ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Clustering the Data by Using Evolutionary Techniques

[1] Bethala Shirisha, [2] H V Ramana Rao
[1][2] Assistant Professor, CMR Institute of Technology, Medchal, Hyderabad.

Abstract: PSO is one of the Population-based Search Algorithm. PSO is a method that optimizes a scenario by iteratively and trying to improve the solutions and it is a method for performing numerical optimization without precise information. PSO is initialized with a population of random solutions called particles and here each individual is treated as a volume less particle in a d-dimensional search space. Differential Evolution is a population-based algorithm. DE is more likely to find a function's true global optimum. Here we applied DE and PSO algorithm on clustering by using same parameters, population size to the efficiency of these two algorithms and finding the best globest using PSO and DE algorithms

Keywords: Differential Evolution, Data Mining, Genetic Algorithms, Knowledge Data Discovery, Particle Swarm Optimization.

1. INTRODUCTION

The amount of data being stored in databases is growing in an exponential manner now-a-days. These databases may contain a rich treasure of knowledge which can help an organization to take a strategic decision or initiate a strategic plan of action. For extraction of knowledge from the stored data, there is a potential drawback of facing a huge overhead in storing the data. These necessities have given rise to a very popular and useful field called Knowledge Discovery from Data (KDD) and Data Mining. The main aim of this KDD is to extract the data from large data sets. Data Mining is the center step of a broader process of KDD.

The knowledge discovered from these large databases should be
(i) accurate
(ii) comprehensible
(iii) interesting

Knowledge Discovery, in a broader sense, involves Data Mining as a core processing activity. The whole Data Mining activity falls under two categories called as Predictive Data Mining and Descriptive Data Mining. Most Data Mining tasks such as classification, association rules, regression, clustering, model visualization and exploratory data analysis come under the above two categories. There are other Data Mining techniques and methods for knowledge extraction in literature. Most methods, generally used for Data Mining tasks, may be classified into Neural Networks, Decision Trees, Statistical Methods, Case-based Reasoning, Bayesian Belief Network, Genetic Algorithm/Evolutionary Programming, Rule Induction, Fuzzy Sets and Rough Sets. In this study, the use of Soft computing techniques in general and Evolutionary Techniques in particular have been explored to address the Data Mining tasks such as classification and clustering. The constituents of soft computing are: Fuzzy Logic (FZ), ACO, Swarm Optimization (SO), Artificial Neural Networks (ANN), Evolutionary Algorithms (EA) including Genetic Programming (GP), Genetic Algorithms (GA), Evolutionary Strategy (ES).

In this paper, the purpose of main component is that the soft computing methods like Differential Evolution and Swarm Optimization have been implemented for developing clustering models for several benchmark datasets and to a real world data set, namely, transportation dataset. The reasons for using soft computing approaches in practice, especially in Data Mining are:

a) Usually, a person can build a new system by gathering the information from particular experts. Soft computing method will get the data of the hidden condition and find the rules that matches with the expert's prediction.
b) Systems often produce results that are not similar from the required ones. This may because of different functions or properties of input while designing the systems. However, soft computing can manage up with this problem.
c) Data mining domain is rapidly expanding and continuously updated with improved techniques so Soft computing can easily be adapted for the new environment so the system designers no need to reform the systems whenever an environment changes.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

d) Abrupt data is one of the characteristic of large databases. The existing methods fail to solve this situation. So Soft computing based method can deal with problem.
e) With improvement of technology, large amount of data is generated. These are intended to solve this huge information, and it is used to get the hidden correlations between the data.

II. EVOLUTIONARY COMPUTATION TECHNIQUES

Many researchers have emphasized the idea of using natural evolutionary process in framing a computerized optimization algorithm. ECTs[1] maintain a population of potential (or candidate) solutions to a problem and not just one solution. The aim of getting good data and removing bad solutions seems to merge well with desired values of one optimization algorithm. Generally, ECTs procedure is to initialize the population of individuals and every individual recur rents a solution to the problem. The quality of every solution is evaluated using a fitness function. Each Individual are modified by using unary transformation (mutation) and higher order transformation (crossover). This procedure is repeated until convergence is reached. The solution what we get is expect to be a near-optimum solution. A pseudo-code for an ECT is to Initialize the population and evaluate the fitness of each individual in the population and Apply selection on the population to form a new population and modify the individuals in the population using some evolutionary operators after that find the fitness value of each individual in the population . Repeat this step till the required criteria is satisfied.

There are five major ECTs:

- GA →Genetic Algorithm
- DE →Differential Evolution
- PSO →Particle Swarm Optimization
- EP→Evolutionary Programming
- GP→Genetic Programming

These five ECTs are population based approaches . ECTs have already been successfully applied to a wide variety of optimization problem, for example: pattern recognition, scheduling, image processing, etc. Thus, ECTs can be referred as global optimization algorithms.

