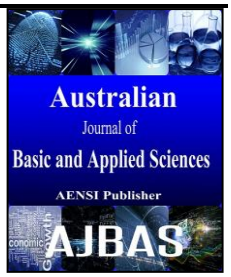
Australian Journal of Basic and Applied Sciences, 9(10) Special 2015, Pages: 364-375

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com



Robust Ensemble Co-Clustering Algorithm (RECCA) For Enzyme Clustering

[1]Ms.R.Rajeswari and [2]Dr. G.GunaSekaran

[1]*PhD Research scholar, St. Peter's University, Chennai, India.*
[2]*Principal, Meenakshi College of Engineering, Chennai, India.*

ARTICLE INFO

Article history:
Received 28 January 2015
Accepted 25 February 2015
Available online 6 March 2015

Keywords:
Bioinformatics, DNA, Protein, Clustering, Coclustering, SS-NMF, NMF, CMRF, SS-CMRF, SRC, TSVM, RECCA, Accuracy, Computation Time, Text, Gene Expression, Image High Order Coclustering, High Order Coclustering, Supervised, Semisupervised.

ABSTRACT

This research work aims to propose a mechanism with improved preprocessing technique for enzyme clustering. A robust ensemble mechanism is proposed in this 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. Simulation results proved that the proposed mechanism RECCA performs better in terms of accuracy and computation time.

To Cite This Article: Ms.R.Rajeswari and Dr. G.GunaSekaran., Robust Ensemble Co-Clustering Algorithm (RECCA) For Enzyme Clustering. ***Aust. J. Basic & Appl. Sci.,*** *9(10): 364-375, 2015*

INTRODUCTION

Recently, there has been a common augment in the amount of data publicly obtainable in wide-reaching manner predominantly in the field of Bioinformatics, where massive amounts of data have been collected in the form of DNA sequences, protein sequences and structures, information on biological pathways, etc. This has shown the way to varied and scattered sources of biological data. Protein function prediction, and especially enzyme function prediction is on the go Bioinformatics research arena due to the exponential augment in the number of proteins being discovered. This is due to the sequenced genomes, to the difficulties in experimentally characterizing enzyme function and mechanisms, and to the potential biotechnological use of newly discovered enzyme functions. With the above mentioned aspects, prediction of protein's function is a firm job typically carried out by labor-intensive experimental work or in a semi-automatic manner by making use of sequence homology. This research dimension is capable enough to profit from clustering techniques, since they permit the creation of groups of similar proteins that can be jointly studied. The style in which biological information is collected in using loads of dissimilar datasets pretenses a research challenge for incorporating clustering algorithms.

As an example, the Protein Data Bank (PDB) is a repository of 3D structural data, has dozens or even hundreds of entries for the same molecule. Inconsistencies and redundancies are probably arise due to the attributes representing a given concept may have different names in different databases. Conflicts between data values will also stay ahead, as diverse sources may have unusual attribute values for the same real-world object, due to different representations, scaling or encoding. In this research work robust ensemble co-clustering is introduced in order to analyze how the integration of various data sources in the form of constraints affects the success of enzyme clustering, which might lead to important information about the functions and structures of the enzymes, as well as functional diversification acquired throughout family evolution and to improve the performance for the same.

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;
- ✓ The lateral way for the collected information to be transformed into constraint sets for the meticulous biological problem;
- ✓ To perform co-clustering in order to improve the performance by reducing the computation time and increasing average accuracy value.

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

Australian Journal of Basic and Applied Sciences, 9(10) Special 2015, Pages: 364-375

Literature Review:

Sugato Basu *et al*.,2004 proposed a probabilistic model for semi-supervised clustering based on Hidden Markov Random Fields (HMRFs) that provides a principled framework for incorporating supervision into prototype-based clustering. Bilenko *et al*.,2004 proposed new methods for the two approaches as well as presents a new semi-supervised clustering algorithm that integrates both of these techniques in a uniform, principled framework. Klein *et al*.,2002 modified a constrained clustering algorithm to perform exploratory analysis on gene expression data using prior knowledge presented in the form of constraints. Also authors studied the effectiveness of various constraints sets. Wagstan *et al*.,2001 demonstrated how the popular k-means clustering algorithm can be profitably modified to make use of this information. In experiments with artificial constraints on six data sets, authors observed some improvements in clustering accuracy.

Erliang Zeng *et al*.,2007 modified a constrained clustering algorithm to perform exploratory analysis on gene expression data using prior knowledge presented in the form of constraints. Authors have studied the effectiveness of various constraints sets. To address the problem of automatically generating constraints from biological text literature, authors considered cluster-based and method similarity-based method. Shahreen Kasim *et al*.,2013 presented a novel computational framework called the "multi-stage filtering-Clustering Functional Annotation" (msf-CluFA) for clustering gene expression data. The framework consists of fuzzy c-means clustering, achieving dominant cluster, improving confidence level components. In protein databases there was a substantial number of proteins structurally determined but without function annotation. Understanding the relationship between function and structure can be useful to predict function on a large scale. Marcelo Boareto *et al*.,2012 have analyzed the similarities in global physicochemical parameters for a set of enzymes which were classified according to the four Enzyme Commission (EC) hierarchical levels. Also by using relevance theory authors have introduced a distance between proteins in the space of physicochemical characteristics.

