# GA Based Clustering of Mixed Data Type of Attributes (Numeric, Categorical, Ordinal, Binary and Ratio-Scaled)

**Rohit Rastogi[1], Pinki Mondal[2], Kritika Agarwal[3], Rachit Gupta[4]** and **Shilpi Jain[5]**

*Submitted in July 2013; Accepted in March, 2015*

***Abstract - Data mining discloses hidden, previously unknown, and potentially useful information from large amounts of data. As comparison to the traditional statistical and machine learning data analysis techniques, data mining emphasizes to provide a convenient and complete environment for the data analysis. Data mining has become a popular technology in analyzing complex data. Clustering is one of the data mining core techniques.***

***Data mining and data clustering, the prominent field of today it is a highly desirable task to apply unsupervised classification analysis on high volume of data sets with combined ordinal, ratio-scaled, binary and nominal with numeric, categorical, with values. However, most already available data merging and grouping through unsupervised classification algorithms are effective for the data with numeric category rather than the mixed data set. So,in this paper we have made efforts to present a new amalgamation techniques for these combined data sets by doing changes in the common cost function, and here we have tofindtraceof the internal cluster dispersion matrix.***

***To obtain correct clustering result the algorithm used is GA that optimizes the new cost function. We can compare and analyze that for high dimensional sets of data having mixed attributes GA-based clustering algorithm is feasible.***

***Core Idea of Our Paper***

***By this paper, we try to describe a technique for estimating the cost function metrics from mixed numeric, categorical and other type databases by using a uncertain grade-of-membership clustering model with the efficiency of Genetic Algorithm. This technique can be applied to the problem of opportunity analysis for business decision-making.***

***This general approach could be adapted to many other applications where a decision agent needs to assess the value of items from a set of opportunities with respect to a reference set representing its business. For processing numeric attributes, instead of generalizing them, a prototype may be developed for experiments with synthetic and real data sets,***

***and comparison with those of the traditional approaches. The results confirmed the feasibility of the framework and the superiority of the extended techniques.***

*[1]Sr. Asst Professor, CSE-Dept-ABES Engg. College, Ghaziabad (U.P.), India, +91-9818992772*

*[2,3,5]B.Tech.CSE-IIIYr., CSE-Dept.-ABES Engineering College, Ghaziabad (U.P.), India*

*[4]B.Tech. IT-Final Yr., IT-Dept-ABES Engg College, Ghaziabad (U.P.), India*

***Index Terms - Clustering algorithms, categorical dataset, numerical dataset, clustering, data mining, pattern discovery, genetic algorithm.***

## 1.0 INTRODUCTION

The basic operation in Data Mining is partitioning of sets of objects present in the database into homogenous clusters or groups is the basic work in data mining. The beneficial way in numerous tasks likeclustering, image processing, sequence analysis, market research, pattern recognition, spatial analysis, economics etc. To implement the operation of partitioning clustering is the most widely used approach. This technique partitions the sets of objects into unsupervised classifiers in such a way that the objects contained in the common cluster are more similar to each other than objects in indifferent clusters.

Data mining and warehousing differs from other traditional applications and analysis of clusters in such a way that it deals with large high dimensional data.According to this attribute, manyunsupervised classification algorithms are discontinued to beused. One more characteristic is that data mining data often contains all types of mixed attributes in real life practical applications. The traditional method to handle categorical, nominal, ratio-scaled or ordinal attributes as numeric with the help of dissimilarity matrices (after calculating Euclidean/Manhattan/Minkowaski distances and applying normalization (standard deviation/ Z-score or min-max normalization on the results) and applying the related algorithms for numeric values, but due to the unordering of many categorical domains it does not always yield useful and meaningful results.

Many already available unsupervised classification algorithms can handle either only numeric attributes or both data types but not efficient when clustering is performed on large sets of data. Few algorithms can perform both well, such as k-prototypes etc.

We give a new cost function for clustering to process large sets of data with mixed numeric and categorical and other values by doing changes in the common used trace of within cluster dispersion matrix. In clustering process, we introduce genetic algorithm(GA) so that the cost function can be minimized. The benefit of high search efficiency is achieved in GA as GA uses search strategy globally and also implements in parallel.

The remaining paper is organized as follows: Some mathematical preliminaries of the algorithm are included in the next section. Then GA is briefly discussed with modified and efficient cost function for all the data sets. In last section there are summaries the discussions.

