ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

Improved Ant Colony Optimization towards Robust Ensemble Co-Clustering Algorithm (IACO-RECCA) for Enzyme Clustering

Ms.R.Rajeswari,
PhD Research scholar,
St. Peter's University,
Chennai, India.

Dr. G.GunaSekaran
Principal,
Meenakshi College of Engineering,
Chennai, India.

Abstract: ***This research work intends to propose a system with Improved Ant Colony Optimization (IACO) based on enhanced preprocessing method for enzyme clustering. A powerful optimization system is proposed in this research work initially deals with the enhanced principal component analysis. At that point the target function for the co-clustering troupe towards application to enzyme clustering is presented. An optimization technique for spectral co-clustering ensemble algorithm is described with constructive mathematical modeling. The proposed algorithm (IACO-RECCA) is capable enough to perform co-clustering with the objective function as the primary component. Simulation results proved that the proposed mechanism IACO-RECCA performs better in terms of accuracy and computation time.***

KEYWORDS: **Bioinformatics, DNA, Clustering, Co clustering, SS-NMF, NMF, CMRF, SS-CMRF, SRC, TSVM, RECCA, Accuracy, Computation Time, Text, Gene Expression, Image High Order Co clustering, High Order Co clustering, Supervised, Semi supervised, Optimization, Ant Colony, Swarm.**

I. INTRODUCTION

As of late, there has been a typical enlarge in the measure of information freely realistic in wide-coming to way transcendently in the field of Bioinformatics, where gigantic measures of information have been gathered as DNA successions, protein groupings and structures, data on natural pathways, and so on. This has demonstrated the best approach to changed and scattered wellsprings of natural information. Protein capacity forecast, and particularly catalyst capacity expectation is on the go Bioinformatics research coliseum because of the exponential expands in the quantity of proteins being found. This is because of the sequenced genomes, to the challenges in tentatively describing protein capacity and instruments, and to the potential biotechnological utilization of new found chemical capacities. With the aforementioned angles, expectation of protein's capacity is a firm employment regularly done by work concentrated exploratory work or in a self-loader way by making utilization of succession homology. This exploration measurement is sufficiently proficient to benefit from clustering methods, since they allow the formation of gatherings of comparable proteins that can be mutually concentrated on. The style in which natural data is gathered in utilizing heaps of disparate datasets affectations an examination challenge for consolidating clustering calculations.

Clashes between information qualities will likewise stay ahead, as various sources may have strange characteristic qualities for the same true question, because of distinctive representations, scaling or encoding. In this examination work strong gathering co-clustering is acquainted all together with break down how the joining of different information sources as requirements influences the accomplishment of compound grouping, which may prompt essential data about the capacities and structures of the proteins, and also utilitarian broadening procured all through family development and to enhance the execution for the same.

The clustering problem consists in ordering a set of data into groups, based on the features of the data samples. Cluster analysis is an unsupervised learning method that is used for the exploration of interrelationships among a collection of patterns, by organizing them into homogenous clusters. In the last decades, ant colonies (and other social insects) have inspired clustering algorithms that mimic the ants' abilities for separating and clustering larvae and dead bodies. While many approaches use the basic concept of picking and dropping data vectors [Shutin and Kubin.,2004], others rely on different properties of social insects, such as chemical communication between the ants or building mechanical structures by self-assembling behavior [Molisch.,2005]. The remarkable contributions of this paper are:

- ✓ The knowledge of whether or not adding information from external sources to the database is able to improve the clustering quality for this application;

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

- ✓ The lateral way for the collected information to be transformed into constraint sets for the meticulous biological problem;
- ✓ To perform optimization using Improved Ant Colony Optimization (IACO) for Robust Ensemble Co-Clustering Algorithm (IACO-RECCA) For Enzyme Clustering
- ✓ To perform co-clustering in order to improve the performance by reducing the computation time and increasing average accuracy value.

II. LITERATURE REVIEW

Rongjian Li et al.,2015 proposed two formulations for evolutionary co-clustering and feature selection based on the fused Lasso regularization. The evolutionary co-clustering formulation is able to identify smoothly varying hidden block structures embedded into the matrices along the temporal dimension. It was very flexible and allows for imposing smoothness constraints over only one dimension of the data matrices. In Kanzawa and Endo. 2012, some types of fuzzy co-clustering algorithms are proposed. It was shown that the common base of the objective functions for quadratic-regularized fuzzy co-clustering and entropy-regularized fuzzy co-clustering is very similar to the base for quadratic-regularized fuzzy nonmetric model and entropy-regularized fuzzy nonmetric model, respectively. Hua Wang et al., 2011 presented a general HOCC framework, named as Orthogonal Nonnegative Matrix Tri-factorization (O-NMTF), for simultaneous clustering of multi-type relational data. The proposed O-NMTF approach employs Nonnegative Matrix Tri-Factorization (NMTF) to simultaneously cluster different types of data using the inter-type relationships, and incorporate intra-type information through manifold regularization, where, different from existing works, we emphasize the importance of the orthogonal ties of the factor matrices of NMTF. Takeuchi. 2008 described that a graph-based co-clustering approach is suitable for extraction of verb synonyms from large scale texts. Their proposed bipartite graph algorithm can produce clusters of verb synonyms as well as noun synonyms taking into account word co-occurrence between verb and its argument. In Honda et al., 2014, a new fuzzy co-clustering model was proposed, which is a fuzzy variant of multinomial mixture density estimation. Multinomial mixtures is a probabilistic model for co-clustering of co occurrence matrices and the proposed method extends multinomial mixtures so that the degree of fuzziness can be tuned in a similar manner to K-L information-based FCM. Bing-Kun Bao et al., 2015 proposed a co-clustering method, called co-clustering via local and global consistency, to not only make use of the relationship between word and document, but also jointly explore the local and global consistency on both word and document spaces, respectively. That method has the following characteristics: 1) the word-document relationships is modeled by following information-theoretic co-clustering (ITCC); 2) the local consistency on both interword and interdocument relationships is revealed by a local predictor; and 3) the global consistency on both interword and interdocument relationships is explored by a global smoothness regularization. Guoping Qiu et al.,2004 presented a method which simultaneously models and clusters large sets of images and their low-level visual features. A computational energy function suited for co-clustering images and their features was constructed and a Hopfield model based stochastic algorithm is then developed for its optimization. Fan Yang et al.,2009 proposed an idea of analyzing both queries and advertisements which occur with queries at the same time. It was a co-clustering algorithm that suggests queries by co-clustering advertisements and queries. It poses the co-clustering problem as an optimization problem in information theory.

III. PROPOSED WORK

The proposed research work initially deals with the enhanced principal component analysis. Then the objective function for the co-clustering ensemble towards application to enzyme clustering is presented. A spectral co-clustering ensemble algorithm is described with constructive mathematical modeling followed with the brief algorithm description. The proposed algorithm is capable enough to perform co-clustering with the objective function as the primary component.