Due to inspiration by the principle of gene transposon proposed by Barbara McClintock, a new immune computing algorithm for automatic clustering named as Gene Transposon based Clone Selection Algorithm (GTCSA) proposed in this Ruochen Liu *et al*.,2012. It does not require a prior knowledge of the number of clusters; an improved variant of the clonal selection algorithm used to determine the satisfied number of clusters and the appropriate partitioning of the data set as well. Clara Higuera *et al*.,2013 proposed an expert system (ES), making the main contribution, to cluster a complex data set of 365 prokaryotic species by 114 metabolic features, information which may be incomplete for some species. Inspired on the human expert reasoning and based on hierarchical clustering strategies, Clara Higuera *et al*.,2013 proposed ES estimates the optimal number of clusters adequate to divide the dataset and afterwards it starts an iterative process of clustering, based on the Self-organizing Maps (SOM) approach, where it finds relevant clusters at different steps by means of a new validity index inspired on the well-known Davies Bouldin (DB) index.

Rosfuzah Roslan *et al*.,2010 aimed at enhancing the overlap between computational predictions. Guoren Wang *et al*.,2010 explored a novel concept of local conserved gene cluster (LC-Cluster). To avoid the exponential growth in subspace search, we further authors have presented two efficient algorithms, namely falconer and e-falconer, to mine the complete set of maximal LC-Clusters from gene expression data sets based on enumeration tree. Thanh-Phuong Nguyen and Tu-Bao Ho.,2012 have presented a novel method to effectively predict disease genes by exploiting, in the semi-supervised learning (SSL) scheme, data regarding both disease genes and disease gene neighbours via protein–protein interaction network. Multiple proteomic and genomic data were integrated from six biological databases, including Universal Protein Resource, Interologous Interaction Database, Reactome, Gene Ontology, Pfam, and InterDom, and a gene expression dataset.

Banerjee *et al*.,2004 introduce a partitional co-clustering formulation that was driven by the search for a good matrix approximation-every co-clustering was associated with an approximation of the original data matrix and the quality of co-clustering was determined by the approximation error. Dhillon *et al*.,2003 presented an innovative co-clustering algorithm that monotonically increases the preserved mutual information by intertwining both the row and column clusterings at all stages. Bin Gao *et al*.,2006 proposed a consistent information theory which generates an effective algorithm to obtain the co-clusters of different types of objects. Inderjit Dhillon *et al*.,2001 presented the novel idea of modeling the document collection as a bipartite graph between documents and words, using which the simultaneous clustering problem can be posed as a bipartite graph partitioning problem. To solve the partitioning problem, authors used a new spectral co-clustering algorithm that uses the second left and right singular vectors of an appropriately scaled word-document matrix to yield good bipartitionings.

Proposed Work:

The proposed research work initially deals with the enhanced principal component analysis. Then the

Australian Journal of Basic and Applied Sciences, 9(10) Special 2015, Pages: 364-375

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.

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 (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:

$$WD_i^2 = (R_i - Q_i)^2\big((n - R_i + 1) + (n - Q_i + 1)\big)$$
$$WD_i^2 = D_i^2(2n + 2 - R_i - Q_i) \quad (3)$$

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).

Objective Function For Co-Clustering Ensemble:

Given t partitions, with the qth partition $\left(\mu^{(q)}, v^{(q)}\right)$ having $k^{(q)}$ row clusters and $\ell^{(q)}$column clusters. T is defined as a consensus function $N^{\{m\times t, n\times t\}} \rightarrow N^{\{m,n\}}$ mapping a set of co-clusterings to an integrated co-clustering:

$$T: \left\{\left(\mu^{(q)}, v^{(q)}\right) | q \in \{1, \ldots, t\}\right\} \rightarrow \{(\mu, v)\} \quad (7)$$

Let the set of partitions $\left\{\left(\mu^{(q)}, v^{(q)}\right) | q \in \{1, \ldots, t\}\right\}$ be denoted by ϕ. If there is no background information about the relative importance of the individual partitions, then a reasonable goal for the consensus solution is to seek a co-clustering that shares the most information with the original co-clusterings.

