Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7778406

AN ENHANCED DATA MINING CLUSTERING APPROACH FOR RETRIEVAL OF INFORMATIONAL PATTERNS IN GRAPHS VIA IF-NMFCM AND PSO-BASED FEATURES

Yashwant Singh Sangwan[1], Akshat Gupta[2], Nelofar Bashir3, Aijaz Ahmad Wani[4], Anish Soni[5], Anil Kumar[6]

[1]Assistant Professor, Department of Computer Science, Govt. Degree College, Hisar (Haryana), INDIA
Email: yssangwan@gmail.com

[2]Department of Computer Science & Engineering., National Institute of Technology, Goa, INDIA
Email: guptaakshat702@gmail.com

[3]Research Scholar, School of Computing and Artificial Intelligence, Department of Computer Science, NIMS University, Jaipur (Rajasthan), INDIA
Email: nelofarbashir111@gmail.com

[4]Lecturer, Department of Computer Sciences, Govt. Degree College Pulwama (J&K), INDIA
Email: aijazsidiquee@gmail.com

[5]Assistant Professor, Department of Computer Science, Akal Degree College, Mastuana Sahib, Sangrur (Punjab), INDIA
Email: soni_anish@yahoo.com

[6]Assistant Professor, Department of Computer Science, Public College, Samana, Patiala (Punjab), INDIA
Email: akjjakhar@gmail.com

Abstract: Finding the interaction flow between various compounds found in chemicals or genes as graph interactions is made possible by the important research field known as graph pattern mining. The task of mining the graphs to extract the interaction flow would be more challenging. The most common data mining method, called clustering, groups the nodes that make up graphs with extensive communication to discover the graphical interaction flow. The process of graph pattern mining can be carried out using a variety of research methodologies. We introduced Flow-Based Algorithms (FBA) for Local Graph Clustering in our previous research approach, which aimed to group various graphs based on flow. However, the presence of numerous irrelevant features, which needed to be avoided for better performance, prevented this method from increasing the clustering accuracy. This issue introduces the Irrelevant Feature aware NMF clustering method, which focuses on the targeted research work (IF-NMFCM). To prevent other associated dataset complexities and increase the cluster accuracy of strategies during this work, it is crucial to remove irrelevant features. Here, principal component analysis is initially used for data pre-processing. Then, using the PSO method or the particle swarm optimization method, feature selection is carried out in order to avoid the irrelevant features. Adopting this approach would produce the ideal set of features, which could increase the accuracy of clustering. Finally, pre-processed data would be represented using a graph after being clustered using the hierarchical NMF clustering method. The proposed research work leads to provide higher results than the current method in terms of improved

accuracy rate, according to the overall analysis of the proposed research method done in a Mat lab simulation environment.
Keywords: Graph mining, Clustering, Pattern mining, Useful information extraction.

INTRODUCTION

The research community that is working on data mining has given a significant amount of attention to the process of developing computational approaches that rely on classification strategies for determining the type of data that is provided in the form of graphs. The use of graph classification has the potential to yield enormous benefits in a variety of contexts. When it comes to issues concerning drug discovery, graph-based classification will be utilized wherever it is able to find the structural characteristics of a compound and their impact on the treatment of a particular illness. In recent times, numerous remarkable classification algorithms for graph-dependent knowledge have been developed [1, 2]. A number of algorithms make the assumption that the fundamental components of a graph—including its nodes, edges, paths, strongly connected components, trees, and sub-graphs—are responsible for the graph's inherent properties. These substructure-dependent algorithms identify the essential components that are present in each graph and make use of those components in order to classify the graphs into their respective categories [3].

However, in recent times, frequent subgraph-based technique has been introduced. In this technique, frequent subgraph mining algorithms are used for sub-graph generation, which is observed to occur several numbers of times in the graph database adequately, and for construction of the feature vectors using the obtained subgraphs [4]. This technique was introduced in recent times. A method of mathematical programming that is known as gBoost or boosting methodology is proposed here as a tool for the collection of additional informative pattern data. In this particular study, an effort is made to construct a prediction-based regulation using gBoost, which will result in decreased iteration. The implementation of this form of boosting and its application to facilitate graph knowledge contributes to the development of branch-and-bound pattern-search rule, which in turn helps in providing support for the DFS code tree. Therefore, the search that is created is then further reused for iterations at a later stage in order to reduce the amount of time needed for computation [5]. When learning multiple connected tasks at the same time, the sample size for each job will increase, which will improve the performance achieved in prediction. Therefore, learning multiple tasks at once is especially helpful when the size of the training sample is relatively small for each job [6].