A. Enhanced Principal Component Analysis

An enhanced weighted version of PCA (EPCA) is introduced where more importance is given to observations whose values are more important. The higher the absolute expression value the more probable is that the meeting minutes are related to the particular topic. To that end, this enhanced PCA uses a new correlation coefficient that gives higher weights to observations that are considered to be more important. Also, the correlation coefficient is sensitive to the presence of outliers and noise in the data. The ranks of the observations are used. In the meeting dataset ranking the observations for each conversation from 1 (highest rank) to n

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

(lowest rank) is taken. The Pearson's correlation coefficient of the ranked data is thus obtained using the Spearman's rank correlation coefficient r_S, which is given by the expression

$$r_s = \frac{\sum_{i=1}^{n}(R_i - R)(Q_i - Q)}{\sqrt{\sum_{i=1}^{n}(R_i - R)^2 \sum_{i=1}^{n}(Q_i - Q)^2}} \quad ---(1)$$

where R and Q are the average ranks. However, for computational purposes, a more convenient expression which assumes there are no ties is

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2}{n^3 - n} \quad ---(2)$$

It is clear from this rewritten form of r_S that the calculation of the distance between two ranks in Spearman's coefficient is given by

$$D_i^2 = (R_i - Q_i)^2,$$

which does not take rank importance into account, because if $(R_i - Q_i)$ is, for instance, (1, 3) or (n- 2,n), the contribution is the same. The following alternative distance measure is proposed:

$$WD_i^2 = (R_i - Q_i)^2\big((n - R_i + 1) + (n - Q_i + 1)\big)$$

$$WD_i^2 = D_i^2(2n + 2 - R_i - Q_i) ---(3)$$

The first term of this product is D_i^2, exactly as in Spearman's coefficient, and represents the distance between R_i and Q_i; the second term is a linear weighting function which represents both the importance of R_i and Q_i. Hence the weighted rank measure of correlation is obtained using

$$r_w = 1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2(2n + 2 - R_i - Q_i}{n^4 + n^3 - n^2 - n} \quad ---(4)$$

which yields values between -1 and +1. The calculation of the distance between two ranks R_i and Q_i is given by $WD_i^2 = (R_i - Q_i)^2(2n + 2 - R_i - Q_i)$ where the second term of the product is a linear weighting function which represents the importance of R_i and Q_i. Hence, the distance measure is

$$W_2D_i^2 = (R_i - Q_i)^2(2n + 2 - R_i - Q_i)^2 \quad ---(5)$$

which reflects more than WD_i^2 the higher importance of agreement on top ranks. It is common to define rank correlation coefficients, such as Spearman's, as a linear function of the distance between the two vectors of ranks. In this research, this corresponds to define a coefficient of the form

$$W_2D_i^2 = A + B\sum_{i=1}^{n}(R_i - Q_i)^2(2n + 2 - R_i - Q_i)^2 \quad ---(6)$$

where the conversations are such that it takes values between -1 and +1. In order to find A and B, we will start by doing a specific data transformation and then compute the Pearson's coefficient on the transformed data. The expression obtained is exactly of the form, from where the constants A and B follow. The transformation consists in substituting the value of observation i in the first variable by the value of $R_i' = R_i(2n + 2 - R_i)$, where R_i is the rank of that observation. It is clear from above that the computation of the new correlation coefficient is equivalent to do a data transformation to each variable as $R_i' = R_i(2n + 2 - R_i)$ and then compute the Pearson's correlation coefficient. R_i represents the rank of each observation value; usually the smallest value has rank 1, the second smallest rank 2, and so on [Uma and Suguna.,2015].

B. IACO-RECCA

Swarm Intelligence [Engelbrecht.,2007, Engelbrecht.,2005, Kennedy et al.,2001], which deals with the collective behavior of small and simple entities, has been used in many application domains. ACO, proposed in the early 90s [Dorigo.,2004, Dorigo et al.,1996, Dorigo et al.,1997, Martens et al.,2007], is one of the most

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

famous meta-heuristic under the umbrella of Swarm Intelligence. Since its inception, ACO has been used to solve many complex problems including those related to data mining [Abraham et al.,2006, Han and Kamber.,2006, Witten and Frank.,2005] as well as other combinatorial optimization problems. ACO is inspired by the food foraging behaviour of biological ants [Eiben and Smith.,2007]. Ants pass on the information about the trail they are following by spreading a chemical substance called pheromone, in their search environment. Ants communicate with each other by means of pheromone. Other ants that arrive in the vicinity are more likely to take the path with higher concentration of pheromone than the paths with lower concentrations. In other words, the desirability of possible paths is proportional to their pheromone concentrations. The pheromone evaporates with time and if new pheromone is not added, the older one dwindles away. This phenomenon has been modelled in the ACO meta-heuristic. The learning of proposed algorithm is based on ACO with following major requirements:

- ✓ The ability to represent a complete solution as a combination of different components.
- ✓ There should be a method to determine the fitness or quality of the solution.
- ✓ A heuristic measure for the solution's components (this is desirable but not necessary).

Suppose, we have a connected graph $G = (V, E)$ where $|V|$ denotes the total number of nodes/ vertices and $|E|$ total number of connecting edges in graph. The simple ant colony optimization meta-heuristic can be used to find the shortest path between a given source node $'V_s'$ and a given destination node $'V_d'$ in the graph $'G'$. The path length is either given by the number of nodes on the path or summation of cost values on edges constituting the path. Each edge of the graph connecting the nodes $'V_i'$ and $'V_j'$ has a variable (artificial pheromone), which is modified by the ants when they visit the nodes.

From a node, when an ant decides which node to move next, it uses two parameters to calculate the probability of moving to a particular node; first, distance to that node and second, amount of pheromone on the connecting edge. Let $d_{i,j}$ be the distance between the nodes $'i'$ and $'j'$, the probability that the ant chooses $'j'$ as the next node after it has arrived at city $'i'$ where $'j'$ is in the set $'S'$ of cities that have not been visited.

$$p_{i,j} = \frac{[T_{i,j}]^{\alpha}[n_{i,j}]^{\beta}}{\sum_{k \epsilon S}[T_{i,j}]^{\alpha}[n_{i,j}]^{\beta}} \quad ---(7)$$

where $T_{i,j}$ is the pheromone value on edge $e(i,j)$ and $n_{i,j}$ is a heuristic value calculated as $1/d_{i,j}$. The parameters α and β are influencing factors of pheromone value and heuristic value respectively. The pheromone at edges is usually initialized with small random values at start. The complete route/tour of an ant from a source node to a destination node is called a solution to the problem at hand. The evaluation of a solution is done using a fitness function. Some best ants (having good solutions) or all ants modify the pheromone values on the edges added to their tour. One possible modification of the pheromone may be done as:

$$T_{i,j} = T_{i,j} + \frac{Q}{L} \quad ---(8)$$

where $'Q'$ is some constant and $'L'$ is the length of the tour, small the value of $'L'$ high the pheromone value added to the previous pheromone value on an edge. With time, concentration
of pheromone decreases due to diffusion affects; a natural phenomenon known as evaporation. This also ensures that old pheromone should not have a too strong influence on the future. So, with evaporation, chances to stuck at local minima is minimized in ACO. This evaporation can be performed as:

$$T_{i,j} = T_{i,j}.\rho \quad \{i.e., \rho \text{ will be between } 0 \text{ and } 1\} \quad ---(9)$$

The first stage in an IACO algorithm starts with designing a problem search space in which the ants conduct the search in order to find the candidate solutions. The search space for IACO (Improved Ant Colony Optimization based RECCA) is defined with the help of input datasets. The overall search space is divided into two parts; "Class Hierarchy sub-graph" and "Antecedent Construction sub-graph". In the same way, an ant is equipped with two types of memories. First type of memory named "class memory" is used to save the classes selected during the tour of the ant in the class hierarchy sub-graph. The second type of memory named "antecedent memory" saves antecedent part of the rule during the ant tour in the antecedent construction subgraph.

