Semi-Supervised Tumor Data Clustering via Spectral Biased Normalized Cuts

R. Rajeswari [1], G. Gunasekaran [2]
Research scholar,
St. Peter's Institute of Higher Education & Research,
St. Peter's University, Avadi, Chennai 600054.
Principal, Meenakshi Engineering College,
Chennai, India.

Abstract:- **Tumor clustering is the important techniques for tumor discovery, which is useful for the diagnosis and treatment of cancer. While different algorithms have been proposed for tumor clustering, few algorithms are useful for tumor discovery. In such case, (SS-NMF) Semi-Supervised Non-Negative Matrix Factorization frame-work for data clustering also been discussed. In SS-NMF, users can provide monitor & supervise for clustering with pair wise constraints on a few data objects through an iterative algorithm, they performed symmetric tri-factorization for the data similarity matrix clusters to infer. Though the quality of clustering on low dimensionality datasets for fewer constraints cannot be achieved efficiently. To overcome this problem, we first review the knowledge of experts as constraints in the clustering process, also propose a Feature Selection based model for Semi Supervised Cluster Ensemble (FSM-SSC) tumor clustering technique applied to dispel the effect of noisy data using K-means algorithm. Then the Double Selection, Based model for Semi-Supervised Cluster Ensemble (DSM-SSCE) that adopts both FS-SSCE and to improve the performance of tumor clustering by considering the prior knowledge of dataset. Enhancement of DSM-SSCE a new clustering technique is (MDSM-SSCE) Modified Double Selection, Based model for Semi-Supervised Cluster Ensemble which adopts multiple tumor clustering solutions. This proposed frame work model SS-SBNC Semi Supervised Spectral Biased Normalized Cut algorithm is applied in machine learning extensively. Spectral Clustering is built upon Spectral graph theory, and has the ability to process the clustering quality of constraints even with low dimensionality subsets and improvise the performance by the use of Biased Normalized Cut algorithm.**

Index Terms: Semi-Supervised Clustering, NMF, FSM-SSCE, DSM-SSCE, MDS-SSCE, SS-SBNC.

I. INTRODUCTION

A tumor may be primary or secondary. If it is the origin, then it is known as primary. If the part of the tumor spreads to another place and grows on its own, then it is known as secondary. The brain tumor affects CSF (Cerebral Spinal Fluid) and causes strokes. The physician gives the treatment for the strokes rather than the treatment for tumors. So the detection of the tumor is important for that treatment. The life expectancy of the person affected by the brain tumor will increase if it is detected at an earlier stage. Normally tumor cells are of two types Mass and Malignant. The detection of the malignant tumor is somewhat difficult to mass tumor. For the accurate recognition of the malignant a 3-D representation of brain and 3-D analyzer tool is required. This paper focuses on the detection of mass tumor. The development platform for the detection is mat lab because it is easy to develop and execute. At the end, we are providing systems that detect the tumor and its shape.

Types of Tumor

Brain tumors are classified based on the type of tissue involved, the location of the tumor, whether it is benign or malignant.

1) Benign brain tumor:

This type of tumor generally do not consist cancer cells and can be removed. Benign brain tumors usually have an obvious border or edge. They don't spread to other parts of the body. However, benign tumors can cause serious health problems.

2) Malignant brain tumor:

This consists of cancer cells and hence also called as brain cancer. They are likely to grow rapidly and can affect nearby healthy brain tissues. This type of tumor can be a threat for life. Now, depending on what is type of cell of tumor, doctor group brain tumors by grades. There are four grades as grade I to grade IV. Cells from low-grade tumors (grades I and II) look more normal and generally grow more slowly than cells from high-grade tumors (grades III and IV). Over time, a low-grade tumor may become a high grade tumor.

Tumor clustering plays an important role in identifying malignancies from cancer gene expression profiles. Although there exist a lot of research works for tumor clustering, most of them adopt single clustering algorithms, such as self-organizing map (SOM), hierarchical clustering (HC), model based clustering, nonnegative matrix factorization (NMF), penalized matrix decomposition (PMD), and so on, to identify different types of tumors from gene expression profiles. For example, the self-organizing feature map is applied by Golub et al. [14] to identify acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) from microarray data. The combination of hierarchical and probabilistic clustering techniques are adopted by Bhattacharjee et al. [7] to distinguish different subtypes of lung adenocarcinoma from cancer gene expression profiles. The two-way hierarchical clustering algorithm is designed by Notterman

et al. [28] to identify adenoma from adenocarcinoma and normal tissues. The model-based clustering approach is applied by Yeung et al., to perform tumor clustering from bio-molecular data. In general, most of the research works in the earlier years only consider simple clustering algorithms, and do not take into account other powerful approaches, such as graph theory, matrix decomposition, biclustering, and so on, which may be useful for identifying the structure of gene expression profiles.

