools, holding small tools, grasping with palm facing the object, holding thin, float objects, holding cylindrical tools and supporting a heavy load respectively.

Data Set	Approach			
	PSO	Firefly	Cuckoo	Bat
sEMG	100 $\pm$ 0.5	99 $\pm$ 0.3	100 $\pm$ 0.7	100 $\pm$ 0.3
Arrhythmia	99 $\pm$ 2.5	98 $\pm$ 2.3	99 $\pm$ 2.6	82 $\pm$ 2.5
MPE	90 $\pm$ 2.7	72 $\pm$ 2.2	98 $\pm$ 1.9	99 $\pm$ 2.4
HD	19 $\pm$ 1.1	24 $\pm$ 1.3	17 $\pm$ 1.1	21 $\pm$ 1.4
Arcene	44 $\pm$ 2.1	44 $\pm$ 2.2	44 $\pm$ 2.2	25 $\pm$ 2
Dorothea	10 $\pm$ 3.2	10 $\pm$ 3.3	9 $\pm$ 3.1	9 $\pm$ 3.1

Table 1. Purity on medical data sets.

VI. CONCLUTION

In this paper, we introduced four main clustering approaches, which are based on swarm intelligence, and analyzed their time complexities. Our analysis showed that the Cuckoo clustering
is the slowest one. Firefly clustering is slow when the number of agents is large. In comparison, the PSO and Bat are relatively faster than the other two approaches. After that, we conducted experiments on synthetic data by considering four aspects (data size, dimensionality, number of clusters and number of agents) to demonstrate our assumption, while we also conducted experiments on three real data sets to further confirm our assumption. Besides the conclusion on
efficiency, we also conclude that there is no significant difference for these four clustering approaches on purity based on the experimental results using both synthetic data and real data. In future, we aim to propose a new clustering algorithm based on swarm intelligence as the execution time of these four existing approaches is still not acceptable. Moreover, we are going to compare newly developed state-of-the-art approaches rather than just four classic swarm intelligence algorithms.

VII.REFERNCES

[1] Martins, J. R., & Ning, A. (2021). Engineering design optimization. Cambridge University Press.

[2] Rao, S. S. (2019). Engineering optimization: theory and practice. John Wiley & Sons.

[3] Pakhira, M. K. (2014, November). A linear time-complexity k-means algorithm using cluster shifting. In 2014 international conference on computational intelligence and communication networks (pp. 1047-1051). IEEE.

[4] Gong, X., Liu, L., Fong, S., Xu, Q., Wen, T., & Liu, Z. (2019). Comparative research of swarm intelligence clustering algorithms for analyzing medical data. IEEE Access, 7, 137560-137569.

[5] Onwubolu, G. C., & Babu, B. V. (2013). New optimization techniques in engineering (Vol. 141). Springer.

[6] Nwankpa, C. E. (2020). Advances in optimisation algorithms and techniques for deep learning. Advances in Science, Technology and Engineering Systems Journal, 5(5), 563-577.

[7] Yeomans, J. S. (2019). A Nature-Inspired Metaheuristic Approach for Generating Alternatives. In Advanced Methodologies and Technologies in Business Operations and Management (pp. 722-733). IGI Global.

[8] Shrivastava, D., Sanyal, S., Maji, A. K., & Kandar, D. (2020). Bone cancer detection using machine learning techniques. In Smart Healthcare for Disease Diagnosis and Prevention (pp. 175-183). Academic Press.

[9] Wikipedia contributors. (2021, August 23). Cuckoo search. Wikipedia. Retrieved July 22, 2021, from https://en.wikipedia.org/wiki/Cuckoo_search.

[10] J. Kennedy and R. Eberhart, ``Particle swarm optimization," in Proc. IEEE ICNN, vol. 4, Nov./Dec. 1995, pp. 1942-1948.

[11] X.-S. Yang, Nature-Inspired Metaheuristic Algorithms. Luniver Press, 2008.

[12] X.-S. Yang and S. Deb, ``Cuckoo search via Lévy heights," in Proc. World Congr. Nature Biologically Inspired Comput. (NaBIC), Dec. 2009, pp. 210-214.

[13] Iris Data Set. Accessed: Sep. 8, 2015. [Online]. Available: https://archive. ics.uci.edu/ml/datasets/Iris.

[14] B. H. Williams, M. Toussaint, and A. J. Storkey, ``A primitive based generative model to infer timing information in unpartitioned handwriting data," in Proc. IJCAI, 2007, pp. 11191124.

[15] Image Segmentation Data Set. Accessed: Sep. 8, 2015. [Online].Available: https://archive.ics.uci.edu/ml/datasets/Image+Segmentation.