Dynamic and Optimized Prototype Clustering for Relational Data based on Multiple Prototypes

A.V. Suryanarayana Raju, Bharath Siva Varma,

***Abstract:** In artificial intelligence related applications such as bio-medical, bio-informatics, data clustering is an important and complex task with different situations. Prototype based clustering is the reasonable and simplicity to describe and evaluate data which can be treated as non-vertical representation of relational data. Because of Barycentric space present in prototype clustering, maintain and update the structure of the cluster with different data points is still challenging task for different data points in bio-medical relational data. So that in this paper we propose and introduce A Novel Optimized Evidential C-Medoids (NOEC) which is relates to family o prototype based clustering approach for update and proximity of medical relational data. We use Ant Colony Optimization approach to enable the services of similarity with different features for relational update cluster medical data. Perform our approach on different bio-medical related synthetic data sets. Experimental results of proposed approach give better and efficient results with comparison of different parameters in terms of accuracy and time with processing of medical relational data sets.*

***Index Terms:** Data clustering, multiple prototypes, artificial intelligence, and prototype based clustering, c-medoids and Ant Colony Optimization (ACO).*

I. INTRODUCTION

Clustering is domain which forms the various clusters based on the dataset and with various methodologies. According to the different methodologies the dataset is selected and arrange the model for processing of datasets for quality of clusters. These models are new articles in the portrayal space. Every model typically speaks to a bunch of items. The principle preferences of model based strategies are that they give a natural summarization of the given information in couple of prototypical examples also, in this way lead to conceivable and interpretable group structures. What's more, they have a low computational intricacy, as a rule in O(N K), with N the quantities of items in the informational collection what's more, K the quantity of models. This low intricacy alone clarifies the prominence of model based methodology in real life applications. The most utilized model based calculations are the K-implies calculation and its varieties (for example K-means++ [2], K-medoids [3], fuzzy C-means [4]), just as the group of Unsupervised Neural Network methodologies, for example, Self- Sorting out Map [5], Neural Gas [6] or Boltzmann Machines [7]. In the event that the objects of a dataset are depicted in vectorial shapes, the meaning of group's models is clear. All things considered a model is a vector characterized in the equivalent vectorial space, generally characterized as the vectorial barycentre of the articles (vectors) having a place with its bunch. Be that as it may, in many case the articles can't be effectively characterized in a vectorial space without lost data or potentially an expensive pre-processing (for example pictures, systems, arrangements, writings). To break down such non-vectorial datasets, it isn't unexpected to portray the information utilizing the relations or the similitude’s between the items, utilizing a uniqueness or separation grid. Therefore, they are once in a while called Relational Data. Social grouping calculations structure a group of strategies adjusted to social information. Some clustering calculations are normally adjusted to manage difference network and can be utilized to dissect social datasets. None of these calculations use models and they doesn't profit by the related favorable circumstances. Specifically, they all have a non-straight computational multifaceted nature. Since the limit between bunches in certifiable informational collections normally covers, delicate grouping strategies, for example, fluffy clustering, are more reasonable than hard grouping for certifiable applications in information examination. So that in this paper, we propose and introduce A Novel Optimized Evidential C-Medoids (NOEC) which is relates to family o prototype based clustering approach for update and proximity of medical relational data. We use Ant Colony Optimization approach to enable the services of similarity with different features for relational update cluster medical data. Perform our approach on different bio-medical related synthetic data sets. Real time evaluation of proposed approach gives better and efficient results with respect to different medical data sets.

Manuscript published on 30 September 2019

* Correspondence Author
A.V. Suryanarayana Raju*, M. Tech(CST) in department of CSE, S.R.K.R Engineering College, AP, India.
Bharath Siva Varma, Assistant Professor in the department of CSE in S.R.K.R. Engineering College, AP, India

II. REVIEW OF LITERATURE

This section describe with different authors opinion regarding the implementation of medical relational data issues using clustering algorithms.Clustering methods or techniques are widely used in medical industries and implementing on medical datasets. It is a very easy process to diagnose any type of disease from various types of patient reports. This will provide fast, adequate, reliable and less costly healthcare delivery to patients. The authors described how they compare the performance of the three clustering algorithms on heart disease. To improve the performance they adopted the Silhouette width measure. The performance is improved and CLARA clustering shows better performance compare with the existing ones such as K-means and PAM.All the experiments are limited to dividing clustering algorithms according to their functionalities. For some algorithms, the user has to select the number of clusters which may lead to the incorrect and unmatched results based on the selected dataset.