The clustering approach is adopted by Leung et al. [23] to identify subgroups in liver cancer. The correlated biclustering approach is used to perform clustering analysis and identify local structures from gene expression profiles. In summary, although different tumor clustering approaches based on single clustering algorithms have been successfully applied to a variety of bio-molecular data, there is a further need to improve their robustness, stability and accuracy. In addition, these approaches have not adequately considered the main role of domain knowledge in facilitating the characterization of cancer gene expression profiles. These techniques can be roughly divided into two steps: ensemble generation and consensus aggregation. The first step seeks to generate diverse clustering solutions using different data perturbation techniques and different basic clustering algorithms, while the task of the second stage is to find a valid consensus aggregation of the clustering solutions in the ensemble to enhance the performance and stability of clustering. In general, compared with single clustering algorithms, cluster ensemble approaches are more robust, stable and effective for tumor clustering from gene expression profiles. There are a number of research works which adopt cluster ensemble approaches to perform tumor clustering. For example, the random subspace based cluster ensemble approach in combination with a cluster stability score is designed by Smolkin et al. to perform tumor clustering from cancer gene expression profiles. The re-sampling based consensus clustering approach is proposed by Monti et al. [26] to per-form cluster analysis on gene expression profiles. Handl et al. [17] provided a survey on the application of cluster ensemble approaches to cancer discovery on bio-molecular data.

The randomized map based cluster ensemble approach is designed by Bertoni et al. [6] to identify the subclasses of different cancer types from microarray data. The graph theory based consensus clustering approach is studied for discovering the underlying structure of cancer data sets. The random projection based fuzzy cluster ensemble framework is designed by Avogadri et al. [4] to perform tumor clustering. The link-based cluster ensemble approach is proposed by I am-on et al. [19] to perform cluster analysis on cancer gene expression profiles. The perturbation based consensus clustering framework is investigated to perform microarray data clustering. A hierarchical clustering based cluster ensemble approach is proposed by Mahata et al. [24] and successfully applied to the melanoma cancer data set and the breast cancer data set to identify different cancer subtypes. The hybrid cluster ensemble frameworks, including triple spectral clustering based consensus clustering (SC^3) and double spectral clustering based consensus clustering (SC^2Ncut), are designed to identify different types of cancers from bio-molecular data. While cluster ensemble approaches have been successfully applied to gene expression profiles, few of them,

(1) Considers incorporating prior knowledge of the characteristics of the profiles into the cluster ensemble framework.

(2) It takes into account the degree of dependencies between the clustering solutions in the ensemble on the accuracy of the cluster validation process.

(3) It considers adopting feature selection techniques to remove noisy genes, and using these techniques to prune redundant clustering solutions in the ensemble at the same time.

In order to deal with the limitations of traditional cluster ensemble approaches and further improve the performance of tumor clustering from cancer gene expression profiles, we apply feature selection and suitable constraints into the cluster ensemble framework, and propose three kinds of feature selection based semi-supervised clustering ensemble frameworks, known as the feature selection based semi-supervised clustering ensemble framework (FS-SSCE), the double selection based semi-supervised clustering ensemble framework (DS-SSCE) and the modified DS-SSCE (MDS-SSCE), for tumor clustering from bio-molecular data. FS-SSCE proposes to adopt different feature selection techniques [8], to perform gene selection, and generate a set of new data sets from the original data set which are as diverse as possible. DS-SSCE not only performs gene selection, but also performs clustering solution selection in the ensemble using feature selection. It also introduces a confidence factor which takes into account prior knowledge of the data set into the process of constructing the consensus matrix. Compared with DS-SSCE, MDS-SSCE considers multiple solution selection strategies and applies an aggregated solution selection function in the process of clustering solution selection. The empirical results on real cancer gene expression profiles show that

(i) FSM-SSCE, DSM-SSCE and MDSM-SSCE work well on bio-molecular data.

(ii) MDSM-SSCE outperforms most of the state-of-the-art tumor clustering approaches.