In this way, the rule consequent and antecedent parts are constructed cooperatively, and collectively represents a complete tour of an ant. In order to use proposed heuristic function, it is mandatory to construct the consequent of the rule before constructing its antecedent part. In other words, the selection of conditions to be included in

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

the rule antecedent part is determined based on the committed class labels. The search space of IACO is modified and two sub-graphs are connected in order to serve this purpose.

The class hierarchy sub-graph is equivalent to the class hierarchical structure given with the problem dataset. The topology of the antecedent construction sub-graph is used. The antecedent construction sub-graph can be considered as a directed acyclic graph, where the vertices are used to represent the terms (attribute–value pair) extracted from the input dataset. For constructing IF–THEN rules, the task of ants is to visit a vertex based on some heuristic (discussed later) and pheromone value found on the edges in the search space. The statistic about the total number of terms present in the dataset can be calculated as:

$$Total\ terms = \sum_{n=1}^{a} b_n \quad ---(10)$$

where $'a'$ is the total number of attributes (excluding the class attribute) and $'b_n'$ is the number of possible values in the domain of attribute $'A_n'$. After a term has been selected (i.e. a vertex is visited), the attribute related to that term is flagged as used. During the tour of an ant, once an attribute is flagged used, no other term related to that attribute can be selected, further. This restriction is necessary; as we cannot allow conditions of the type "Weather = Cloudy AND Weather = Sunny" (both at the same time). In this way, an ant can pick any of the vertexes (in the antecedent construction sub-graph) and there is no fixed ordering in which the vertices are required to be visited.

When an ant starts it tour, the class memory is activated and the antecedent memory is deactivated. Ant starts it tour from the start vertex (in class hierarchy sub-graph) which is originally the root class label of the class hierarchy. As, the proposed algorithm handles single path hierarchical classification algorithm, only one node is allowed to be selected at every single level of the class hierarchy sub-graph. The selection of the nodes in the class hierarchy sub-graph is based on the heuristic values associated with the edges (discussed later). Once the ant reaches the node with the label "Switch Ant Memory", the antecedent memory is activated and class memory is switched off.

Each and every ant can select different sets of class labels as done in ACO. However, this is against the main essence of cooperative learning of ACO. For example, each ant learning a separate goal, cannot meaningfully update the common search environment (by means of pheromone). In order to ensure that entire swarm learns the same set of class labels (i.e. common goal) during an iteration of the IACO algorithm (discussed later), only first ant traverses through the class hierarchy sub-graph and the class labels it selects is simply copied to the class memory part of all the remaining ants. Thus, all the ants search for the terms to be added to the antecedent of the rule that best describes the selected class labels. All the ants are then placed at node "Switch Ant Memory" and the antecedent memories of all the ants are activated while the class memories of all the ants are deactivated. Now, all the ants traverse the antecedent construction sub-graph and fill up their antecedent memories.

1. Spectral co-clustering ensemble algorithm for enzyme clustering

In this work, the final ensemble step can be formulated as a partition problem on a bipartite graph. For convenience of discussion, we use small-bold letters such as u, v as vectors. Capital-bold letters such as **M, E, L** will denote matrices, and capital letters such as V, R will denote vertex sets. Denote the bipartite graph $G = (V_r, V_c, \boldsymbol{E})$ containing two sets of vertices including row labeling vertices V_r and column labeling vertices V_c respectively. It is easy to verify that the adjacency matrix M of the bipartite graph can be written as

$$M = \begin{bmatrix} O & E \\ E^T & O \end{bmatrix} \quad ---(11)$$

where

$$E = \begin{bmatrix} C_{rr} & C_{rc} \\ C_{cr} & C_{cc} \end{bmatrix} \quad ---(12)$$

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

C_{rr} denotes the edge-weights between row labeling vertices that are both in V_r . C_{rc} denotes the edge-weights between labeling vertices with one in V_r and the other in V_c . C_{cc}; C_{cr} are defined similarly and $C_{rc} = C_{cr}^T$. Let $|E|_{ij}$ denote the (i,j)th element of E. $|E|_{ij}$ is the edge weight between two vertices. More specifically,

$$|E|_{ij} = \frac{\sum_{\alpha=1}^{k(i)} \sum_{\beta=1}^{k(j)} O_{\alpha,\beta} \log\left(\frac{|O|.O_{\alpha,\beta}}{O_\alpha^i O_\beta^j}\right)}{\sqrt{\left(\sum_{\alpha=1}^{k(i)} O_\alpha^i \ log \frac{O_\alpha^i}{|O|}\right)\left(\sum_{\beta=1}^{k(j)} O_\beta^i \ log \frac{O_\beta^i}{|O|}\right)}} \quad ---(13)$$

if the ith and jth vertices are both the row labeling vertices for enzyme clusters;

$$|E|_{ij} = \frac{\sum_{\alpha=1}^{\ell(i)} \sum_{\beta=1}^{\ell(j)} O_{\alpha,\beta} \log\left(\frac{|F|.F_{\alpha,\beta}}{F_\alpha^i F_\beta^j}\right)}{\sqrt{\left(\sum_{\alpha=1}^{\ell(i)} F_\alpha^i \ log \frac{F_\alpha^i}{|F|}\right)\left(\sum_{\beta=1}^{\ell(j)} F_\beta^i \ log \frac{F_\beta^i}{|F|}\right)}} \quad ---(14)$$

if the ith and jth vertices are both the column labeling vertices for enzyme clusters. Otherwise $|E|_{ij} = 0$

According to the bipartite graph $G = (V_r, V_c, \boldsymbol{E})$ given above, now we define the co-clustering partition matrix **Y** as

$$Y = \begin{bmatrix} Y_r \\ Y_c \end{bmatrix} \quad ---(15)$$

where $\boldsymbol{Y_r}$ is the partition on row labeling vertex set V_r and $\boldsymbol{Y_c}$ is the partition on column labeling vertex set V_c . Thus, the laplacian matrix L can be defined as

$$L = D - M \quad ---(16)$$

where

$$D = \begin{bmatrix} D_r & O \\ O & D_c \end{bmatrix} \quad ---(17)$$

D_r and $D_c$are diagonal matrices such that $|D_r|_{ii} = \sum_j E_{ij}, |D_c|_{jj} = \sum_i E_{ij}$. Note that the key step is to find the minimum cut vertex partitions on the bipartite graph. The normalized-cut objective function can be expressed as

$$\min_{\boldsymbol{Y}} \boldsymbol{tr}(\boldsymbol{Y}^T \boldsymbol{L} \boldsymbol{Y}) \quad ---(18)$$

