Incremental Clustering using Genetic Algorithm and Particle Swarm Optimization

Neha Chopade
Research Scholar
Sri Satya Sai University of
Technology and Medical Sciences
Sehore, India

Jitendra Sheetlani, PhD
Associate Professor
Sri Satya Sai University of
Technology and Medical Sciences
Sehore, India

ABSTRACT
There are many supervised clustering algorithms based on static datasets for finding their optimal clusters. Clustering is the task of organizing data into clusters such that the data objects that are similar to each other. For finding clusters of data stream of chunks, i.e. for dynamic clustering we proposed a incremental clustering algorithm which is a combination of genetic algorithm and particle swarm optimization. In this paper, first we convert diabetes dataset into rough sets by applying appropriate algorithm, then after conversion rough sets are taken as input for genetic algorithm and after processing fitted chromosomes are generated. These fitted chromosomes are taken as input for particle swarm optimization which results in producing optimized clusters without redundancy. In this paper results are also presented and their comparison from existing approach is also given.

Keywords
Data mining, PSO,ACO GA, fuzzy logic etc.

1. INTRODUCTION
Data Mining (DM) mention to the process of discovering hidden helpful patterns in the information. Variety of terms has been utilized to define this same process such as knowledge discovery, information extraction, etc. DM search big quantity of data trying to detect consistent patterns and/or relations between variables, and then validates the findings by applying the detected pattern to a new data. DM is useful in tasks as classification, clustering, association, and time series analysis [1].

2. INCREMENTAL CLUSTERING ALGORITHM
The clustering problem consists of partitioning n points in space into k clusters thus like to minimize the maximum cluster size. The clustering method most frequently used for IR is hierarchical agglomerative clustering (HAC) [2]. This method initially assigns the n original points to n clusters, and then constantly merges couple of clusters, using the document TF-IDF vectors, until only k clusters are left. The boundaries of the cluster space generalize the documents that have been seen into a model of the categories. Unfortunately, straightforward clustering isn't feasible for maintenance of Web agent user profiles. It's useless to store all the relevant documents and periodically apply a clustering algorithm to build the components of the profile. Instead, only the cluster representations should be kept in the profile, and document representations should be discarded instantly later than the cluster representations have been modified. Algorithms for solving this new problem of incremental clustering have been studied in [3]

3. PARTICLE SWARM OPTIMIZATION (PSO)
PSO is a population- based search algorithm that is initializing with random population solutions, known as particles. Like another developmental computation method, each particle in this algorithm, called PSO is also associated with a velocity. Particles fly via the find area with velocities that are dynamically adjusted as per their historic behaviors. The particles therefore have the propensity to fly towards the improved find region, each in excess of the course of the procedure of investigation. In PSO, a number of simple entities the particles are placed in the search space of some function or issue, and each of these evaluate the object function at its current location. Thereafter, each particle then determines its movement through the search space by combining some aspect of the history of its own current and best (best-fitness) locations with those of one or more members of the swarm, with some random perturbations. The next step release takes vicinity in despite of all particles have moved. Ultimately the swarm as a whole, flock of birds together foraging for food, is likely to move close to an optimum of the fitness function.

The particle swarm (PS) is really more than now a set of particles. A particle by itself has almost doesn't solve any issues; growth takes place only when they i.e. the particles interact. Populations are organized according to some sort of communication structure or topology. This is often thought of as a social network. The topology normally comprise of bidirectional edges connecting pair of particles. It's like the alphabet j appearing in i's neighborhood, and as i in j neighbor. All particles communicate with another particle and is affected through the best point found through any element of its topological neighborhood.

Each individual in the particle swarm is composed of three D-dimensional vectors, where D is the dimension of the detect space. These are the existing position x_i the earlier best position pi, and the velocity vi.