In this paper, we emphasize the use of GA, PSO[2], and DE for developing the data clustering problems.

2.1 Genetic Algorithm

Genetic Algorithm (GA) is an iterative optimization procedure, first proposed by J.H.Holland. Individuals in GA are chromosomes and each chromosome consists of a string of cells called genes ,the value of each gene is called allele. GA works with a number of solutions collectively known as population in each iteration. A flowchart of the working principle of a simple GA is shown in Fig 2.

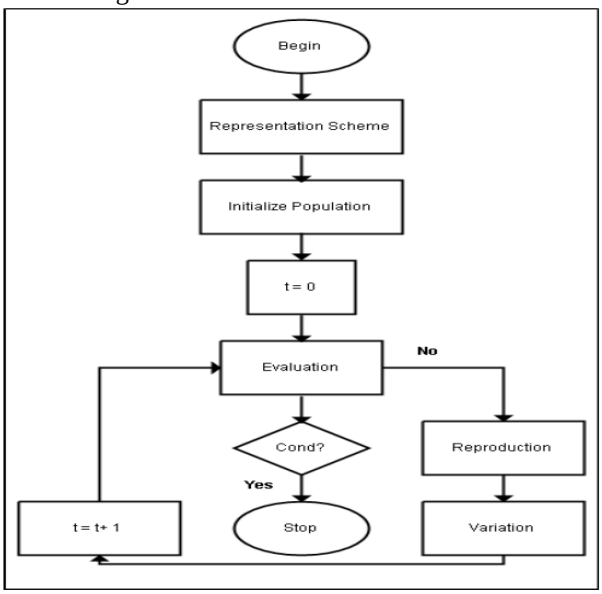

Fig 2 Flowchart for working principle of GA

a) Genetic recurrentation of a solution

In GA, the first step is recurrentation of a solution. There are number of ways the solution vector may be recurrented. Recurrentations using discrete values, real no's, or a combination of few of them are generally chosen.[3] Depending on the problem, mixture of real and discrete variables are also used for the recurrentation.

b) Initialization of Population

Usually, an unbiased set of random solutions are initialized in a pre-defined search space bound by its lower and upper limit. However, if some problem information is available then biased distribution can be created in any suitable part of search space.

c) Evolution of a Solution

This step is a crucial step for working of GA. In this each solution is calculated and a fitness value is assigned to the solution[4]. The condition for feasibility or infeasibility solution is checked in this step by computing the constraint functions gj(x) and hk(x).

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

d) Reproduction Operator

This phase is also called as selection phase. In this phase, a mating pool of good strings is created in a population. Individual solutions are selected through a fitness-based process where fitter solution is typically more likely to be selected. In each new generation, an amount of the existing population is selected to a variety of new generations[5].

e) Variation Operators

This step allows creating new solutions to the existing population. In a crossover operator, two solutions are picked from the mating pool at random and an information exchange is made between the solutions to create one or more offspring solutions and these are the two main operators which are mostly used for this purpose.[6]

2.2 Particle Swarm Optimization

Particle swarm optimization (PSO) is developed by Dr. Kennedy and Dr. Eberhart in the year 1995 and it is a population based stochastic optimization technique , they inspired by the performance of fish schooling or bird flocking. PSO is applied in functional optimization, fuzzy system control, ANN training and other areas where GA can be applied. PSO has no evolution operators such as crossover and mutation whereas GA has that operators. When compared to GA, the PSO[7] is very easy to implement and there are few parameters to change. Technique based on the movement and intelligence of swarms (i.e., Bird Flocking or Fish Schooling)

•Particle : Very small portion of area
•Swarm : A group of moving insects
•Optimization : Choosing the best solution from the available alternatives

Assume the following situation that the group of birds are searching for the food and there is only one piece of food and birds are searching for that area but they don't know where the food is located. But they know how far the food is in each iteration[8]. So what's the best strategy to find the food? The effective one is to follow the bird which is nearest to the food.

PSO has learned from this situation which is used to solve the problems. In PSO, every solution is nothing but a "bird" in the search space. We call it "particle". The particles will fly through the problem space by following the recent particles. All the particles will have different velocities which gives direction to the flying particles and are having some fitness values which are evaluated by using fitness function.

PSO is initialized with a group of random particles and then searches for the most favourable situation by updating the generations. In each iteration, every particle i.e solution is updated by following two "best" values. The first one is the best solution (fitness)which is achieved. And all the fitness value is stored. The generated value is called prebest. Another "best" value is tracked by the particle swarm optimizer which is obtained by any particle in the population and this best value is called globest which is nothing but Global test.[9] When a particle takes part of the population as its topological neighbours, then the best value is lbest which is nothing but local best.