## 2.0 BETTERMENT BY THE USE OF GENETIC ALGORITHM

With the basic features of GA like encoding, crossover, mutation, appropriate fitness function and reproduction with survivor selection, the GA can be able to design better clustering and unsupervised classification operations.

The proposed approach can be described with experiments and their results. The algorithm can be run on real-life datasets to test its clustering performance against other algorithms. At the same time, its properties are also empirically studied. One observation from the above analysis is that our algorithm's computation complexity is determined by the component clustering algorithms. So far, many efficient clustering algorithms for large databases are available, it implicate that our algorithms will effective for large-scale data mining applications, too.

## 3.0 COST FUNCTION, INITIAL POPULATION AND SELECTION

**Initial Population**

The size of the initial population is an important issue because a large population can effectively sample the parameter space. However the larger the population, the higher the computational cost. A compromise must be found. An interesting means to both have an effective sampling and a reasonable computational cost is to decrease the population after the first iteration of the process. For instance, take an initial population of 1000 chromosomes, then choose the 500 better chromosomes and work with this population's size until the end.

**Cost Function**

The cost function represents the problem we want to solve. For instance, the cost function of the well-known traveling salesman is the distance the salesman has to cover to visit all the towns exactly once. Most search problems may be posed as the search for the optimal value of a function satisfying a set of constraints. The function represents the relationships between the different parameters which we seek to optimize. When those relationships are well-defined and simple enough to be modeled mathematically (e.g. by convex functions), the analytical methods (e.g. Lagrange multipliers) of mathematics should be applied to the problem. When those relationships are so complex as to appear unpredictable or random, the model itself may be ill-posed and only random or exhaustive search offers any hope of an answer. Genetic algorithms are well suited for the real-world problems which lie between these two extremes.

**Selection**

We have now a set of chromosomes. In order to enhance the average fitness of the population, we will generate a new population from the previous one according to the quality of each chromosome: the higher fitness value of a chromosome, the higher its probability to be included in the new generation. The standard way to do this is called the casino roulette method:

## 4.0 COST FUNCTION FOR NUMERIC DATA CLUSTERING

The trace of the within cluster dispersion matrix is the widely used cost function.The cost function is defined as:

$$C(W) = \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij}^2 (\mathrm{d}(x_j, x_i))^2 \ , w_{ij} \epsilon \{0,1\} \quad (1)$$

Here, $w_{ij}$ is the degree of membership of $x_j$ belonging to cluster i. **W** is a $k \times n$ order partition matrix. The function d(.) is a measure of dissimilarity often defined as the Euclidean distance. For data set having real attributes, i.e., X $\subset R^m$, we have

$$\mathrm{d}(x_j, x_i) = \left(\sum_{l=1}^{m} | x_{jl} - x_{il} | 2\right)½ \quad (2)$$

Since, $w_{ij}$ indicates $x_j$ belonging to cluster i, and $w_{ij} \epsilon [0,1]$, we call **W** to be a hard k-partition.

## 5.0 COST FUNCTION FOR MIXED DATA CLUSTERING

**5.1Max-Min Normalization for numeric data**

For clustering the numeric data, first we will normalize numeric data so as to prevent the dominance of particular attribute. For which the normalization formula is as follows:-

$$\mathrm{n_i} = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \times (Rh - Rl) + Rl \quad (3)$$

Where, $x_i$ is the i-thobject.Rh and Rlare the higher and lower ranges respectively. N is the new normalized matrix containing all types of data.

**5.2 Normalizing ratio-scaledvalues:-**

First, we will take log of the ratio-scaled values, given as

$$f(n) = \log(n) \quad (4)$$

**5.3Normalizing ordinal values:-**

First we assign ranks to the values as, better the value higher the rank and vice versa. Now, based on their ranks we will normalize them. Give 1 to the highest rank and 0 to the lowest one and other ranks get the value as:

$$\prec(\mathrm{r}) = \frac{1}{no.of different ordinal values - 1} \times (\mathrm{r}\text{-}1) \quad (5)$$

**5.4 Normalizing categorical values:-**

If the two values match put value 1 and otherwise 0.