The i^{th} particle is describe like $X_i = (x_i1, x_i2, \ldots, x_iD)$. At all generation, all particles is modify by the following two 'best' values. The initial is a best earlier location (the location provide the best fitness value) a particle has attain so far. This value is known pBest. The pBest of the i_{th} particle is described as Pi = (pi1, pi2,. . . , piD). At all iterations the P particle vector with the best fitness in the neighborhood, designated l or g, and the P vector of the existing particle are combined to adjust the velocity along all dimensions, and which velocity is then exploited to calculate a newest position for the particle. The adjustment portion to the velocity influenced by the person earlier best position (P) is measured as the cognition factor, and the portion influenced by the best in the vicinity is the social element. With the computation of the inertia factor

ω, (bring in for balancing the global and the local find out), velocity and position update equations are:

$$V_i=\omega \times v_i+n_1 \times r_{and}\ () \times (p_i-x_i)+n_2 \times r_{and}\ () \times (p_g-x_i)$$
$$X_i=x_i+v_i$$

Where rand() and rand() are two random numbers separately produced within the range [0,1] and η1 and η2 are two learning element that control the influence of the cognitive and social elements. . In (1.1), if the total on the right side increases a constant value, then the velocity on that dimension is assigned to be ±Vmax. Thus, particles' velocities are clamped to the range [-Vmax, Vmax] that serves as a restraint to control the global examination capability of particle swarm. Thus, the probability of particles leaving the search space is reduced. Presented the basic scheme of PSO algorithm [4].

4. ANT COLONY OPTIMIZATION (ACO)

ACO method is induced by the conduct of actual ant colonies. When detect for food, ant's initially explore the encircling space, leaving chemical evidence on the path that it took in order to be followed by other ants. As an ant search food, it evaluates the quality and the quantity of it and transmits some of them back to the nest dropping pheromone in amounts proportional to the quality and the quantity of the Founded food. The pheromone trial will probabilistically guide the ants to the food source. Eventually, this sort of behavior will lead the convergence of the ant taking the shortest path to the best food source.

The key step of ACO algorithms are as follows:

1) Pheromone trail starting.
2) Solution building utilize of pheromone trail: every ant constructs an entire strategy to the issue pursuance to a probabilistic model.
3) Solution evaluation: evaluate the value of the Result depend on a problem specific fitness function.

- An atmosphere that represents the problem domain in such a way that it's suitable for the ants to navigate and construct an explanation for the issue.
- A problem established heuristic evaluation function (η), which represents a value of element for the exceptional steps that build the result.
- A rule for pheromone updating (τ), which takes under consideration the evaporation and the reinforcement of the trails. A probabilistic transition rule depend on the quality of the heuristic characteristic (η) and on the strength of the pheromone trail (τ) that is used to iteratively build the result. A FF thru that the build solution is evaluated.
- A clear specification of when the algorithm converges to a solution [1].

5. GENETIC ALGORITHM

In the past, evolutionary algorithms have been relating in many real life issues. GA is an evolutionary algorithm. GA has emerged as a robust optimization and practical, approach and search technique. A GA is a search algorithm that is inspired by the way nature evolves species using natural selection of the fittest individual's.

The possible solutions to problem being solved are represented by a chromosomes population. A chromosome is a binary digits string and all digits that create up a chromosome is known gene. This preliminary population can be totally random or may be created manually using processes such as greedy algorithm. The pseudo code of a primary algorithm for GA is as follow [5]:-

```
Initialize (population)
Evaluate (population)
While (stopping condition not satisfied)
{
Election (population)
Crossover (population)
Mutate (population)
Evaluate (population)
}
```

6. FUZZY LOGIC

Fuzzy Logic is suitable for the intrusion detection (ID) difficulty for two main reasons. First, many quantitative functions are comprised in ID. Security-associated data categorizes the statistical capacity into four kinds: categorical, ordinal, binary express and linear categorical. Both linear and ordinal definite capacity is quantitative characteristic that could potentially be viewed as fuzzy variables. Two instances of ordinal measurements are the CPU usage time and the connection period. A case of a linear categorical measurement is the no. of dissimilar TCP/UDP services initiated via the source host. The second motivation for exploiting fuzzy logic to address the ID issue is that protections itself comprise fuzziness. Given a quantitative measurement, an interval may also be exploited to signify a normal value. Then, any values falling outside the interval will be considered anomalous to the similar degree regardless of their distance to the gap. The similar applies to values inside the interval. The use of fuzziness in representing these quantitative features enables to clean the abrupt separation of normality and abnormality and presents a measure of the normality degree or abnormality of a particular measure [6].