In order to quantify the statistical information shared between two co-clusterings, mutual information is used as a symmetric measure in our work. Here, the objective function is proposed by adapting the original definition given in [Strehl and Ghosh.,2003] to handle the problem of co-clustering ensemble:

$$(\mu,v)^{(k,l,-opt)} = \arg\max \sum_{q=1}^{t} \phi^{(NMI)}\{(\mu,v),\mu^{(q)},v^{(q)}\} \quad (8)$$

where $(\mu,v)^{(k,l,-opt)}$ is the optimal combined co-clustering and it is the one that has maximal average mutual information with all individual partitions in ϕ given that the number of consensus row clusters desired is k and the number of column clusters is l. In detail, the average normalized mutual information $(ANMI)$ between a single co-clustering (μ,v) and a set of t co-clusterings can be defined as

$$\phi^{(ANMI)}(\phi,(\mu,v)) = \frac{1}{t}\sum_{q=1}^{t} \phi^{(NMI)}\left((\mu,v),\mu^{(q)},v^{(q)}\right) \quad (9)$$

As mentioned before, mutual information (MI) is a symmetric measure and can be used to quantify the statistical information shared between two distributions. Thus, MI provides a sound indication of the shared information between a pair of co-clusterings. The normalized mutual information (NMI) is defined as

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \quad (10)$$

where X and Y denote two vectors, $I(X,Y)$ denotes the mutual information between X and Y. $H(X)$denotes the entropy of X and $H(X) = I(X,X)$. Suppose there are two co-clusterings(X_r,X_c) $and\ (Y_r,Y_c), i.e., (X_r,Y_r),(X_c,Y_c)$ denote the row and column cluster labeling variables respectively. Then, the NMI between two co-clusterings can be defined as

$$NMI\big((X_r,X_c),(Y_r,Y_c)\big) = NMI\big((X_r,Y_r)\big) + NMI\big((X_c,Y_c)\big) = \frac{I(X_r,Y_r)}{\sqrt{H(X_r)H(Y_r)}} + \frac{I(X_c,Y_c)}{\sqrt{H(X_c)H(Y_c)}} \quad (11)$$

It is clear that $NMI(X_r,X_r) = NMI\,(Y_c,Y_c) = 1$, as desired. According to Eqs. (11), (9) can be further rewritten as

$$\phi^{(ANMI)}(\phi,(\mu,v)) = \frac{1}{t}\sum_{q=1}^{t} \phi^{(NMI)}\left((\mu,v),\mu^{(q)},v^{(q)}\right) = \frac{1}{t}\sum_{q=1}^{t}\left(\frac{I(\mu,\mu^{(q)})}{\sqrt{H(\mu)H(\mu^{(q)})}} + \frac{I(v,v^{(q)})}{\sqrt{H(v)H(v^{(q)})}}\right) \quad (12)$$

Eq. (12) needs to be estimated by the sampled quantities provided by the co-clusterings. Then, the normalized mutual information estimate $\phi^{(NMI)}$can be defined as

$$\phi^{(NMI)}\left((\mu^i,v^i),(\mu^j,v^j)\right) = \phi^{(NMI)}\left((\mu^i,\mu^j)\right) + \phi^{(NMI)}\left((v^i,v^j)\right) = \frac{\sum_{\alpha=1}^{k(i)}\sum_{\beta=1}^{k(j)} O_{\alpha,\beta}\log\left(\frac{|O|.O_{\alpha,\beta}}{O_\alpha^i O_\beta^j}\right)}{\sqrt{\left(\sum_{\alpha=1}^{k(i)} O_\alpha^i\ log\frac{O_\alpha^i}{|O|}\right)\left(\sum_{\beta=1}^{k(j)} O_\beta^i\ log\frac{O_\beta^i}{|O|}\right)}} + \frac{\sum_{\alpha=1}^{\ell(i)}\sum_{\beta=1}^{\ell(j)} O_{\alpha,\beta}\log\left(\frac{|F|.F_{\alpha,\beta}}{F_\alpha^i F_\beta^j}\right)}{\sqrt{\left(\sum_{\alpha=1}^{\ell(i)} F_\alpha^i\ log\frac{F_\alpha^i}{|F|}\right)\left(\sum_{\beta=1}^{\ell(j)} F_\beta^i\ log\frac{F_\beta^i}{|F|}\right)}} \quad (13)$$

where $|O|$ and $|F|$ denote the number of objects and features in a cocluster respectively. (O_α^i,F_α^i) denotes the number of objects and features in co-cluster CO_α according to (μ^i,v^i) and (O_α^j,F_α^j) denotes the number of objects and features in co-cluster CO_β according to (μ^j,v^j).