$$\delta(a, b) = \begin{cases} 0 & a = b \\ 1 & a \neq b \end{cases} \quad (6)$$

## 6.0 RE-DEFINING COST FUNCTION

When Nhas attributes with numeric and mixed values, assuming that eachobject is denoted by $n_i$=[ $n_{i1}^r, \ldots, n_{it}^r, n_{i,t+1}^c, \ldots\ldots, n_{im}^c, n_{i,m+1}^b, \ldots\ldots, n_{i,y}^b, n_{i,y+1}^o, \ldots \ldots,$ $n_{iu}^o, n_{i,q+1}^{rs}, \ldots \ldots, n_{is}^{rs}$,],the dissimilarity between two mixed-typeobjects $n_i$and $n_j$can be measured by the following Eq.(7)

$$d(n_i, n_j) = [\left(\sum_{l=1}^{t} |n_{il}^r - n_{jl}^r|)2 + \prec_1 . (\sum_{l=t+1}^{m} |n_{il}^c , n_{jl}^c|)2 + \prec_2 . \left(\sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b \right|)2 + \prec_3 . (\sum_{l=y+1}^{u} |n_{il}^o - n_{il}^o \Big|\Big) 2\Big) + \prec_4 . \sum_{l=y+1}^{u} |n_{il}^{rs} - n_{il}^{rs}|2)]½$$

Whereall the terms are squared Euclidean distance measure on the mixed attributes.

Using Eq. (7) for mixed-type objects, we can modify the cost function of Eq. (1) for mixeddata clustering. In addition, modifythecost function to extend the hard k-partitioning as:

$$C(W) = \sum_{l=1}^{K}\left(\sum_{j=1}^{n} w_{ij}^2 \sum_{l=1}^{t} | x_{jl}^r - p_{il}^r|2 + \prec_1 \sum_{j=1}^{n} w_{ij}^2 \sum_{l=t+1}^{m} |x_{jl}^c - p_{il}^c|2 + \prec_2 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b |2 + \prec_3 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=y+1}^{u} |n_{il}^o - n_{il}^o |2) + \prec_4 . \sum_{l=y+1}^{u} |n_{il}^{rs} - n_{il}^{rs} |2), \quad w_{ij} \varepsilon [0,1] \quad (8)$$

Let $C_i^r = \sum_{j=1}^{n} w_{ij}^2 \sum_{l=1}^{t} | x_{jl}^r - p_{il}^r | 2$

$$C_i^c = \prec_1 \sum_{j=1}^{n} w_{ij}^2 \sum_{l=t+1}^{m} |x_{jl}^c - p_{il}^c |2 C_i^b$$

$$= \prec_2 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b |2 \; C_i^o$$

$$= \prec_3 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=y+1}^{u} |n_{il}^o - n_{il}^o |2$$

(9)

We rewrite Eq.(8) as:

$$C(W) = \sum_{i=1}^{k} (C_i^r + C_i^c + C_i^b + C_i^o) \qquad (10)$$

## 7.0 GA-BASED CLUSTERING ALGORITHM FOR MIXED DATA

Clustering is a fundamental and widely applied method in understanding and exploring a data set. Interest in clustering has increased recently due to the emergence of several new areas of applications including data mining, bioinformatics, web use data analysis, image analysis etc. To enhance the performance of clustering algorithms, Genetic Algorithms (GAs) is applied to the clustering algorithm. GAs are the best-known evolutionary techniques. The capability of GAs is applied to evolve the proper number of clusters and to provide appropriate clustering. This paper present some existing GAbased clustering algorithms and their application to different problems and domains.

## 8.0 GENETIC ALGORITHM

In the field of artificial intelligence, a genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a meta heuristic) is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. Genetic algorithms find application in bioinformatics, phylogenetics, computational,science, engine ering, economics, chemistry,manufacturing, mathematics, physi cs, pharmacometrics and other fields.

## 9.0 GA-BASED CLUSTERING ALGORITHM

### 9.1 Algorithm:

Step1. Begin
Step2. Define pop-size as desired population size
Step3.Randomly initializes pop-size population
Step4.While (Ideal best found or certain number of generations met)
O Evaluate fitness
O While (number of children=population size)
O Select parents
O Apply evolutionary operators to create children
O End while
Step5. End While
Step6. Return Best solution
Step7. End

First of all, the following three problems should be solved to employ GA:
(1) Encoding of the clustering solution into the gene string.
(2) Designing of a reasonable fitness function.
(3) Selection or designing genetic operators including their parameters that guarantees fast convergence.