One way to solve the partition problem of the bipartite graph is to compute the left and right eigenvectors of the matrix A defined as

$$A = D_r^{-1/2} E D_c^{-1/2} \quad ---(19)$$

After the left and right eigenvectors of matrix **A** are obtained, the left and right eigenvectors of the second to the $(\omega + 1)th$ eigenvalues are selected as $U = [u_2, u_3, \dots, u_{\omega+1}]$ and $V = [v_2, v_3, \dots, v_{\omega+1}]$ respectively. Here, the $\omega = log_2 k$ singular vectors $u_2, u_3, \dots, u_{\omega+1}$, and $v_2, v_3, \dots, v_{\omega+1}$ often contain k-modal information about the original co-clustering labelling. Thus, the k-dimensional data matrix can be written as

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

$$X = \begin{bmatrix} D_r^{-1/2} & U \\ D_c^{-1/2} & V \end{bmatrix} \quad ---(20)$$

At last, the classical k-means algorithm is preformed on X, and the final consensus co-clustering result is obtained.

IV. ABOUT THE DATASET

Several datasets [Yanhua Chen et al.,2010] have been taken for the performance analysis. The datasets for text pair wise co clustering is shown in Table 1. The datasets for Text High-Order (Word-Document-Category) co clustering is presented in Table 2. The datasets for gene expression pair wise (Condition-Gene) co clustering is given in Table 3. The datasets for Image High-Order (Color-Image-Texture) co clustering is depicted in Table 4.

TABLE 1. Data Sets for Text Pairwise (Document-Word) Coclustering

Name	Datasets	Data Structure	No. of clusters	No. of documents
CT1	oh15	Adenosine-Diphosphate, Blood-Vessels	2	154
CT2	oh15	Aluminium, Blood-Coagulation-Factors	2	122
CT3	re0	Interest, reserves	2	261
CT4	re0	housing, jobs	2	55
CT5	re0	housing, interest, jobs	3	274
CT6	oh15	Aluminium, Blood-Vessels, Leucine	3	207
CT7	re0	cpi, housing, ipi, lei, retail	5	144
CT8	re0	bop, cpi, gnp, housing, interest, ipi, jobs, lei, money	10	1150

TABLE 2. Data Sets for Text High-Order (Word-Document-Category) Coclustering

Name	Datasets	Data Structure	No. of clusters	No. of documents
HT1	oh15, re0	{ Adenosine-Diphosphate, Aluminium, Cell-Movement}, {cpi,money}	2	899
HT2	oh15, re0	{Blood-Coagulation-Factors, Enzyme-Activation, Staphylococcal-Infections}, {jobs,reserves}	2	461
HT3	oh15, re0	{Aluminium, Blood-Coagulation-Factors, Blood-Vessels}, {housing,retail}	2	256
HT4	oh15, re0	{Aluminum, Cell-Movement, Staphylococcal-Infections}, {cpi, jobs}	2	391
HT5	WAP, re0	{media, film, music}, {cpi, jobs}	2	404
HT6	Newsgroup	{rec.sport.baseball, rec.sport.hockey}, {talk.politics.guns, talk.politics.mideast,talk.politics.misc}	2	500
HT7	Newsgroup	{comp.graphics, comp.os.ms-windows.misc}, {rec.autos,rec.motorcycles}, {sci.encrypt, sci.electronics}	3	300
HT8	Newsgroup	{ comp.graphics, comp.os.ms-windows.misc}, {sci.electronics, sci.med}	2	3932
HT9	Newsgroup	{rec.autos, rec.motorcycles, rec.sport.baseball}, {sci.crypt, sci.electronics, sci.space}	2	5942

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

TABLE 3. Data Sets for Gene Expression Pair wise (Condition-Gene) Co clustering

Name	Datasets	Data Structure	No. of clusters	No. of documents
BT1	ALL/AML	ALL, AML	2	72
BT2	Breast Cancer	Relapse, Non-relapse	2	97
BT3	Central Nervous	Class1, Class2	2	60
BT4	Colon Tumor	Positive, Negative	2	62
BT5	Lung Cancer	MPM, ADCA	2	181
BT6	Ovarian Cancer	Cancer, Normal	2	253
BT7	ALL/MLL/AML	ALL,MLL,AML	3	72

TABLE 4. Data Sets for Image High-Order (Color-Image-Texture) Coclustering

Name	Datasets	No. of Modalities	No. of clusters	No. of documents
IT1	eggs,decoys	3	2	200
IT2	dawn,foliage	3	2	200
IT3	decoys,dawn	3	2	200
IT4	decoys,firearms,cards,buses	3	4	400
IT5	abstract,dawn,foliage,waves	3	4	400
IT6	eggs,decoys,dawn,foliage	3	4	400
IT7	eggs,decoys,buses,abstract,texture,dawn	3	6	600

V. RESULTS AND DISCUSSIONS

Performance of IACO-RECCA is made a comparison with Semi supervised Non-negative Matrix Factorization (SS-NMF) [Yanhua Chen et al.,2010], Non-negative Matrix Factorization (NMF) [Xu et al.,2003], Combinatorial Markov Random Field (CMRF) [Bekkerman and Jeon.,2007], Semi supervised Combinatorial Markov Random Field (SS-CMRF) [Bekkerman and Sahami.,2006], Spectral Relational Clustering (SRC) [Long et al.,2006] and Transductive Support Vector Machines (TSVM) [Joachims.,1999] in terms of accuracy and computation time. Figure 1 uses the Text Pairwise (Document-Word) Coclustering datasets depicted in Table 1. Figure 2 uses the Gene Expression Pairwise (Condition-Gene) Coclustering datasets depicted in Table 3. Figure 3 uses the Text High-Order (Word-Document-Category) Coclustering datasets depicted in Table 2. Figure 4 uses the Image High-Order (Color-Image-Texture) Coclustering datasets depicted in Table 4. The experiments are performed on a Windows 8.1 machine with Intel Core i3 processors and 4 GB DDR III RAM. The experiments on algorithms are evaluated using MATLAB R2012a.