After finding the values of prebest and globest, the particle updates its velocity and positions with following equation (1) and (2).

pv[ ] = pv[ ] + c1 * rand() * (prebest[] - current[]) + c2 * rand() * (globest[] - current[])---------- (1)
current[] = current[] + pv[ ]---------- (2)
current[] is current particle (solution), pv[ ] is the particle velocity, globest[] is global best. rand () is a random number between (0,1). c1,c2 are learning factors. usually c1 = c2 = 2.

For each particle we calculate the fitness value and If the new generated fitness value is better than the previous best fitness value (prebest) then set current fitness value as the new prebest. Choose the particle with the best fitness value as globest among all the particles.

For each particle calculate the particle velocity according to the equation (1) and then Update the particle position according to equation (2) More iterations or less error criteria cannot reach particles velocities on each dimension and they are clamped to a maximum velocity Vmax. If the sum of accelerations may cause the velocity on that element and it may exceed Vmax, the parameter which is specified by the user.[11] Then the velocity on that dimension is limited to Vmax.

2.2.1 Pros of PSO algorithm:

1. PSO can apply on both engineering use and scientific research.
2. PSO adopts the real number code, and it is decided directly by the solution.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

3. PSO is not having modified and same calculations. After evaluating all iterations, the most optimist particle and the speed can transmit information onto the other particles .

2.2.2 Cons of PSO algorithm:

1. This method cannot solve otimization problems which are on non-coordinate system such as moving rules of particles.
2. The technique suffer from limited optimism and it can have the a lesser amount of speed and the direction.

Here the important thing is random values are more in PSO and the multiplicity of population is not enough. To overcome the drawbacks faced by PSO and new algorithm DE (Differential Evolution) has been proposed.

2.3 Differential Evolution

DifferentialEvolution(DE) algorithm is a new heuristic approach. It contains three cons; for determining the correct least value from the multimodal search space. It is nothing but a population-based heuristic recently proposed by R.Stornand K.Price in 1995. It is a simple and easy to use this evolution plan, which rapid and strong at numerical optimization. DE algorithm is proposed mainly for the numeric optimization problems and it is a population based algorithm like genetic algorithms using the similar operators; mutation, selection and crossover.

Why use Differential Evolution?

• DE is used to find the estimated solutions to the given problem.
• Global optimisation is required in the fields of engineering, statistics and finance[12].
• But many problems have some objective functions which are non-linear, noisy, flat.
• The problems which are not possible to solve analytically.

Evolutionary Algorithms

It includes Genetic Algorithms, Evolutionary Programming and Strategies.

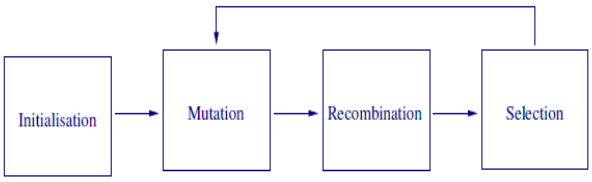

Fig 3: Evolutionary Algorithm Procedure

Notation:

Assume if we want to optimise a function with D1 real parameters then we have to select the size of population N1 (it must be at least 5) and the parameter vectors have the form:

$$X_{i,G} = [X_{1,i,G}, X_{2,i,G}, \ldots\ldots X_{D,i,G}] \quad i = 1,2,\ldots N1.$$

G ← generation number

a) Initialisation:

Identify lower and upper bound of each parameter and from that randomly select the parameter values ranging from $[xj^L, xj^U]$

$$x_j^L \leq x_{j,i,1} \leq x_j^U$$

b) Mutation:

• Each of the N1 parameter vectors will undergoes all the remaining three stages.(Initialisation, Selection and Recombination) and improves the search space. From the known vector Xi, G at random choose all the 3 vectors that Is (Xv1,G), (Xv2,G) and (Xv3,G) such that the indices i, v1, v2 and v3 are distinct.[13] and sum the difference of two vectors to the 3rd is

$$V_{i,G+1} = Xv_{1,G} + F(X_{v2,G} - Xv_{3,G})$$

• mutation factor → F is constant and the value is from 0 to 2. $V_{i,G+1}$ →donor vector.

c) Recombination:

It generates booming solutions from before iterations i.e generations. And we get the trial vector value Vi,G+1 from the current values of target vector xi and G is the element of donor vector Vi,G+1 and values of the donor vector enter the trial vector with the probability and it is CR.

$$u_{j,i,G+1} = \begin{cases} v_{j,i,G+1} & \text{if } \text{rand}_{j,i} \leq CR \text{ or } j = \text{I}_{\text{rand}} \\ x_{j,i,G} & \text{if } \text{rand}_{j,i} > CR \text{ and } j \neq \text{I}_{\text{rand}} \end{cases}$$

$$i = 1, 2, \ldots, N; \; j = 1, 2, \ldots, D$$

$$\text{rand}_{j,i} \sim U[0, 1]$$

I_{rand} → random integer ranging from [1, 2, ...,D]