**9.2 Encoding:** From Eq.(1) and (8), it is well known that the purpose of clustering is to obtain a (fuzzy) partition matrix **W.** Then using the fitness function (stated below) we can improve the chances of a particular data point to be chosen.Then after selecting that particular cluster we can further subdivide the data points in the cluster, based on their fitness values.
Note that since we process data having mixed attributes, in parallel to numeric parameter mixed parameters are also there in gene string. Therefore, not ordered for binary attributes and can be directly encoded rather than doing normalization first.

**9.3 Fitness function**: We are taking the fitness function such that fitness value is inversely proportional to the cost function value, i.e., the fuzzy clustering partition is better when the cost function is smaller. So GA asks for a larger fitness value. Hence, fitness function is defined with the use of clustering

cost function. Exponentially increased cost function will sharply reduce the fitness function.

$f(g) = \frac{1}{1+e^{C(W)}}$(11)

**9.4 Genetic operators**: A genetic operator is an operator used in genetic algorithms to maintain genetic diversity, known as mutation and to combine existing solutions into others, crossover. The main difference between them is that the mutation operators operate on one chromosome, that is, they are unary, while the crossover operators are binary operators. Genetic variation is a necessity for the process of evolution. Genetic operators used in genetic algorithms are analogous to those in the natural world: survival of the fittest, orselection; reproduction (crossover, also called recombination); and mutation.

**Types of Operators**

1. Selection (genetic algorithm)
2. Crossover (genetic algorithm)
3. Mutation (genetic algorithm) and selection probability is:

$P_s(g_{(i)}) = \frac{f(g_i)}{\sum_{i=1}^{n} f(g_i)}$ (12)

Operation probabilities for crossover and mutation are assigned as Eq. (13)

$$P_c(g_i, g_j) = \begin{cases} \frac{\alpha_1(f_{max} - f')}{f_{max} - \bar{f}} f' \geq \bar{f} \\ \alpha_2 otherwise \end{cases}$$

$$P_m(g_i) = \begin{cases} \frac{\alpha_3(f_{max} - f(g_i))}{f_{max} - \bar{f}} f(g_i) \geq \bar{f} \\ \alpha_4 otherwise \end{cases} \quad (14)$$

where, $f_{max} = max_{l=1}^{N}\{f(g_l)\}$

$\bar{f} = \sum_{l=1}^{N} f(g_l)$, $f' =, f(g_j)\}$,and$\alpha_i \epsilon [0, 1]$

Apart from the operators mentioned above, a new operator for the clustering algorithm is defined.

Gradient operator: Changes in the existing weights are done as per the formula:

It includes two steps iteration as:

$$w_{ij} = \sum_{l=1}^{k} \frac{(d(x_j, x_i))2}{(d(x_j, x_i))^2}, i, \text{j} \quad (15)$$

**10.0 A REAL-LIFE PRACTICAL SAMPLE DATA TABLE OF MIXED DATA TYPES**

We are representing the real life concept of our approach by taking the data of 5 employess working in a company. Here we will use every kindof data (related to all data types) to show that our method works for every kind of data.In this example :

**We are taking the weighted matrix ($W_{ij}$) as:**

**Table 1**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **1** | **0** | | | | |
| **2** | **0.4** | **0** | | | |
| **3** | **0.2** | **0.2** | **0** | | |
| **4** | **0.1** | **0.3** | **0.2** | **0** | |
| **5** | **0.5** | **0.2** | **0.1** | **0.4** | **0** |

Test-1 cantains salary of an employee (numeric data)

Test-2 shows whether the employee is male or female(binary data- Male=1/ Female=0)

Test-3 shows the department to which employee belong (categorical data)

Test-4 depicts the ability of an employee (ordinal values) Exc.-Excellent, Fair or Good

Test-5 shows avg. credit points alloted according totheir performance (ratio-scaled values)

Last Column shows the log value of ratio-scaled data type.