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 $G = (V_r, V_c, E)$ 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

$$M = \begin{bmatrix} O & E \\ E^T & O \end{bmatrix} \quad (14)$$

where

$$E = \begin{bmatrix} C_{rr} & C_{rc} \\ C_{cr} & C_{cc} \end{bmatrix} \quad (15)$$

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 and can be obtained according to Eq.(13). More specifically,

$$|E|_{ij} = \frac{\sum_{\alpha=1}^{k(i)} \sum_{\beta=1}^{k(j)} O_{\alpha,\beta} \log\left(\frac{|O|.O_{\alpha,\beta}}{O_\alpha^i O_\beta^j}\right)}{\sqrt{\left(\sum_{\alpha=1}^{k(i)} O_\alpha^i \log\frac{O_\alpha^i}{|O|}\right)\left(\sum_{\beta=1}^{k(j)} O_\beta^i \log\frac{O_\beta^i}{|O|}\right)}} \quad (16)$$

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

$$|E|_{ij} = \frac{\sum_{\alpha=1}^{\ell(i)} \sum_{\beta=1}^{\ell(j)} O_{\alpha,\beta} \log\left(\frac{|F|.F_{\alpha,\beta}}{F_\alpha^i F_\beta^j}\right)}{\sqrt{\left(\sum_{\alpha=1}^{\ell(i)} F_\alpha^i \log\frac{F_\alpha^i}{|F|}\right)\left(\sum_{\beta=1}^{\ell(j)} F_\beta^i \log\frac{F_\beta^i}{|F|}\right)}} \quad (17)$$

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

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

$$Y = \begin{bmatrix} Y_r \\ Y_c \end{bmatrix} \quad (18)$$

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

$$L = D - M \quad (19)$$

where

$$D = \begin{bmatrix} D_r & O \\ O & D_c \end{bmatrix} \quad (20)$$

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

$$\min_Y tr(Y^T L Y) \quad (21)$$

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

$$A = D_r^{-1/2} E D_c^{-1/2} \quad (22)$$

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 $U = [u_2, u_3, \ldots, u_{\omega+1}]$ and $V = [v_2, v_3, \ldots, v_{\omega+1}]$ respectively. Here, the $\omega = log_2 k$ singular vectors $u_2, u_3, \ldots, u_{\omega+1}$, and $v_2, v_3, \ldots, v_{\omega+1}$ often contain k-modal information about the original co-clustering labeling. Thus, the k-dimensional data matrix can be written as

$$X = \begin{bmatrix} D_r^{-1/2} & U \\ D_c^{-1/2} & V \end{bmatrix} \quad (23)$$

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

Algorithm description:

According to the above inference, we design an algorithm based on spectral method for enzyme co-clustering ensemble. The algorithm procedure is descried step by step as follows.

Algorithm (Spectral Co-Clustering Ensemble)
Input:
$Original\ data\ matrix\ X_{mn}, num.of\ row\ clusters\ k, num.of\ column\ clusters\ \ell\ (i.e., K \times \ell\ co\ clusters\ in\ total)$

01. Divide X_{mn} into k row clusters and ℓ column clusters by the co-clustering algorithms and the base co-clustering labeling are obtained.
02. Compute pairwise similarities of the base co-clustering labelling according to Eqs. (10) and (11). Construct the adjacency matrices **M.**
03. Construct the diagonal matrices D_r, D_c where $|D_r|_{ii} = \sum_j E_{ij}$ and $|D_c|_{jj} = \sum_i E_{ij}$
04. Calculate **A** as defined in Eq. (16).

05. Perform singular value decomposition (SVD) on matrix **A**. Compute $\omega = log_2 k$ singular vectors of **A**, $u_2, \ldots u_{\omega+1}$ and $v_2, \ldots v_{\omega+1}$. Denote the left and right eigenvectors of the 2nd to the $(\omega + 1)th$ eigenvalues as U and V respectively.

06. Construct $X_r = D_r^{-1/2} U$ and $X_c = D_c^{-1/2} V$.

07. Run k-means algorithm on the x-dimensional data $X_r$to get the row labelings partition matrix Y_r; Similarly get Y_c from X_c .

Output:

$The\ final\ consensus\ co-clustering\ result.$

It can be observed that the main computational cost is to perform SVD on the matrix **A** on Step 5. Consider Lanczos algorithm to compute the eigenvectors [Shi *et al.*, 2010]. The complexity of our algorithm is $O(eN(|m| + |n|)^2$, where e is the number of eigenvectors desired, N is the number of Lanczos iteration steps and $(|m| + |n|)^2$ is the upper bound of the nonzero entries of matrix M. More performance in detail are recorded in the next section.

Several datasets have been taken for the performance analysis. The datasets for text pairwise coclustering is shown in Table 1. The datasets for Text High-Order (Word-Document-Category) coclustering is presented in Table 2. The datasets for gene expression pairwise (Condition-Gene) coclustering is given in Table 3. The datasets for Image High-Order (Color-Image-Texture) coclustering is depicted in Table 4.

About The Dataset:

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

Table 3: Data Sets for Gene Expression Pairwise (Condition-Gene) Coclustering

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

RESULTS AND DISCUSSIONS

Performance of RECCA is made a comparison with Semisupervised 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), Semisupervised 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.