- I_{rand} make sure that $V_{i,G+1} \neq X_{i,G}$

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

- End vector $X_{i,G}$ is compare with $V_{i,G+1}$ the one with low generated value is updated to the next iteration.

$$x_{i,G+1} = \begin{cases} u_{i,G+1} & \text{if } f(u_{i,G+1}) \le f(x_{i,G}) \\ x_{i,G} & \text{otherwise} \end{cases} \quad i = 1,2,\ldots,N$$

• The three steps will continue until stopping condition is reached

A major drawback of DE, like many of its other counter parts is its dependence on parameters. To overcome this drawback of Differential Evolution(DE).

III. COMPARISION OF DE AND PSO

The main goal of an optimization problem is to discover the better solution from given set of conditions. This approach fails to find out the global optima for multimodal optimization problems. This problem involves a fitness function, which needs to be increased or decreased mathematically. The problem is being recurrented as an intelligent search problem [14].This technique of searching is known as evolutionary based searching. The previous optimization techniques utilize derivative approach to find the optima. All evolutionary based algorithms follow a single approach for solving an optimization problem. It will initialize the population with all the vectors or particles and calculate the fitness function for each particle or vector and based on new generated best fitness value, update the position of the vectors or particles.

IV. CONCLUSION

This paper investigated the descriptive method of Data Mining task known as clustering for various benchmark datasets and one real world dataset namely Transportation dataset using evolutionary computation techniques. The focus is given to extract information through grouping of the data objects in the dataset. The investigation was started with traditional k-means algorithm and improving that by alleviating the difficulties seen in k-means. Few evolutionary computation techniques like PSO and DE have been used to address all research problems in this work. The clustering was viewed as an optimization problem of research with regard to finding intra cluster or/and inter cluster distance. Thus evolutionary computation techniques are used to address this issue. Thereafter the work focused was given to dynamic clustering, which can find optimal number of clusters at run time. A new method of dynamic clustering using adaptive DE has been proposed and the results are accurate when compared to other approaches.

V. FUTURE SCOPE

Although current work shows competitive clustering results for datasets under investigation, it remains to be seen how it behaves for high dimensional datasets. As a further study, dimensionality reduction techniques can be studied so that the proposed methods can become more result oriented.
Since PSO and DE inherently possess parallel characteristics, a suitable parallel algorithm may be evolved and tried for this problem. These parallel algorithms may be suitable for Multi processor/Distributed environments.

REFERENCES

[1] F. Heppner and U. Grenander. A stochastic nonlinear model for coordinated bird flocks. The Ubiquity of Chaos. AAAS Publications, Washington, DC, 1990.

[2] A.P. Engelbrecht. Fundamentals of Computational Swarm Intelligence. John Wiley & Sons, Chichester, UK, 2005.

[3] M. Clerc. Particle Swarm Optimization. ISTE, London, UK, 2006.

[4] R. Eberhart, M. Clerc and J. Kennedy The particle swarm-explosion, multidimensional complex space in IEEE Transactions on Evolutionary Computation, 6(1):58-73, 2002.

[5] J. Kennedy, and R. Eberhart. Swarm Intelligence. Morgan Kaufmann, San Francisco, CA, 2001.

[6] Liu, J.; Lampinen, J. (2005). "A fuzzy adaptive differential evolution algorithm". Soft Computing. http://en.wikipedia.org/wiki/Differential_evolution - cite_ref-qin05selfadaptive_13-0

[7] R. Mendes, J. Kennedy, and J. Neves. The fully informed particle swarm: simpler, maybe better. IEEE Transactions on Evolutionary Computation.

[8] Nowak, J. Szamrej, and B. Latane. From private attitude to public opinion: A dynamic theory of social impact. Psychological Review.

[9] Civicioglu, P. (2012). "Transforming geocentric cartesian coordinates to geodetic coordinates by using differential search algorithm". http:// en. wikipedia. org/ wiki/ Differential_evolution - cite_ref-brest 06 self adapting_15-0

[10] J. Kennedy and R. Eberhart. PSO In the Proceedings of IEEE International Conference on the Networks, pages 1944-1949, IEEE Press, Piscataway, NJ, 1995.

[11] J. Kennedy and R. Eberhart. A discrete binary version of the particle swarm algorithm. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, pages 4104-4108, IEEE Press, Piscataway, NJ, 1997.

[12] W. T. Reeves. Particle systems--A technique for modeling a class of fuzzy objects. ACM Transactions on Graphics.

[13] C. W. Reynolds. Flocks, herds, and schools: A distributed behavioral model. ACM Computer Graphics.

[14] R. Poli. Analysis of the publications on the applications of particle swarm optimisation. Journal of Artificial Evolution and Applications.