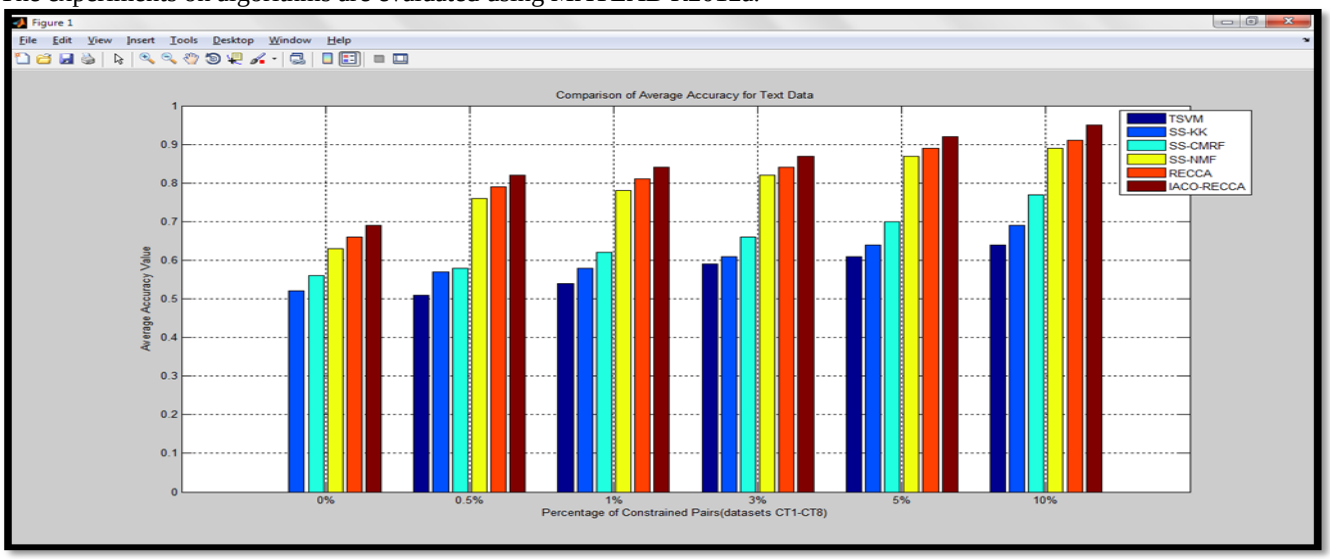

Fig 1. Comparison of Average Accuracy for Text Data

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

Figure 1 shows the performance evaluation of average accuracy for text data. It is evident that the proposed IACO-RECCA mechanism using Enhanced PCA outperforms other mechanisms in terms of document clustering performance with least prior knowledge. The performance values are depicted in Table 5.

TABLE 5: Comparison of Average Accuracy for Text Data

Algorithms / Percentage of Constrained Pairs	TSVM	SS-KK	SS-CMRF	SS-NMF	RECCA	IACO-RECCA
0%	0.00	0.52	0.56	0.63	0.66	0.69
0.5%	0.51	0.57	0.58	0.76	0.79	0.82
1%	0.54	0.58	0.62	0.78	0.81	0.84
3%	0.59	0.61	0.66	0.82	0.84	0.87
5%	0.61	0.64	0.7	0.87	0.89	0.92
10%	0.64	0.69	0.77	0.89	0.91	0.95

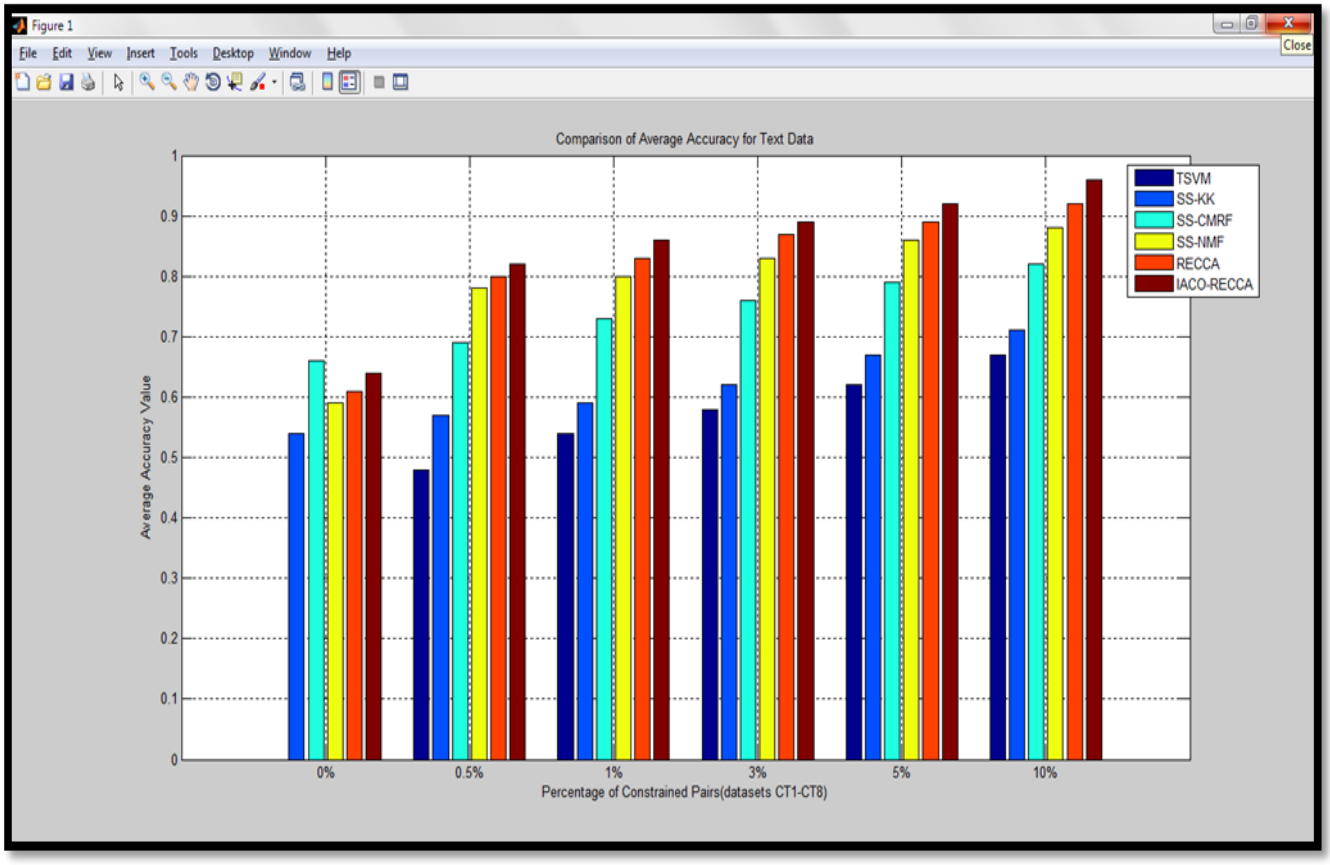

Fig 2. Comparison of Average Accuracy for Gene Expression Data

Figure 2 presents the performance evaluation of average accuracy for gene expression data. It is most visible that the proposed IACO-RECCA mechanism using Enhanced PCA outperforms other mechanisms in terms of increasing percentage of pair wise constraints for semi supervised condition co clustering. The performance values are depicted in Table 6.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

TABLE 6. Comparison of Average Accuracy for Gene Expression Data

Algorithms / Percentage of Constrained Pairs	TSVM	SS-KK	SS-CMRF	SS-NMF	RECCA	IACO-RECCA
0%	0.00	0.54	0.66	0.59	0.61	0.64
0.5%	0.48	0.57	0.69	0.78	0.80	0.82
1%	0.54	0.59	0.73	0.8	0.83	0.86
3%	0.58	0.62	0.76	0.83	0.87	0.89
5%	0.62	0.67	0.79	0.86	0.89	0.92
10%	0.67	0.71	0.82	0.88	0.92	0.96