**Table 2**

| Object-id | Test-1 | Test-2 | Test-3 | Test-4 | Test-5 | Log |
|---|---|---|---|---|---|---|
| **1** | **25K** | **M** | **Code-A** | **Exc.** | **445** | **2.65** |
| **2** | **40K** | **F** | **Code-B** | **Fair** | **22** | **1.34** |
| **3** | **55K** | **M** | **Code-C** | **Good** | **164** | **2.21** |
| **4** | **27K** | **M** | **Code-A** | **Exc.** | **1210** | **3.08** |
| **5** | **53K** | **F** | **Code-B** | **Fair** | **38** | **1.58** |

**The Table 2 is converted into the normalized matrix using the above equations.(3),Eq.(4),Eq.(5),Eq.(6)**

**Table 3**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **1** | **0** | **0** | **0** | **0** | **0** |
| **2** | **0.5** | **1** | **1** | **1** | **1.31** |
| **3** | **1** | **0** | **1** | **0.5** | **0.44** |
| **4** | **0.0666** | **0** | **0** | **0** | **0.43** |
| **5** | **0.9333** | **1** | **1** | **1** | **1.07** |

**We calculate the value of the expression (stated below) to be further used in Eq. (8)**

$$\sum_{j=1}^{n} d(x_j, x_i)$$

**Table 4**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **1** | **0** | | | | |
| **2** | **4.9961** | **0** | | | |

| | | | | | |
|---|---|---|---|---|---|
| **3** | **2.4436** | **2.2569** | **0** | | |
| **4** | **0.1893** | **3.962** | **2.1213** | **0** | |
| **5** | **5.0159** | **0.2453** | **1.6153** | **4.1607** | **0** |

**Now for Eq. (8) we are calculating the value of the expression:**

$$\sum_{j=1}^{n} w_{ij}^{2} \left(d(x_j, x_i)\right) 2$$

**Table 5**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **1** | **0** | | | | |
| **2** | **0.0799** | **0** | | | |
| **3** | **0.0977** | **0.0903** | **0** | | |
| **4** | **0.0018** | **0.3566** | **0.0848** | **0** | |
| **5** | **1.2539** | **0.0098** | **0.0165** | **0.6571** | **0** |

**Now we will calculate the expression:**

$$\sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij}^{2} \left(d(x_j, x_i)\right) 2$$

**Table 6**

| | |
|---|---|
| **1** | **1.3455** |
| **2** | **0.5366** |
| **3** | **0.2013** |
| **4** | **1.1063** |
| **5** | **1.9373** |

**Now using the Eq. (11) we find the fitness value of the above calculated values (above 5 tuples):**

**Table 7**

| | |
|---|---|
| **1** | **0.2066** |
| **2** | **0.3689** |
| **3** | **0.4498** |
| **4** | **0.2497** |
| **5** | **0.1259** |

**First we arrange the above values in ascending order and label each one of them and then using Eq.(12) calculate the selection probability** $P_s(g_{(i)})$

**Table 8**

| | |
|---|---|
| **1** | **0.1474** |
| **2** | **0.2633** |
| **3** | **0.3204** |
| **4** | **0.1782** |
| **5** | **0.0898** |

## 11.0 ANALYSIS ON OUR EXPERIMENTAL RESULTS

By the above calculated tables, we can easily verify the dissimilarity matrices of our real life experimental data shown in tabular structure,

We can comfortablydecide the set of clusters based on the fitness values.We are taking the threshold value for our method to be 0.22. Data item 1 and 5whose fitness value lie below the threshold value can be grouped together in the cluster and the other three tuples can be grouped in another.

Now these clusters can be improved using GA and using the selection probability.

**Result:**

So there can be two clusters:

**C1:- data items 1 and 5.**

**C2:- data items 2,3 and 4**

## 12.0 CONCLUSION

Here, to cluster large sets of data we have presented GA and the performance can be evaluated using large data sets of data. The proposed outcomes can be used to demonstrate the importanceof the algorithms in finding structures in data.

This paper puts an emphasis on the issue that uses the GA to solve the clustering problem. Though the application is specific for the business, our approach is general purpose and could be used with a variety of mixed-type databases or spreadsheets with categorical, numeric and other data values, and temporal information. With improved metrics, artificial intelligence algorithms and decision analysis tools canyield more meaningful results and agents can make better decisions.

This approach, then, can ultimately lead to vastly improved decision-making and coordinating among business units and agents alike. If a class attribute is involved in the data, relevance analysis between the class attribute and the others (or feature selection) should be performed before training to ensure the quality of cluster analysis. Moreover, most variants of the GA use Euclidean-based distance metrics. It is interesting to investigate other possible metrics like the Manhattan distance or Cosine-correlation in the future. To faithfully preserve the topological structure of the mixed data on the trained map, we integrate distance hierarchy with GA for expressing the distance of categorical values reasonably.

## 13.0 ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their valuable suggestions. They would also like to thank Prof. A.K. Sinha

(Dean CRAP-ABES-Engineering College, Ghaziabad) for his involvement and valuable suggestions on soft-computing in the early stage of this paper.