7. LITERATURE SURVEY

Author Name	Year	Approach
Alexander Dockhorn	2016	The clustering technique based totally on edge-lengths of the dendrogram or based totally on region vicinity estimates efficiently detects arbitrary density clusters and shape [7].
Patino Galván	2016	They defined model of educational assessment with adaptive features to the region of information and the students to expect behaviors of instructional overall performance and support the decision making inside the educational context [8].
Mustakim Al Helal	2016	A serious evaluation of a few algorithms and most importantly performance measure that is valid for imbalanced data [9].
M. Omair Shafiq	2016	Analyzes and equate several popular classifier algorithms which have been most usually using in detecting credit score card fraud. The experimental

		consequences endorse that (1) even though finished type accuracy value is 98.25% but fraud detection success value is under 50%, (2) meta and tree classifiers perform better than different other group of classifiers [10].
Doreswamy,	2016	An approach where our combine a clustering approach and a stochastic approach to elect effective features from the highest dimensional breast cancer set of data in rapid time. The results obtained were established to be highly supportive in nature. The feature subset produce exploiting PSO based fast K-means algorithm on KDD cup 2008 data set produced an accuracy of 99.39% and its complexity of time was found to be O(log(k)) [11].
V. Shanmugarajeshwari,	2016	Classification techniques are described and used for educational data mining. The classification process is depending on C5.0 algorithm with good classification accuracy [12].
Pavel Kromer	2016	A recent GA for fixed-length subset election to find feature subsets on the basis of their entropy, estimated by a fast data compression method. The reasonability of this newest fitness criterion and the usefulness of elected feature subsets for practical data mining is evaluated using well-known data sets and several widely-used classification algorithms [13].
Yusuke Nojima	2016	Parallel distributed implementation of fuzzy GBML for fuzzy classifier design from large data [14].

8. PROPOSED WORK

In this paper, an incremental clustering technique is proposed which is combined with genetic algorithm and particle swarm optimization. First we have converted data into rough sets. Then after conversion, genetic algorithm followed by particle swarm optimization is applied.

The fundamental thought of this algorithm is as follows: Let T_{th} denotes a threshold of dissimilarity between data objects. We initially give a price of T_{th} then decide upon an object randomly from the given datasets, let or not it's the middle of a cluster, and decide upon an additional object from the given datasets again, compute distance between the chosen knowledge object and the present cluster center, If this distance is larger than T_{th} then form a brand new cluster and chosen object would be the center of the cluster or else staff the article into current cluster and update its centroid. Choose an object again from the datasets, repeat the process until all objects are clustered.

Flow chart is given below for converting data into rough sets.

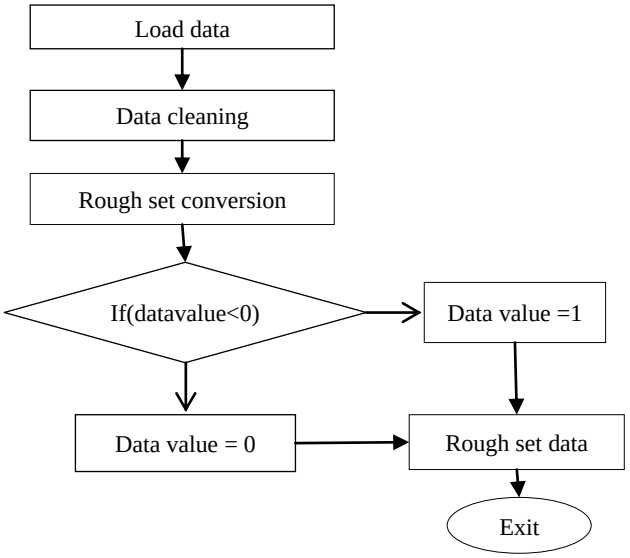


Fig1. Conversion of data into rough sets

After conversion genetic algorithm is applied to rough sets followed by particle swarm optimization.