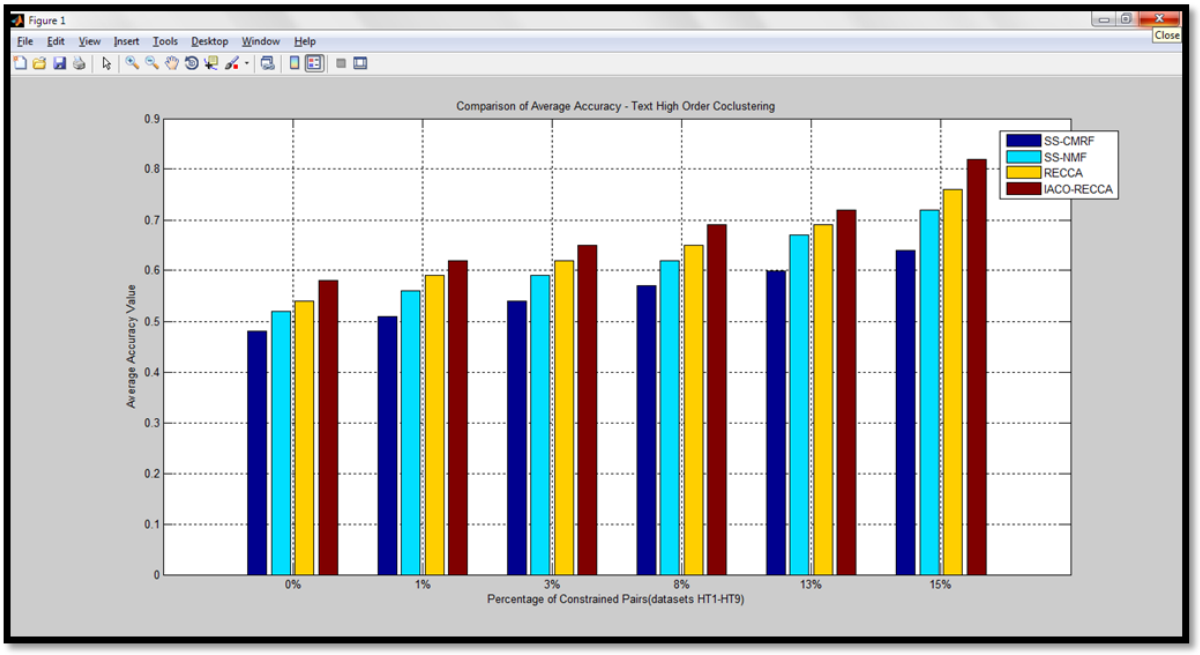

Fig 3. Comparison of Average Accuracy – Text High Order Co clustering

Figure 3 presents the performance comparison of average accuracy for text high order co clustering. It is most obvious that the proposed IACO-RECCA mechanism using Enhanced PCA outperforms other mechanisms. The performance values are depicted in Table 7.

TABLE 7. Comparison of Average Accuracy – Text High Order Co clustering

Algorithms / Percentage of Constrained Pairs	SS-CMRF	SS-NMF	RECCA	IACO-RECCA
0%	0.48	0.52	0.54	0.58
1%	0.51	0.56	0.59	0.62
3%	0.54	0.59	0.62	0.65

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

8%	0.57	0.62	0.65	0.69
13%	0.6	0.67	0.69	0.72
15%	0.64	0.72	0.76	0.82

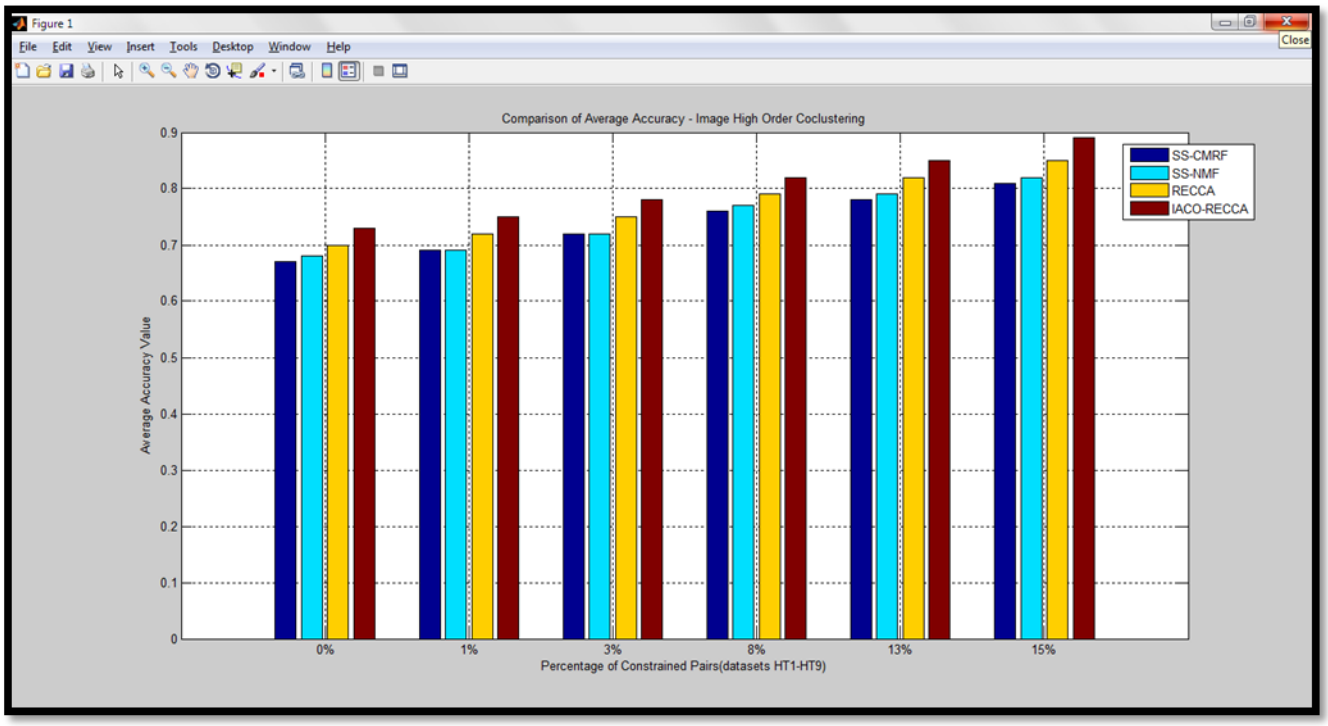

Fig 4. Comparison of Average Accuracy – Image High Order Co clustering

Figure 4. Presents the performance comparison of average accuracy for image high order co clustering. It can be perceived that the proposed IACO-RECCA mechanism using Enhanced PCA outperforms other mechanisms. The performance values are depicted in Table 8.