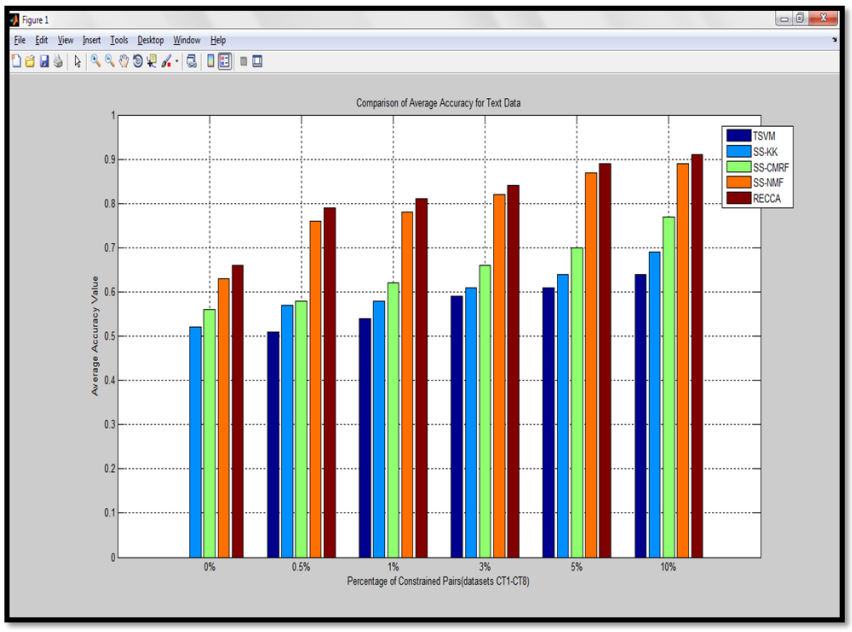


Fig. 1: Comparison of Average Accuracy for Text Data

Figure 1 shows the performance evaluation of average accuracy for text data. It is evident that the proposed 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
0%	0	0.52	0.56	0.63	0.66
0.5%	0.51	0.57	0.58	0.76	0.79
1%	0.54	0.58	0.62	0.78	0.81
3%	0.59	0.61	0.66	0.82	0.84
5%	0.61	0.64	0.7	0.87	0.89
10%	0.64	0.69	0.77	0.89	0.91

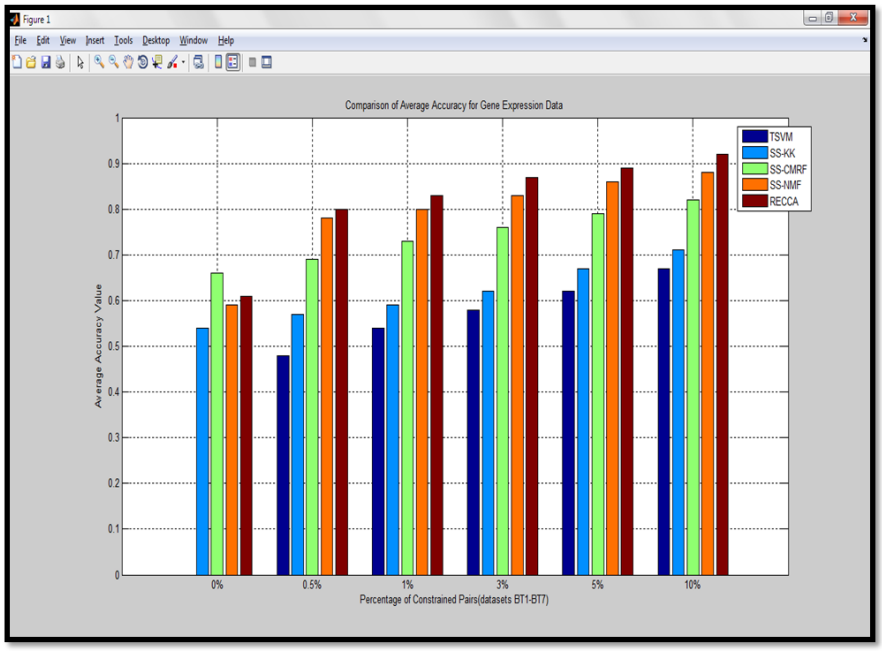


Fig. 2: Comparison of Average Accuracy for Gene Expression Data

Figure 2 presents the performance evaluation of average accuracy for gene expression data. It is most visible that the proposed RECCA mechanism using Enhanced PCA outperforms other mechanisms in terms of increasing percentage of pairwise constraints for semisupervised condition coclustering. The performance values are depicted in Table 6.