Among the various clustering algorithms, Hierarchical andDensity-based clustering (DBC) selects the number of clusters dynamically I,e by themselves. The authors improved the clustering algorithms with the data points from the various types of groups and also compare the similar items compare with other groups. One more issue in clustering algorithms is the processing of huge medical and normal datasets. This will take more time compared with traditional clustering algorithms. Many machine learning and deep learning algorithms are available to improve the performance of the clustering algorithms. Processing of issue-based clustering algorithms is complicated to get the accurate results to overcome this various clustering algorithm adopted with big data and map reduce algorithms.The authors proposed the multi-ant colony method for processing of clustering algorithms which consists of random and individual ant colonies and a queen ant agent. Every ant in this will perform differently with their moving speed. To improve the results of ant colonies a hypergraph is adopted. Kuo et al. [11] described the enhanced algorithm called ant K-means (A) algorithm.

III. BACKGROUND WORK

A few sorts of information can't be depicted as vectorial information with known qualities. These items can speak to basically anything, for example, Tweets, vehicles, groupings of protein, music scores, and so on. An informational index $O=\left\{o^1,o^2,.......,o^N\right\}$ is then usually spoken to by a social lattice $R=[relation(o^i,o^j)]$ with $1\le i,j\le N$. The social lattice frequently appears as a uniqueness grid D, where the qualities can be deciphered as a divergence or a separation d between articles. Little qualities speak to comparative information and the other way around. The insignificant imperatives on a difference measure $d:(i,j)\longrightarrow d(o^i,o^j)$ are given by the separation properties: non-pessimism, symmetry and reflexivity. In this way, a divergence network D for a N-components informational index is: square, symmetric, non-negative and empty (for example d(i; i) = 0 for all i). Note that d need not fulfill the triangle imbalance. In this paper, D is no required to be a lattice dependent on the Euclidean separation. We think about that the informational index O incorporates objects oi from an (obscure) d-dimensional pseudo-Euclidean info space E∗, oi has no more a vectorial portrayal.

Relational k-means clustering (RKMA):

RKMA proposes the dynamic output to define the prototypes for relational data representation such as dissimilarity matrix D. The linear combination is defined by the prototypes k and objects oi are initialize in place of a vector in the data space:

$$\mu^k\sum_{i=1}^{N}\alpha_i^k.o^i, with\sum_{i=1}^{N}\alpha_i^k=1$$

To describe object dissimilarity with different relational objects.

Algorithm 2 RKMA

```
Step: 1 Initialize D, K
Step: 2 allocate every item to a cluster randomly
Step: 3 compute the $\alpha^{x}$ using 5
Step: 4 while the convergence is not attained do
Step: 5 allocate every item to its nearest prototype using 3
Step: 6 Update $\alpha^{x}$ using (5)
End while
```

Algorithm 1 Standard formulation relates to different object relations.

A standard definition would give us the accompanying calculation: With this methodology, the meaning of the models is exceptionally exact and we acquire a decent portrayal of the information structure. Be that as it may, the models are depicted by a vector of coefficients with N esteems. As each article must be analyzed to every model in each progression, the computational multifaceted nature is in any event in O (N2), which is generally unreasonably moderate for generally present day applications.

IV. PROPOSED SYSTEM DESIGN IMPLEMENTATION

Basic implementation procedure of the proposed approach discussed in this section, we also presents novel evidential c-mediods approach with multiple weights medoids. This approach computes weights based on medoids membership degree of different objects relates to specific class labels with respect to dissimilar objects.

4.1. Basic Preliminary Functions

Main objective function of proposed approach let us consider $X=\left\{x_i\,|\,i=1,2,.....,n\right\}$ be the different objects $\tau(x_i,x_j)\triangleq\tau_{ij}$ with respect to dissimilar objects x_i and x_j. Pair wise communication with dissimilar for analyzing data set. Objective function with dissimilar objects

$$J_{NOEC}(M,V)=\sum_{i=1}^{n}\sum_{A_j\subset\Omega,A_j\neq\phi}|A_j|^{\alpha}m_{ij}^{\beta}d_{ij}+\sum_{i=1}\delta^2m_{i\phi}^{\beta}$$