TABLE 8. Comparison of Average Accuracy – Image High Order Coclustering

Algorithms / Percentage of Constrained Pairs	SS-CMRF	SS-NMF	RECCA	IACO-RECCA
0%	0.67	0.68	0.70	0.73
1%	0.69	0.69	0.72	0.75
3%	0.72	0.72	0.75	0.78
8%	0.76	0.77	0.79	0.82
13%	0.78	0.79	0.82	0.85
15%	0.81	0.82	0.85	0.89

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

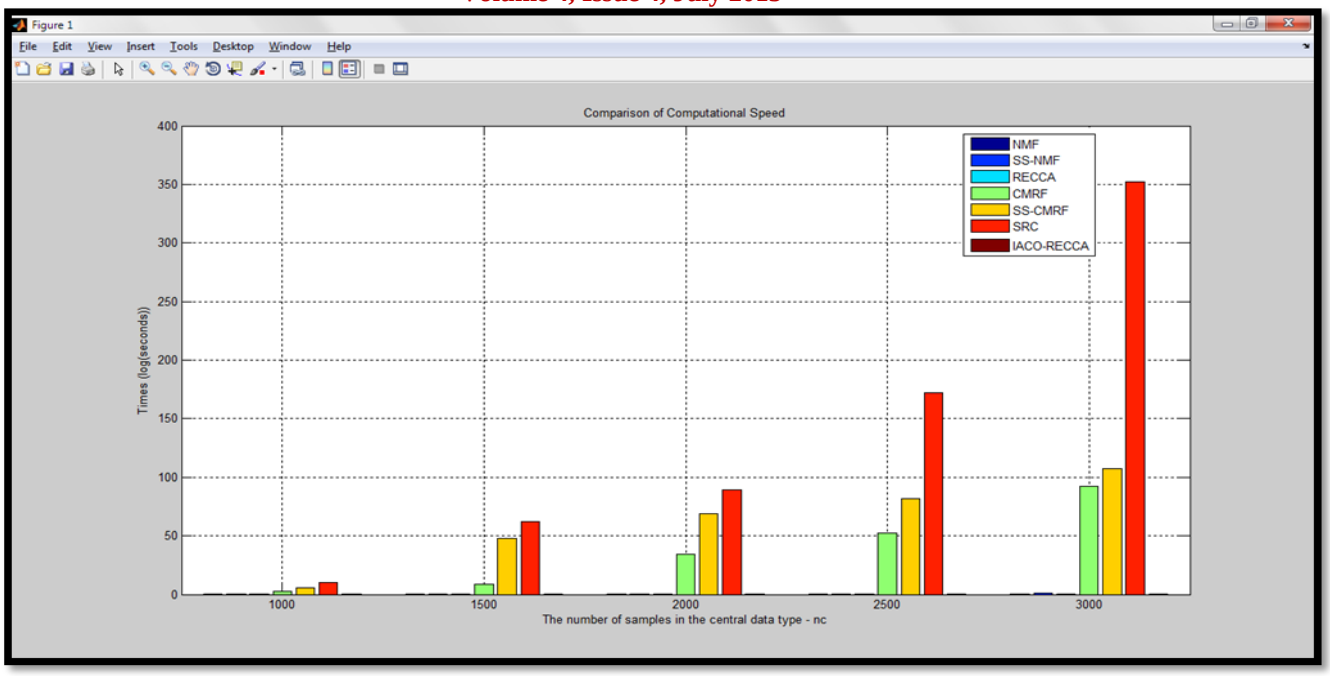

Fig 5. Comparison of Computational Speed - In Log (Seconds) For Increasing N_c

Figure 5 presents the performance of computational time (number of samples in the central data type - N_c) and the results proved that the proposed IACO-RECCA mechanism using Enhanced PCA approach delivers significant better performance over other methods. The performance values are depicted in Table 9.

TABLE 9. Comparison of Computational Speed - In Log(Seconds) For Increasing N_c

Algorithms / Percentage of Constrained Pairs	NMF	SS-NMF	RECCA	CMRF	SS-CMRF	SRC	IACO-RECCA
1000	0.05	0.23	0.21	3	6	10	0.19
1500	0.2	0.45	0.38	9	48	62	0.35
2000	0.1	0.56	0.52	34	69	89	0.48
2500	0.4	0.62	0.57	52	82	172	0.53
3000	0.52	0.84	0.74	92	107	352	0.70

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

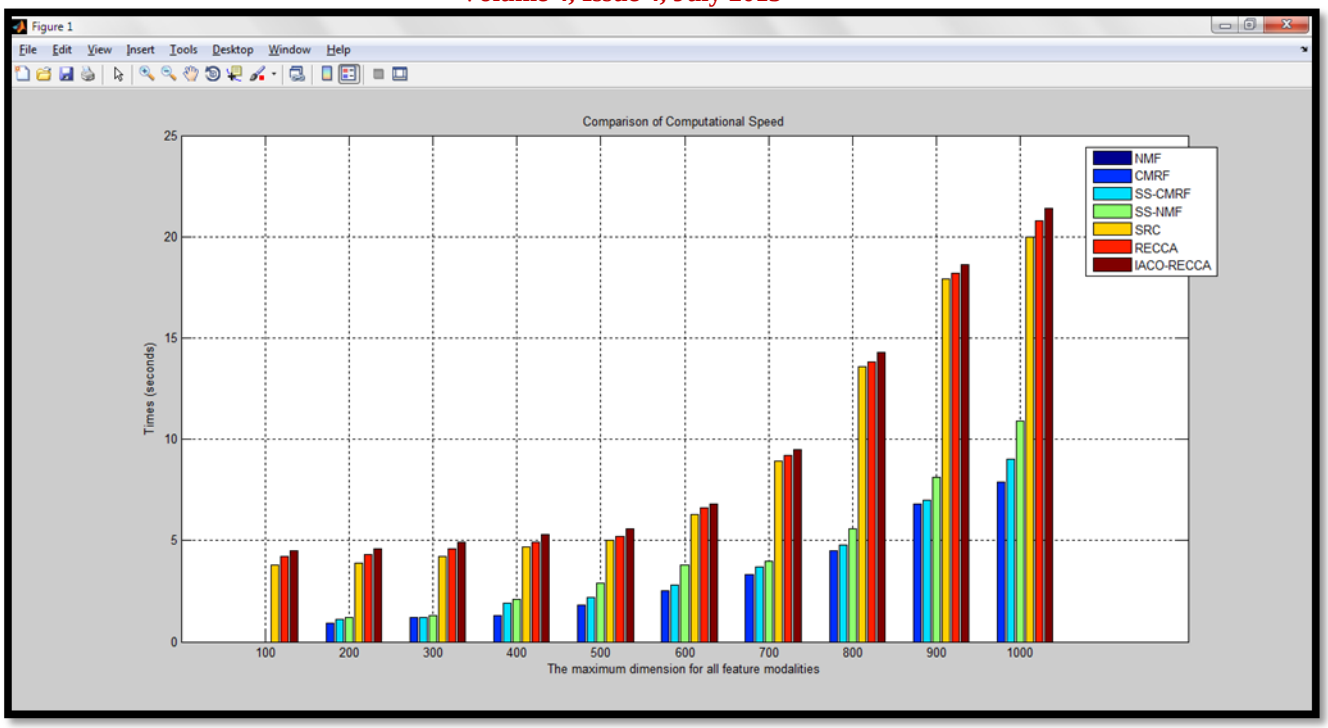

Fig 6. Comparison of Computational Speed - In Log(Seconds) For Increasing N_p

Figure 6 presents the performance of computational time (the maximum feature dimension for all feature modalities - N_p) and the results proved that the proposed IACO-RECCA mechanism using Enhanced PCA approach delivers significant better performance over other methods. The performance values are depicted in Table 10.