Table 6: Comparison of Average Accuracy for Gene Expression Data

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

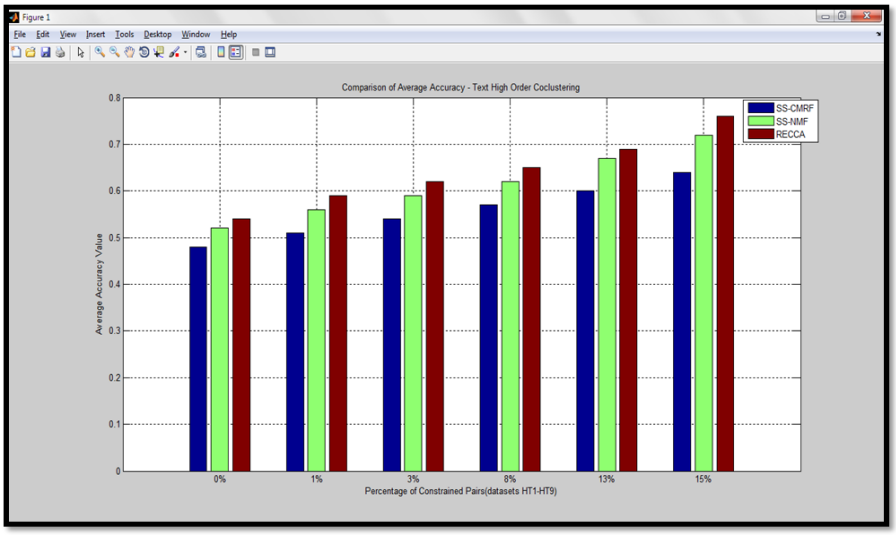


Fig. 3: Comparison of Average Accuracy – Text High Order Coclustering

Australian Journal of Basic and Applied Sciences, 9(10) Special 2015, Pages: 364-375

Figure 3 presents the performance comparison of average accuracy for text high order coclustering. It is most obvious that the proposed 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 Coclustering

Algorithms Percentage of Constrained Pairs	SS-CMRF	SS-NMF	RECCA
0%	0.48	0.52	0.54
1%	0.51	0.56	0.59
3%	0.54	0.59	0.62
8%	0.57	0.62	0.65
13%	0.6	0.67	0.69
15%	0.64	0.72	0.76

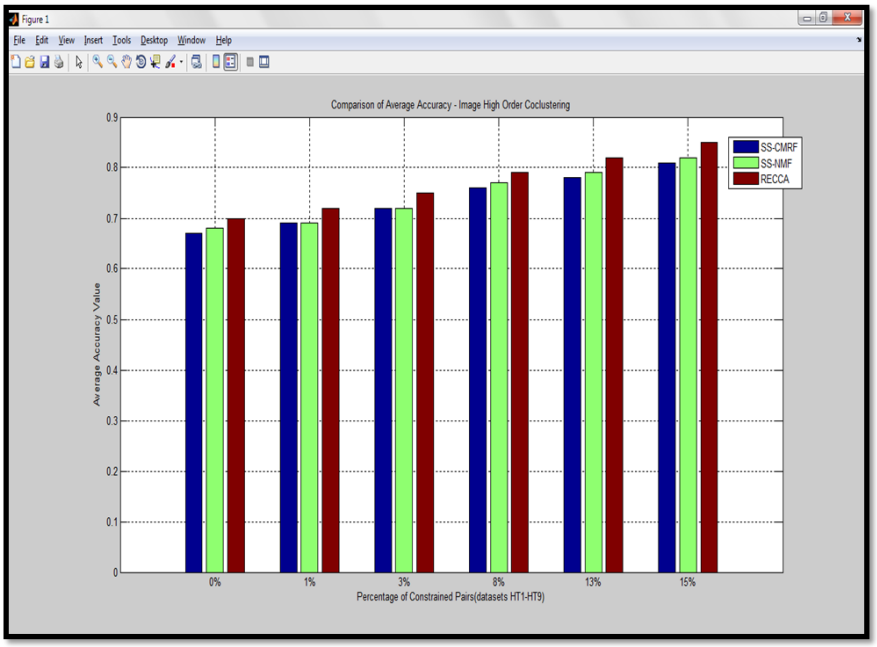


Fig. 4: Comparison of Average Accuracy – Image High Order Coclustering

Figure 4. presents the performance comparison of average accuracy for image high order coclustering. It can be perceived that the proposed 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
0%	0.67	0.68	0.7
1%	0.69	0.69	0.72
3%	0.72	0.72	0.75
8%	0.76	0.77	0.79
13%	0.78	0.79	0.82
15%	0.81	0.82	0.85