Multi objective weight measure functions with respect to dissimilar objects with different forms with specific labels. For multi objective function, let us consider $V^{\Omega}=\left\{v_{ki}^{\Omega}\right\}_{c\times n}$ be the multiple objects with specific class labels. Dissimilar multiple weight objects x_i with cluster $A_j=\left\{w_k\right\}$ would like that as follows:

$$d(x_i,A_j)\triangleq d_{ij}=\sum_{i=1}^{n}(v_{kl}^{\Omega})^{\psi}\tau(i,l)$$

Parameter controls the smoothness of the dissemination of model loads. The loads of uncertain class Aj (jAj j > 1) can be inferred by the included specific classes. In the event that item xi has comparable loads for specific classes !m and !n, it is most likely that xi lies in the covering territory between two classes. Along these lines the change of the loads of article xi for all the included specific classes of Aj ,Varji, could be utilized to express the loads of xi for Aj (meant by v2 ji , and V is utilized to signify the comparing weight matrix2). The littler Varji is, the higher v2ji is. Be that as it may, we should focus on the anomalies. They may hold comparable little loads for each specific class, however have no commitment to the loose classes by any means. The base of xi's loads for all the related specific classes could be mulled over to recognize the exceptions. Based on weight measure functions similar proportional functions, i.e

$$v_{ji}^{2^{\Omega}} = \frac{[\min\left(\left\{v_{ki}^{\Omega}; w_k \in A_j\right\}\right)] / Var\left(\left\{\left\{v_{ki}^{\Omega}; w_k \in A_j\right\}\right\}\right)}{\sum_t [\min\left(\left\{v_{ki}^{\Omega}; w_k \in A_j\right\}\right)] / Var\left(\left\{\left\{v_{ki}^{\Omega}; w_k \in A_j\right\}\right\}\right)}$$

Dissimilar specific objects impressive class labels could be as follows:

$$d_{ij} = \sum_{l=1}^{n} \left(v_{jl}^{2^{\Omega}}\right)^{\psi} \tau(i,l), A_j \subseteq \Omega, A_j \neq \phi$$

4.2. Ant Colony Optimization

The issue of finding ideal group assignments of items and agents of classes is presently defined as a compelled advancement issue, for example to find ideal estimations of M and V subject to a lot of obliges. As in the past, the strategy for Lagrange multipliers could be used to infer the arrangements. The Lagrangian capacity is built as

$$L_{NOEC} = J_{NOEC} - \sum_{i=1}^{n} \lambda_i \left(\sum_{A_j \subseteq \Omega, A_j \neq \phi} m_{ij} - 1 \right) - \sum_{k=1}^{c} \beta_k \left(\sum_{i=1}^{n} v_{ki}^{\Omega} - 1 \right)$$

where i and k are Lagrange multipliers. By computing the first request incomplete subordinates of LNOEC as for mij ,vki, i and k and letting them to be 0, the update conditions of mij and vki could be determined. It is anything but difficult to get that the update conditions for mij are equivalent to the use of SNOEC, then again, actually for this situation dij ought to be determined with target work. The update procedure for the model loads vki is difficult to get since it is a non-straight streamlining issue. Some specifically methods might be received to take care of this issue. Here we utilize a straightforward estimate plan to refresh vki. Basic representation of ant colony optimization as follows:

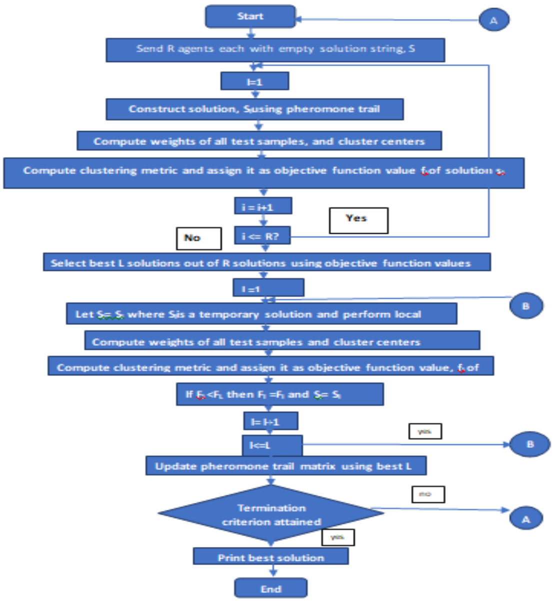
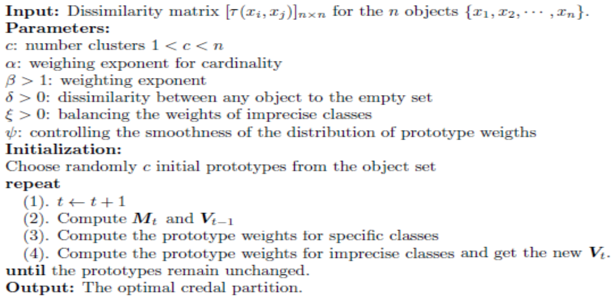