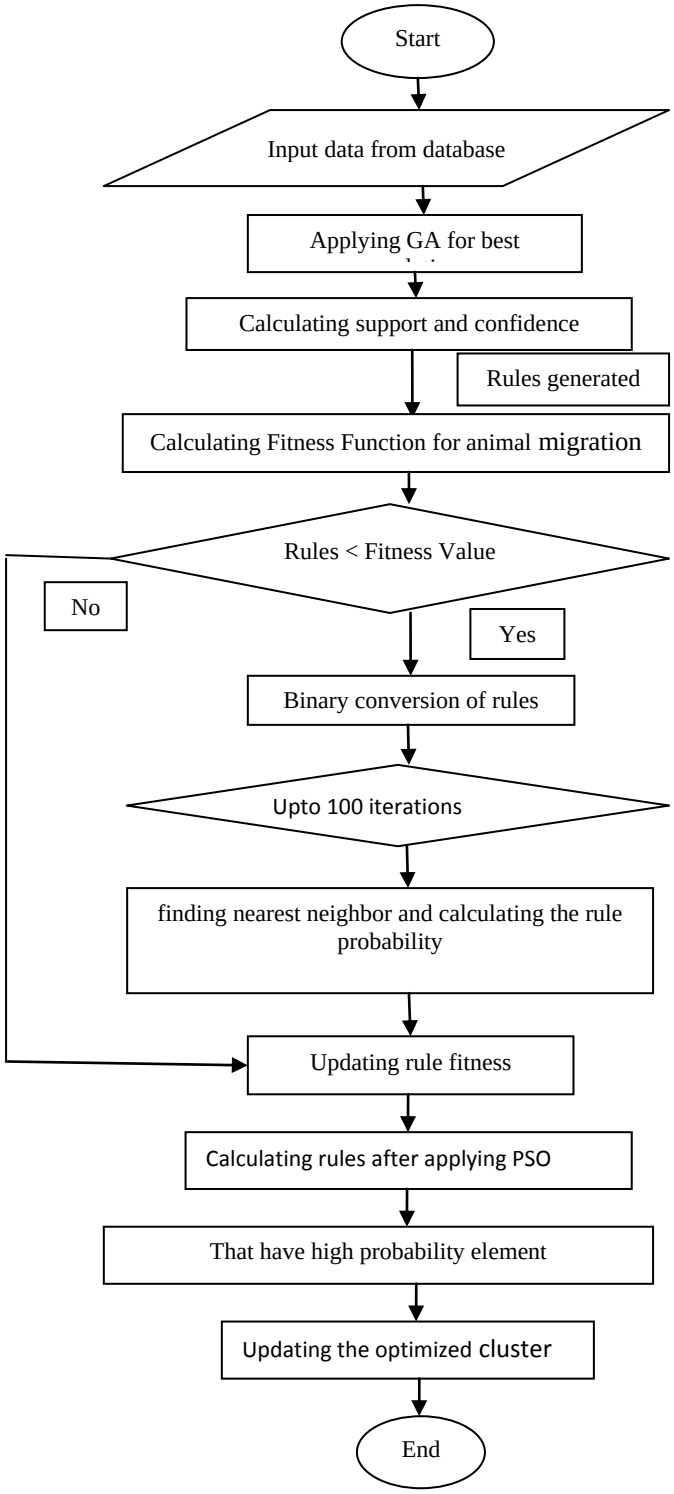


Fig.2. Proposed work

GA

Generate population with the help of genetic algorithm and after that over the fitted offspring apply particle swarm optimization. The genetic process is represented with the help of a flowchart. The genetic algorithm is being shown below so as to clarify the whole functioning of the genetic algorithm:-

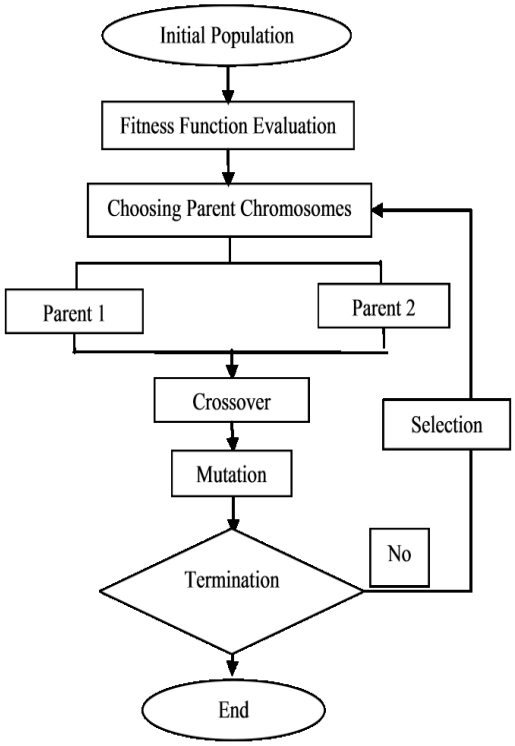


Fig.3. Functioning of the genetic algorithm

PSO execution:

1. Initialize swarms- total data available in cache with count and important parameters.
2. Fitness value- initialize factors like w=0.5, c1=0.1 and c2=0.01. Fitness value depends on three factors- time of access (T), weight which is count of link visited ($w_,$), α which is number of keywords matching the search in the database.
    a. Fit_val = T*w+ $w_,$ * c1 + α * c2
3. For the search:
    a. Initialize the particle's position, velocity, cost, best position and best cost.
    b. Particle_position = it depends upon the location at which the data is stored. Thus giving a position for the data.
    c. Particle_cost = (hit_prob(0.5-1) or miss_prob(0-0.5)) * α
    d. Initialize minVel = 0 and MaxVel = maximum access time for the data.
    e. Particle_velocity = w* p_vel + c1*(p_best-p_curr) + c2*(g_best-p_curr)
    f. Update the data in the segments based on the position, velocity and cost and replace the pages with maximum use into the partition1. For the 2nd half, remove the less fit data and thus, making the cache free from any kind of unnecessary pollution.
4. End.

9. RESULT ANALYSIS

Base Result

Elapsed time is 8.668751 seconds.

Dataset compression rate 73.382821

Training Error rate 26.086957

Propose Result

Elapsed time is 7.975316 seconds.

Dataset compression rate 99.045599

Training Error rate 6.574761

Comparison between base result and proposed result:

Elapsed time comparison

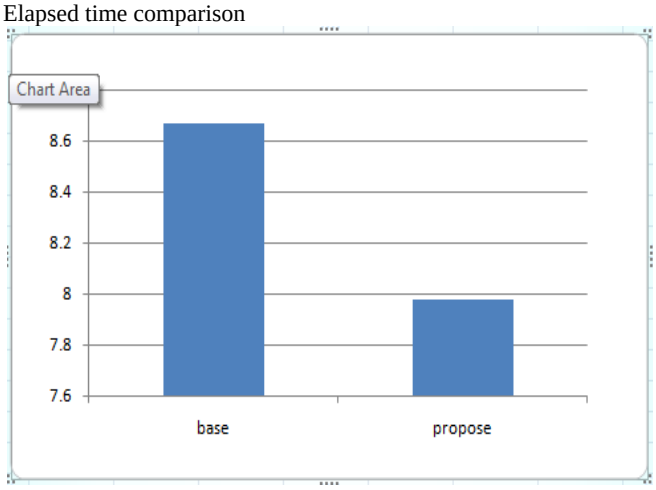


Fig.4 Elapsed time

Data comp

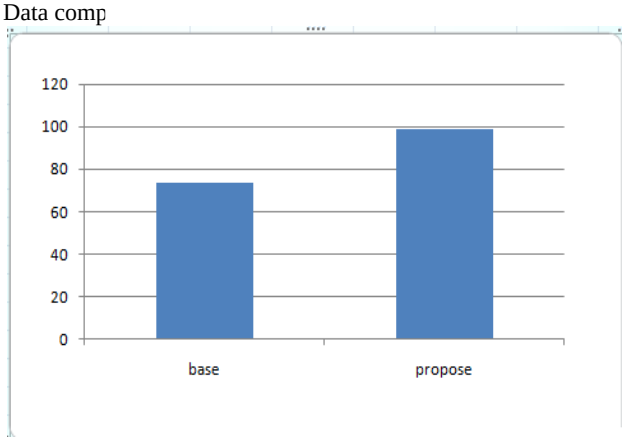


Fig.5 Data compression rate

Training

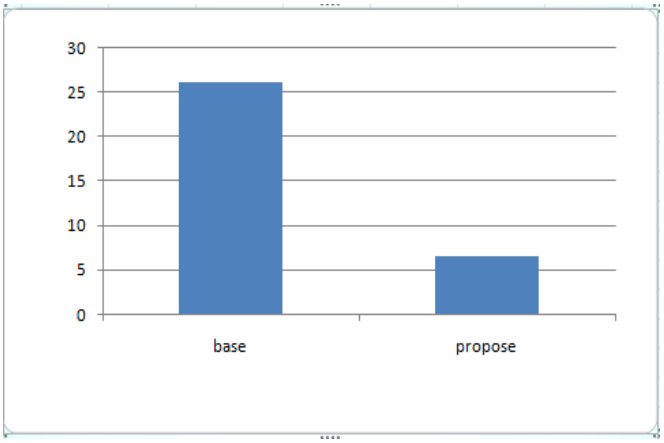


Fig.6 Training error rate

Selection of Data File

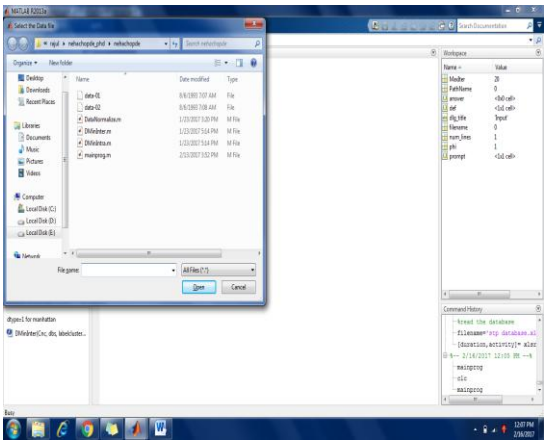


Fig.7 Selection of Data File

Asking Data Points

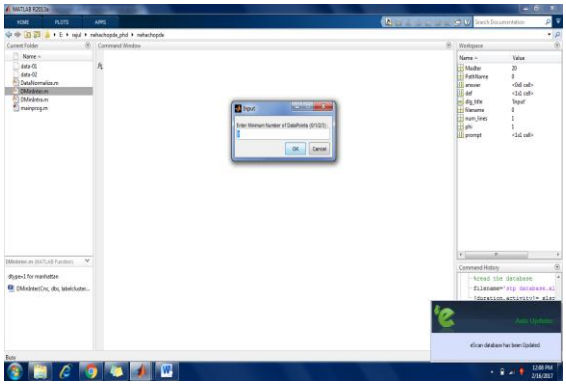


Fig.8 Asking Data Points

Asking Distance Matrix

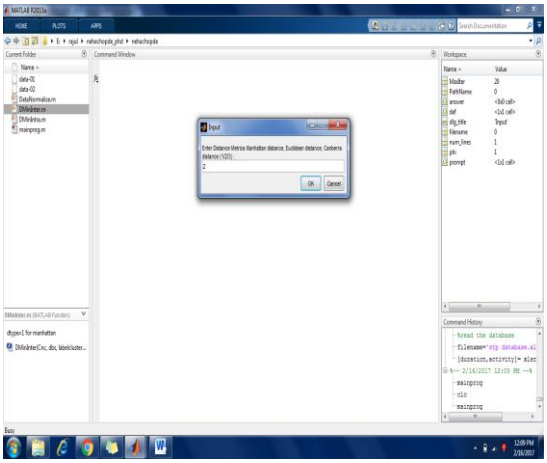


Fig.9 Asking Distance Matrix

Asking Original data/Normalized data

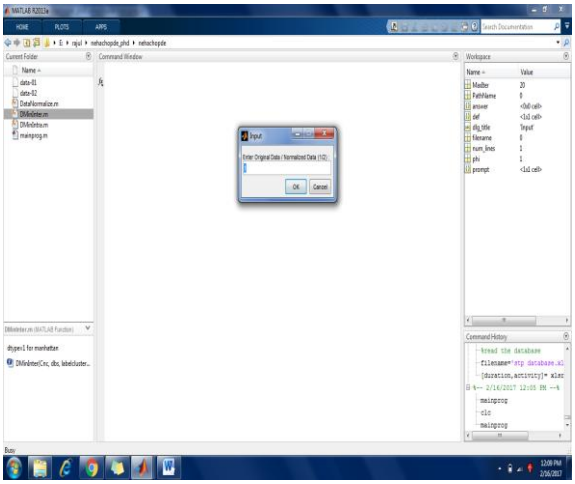
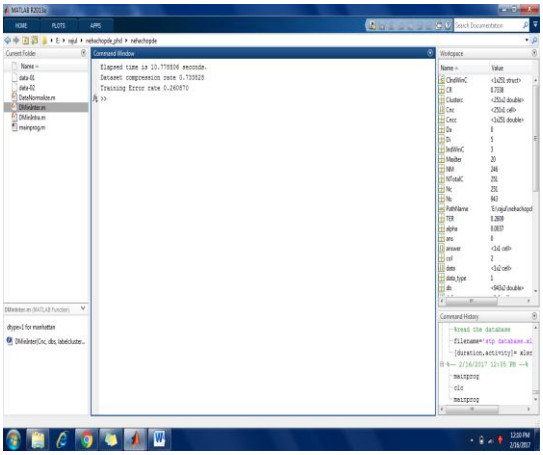