TABLE 10. Comparison of Computational Speed - In Log (Seconds) for Increasing N_p

Algorithms / Percentage of Constrained Pairs	NMF	CMRF	SS-CMRF	SS-NMF	SRC	RECCA	IACO-RECCA
100	0	0	0	0	3.8	4.2	4.5
200	0.2	0.9	1.1	1.2	3.9	4.3	4.6
300	0.2	1.2	1.22	1.3	4.2	4.6	4.9
400	0.1	1.3	1.9	2.1	4.7	4.9	5.3
500	0.2	1.8	2.2	2.9	5	5.2	5.6
600	0.3	2.5	2.8	3.8	6.3	6.6	6.8
700	0.4	3.3	3.7	4	8.9	9.2	9.5
800	0.2	4.5	4.8	5.6	13.6	13.8	14.3
900	0.4	6.8	7	8.1	17.9	18.2	18.6
1000	0.2	7.9	9	10.9	20	20.8	21.4

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

VI. CONCLUSION

This paper presented a mechanism with improved optimization technique namely IACO-RECCA for enzyme clustering. Initially the proposed work IACO-RECCA deals with the enhanced principal component analysis for preprocessing. The objective function for the co-clustering ensemble towards application to enzyme clustering is presented and also described. The objective function plays a major role which can perform co-clustering. Simulation results show that the proposed mechanism IACO-RECCA performs better in terms of accuracy and computation time. Regarding the future direction of this work, RECCA can be hybrid with optimization techniques for the much better performance of accuracy and computation time.

REFERENCES

[1] Abraham, Grosan, Ramos, Swarm Intelligence in Data Mining. Studies in Computational Intelligence, vol. 34, Springer, Heidelberg, New York, 2006.

[2] Bing-Kun Bao; Weiqing Min; Teng Li; Changsheng Xu, "Joint Local and Global Consistency on Interdocument and Interword Relationships for Co-Clustering," Cybernetics, IEEE Transactions on , vol.45, no.1, pp.15,28, Jan. 2015

[3] Dorigo, Gambardella, Ant colony system: a cooperative learning approach to the travelling salesman problem, IEEE Transaction on Evolutionary Computation 1 (April (1)) (1997) 53–66.

[4] Dorigo, Maniezzo, Colorni, Ant system: optimization by a colony of cooperating agents, IEEE Transactions on Systems, Man, and Cybernetics, Part B 26 (February (1)) (1996) 29–41.

[5] Dorigo, Stutzle, Ant Colony Optimization, MIT Press, Cambridge, MA, 2004.

[6] Eiben, Smith, Introduction to Evolutionary Computing. Natural Computing Series, 2nd ed., Springer, 2007.

[7] Engelbrecht, Computational Intelligence, An Introduction, 2nd ed., John Wiley & Sons, Hudson County, New Jersey, United States, 2007.

[8] Engelbrecht, Fundamentals of Computational Swarm Intelligence, John Wiley & Sons, Hudson County, New Jersey, United States, 2005.

[9] Fan Yang; Bin An; Xizhao Wang, "Co-clustering for queries and corresponding advertisement," Machine Learning and Cybernetics, 2009 International Conference on , vol.4, no., pp.2296,2299, 12-15 July 2009

[10] Guoping Qiu, "Image and feature co-clustering," Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on , vol.4, no., pp.991,994 Vol.4, 23-26 Aug. 2004

[11] Han, Kamber, Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann Publishers, Middlesex County, Massachusetts, United States, 2006.

[12] Honda, Oshio, Notsu, "FCM-type fuzzy co-clustering by K-L information regularization," Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on , vol., no., pp.2505,2510, 6-11 July 2014

[13] Hua Wang; Feiping Nie; Heng Huang; Ding, C., "Nonnegative Matrix Tri-factorization Based High-Order Co-clustering and Its Fast Implementation," Data Mining (ICDM), 2011 IEEE 11th International Conference on , vol., no., pp.774,783, 11-14 Dec. 2011

[14] Kanzawa, Endo, "On FNM-based and RFCM-based fuzzy co-clustering algorithms," Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on , vol., no., pp.1,8, 10-15 June 2012

[15] Kennedy, Eberhart, Shi, Swarm Intelligence, Morgan Kaufmann/Academic Press, Middlesex County, Massachusetts, United States, 2001.

[16] Martens, de Backer, Haesen, Vanthienen, Snoeck, Baesens, Classification with ant colony optimization, IEEE Transactions on Evolutionary Computation 11 (October (5)) (2007) 651–665.

[17] Molisch, “Ultrawideband propagation channels—Theory, measurement, and modeling,” IEEE Trans. Veh. Technol., vol. 54, no. 5, pp. 1528–1545, Sep. 2005.

[18] Rongjian Li; Wenlu Zhang; Yao Zhao; Zhenfeng Zhu; Shuiwang Ji, "Sparsity Learning Formulations for Mining Time-Varying Data," Knowledge and Data Engineering, IEEE Transactions on , vol.27, no.5, pp.1411,1423, May 1 2015

[19] Shutin and Kubin, “Cluster analysis of wireless channel impulse responses with hidden Markov models,” in Proc. IEEE ICASSP, May 2004, pp. 949–952.

[20] Takeuchi, "Extraction of Verb Synonyms using Co-clustering Approach," Universal Communication, 2008. ISUC '08. Second International Symposium on , vol., no., pp.173,178, 15-16 Dec. 2008.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 4, July 2015

[21] Uma, Suguna, "Human Interaction Pattern Mining Using Enhanced Principal Component Analysis", International Journal of Informative & Futuristic Research (IJIFR), Volume 2, Issue 7, March 2015, 2279-2289.

[22] Witten, Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann, Middlesex County, Massachusetts, United States, 2005.

[23] Bekkerman, Jeon, "Multi-Modal Clustering for Multimedia Collections," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1-8, 2007.

[24] Bekkerman, Sahami, "Semi-Supervised Clustering Using Combinatorial MRFs," Proc. 23rd Int'l Conf. Machine Learning (ICML) Workshop Learning in Structured Output Spaces, 2006.

[25] Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," Proc. 16th Int'l Conf. Machine Learning, pp. 200-209, 1999.

[26] Long, Wu, Zhang, Yu, "Spectral Clustering for Multi-Type Relational Data," Proc. 23rd Int'l Conf. Machine Learning, pp. 585-592, 2006.

[27] Xu, Liu, Gong, "Document Clustering Based on Non-Negative Matrix Factorization," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 267-273, 2003.

[28] Yanhua Chen, Lijun Wang, Ming Dong, Member,"Non-Negative Matrix Factorization for Semi supervised Heterogeneous Data Co clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, October 2010.