II. LITERATURE SURVEY:

In this section, we provide a review of related works on using user provided information to improve data clustering. We first discuss some algorithms in which prior knowledge is in the form of labeled data. Next, we describe other algorithms for which pair wise constraints are required to be known a priori.

SS-constrained-Kmeans and SS-seeded-Kmeans [3] are the two well-known algorithms in semi-supervised clustering with labels. The SS-constrained-Kmeans seeds the k-means algorithm with the given labeled data and keeps that labeling unchanged through-out the algorithm. Moreover, it is appropriate when the initial seed labeling is noise-free, or if the user does not want the labels of the seed data to change. On the other hand, the SS-seeded-

Kmeans algorithm changes the given labeling of the seed data during the course of the algorithm. Also, it is applicable in the presence of noisy seeds, since it does not enforce the seed labels to remain unchanged during the clustering iterations and can therefore abandon noisy seed labels after the initialization step. Semi-supervised clustering with labels has been successfully applied to the problem of document clustering. Hotho et al. [25] proposed incorporating background knowledge into document clustering by enriching the text features using WordNet.1 In Jones et al. [30], some words per class and a class hierarchy were sought from the user in order to generate labels and build an initial text classifier for the class. A similar technique was the user is made to select interesting words from automatically selected representative words for each class of documents. These user identified words were then used to re-train the text classifier. Active learning approaches have also found applications in semi-supervised clustering. Godbole et al. [18] has proposed to convert a user recommended feature into a mini-document which is then used to train an SVM classifier. This approach has been extended by Raghavan et al., which adjusts SVM weights of the key features to a predefined value in binary classification tasks.

Recently, Huang and Mitchell [26] presented a probabilistic generative model to incorporate A number of previous works adopt feature selection approaches to choose an optimal gene subset in the task of cancer classification. For example, Mundra and Rajapakse [27] integrated the minimum-redundancy maximum relevancy filter into the support vector machine recursive feature elimination approach to select an optimal gene subset and improve the accuracy of cancer classification. Sharma et al. [32] proposed a top-r feature selection approach to per-form gene selection with respect to classification accuracy from microarray data. Chiang and Ho [10] proposed a feature selection approach in combination with a radial basis function based neural network to perform gene selection and improve the performance of cancer classification. The multi-criterion fusion based recursive feature elimination (MCF-RFE) algorithm to select an optimal gene subset from gene expression data sets. A feature selection approach based on an efficient margin based sample weighting algorithm to improve the performance of gene selection investigated how to use model-based entropy to perform feature selection from gene expression data. Mao and Tang [25] adopted a recursive Mahalanobis separability measure to find an optimal gene subset. Cheng et al. [9] designed the Fisher-Markov selector based on Markov random field optimization techniques to perform gene selection from micro-array gene expression data sets. While these approaches are often used in the task of cancer classification, few of the previous works consider applying feature selection to perform tumor clustering, and none of these works adopt feature selection to perform clustering solution selection in an ensemble. Compared with these works, we not only incorporate feature selection into the ensemble framework for tumor clustering, but also adopt this technique to perform clustering solution selection. Furthermore, we take into account multiple clustering solution selection strategies (CSSs) instead of a single clustering solution selection strategy.

III. PROPOSED METHODOLOGY:

We propose a novel semi-supervised model clustering ensemble frame works DSM-SSCE and MDSM-SSCE to perform data clustering on bio-molecular data. The incorporation of expert's knowledge as constraints in to the cluster ensemble framework. The adoption of feature selection techniques in an ensemble to remove noisy genes in the gene dimensions. The adoption of multiple feature selection techniques to perform clustering solution selection in the ensemble.

1. *Feature Selection based model for Semi Supervised Cluster Ensemble (FSM-SSC)*

Feature selection, or subset selection, is the method of reducing dimensionality in machine learning. It is important for different reasons: first total computation can be reduced if we can reduce the dimensionality. Secondly all the features may not be helpful to classify the data; some may be redundant and irrelevant from the classification point of view. Thus it is needed to determine automatically relevant subset of features. In order to address the above mentioned problems, feature selection is needed both for unsupervised as well as supervised classification problems. But most of these techniques pose the feature selection problem as a single objective optimization technique. They have mostly optimized a single cluster quality measure. In recent years there are some approaches which use multi objective optimization to solve the unsupervised feature selection problem. A multiobjective wrapper based approach to solve the unsupervised feature selection problem is developed by Morita et al.*K*-means clustering technique is utilized as the underlying partitioning method and authors have varied the number of clusters in a range.