Start
A
Send R agents each with empty solution string, S
I=1
Construct solution, S using pheromone trail
Compute weights of all test samples, and cluster centers
Compute clustering metric and assign it as objective function value f of solution s
i = i+1
Yes
No
i <= R?
Select best L solutions out of R solutions using objective function values
I =1
B
Let S= S where S is a temporary solution and perform local
Compute weights of all test samples and cluster centers
Compute clustering metric and assign it as objective function value, f of
If F <F then F =F and S= S
I= I+1
yes
I<=L
B
Update pheromone trail matrix using best L
no
Termination criterion attained
A
yes
Print best solution
End

Figure 1 Ant colony optimization with respect to multiple objective weights.

Basic algorithmic procedure for proposed approach with multi attributes weights to be represented as matrix weight V.



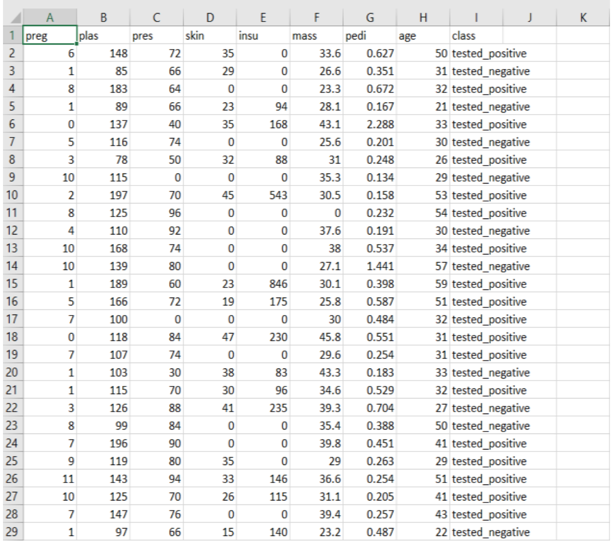
```
Input: Dissimilarity matrix $[\tau(x_i, x_j)]_{n\times n}$ for the $n$ objects $\{x_1, x_2, \cdots, x_n\}$.
Parameters:
$c$: number clusters $1 < c < n$
$\alpha$: weighing exponent for cardinality
$\beta > 1$: weighting exponent
$\delta > 0$: dissimilarity between any object to the empty set
$\xi > 0$: balancing the weights of imprecise classes
$\psi$: controlling the smoothness of the distribution of prototype weigths
Initialization:
Choose randomly $c$ initial prototypes from the object set
repeat
   (1). $t \leftarrow t + 1$
   (2). Compute $\boldsymbol{M}_t$ and $\boldsymbol{V}_{t-1}$
   (3). Compute the prototype weights for specific classes
   (4). Compute the prototype weights for imprecise classes and get the new $\boldsymbol{V}_t$.
until the prototypes remain unchanged.
Output: The optimal credal partition.
```

Algorithm 2 Proposed algorithms with calculation of multiple weights

The optimization procedure consists 3 steps: group task update, model loads of specific classes update and afterward model loads of uncertain classes update. The first two stages improve the target capacity esteem by the utilization of Lagrangian multiplier technique. The third step attempts to find great delegate objects for uncertain classes. On the off chance that the technique to decide the loads for uncertain classes is of useful importance, it will likewise keep the target capacity expanding. Truth is told the methodology of refreshing the model loads is like the possibility of one-advance Gaussian-Seidel emphasis strategy, where the calculation of the new factor vector utilizes the new components that have just been registered, and the old components that have not yet to be progressed to the following cycle.

V. EXPERIMENTAL EVALUATION

In this section, we describe the experimental evaluation of proposed approach with existing relational k-medoids clustering approach with respect to synthetic medical relational data with respect to different attributes in relation of time, accuracy and others with different libraries implemented in JAVA and NETBEANS tool tested in windows operating system. Example data sets with respect to different attributes shown in figure 2.