Fig.10 Asking Original data/Normalized data



10. CONCLUSION

Data is processed into useful term in data mining and various methods have been introduced for processing the data. Genetic algorithm and PSO algorithm are used in our proposed work which shows the improvement with the base approaches. Clustering is also performed in this paper which forms the grouping of the same form of data to make the selection of the data easy and fast. We convert our data into rough sets by applying some procedures over the dataset. Generate population with the help of genetic algorithm and after that over the fitted offspring we apply particle swarm optimization. Then we applied PSO algorithm for in which we calculate the particle velocity. At the end we get the optimized result of the data after performing 100 iterations over it. We defined the fitness value and initialize the factors like w=0.5, c1=0.1 and c2=0.01. We initialized the probability of hit between 0.5 to 1 and probability of miss between 0 to 0.5. This method is improved form of the existing approach in the form of training error rate, elapsed time and data compression rate. We apply our proposed algorithm on diabetes dataset. Our simulator is MATLAB which shows that our results are better than in the existing technique.

11. REFERENCES

[1] Ahmed Sameh, Khalid Magdy "Data Mining Ant Colony for Classifiers" International Journal of Basic & Applied Sciences IJBAS-IJENS Vol:10 No:03, 101303-4646 IJBAS-IJENS © June 2010 IJENS

[2] Rasmussen, E. Clustering Algorithms. Frakes, W., Baeza-Yates, R. (eds.), Information Retrieval: Data Structures and Algorithms Prentice-Hall, 1992.

[3] Charikar, M., Chekuri, C, Feder, T., Motwani, R. Incremental Clustering and Dynamic Information Retrieval Proceedings of the 29th ACM Symposium on Theory of Computing, 1997.

[4] Sunita Sarkar, Arindam Roy and Bipul Shyam Purkayastha " Application of Particle Swarm Optimization in Data Clustering: A Survey" International Journal of Computer Applications (0975 – 8887) Volume 65– No.25, March 2013.

[5] Chayanika Sharma , Sangeeta Sabharwal, Ritu Sibal "A Survey on Software Testing Techniques using Genetic Algorithm" IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 1, January 2013 ISSN (Print): 1694-0784 | ISSN (Online): 1694-0814.

[6] Harshna, Navneet kaur "Survey Paper of Fuzzy Data Mining using Genetic Algorithm for Intrusion Detection" International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013 1687 ISSN 2229-5518.

[7] Alexander Dockhorn, Christian Braune, and Rudolf Kruse "Variable Density Based Clustering" 978-1-5090-4240-1/16/$31.00 ©2016 IEEE.

[8] Patiño Galván "Educational Evaluation and Prediction of School Performance through Data Mining and Genetic Algorithms" FTC 2016 - Future Technologies Conference 2016 6-7 December 2016 | San Francisco, United States, 978-1-5090-4171-8/16/$31.00 ©2016 IEEE

[9] Mustakim Al Helal, Mohammad Salman Haydar and Seraj Al Mahmud Mostafa "Algorithms Efficiency Measurement on Imbalanced Data using Geometric Mean and Cross Validation" 2016 International Workshop on Computational Intelligence (IWCI) 12-13 December 2016, Dhaka, Bangladesh, 978-1-5090-5769-6/16/$31.00 ©2016 IEEE.

[10] M. Omair Shafiq "Event Segmentation using MapReduce based Big Data Clustering" 2016 IEEE International Conference on Big Data (Big Data), 978-1-4673-9005-7/16/$31.00 ©2016 IEEE.

[11] Doreswamy, Umme Salma M "PSO Based Fast K-means Algorithm for Feature Selection from High Dimensional Medical data set"2016 IEEE.

[12] V. Shanmugarajeshwari, R. Lawrance "Analysis of Students' Performance Evaluation using Classification Techniques" 978-1-4673-8437-7/16/$31.00 ©2016 IEEE.

[13] Pavel Kromer ¨ Jan Platos "Genetic Algorithm for Entropy-based Feature Subset Selection" 978-1-5090-0623-6/16/$31.00 c 2016 IEEE.

[14] Yusuke Nojima and Hisao Ishibuchi "Effects of Parallel Distributed Implementation on the Search Performance of Pittsburgh-style Genetics-based Machine Learning Algorithms" 978-1-5090-0623-6/16/$31.00 c 2016 IEEE.