2. *Single Selection Based Semi-Supervised Clustering Ensemble (SSBM-SSC)*

The original data set X is given in which it have gene expression profile with n samples and m genes(X ¼ fx1; x2; . . . ; xng, xi 2 X,xi ¼ fxi1; xi2; . . . ; ximg), first FSM-SSCE applies different feature selection approaches to the gene expression profile X, projects X onto a set of low dimensional spaces, and generates B new data sets. To perform clustering on the B new data sets to obtain B corresponding clustering solutions, the pair wise constrain based K-means algorithm is used, which represents the expert's knowledge is adopted. Finally, by using the selected clustering solutions FS-SSCE constructs a consensus matrix, and adopts the pair wise constraint based K-means algorithm to obtain the final result.

In the first stage, to generate a set of new data sets and to remove noisy genes FS-SSCE adopted for feature selection. It is known that, the feature selection approaches are divided into two most important types: supervised feature selection approaches and unsupervised feature selection approaches. To implement the pair wise constrained clustering framework known as PC-K means to estimate the labels Y of the cancer samples, which takes

into account a limited number of must-link and cannot-link constraints between pairs of cancer samples specified by the experts.

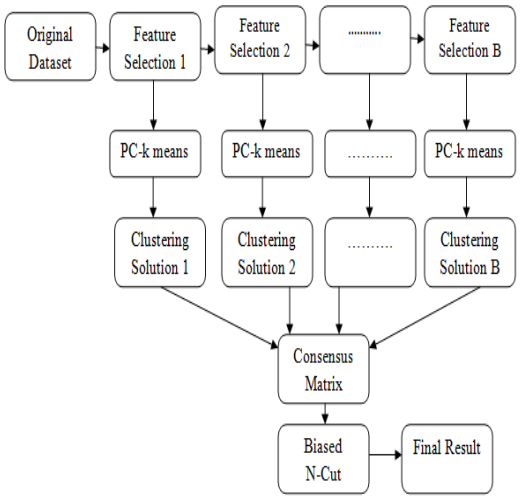

Fig.1Provides an overview of the feature slection based Semi-supervised model for clustering ensemble framework.

In this dissertation, we propose to view clustering solutions as new attributes of the original data set, and adopt feature selection approaches, such as feature selection based on mutual information maximization, mutual information feature selection, max-relevance min-redundancy, joint mutual information, double input symmetrical relevance, conditional infomax feature extraction, interaction capping and conditional redundancy, to perform clustering solution selection in DSM-SSCE.

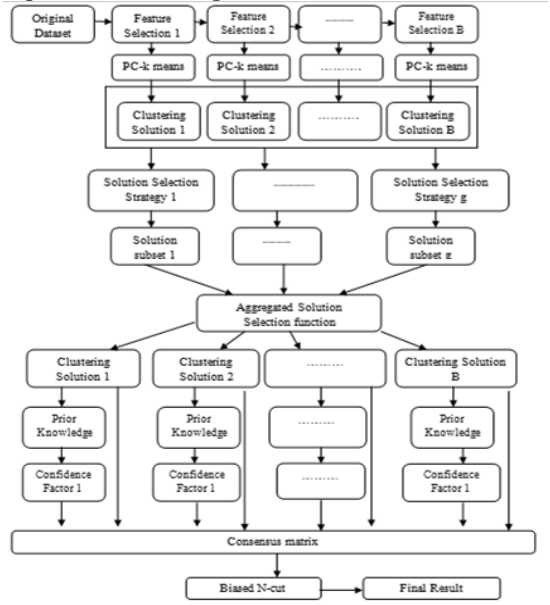

Fig. 2. An overview of the double selection based semi-supervised clustering ensemble framework

3. Double Selection Based Model for Semi-Supervised Clustering Ensemble (DSBM-SSC)

An overview of the double selection based semi-supervised clustering

Ensemble framework is shown in fig.2. Compared with FS-SSCE, DS-SSCE adopts feature selection techniques to perform clustering solution selection and choose a set of representative clustering solutions.

Traditional clustering solution selection strategies first adopt the normalized mutual information or the adjusted Rand index (ARI) to measure the differences between any pair of clustering solutions (Ib; Is) (b; s 2 B, B is the number of clustering solutions) in the ensemble.