	A	B	C	D	E	F	G	H	I	J	K
1	preg	plas	pres	skin	insu	mass	pedi	age	class		
2	6	148	72	35	0	33.6	0.627	50	tested_positive		
3	1	85	66	29	0	26.6	0.351	31	tested_negative		
4	8	183	64	0	0	23.3	0.672	32	tested_positive		
5	1	89	66	23	94	28.1	0.167	21	tested_negative		
6	0	137	40	35	168	43.1	2.288	33	tested_positive		
7	5	116	74	0	0	25.6	0.201	30	tested_negative		
8	3	78	50	32	88	31	0.248	26	tested_positive		
9	10	115	0	0	0	35.3	0.134	29	tested_negative		
10	2	197	70	45	543	30.5	0.158	53	tested_positive		
11	8	125	96	0	0	0	0.232	54	tested_positive		
12	4	110	92	0	0	37.6	0.191	30	tested_negative		
13	10	168	74	0	0	38	0.537	34	tested_positive		
14	10	139	80	0	0	27.1	1.441	57	tested_negative		
15	1	189	60	23	846	30.1	0.398	59	tested_positive		
16	5	166	72	19	175	25.8	0.587	51	tested_positive		
17	7	100	0	0	0	30	0.484	32	tested_positive		
18	0	118	84	47	230	45.8	0.551	31	tested_positive		
19	7	107	74	0	0	29.6	0.254	31	tested_positive		
20	1	103	30	38	83	43.3	0.183	33	tested_negative		
21	1	115	70	30	96	34.6	0.529	32	tested_positive		
22	3	126	88	41	235	39.3	0.704	27	tested_negative		
23	8	99	84	0	0	35.4	0.388	50	tested_negative		
24	7	196	90	0	0	39.8	0.451	41	tested_positive		
25	9	119	80	35	0	29	0.263	29	tested_positive		
26	11	143	94	33	146	36.6	0.254	51	tested_positive		
27	10	125	70	26	115	31.1	0.205	41	tested_positive		
28	7	147	76	0	0	39.4	0.257	43	tested_positive		
29	1	97	66	15	140	23.2	0.487	22	tested_negative		

Figure 2 Bio-medical data sets with different attributes.

Prototype cluster formation with different data attributes shown in figure 3.

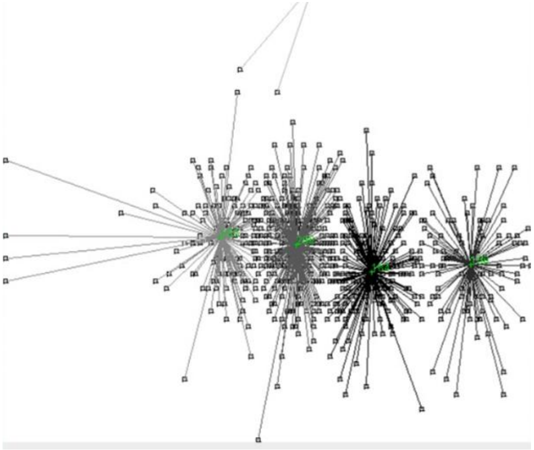


Figure 3 Optimized cluster representation with different attributes

Time performance for our suggested strategy shown in figure 4, different information sets like incident, diabetic issues, with multi features in recent feature selection with arbitrarily improvement real-time information streams.

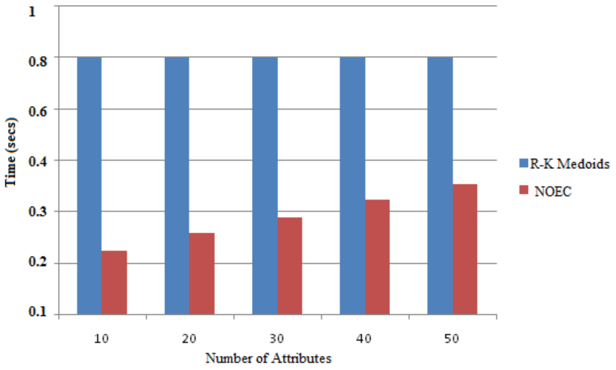


Figure 4 Performance of time comparison with different attributes.