**REFERENCES**

[1]. LI Jie, GAO Xinbo, JIAO Li-cheng, "A GA-Based Clustering Algorithm for Large Data Sets withMixedNumeric and Categorical Values",National Key Lab. of Radar Signal Processing, Xidian Univ., Xi'an 710071, China

[2]. M. R. Anderberg. Cluster Analysis for Applications. Academic Press, New York,1973.

[3]. B. Everitt. Cluster Analysis. Heinemann Educational Books Ltd., 1974.

[4]. Zhexue Huang, Michael K.Ng. A fuzzy *k*-modes algorithm for clustering categorical data. IEEE Trans. on Fuzzy Systems, 7(4): 446-452, August, 1999.

[5]. Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data Mining. Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Dept. of Computer Science, the University of British Columbia, Canada, pp.1-8.

[6]. R. Krovi. Genetic Algorithm for Clustering: A Preliminary Investigation. IEEE press, Pp.504-544.

[7]. J. H. Holland. Adoption in Natural and Artificial System. Ann Arbor, MI: Univ. Mich. Press, 1975.

**TECHNICAL PROFILE**

Mr. Rohit Rastogi received his B.E. degree in Computer Science and Engineering from C.C.S.Univ. Meerut in 2003, the M.E. degree in Computer Science from NITTTR-Chandigarh (National Institute of Technical Teachers Training and Research-affiliated to MHRD, Govt. of India), Punjab Univ. Chandigarh in 2010.

He was Asst. Professor at IMS College, Ghaziabad in computer Sc. Dept. His research interests include Data ware Housing and Data Mining, Design Analysis of Algorithm, Theory of Computation & Formal Languages and Data Bases.

He is a Sr. Asst. Professor of CSE Dept. in ABES Engineering. College, Ghaziabad (U.P.-India), affiliated to Gautam Buddha Tech. University and Mahamaya Tech. University (earlier Uttar Pradesh Tech. University) at present and is engaged in Clustering of Mixed Variety of Data and Attributes with real life application applied by Genetic Algorithm, Pattern Recognition and Artificial Intelligence.

He has served as the technical reviewer of 7 papers in IIIrd International Conference on Computing, Communications and Informatics (IC3'2014) at GCET, Greater Noida, NOIDA, India on September, 24-27, 2014 And Worked as the reviewer for the SPICES-2015 at NIT Kerala, Kojhicode for international conf. of Signal Processing and Communication…Currently working as the reviewer in the technical reviewer committee for the INDIA-2015 is Second International Conference on Information System Design and Intelligent Applications organized by Faculty of Engineering, Technology and Management, University of Kalyani, Kalyani-741235, West Bengal, India.

Currently designated reviewer on the technical program committee for the International Conference on Computing in Mechanical Engineering (ICCME-2015) (ICCME-2015). The proceedings of ICCME'15 will be published by Springer as a special volume in the Lecture Notes in Mechanical Engineering (ISSN: 2195-4356). All accepted papers will also be archived in the SpringerLink digital Library.He is UGC-NET -2014 qualified.

He has mentored around 20 Live Projects in Digital Logic Design at Graduation level like Automatic street Light Controller, Darkness detector, Visitor counter and Car Parking system etc.

He is CSI-student Coordinator of ABES-EC CSI student Chapter and life member of ISTE.

He keeps himself engaged in various competitive events, activities, webinars, seminars, workshops, projects and various other teaching Learning forums.

He has been awarded in different categories by ABES-EC, Gzb. College management for improved teaching, significant contribution, human value promotions and long service etc.

He has authored/co-authored, participated and presented research papers in various Science and Management areas in around 40 International Journals and International conferences including prestigious IEEE and Springer and 10 national conferences including SRM Univ., Amity Univ. and Bharti Vidyapeetha etc. He has guided five ME students in their thesis work and students of UG and PG in around 100 research papers. He has developed many commercial applications and projects and supervised around 30 B.E. students at graduation level projects.

His research interests include Data ware Housing and Data Mining, Design Analysis of Algorithm, Theory of Computation & Formal Languages and Data Bases. At present, He is engaged in Clustering of Mixed Variety of Data and Attributes with real life application applied by Genetic Algorithm, Pattern Recognition and Artificial Intelligence.

Also, He is preparing some interesting algorithms on Swarm Intelligence approaches like PSO, ACO and BCO etc..