The diversity g of all the clustering solutions in the group is sorted in ascending order. As a final point, mainly CSSs are defined based on the diversity, which include strategies which select clustering solutions with low diversity, high diversity and medium diversity. In addition, random selection strategy and adaptive clustering group selection strategy (ACES) based on an adaptive factor proposed by Azimi and Fern have also been used. While CSSs have been successfully applied to perform clustering solution selection. Most of them do not apply optimization in the process of clustering solution selection, which will prevent them from obtaining the best result.

In Fig. 2, DSM-SSCE first obtains a set of clustering solutions E ¼ fI1; I2; . . . ; IBg by PC-K means from different data sets in the low dimensional space .Then, similar feature selection approaches are adopted to perform clustering solution selection in E. DSM-SSCE will obtain g clustering solution subsets E1; E2; . . .; Eg, and each subset includes a small number of clustering solutions. DSM-SSCE adopts direct combination as the aggregated solution selection function, which directly combines the clustering solutions in each subset, obtains a new set E0 of clustering solutions, and constructs an adjacency matrix Ab with entries abij for each clustering solution Ib in E0 (where b 2 f1; . . .; B0g, and B0 is the number of clustering solutions in E0).

Subsequently, DSM-SSCE computes an assurance factor based on prior knowledge of the original data set which is represented by a set of pair wise constraints for each clustering solution Ib. In particular, it constructs a matrix P based on the must-link constraints and the cannot-link constraints.

Most of the pair wise constraints satisfies a clustering solution Ib, to have a high Confidence factor ab. Otherwise, ab will be small. Next, DSM-SSCE constructs a consensus matrix A0 by considering all the membership matrices of the clustering solutions and the corresponding confidence factors.

In Conclusion, the Biased Normalized Cut algorithm is adopted to partition the consensus matrix A0 and obtain the final result. In synopsis FS-SSCE is a special form of DS-SSCE without taking into account the clustering solution selection process in the group, and without considering the confidence factor in the process of constructing the consensus matrix.

4. The Modified Double Selection Based Semi-Supervised Clustering Ensemble (MDSBM-SSC)

Compared with DSM-SSCE, MDS-SSCE applies multiple clustering solution selection strategies for choosing clustering solutions, generates multiple solution subsets, and adopts a refined subset of clustering solutions using feature selection based on max-relevance min-redundancy, which is used as the aggregated solution selection function instead of adopting direct combination. Explicitly, MDS-SSCE first adopts NMI to calculate the similarity between two clustering solutions in the new set of clustering solutions E0 ¼ fI1; I2; . . . ; IB0g, and selects the most representative clustering solution I.

It selects the clustering solution Ib from the unvisited clustering solution set (E0 * E*) with the best tradeoff between relevance and redundancy in a step by step manner, until the number of clustering solutions satisfies the requirement which is pre-specified by the user. The subsequent steps of MDSM-SSCE are the same as those of DSM-SSCE.

The time complexity of the proposed algorithm MDSSSCE is related to the computational cost of different feature selection algorithms ($O(nm^2)$), the PC-Kmeans algorithm (O(n)), the construction of the consensus matrix ($O(n^2)$) and the Biased Normalized Cut algorithm ($O(n^3)$). As a result, the time complexity of MDS-SSCE is $O(nm^2+n^3)$.

5. Spectral Clustering

Spectral clustering involves constructing an affinity matrix from the data and requires, in the original version, the prior knowledge of the number of clusters. In the incremental learning the update process is performed using each data vector at a time, and the centroid is updated by rule defined as:

$$mi(t + 1) = mi(t) + hci[x(t) - mi(t)$$

where t denotes time, x(t) is the input vector, mi(t) is the centroid to update and hci is the eighborhood function [3]. In batch training algorithm, at each step the whole data set is partitioned along the Neurons of the SOM map without performing any update. After partitioning the data set, all neurons are updated according to the following rule:

$$mi(t + 1) = Pni=1\ hcixjPni=1\ hci$$

At the end of each training step, the new centroid mi(t + 1) is therefore a weighted average of the data vectors belonging to the i − th cluster. The SOM software used in our experiments is the SOM Matlab Toolbox [8], that is more recent than SOM PAK [3] used in [4]. The SOM Matlab Toolbox allows to choose between incremental or batch training, contrarily to SOM PAK where only incremental algorithm is implemented.