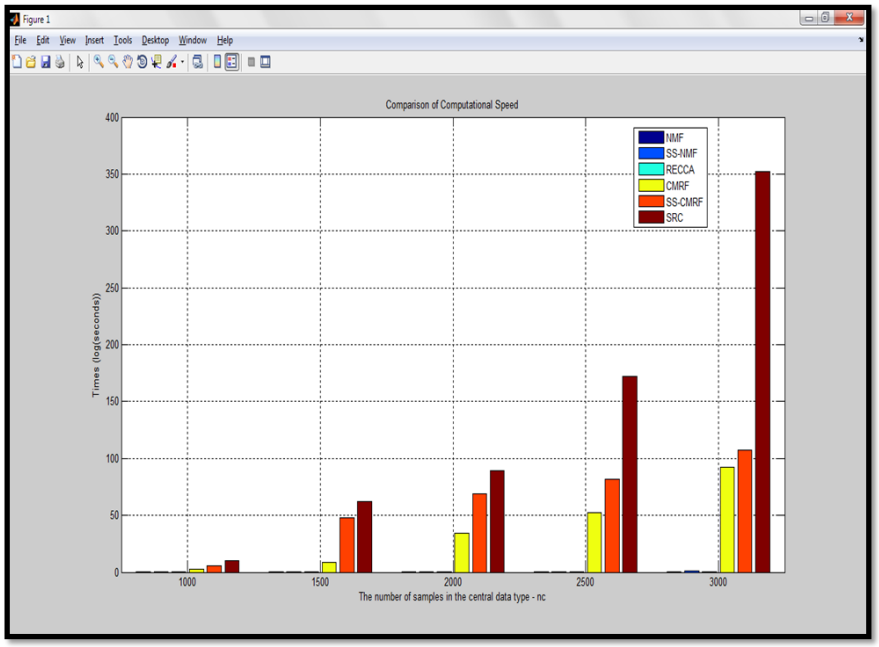


Fig. 5: Comparison of Computational Speed - In Log(Seconds) For Increasing N_c

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 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
1000	0.05	0.23	0.21	3	6	10
1500	0.2	0.45	0.38	9	48	62
2000	0.1	0.56	0.52	34	69	89
2500	0.4	0.62	0.57	52	82	172
3000	0.52	0.84	0.74	92	107	352

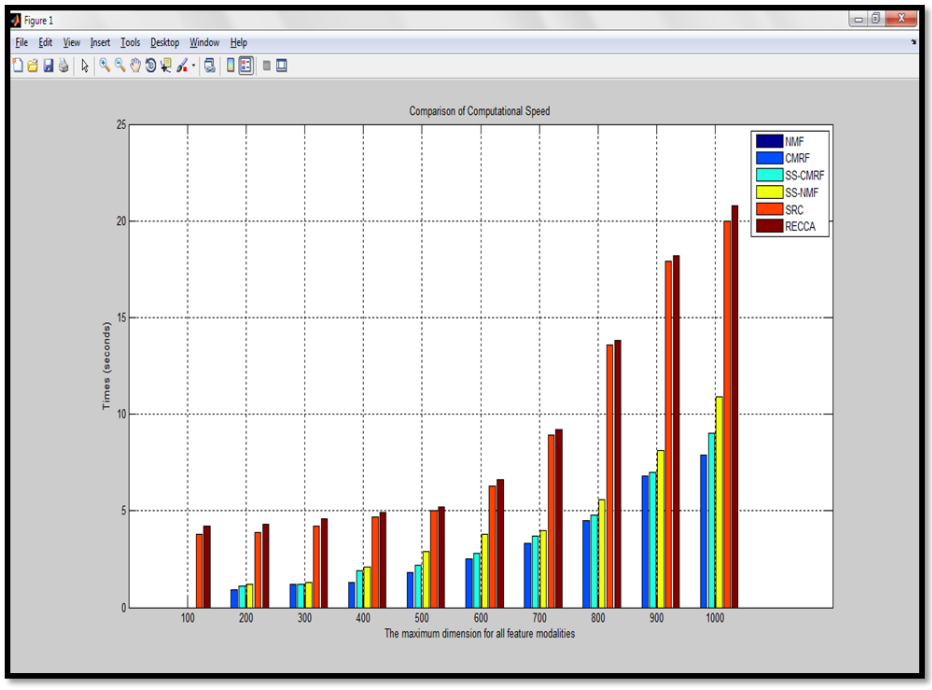


Fig. 6: Comparison of Computational Speed - In Log(Seconds) For Increasing N_p

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 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
100	0	0	0	0	3.8	4.2
200	0.2	0.9	1.1	1.2	3.9	4.3
300	0.2	1.2	1.22	1.3	4.2	4.6
400	0.1	1.3	1.9	2.1	4.7	4.9
500	0.2	1.8	2.2	2.9	5	5.2
600	0.3	2.5	2.8	3.8	6.3	6.6
700	0.4	3.3	3.7	4	8.9	9.2
800	0.2	4.5	4.8	5.6	13.6	13.8
900	0.4	6.8	7	8.1	17.9	18.2
1000	0.2	7.9	9	10.9	20	20.8

Conclusion:

This paper presented a mechanism with improved preprocessing technique for enzyme clustering. Initially the proposed work 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 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

Banerjee, Dhillon, Ghosh, Merugu, Modha, 2004. “A Generalized Maximum Entropy Approach to Bregman Co- Clustering and Matrix Approximation,” Proc. 10th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp: 509-514.