NOEC works regularly better than its opponents with all different choice dimensions, while Relational K-Medoids approach gives least efficiency on category circumstances. Realize that a larger choice results in an improved excellence with better time efficiency results representation.

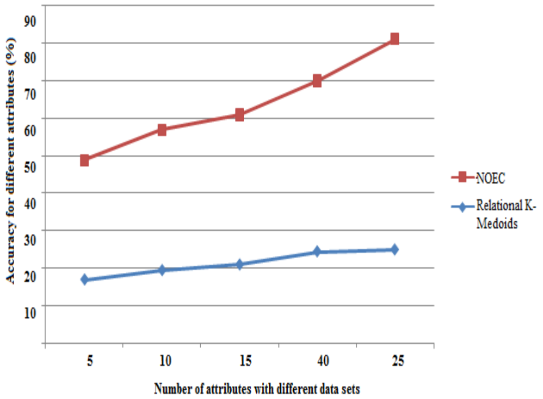


Figure 5 Performance of proposed approach with respect to dissimilar multiple weighted objects

Figure 5 show comparison results with respect to accuracy of the proposed and conventional approaches. Based on above results our proposed approach gives better performance with respect to time and accuracy.

VI. CONCLUSION

In this paper, we propose a novel optimized evidential C-medoids approach (NOEC). Main contribution of this approach is to handle multi-attribute weight calculation for different attributes. Proposed approach applied for multiple weighted medoids with representative medoids classes. Experimental results of proposed approach with credential partition with respect to weighted capture of architectural maintained to improve quality of classes. Further improvement of proposed approach is to support corresponding attribute relation for different attributes in medical data sets.

REFERENCES

1. Parisa Rastin, Basarab Matei, “Prototype-based Clustering for Relational Data using Barycentric Coordinates”, 978-1-5090-6014-6/18/$31.00 ©2018 IEEE.
2. Kuang Zhou, Arnaud Martin, Quan Pan, Zhun-ga Liu, “ECMdd: Evidential c-medoids clustering with multiple prototypes”, arXiv:1606.01113v1 [cs.AI] 3 Jun 2016.
3. L. Kaufman and P. Rousseeuw, Clustering by means of medoids. North-Holland, 1987.
4. J. C. Dunn, “A fuzzy relative of the isodata process and its usein detecting compact well-separated clusters,” Journal of Cybernetics,vol. 3, no. 3, pp. 32–57, 1973.
5. T. Kohonen, Ed., Self-organizing Maps. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1997.
6. M. Liu, X. Jiang, A. C. Kot, A multi-prototype clustering algorithm, Pattern Recognition 42(5) (2009) 689{698.
7. C.-W. Tao, Unsupervised fuzzy clustering with multi-center clusters, Fuzzy Sets and Systems128 (3) (2002) 305{322.
8. D. Ghosh, A. Shivaprasad, et al., Parameter tuning for multi-prototype possibilistic classifierwith reject options, in: Fuzzy Systems (FUZZ), 2013 IEEE International Conference on, IEEE,1{6, 2013.
9. T. Luo, C. Zhong, H. Li, X. Sun, A multi-prototype clustering algorithm based on minimumspanning tree, in: Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh InternationalConference on, vol. 4.
10. S. Ben, Z. Jin, J. Yang, Guided fuzzy clustering with multi-prototypes, in: Neural Networks(IJCNN), The 2011 International Joint Conference on, IEEE, 24.
11. R. J. Kuo, H. S. Wang, T.-L. Hu, and S. H. Chou, “Application of ant K-means on clustering analysis,” Computers and Mathematics with Applications, vol. 50, no. 10-12, pp. 1709–1724, 2005.
12. Kalluri Mohana Priyanka, DS Bhupal Naik, A "Performance Analysis of Harmony Search Optimization" in K-Means Clustering, IJDCST @ Jan,-2017, Issue- V-5, I-1, SW-06.

AUTHORS PROFILE

A.V. Suryanarayana Raju is pursuing M. Tech.(CST) in department of CSE in S.R.K.R Engineering College, India. He did his B. Tech (CSE) in Vishnu Institute of Technology. This is the first paper that is going to be published by him.

Sri. P. Bharath Siva Varma is an Assistant Professor in the department of CSE in S.R.K.R. Engineering College, India. He did his B. Tech (IT) in Andhra University and M. Tech (CSE) in JNTU kakinada.