6. BIASED NORMALIZED CUTS :

To incorporating the prior information given to us about the image to define the notion of biased normalized cuts. Recall that our problem is: we are given a region of interest in the image and we would like to segment the image so that the segment is biased towards the specified region. A region is modeled as a subset T ⊆ V, of the vertices of the image. We would be interested in cuts (S, $\bar{S}$) which not only minimize the Biased Normalized Cut value but, at the same time, have sufficient correlation with the region specified by T. To model this, we will first associate a vectors T to the set T .

A. We have defined it in a way such that $Pi\in V\ sT(i)di = 0$ and $Pi\in V\ sT(i)\ 2di = 1$.

This notion of biased normalized cuts is quite natural and motivated from the theory of local graph partitioning where the goal is to find a low-conductance cut well correlated with a specified input set. The correlation is specified by a parameter κ ∈ (0, 1). This allows us to explore the possibility of image segments which are well-correlated with the prior information which may have much less Biased Normalized Cut value than T itself and, hence, refine the initial guess. In particular, we consider the spectral relaxation in Figure 3 to a κ-biased Biased Normalized Cut around T. Note that x = sT, is a feasible solution to this spectral relaxation. Also note that if v2 satisfies the correlation constraint with sT, then it will be the optimal to this program. What is quite interesting is that one can characterize the optimal solution to this spectral program under mild conditions on the graph G and the set T and, as it turns out, one has to do very little effort to compute the optimal solution if one has already computed the spectrum of LG.

B. Algorithm 1: Biased Normalized Cuts (G, ω,s_T,γ)

Require: Graph G = (V,E), edge weight function ω, seed s_T and a correlation parameter $\gamma \in (-\infty, \lambda_2(G))$

Step1: $A_G(i,j) \leftarrow \omega(i,j)$, $D_G(i,i) \leftarrow \sum_j \omega(i,j)$

Step 2: $L_G \leftarrow D_G - A_G$, $L_G \leftarrow D_G^{-1/2} L_G D_G^{-1/2}$

Step3: Compute $u_1, u_2, \ldots u_k$ the eigen vectors of L_G corresponding to the K smallest eigen values $\lambda_1, \lambda_2, \ldots \lambda_k$.

Step 4: $\omega_i \leftarrow \frac{u_i^T D_G S_T}{\lambda i - \gamma}$, for i=2,…,K.

Step 5: Obtain the biased normalized cut, $x * \alpha \sum^{k}_{i=2} \omega_i u_i$

Our method can be faster than the min-cut/max-flow cut based approaches in an interactive setting as these eigenvectors need to be computed just once. In addition the real valued solution like the one shown in proposed method might provide the user better guidance than a hard segmentation produced by a min-cut algorithm.

C. Performance Evaluation & Comparison of Cluster Ensemble Approaches

We now compare MDS-SSCE with state-of-the-art tumor clustering approaches with respect to NMI on all the data sets. The approaches include the PC-Kmeans algorithm (PCK), the PC-Kmeans based semi-supervised clustering group approach (PCKCE), the link-based cluster group approaches based on average linkage with connected-triple-based similarity (LCE(CTS)),

simRankbased similarity (LCE(SRS)), and approximate SimRankbased similarity (LCE(ASRS)).

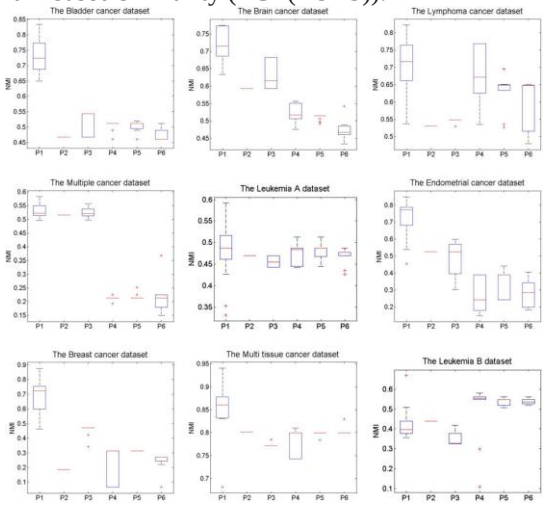

Fig.4:Comparison of different pproaches (where P1, P2, P3, P4, P5 and P6 denote the approaches MDS-SSCE, PCK, KCE, LCE(CTS), LCE (SRS) and LCE(ASRS) respectively.

We also perform t-test with a confidence level of 95 percent to study the difference between MDS-SSCE and other cluster group approaches.