Bekkerman, Jeon, 2007. “Multi-Modal Clustering for Multimedia Collections,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp: 1-8.

Bekkerman, Sahami, 2006. “Semi-Supervised Clustering Using Combinatorial MRFs,” Proc. 23rd Int’l Conf. Machine Learning (ICML) Workshop Learning in Structured Output Spaces.

Bilenko, Basu, Mooney, 2004. "Integrating Constraints and Metric Learning in Semi-Supervised Clustering", 21st Int. Conf. on Machine Learning, pp: 11-18.

Bin Gao, Tie-Yan Liu, Wei-Ying Ma, 2006. “Star-Structured High-Order Heterogenous Data Co-Clustering Based on Consistent Information Theory,” Proc. Sixth IEEE Int’l Conf. Data Mining, pp: 880-884.

Clara Higuera, Gonzalo Pajares, Javier Tamames, Federico Morán, 2013. Expert system for clustering prokaryotic species by their metabolic features, Expert Systems with Applications, 40(15): 6185-6194, ISSN 0957-4174,

Dhillon, Mallela, Modha, 2003. “Information-Theoretic Co-Clustering,” Proc. Ninth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp: 89-98.

Erliang Zeng, Chengyong Yang, Tao Li, Narasimhan, 2007. "On the Effectiveness of Constraints Sets in Clustering Genes," Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on, pp: 79,86, 14-17.

Guoren Wang, Yuhai Zhao, Xiangguo Zhao, Botao Wang, Baiyou Qiao, 2010. "Efficiently mining local conserved clusters from gene expression data", Neurocomputing, 73(7-9): 1425-1437.

Inderjit Dhillon, 2001. "Co-clustering documents and words using bipartite spectral graph partitioning", KDD '01 Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp: 269-274.

Joachims, 1999. “Transductive Inference for Text Classification Using Support Vector Machines,” Proc. 16th Int’l Conf. Machine Learning, pp: 200-209.

Klein, Kamvar, Manning, 2002. "On the Efffectiveness of Constraints Sets in Clustering Genes From Instance-Level Constraints to Space-Level Constraints: Making the most of prior knowledge in data clustering", 19th Int. Conf. on Machine Learning, pp: 307-313.

Long, Wu, Zhang, Yu, 2006. “Spectral Clustering for Multi-Type Relational Data,” Proc. 23rd Int’l Conf. Machine Learning, pp: 585-592.

Marcelo Boareto, Michel Yamagishi, Nestor Caticha, Vitor Leite, 2012. "Relationship between global structural parameters and Enzyme Commission hierarchy: Implications for function

prediction", Computational Biology and Chemistry, 40: 15-19, ISSN 1476-9271,

Rosfuzah Roslan, Razib Othman, Zuraini Shah, Shahreen Kasim, Hishammuddin Asmuni, Jumail Taliba, Rohayanti Hassan, Zalmiyah Zakaria, 2010. "Utilizing shared interacting domain patterns and Gene Ontology information to improve protein–protein interaction prediction", Computers in Biology and Medicine, 40(6): 555-564

Ruochen Liu, Licheng Jiao, Xiangrong Zhang, Yangyang Li, 2012. "Gene transposon based clone selection algorithm for automatic clustering", Information Sciences, 204(30): 1-22.

Shahreen Kasim, Safaai Deris, Razib Othman, 2013. "Multi-stage filtering for improving confidence level and determining dominant clusters in clustering algorithms of gene expression data", Computers in Biology and Medicine, 43(9): 1120-1133.

Shi, Fan, Yu, 2010. "Efficient semi-supervised spectral co-clustering with constraints", IEEE International Conference on Data Mining, pp: 1043-1048.

Strehl, Ghosh, 2003. "Cluster ensembles – a knowledge reuse framework for combining multiple partitions", Journal Machine Learn. Res., pp: 583-617.

Sugato Basu, Mikhail Bilenko, Raymond J. Mooney, 2004. "A Probabilistic Framework for Semi-Supervised Clustering", Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), pp: 59-68, Seattle, WA.

Thanh-Phuong Nguyen, Tu-Bao Ho, 2012. "Detecting disease genes based on semi-supervised learning and protein–protein interaction networks", Artificial Intelligence in Medicine, 4(1): 63-71.

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

Wagstan, Cardie, Rogers, Schroedl, 2001. "Constrained K-Means Clustering with Background Knowledge", 18th Int. Conf. on Machine Learning, pp: 577-584.

Xu, Liu, Gong, 2003. “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.

Yanhua Chen, Lijun Wang, Ming Dong, Member, 2010. "Non-Negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering", Ieee Transactions On Knowledge And Data Engineering, 22.