The Performance of MDSM-SSCE, DSM-SSCE and FSM-SSCE in Terms of the Average and the Standard Deviation Values of NMI on the Different Data Sets (the Bold Values Denote Better Results)

IV. SOLUTIONS OF THE PRIOR KNOWLEDGE EFFECT:

In order to study the effect of prior knowledge, we compare MDS-SSCE using prior knowledge with MDS-SSCE without using prior knowledge (MDS-SSCE(no)) with respect to NMI on the Lymphoma cancer data set, the Bladder cancer data set, the Endometrial cancer data set and the Multi-tissue data set. Fig. 6 provides the comparison results of MDS.

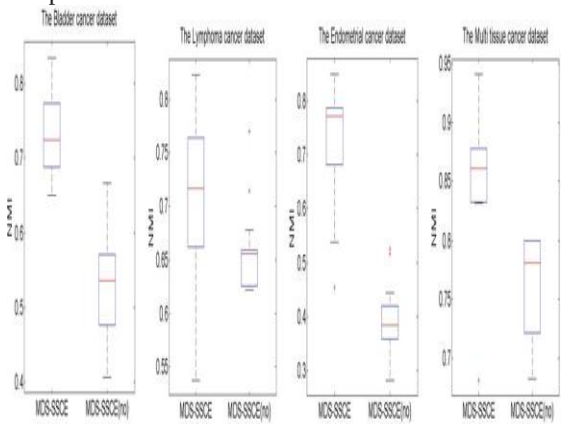

Fig.3The Effect of effect of prior knowledge

The performance of MDS-SSCE, PCK,PCKCE, LCE(CTS), LCE(SRS) and LCE(ASRS), referred to as P1, P2, P3, P4, P5, and P6, with respect to NMI on the data sets. It is observed that MDS-SSCE significantly outperforms their competitors. For example, MDSM-SSCE obtains the best average NMI value 0.86 on the Multi tissue cancer data set, and successfully assigns most of the samples to the corresponding classes which include the breast, the prostate, the lung and the colorectal. The possible reasons are as follows: (1) Compared with other cluster group approaches, the experts' knowledge in the form of pair wise constraints is now considered, which provides more information to facilitate clustering.

(2) Compared with single PC-Kmeans, MDSM-SSCE integrates multiple clustering solutions to improve the performance of single PC-Kmeans and obtains more accurate results.

(3) Compared with PCKCE, feature selection techniques are adopted to avoid the effect of noisy genes. In addition, an optimal subset of clustering solutions in the group is selected. In general, MDSMSSCE outperforms most of the state-of-the-art tumor clustering approaches when applied to bio-molecular data.

SSCE and MDS-SSCE(no). It can be seen that MDS-SSCE clearly outperforms MDS-SSCE(no) on all the data sets. For example, the average NMI value obtained by MDS-SSCE is 0.7301, which is 20 percent greater than the NMI value of 0.5223 in the case of MDS-SSCE(no). The possible reason is that the prior knowledge provided by the experts plays an important role.

The t-Test Results between MDSM-SSCE and PCK, PCKCE, LCE(CTS), LCE(SRS), LCE(ASRS) on the Cancer Data Sets (the Value 1 Indicates that the Null Hypothesis that There Is No Significant Difference with a Confidence Level of 95 percent between the Two Means Can Be Rejected; Otherwise, the Value is 0 role in elucidating the characteristics of gene expression profiles.

IV. CONCLUSION AND FUTURE WORK

This paper explores the clustering problem based on bio-molecular data. The major contribution of the paper is an advanced modified double selection model based semi supervised clustering group framework known as MDSM-SSCE for performing cancer profile clustering in the gene expression. When Compared with traditional clustering approaches, the proposed framework model is featured by the following solutions: (1) The incorporation of experts knowledge into the cluster group framework as Constraints. (2) To remove noisy genes in the gene dimension the adoption of feature selection techniques. (3) To perform clustering solution selection in the group by adopting multiple feature selection techniques. (4) To improvise the performance and to obtain quality constraints with low dimensionality subsets by adopting the aggregated solution selection. The proposed framework works well on bio-molecular data with the experimental results on cancer gene expression profiles and outperforms with all most of the state-of-the-art clustering approaches. In future, we shall consider other ways to make use of background knowledge in gene expression analysis.

V. REFERENCES:

[1] J. Azimi and X. Fern, "Adaptive cluster ensemble selection," in Proc. Int. Joint Conf. Artif. Intell., 2009, pp. 992–997.
[2] A. A. Alizadeh, M. B. Eisen, and R. E. Davis et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," Nature, vol. 403, pp. 503–511, 2000.
[3] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub,and S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," Nature Genetic, vol. 30, no. 1, pp. 41–47, 2002.
[4] R. Avogadri and G. Valentini, "Fuzzy ensemble clustering based on random projections for DNA microarray data analysis," Artif. Intell. Med., vol. 45, no. 2-3, pp. 173–183, 2009.
[5] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in Proc. SIAM Int. Conf. Data Mining, 2004, pp. 1–8.
[6] A. Bertoni and G. Valentini, "Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses," Artif. Intell. Med., vol. 37, no. 2, pp. 85–109, 2006.
[7] A. Bhattacharjee, W. G. Richards, and J. Staunton et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinomas sub-classes," Proc. Nat. Acad. Sci. USA, vol. 98, no. 24, pp. 13790–13795, 2001.
[8] G. Brown, A. Pocock, M.-J. Zhao, and M. Luj_an, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," J. Mach. Learn. Res., vol. 13, pp. 27–66, 2012.
[9] Q. Cheng, H. Zhou, and J. Cheng, "The Fisher-Markov selector: Fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data," IEEE Trans. PatternAnal.Mach.Intell.,vol.33,no.6,pp.1217–1233, Jun.2011.
[10] J.-H. Chiang and S.-H. Ho, "A combination of rough-based feature selection and RBF neural network for classification using gene expression data," IEEE Trans. Nanobiosci., vol. 7, no. 1, pp. 91–99, Mar. 2008.
[11] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: A comparative study," BMC Bioinformatics, vol. 9, article 497, 2008.
[12] L. Dyrskjot, T. Thykjaer, and M. Kruhoffer, et al., "Identifying distinct classes of bladder carcinoma using microarrays," Nature Genetic, vol. 33, no. 1, pp. 90–96, 2003.
[13] X. Z. Fern and W. Lin, "Cluster ensemble selection," Statist. Anal. Data Mining, vol. 1, no. 3, pp. 787–797, 2008.
[14] T. R. Golub, D. K. Slonim, and P. Tamayo, et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression," Science, vol. 286, no. 5439, pp. 531–537, 1999.
[15] J. Gu, W. Feng, J. Zeng, H. Mamitsuka, and S. Zhu, "Efficient semi-supervised MEDLINE document clustering with mesh semantic and global content constraints," IEEE Trans. Cybern., vol. 43, no. 4, pp. 1265–1276, Aug. 2013.
[16] J. Hartigan, Clustering Algorithms. Hoboken, NJ, USA: Wiley, 1975.
[17] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis bioinformatics," Bioinformatics, vol. 21, no. 15, pp. 3201–3212, 2005.
[18] Y. Hoshida, J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Subclass mapping: Identifying common subtypes in independent disease data sets," PLoS ONE, vol. 2, no. 11, p. e1195, 2007.
[19] N. Iam-on, T. Boongoen, and S. Garrett, "LCE: A link-based clus-ter ensemble method for improved gene expression data analy-sis," Bioinformatics, vol. 26, no. 12, pp. 1513–1519, 2010.
[20] A. Jakulin, "Machine learning based on attribute interactions," PhD thesis, Dept. of Computer Science, Univ. of Ljubljana, Ljubljana, Slovenia, 2005.
[21] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," Nucleic Acids Res., vol. 40, no. (Database issue), pp. D109–D114, 2012.
[22] Bansal N, Blum A, Chawla S (2002) Correlation clustering. Proceedings of the 43rd symposium on foundations of computer science, pp 238–247
[23] Bar-Hillel A, Hertz T, Shental N, Weinshall D (2003) Learning distance functions using equivalence relations. Proceedings of the 20th international conference on machine learning, pp 11–18
[24] Basu S, Banerjee A, Mooney RJ (2002) Semi-supervised clustering by seeding. Proceedings of the 19th international conference on machine learning, pp 27–34
[25] Basu S, Bilenko M, Mooney RJ (2004) A probabilistic framework for semi-supervised clustering. Pro-ceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining, pp 59–68
[26] Blum A, Mitchell TM (1998) Combining labeled and unlabeled data with co-training. Annual workshop on computational learning theory, Proceedings of the 11th annual conference on Computational learning theory, pp 92–100
[27] Boley D (1998) Principal direction divisive partitioning. Data Mining Knowledge Discovery .