The improvement of generalization performance is the primary goal of MTL, which is also referred to as multi-task learning. Learning multiple tasks at the same time that are connected to one another allows for this performance to be geared toward providing classification or supervised regression. MTL has been generating interest levels and intense analytical levels of interest in relation to the processing of data within the machine learning groups in recent times [7]. It has been determined that learning several interconnected tasks at the same time typically results in an improvement in modeling accuracy. In lieu of different multiple types of cancer that each have their own data set, the issues related to the status prediction of cancer have been taken into consideration and based on the microarray data sets [8]. Each data set contains

various patient-specific microarray information that was collected from a variety of sources. Patients who fall into this category are those who may or may not be given a diagnosis of the cancer that was used in the prediction. It would appear that the various types of cancer can be quite distinct from one another, such as prostate and breast cancer, while other types of cancer share certain characteristics with ovarian and breast cancer. The models that have been developed are learning models that have taken into consideration the fact that different cancer types share similarities and characteristics. On the other hand, a few additional learning models are likely going to take into consideration a variety of cancers.

In the case of MTL [9], a selection of common subsets of features has been considered for all of the relevant tasks using the methods that are currently used in feature selection. Recently, there has been a lot of interest shown in regards to Graph classification because of the variations that are on the rise in the application while including those aspects and objects that are structurally complex and the relationships that are therein. This is due to the fact that Graph classification includes all of these aspects and objects. Although in the case of graphs, these aspects have provided a means for learning, the space on account of data features is probable, and it necessitates careful exploration in order to render favor classes whose cost is higher [10].

RELATED WORK

Vogelstein and colleagues [11] conducted a statistical study by utilizing a novel-based graph/class model. It presented two different approaches as possible methods for estimating the signal-sub graph. The first one utilized only the information that was on the label of the vertex, whereas the second one made use of a structure that is graph-based. An algorithm called igBoost was presented by Pan et al. [12] for handling imbalanced class distributions and noise. This algorithm boosts imbalanced graphs.

The Correspondence-based Quality Criterion was proposed by Thomaet al. [13] as a method for performing effective feature selection among frequent subgraphs. God bole et al. [15] implemented the methodologies for improving the discriminative classifiers that were already there for the purpose of making it easier to make predictions regarding multi-labeled aspects and objects. Techniques that were discriminative were used in this study, and support vector machines were used to improve performance in relation to uni-labeled text classification issues.

The authors Liu et al. [16] discussed the joint feature selection problem across a group of related tasks of applications. These applications included many different subfields of computer vision and biomedical informatics.

A novel-based element determination system was presented by Feiet al. [17] in relation to the diagram. It planned and carried out an element determination strategy known as the structure-based attribute selection methodology, which involved the spatial dispersion of positioning highlights and the participants' commitments to the order.

By presenting statistical methods such as correlation-based feature selection (CFS), Bach et al. [18] were able to reduce the size of the corpus, and this method has since seen widespread application for the purpose of feature selection. The Non-Polynomial (NP) hard nature of the techniques that were used led to the selection of the features for optimal use so that they could

be used. [19] Karaboga et al. note that Feature Selection is a technique that has seen extensive use in the creation of ensembles. When performing ensemble constructions using Feature Selection, the results will be improved when FS is optimized.

EFFICIENT GRAPH PATTERN MINING APPROACH

Mining graph patterns is currently a very popular research field in many different fields, and it assists researchers in locating the interaction flow between various compounds that are present in a chemical or gene as a graph interaction. Principal component analysis is the method that is used to begin the process of data pre-processing here. The particle swarm optimization method, also known as the PSO method, is then utilized in the feature selection process in order to eliminate any superfluous characteristics. If this method were used, the outcome would be an optimal set of features, which has the potential to result in increased clustering accuracy. And then finally, a graph-based representation of the pre-processed data would be created, and this representation would be clustered using a hierarchical NMF clustering method after it had been pre-processed.

3.1 Analysis of the most important components in data pre-processing

The term "principal component analysis" refers to a process that involves the elimination of a number of variables. When data are present in numerous variables (possibly a massive number of variables), and it has been observed that there is some degree of redundancy present in these variables, it is extraordinarily helpful to have this information. Under these conditions, redundancy indicates that some of the variables demonstrate correlation with one another, which may be because they are measuring the same construct. This could be the case because some of the variables are measuring the same thing. As a result of this redundant act, the conclusion that it must be possible to reduce the extracted variables into primary components (synthetic variables) that are significantly fewer in number and will be responsible for the majority of the differences in the extracted variables has been reached. According to the jargon used in the industry, a primary component is a linear mixture of variables that have been extracted and weighed to their optimum levels.

PCA is used to fit the data, and one tool that is utilized for this is the n-dimensional ellipsoid. It can be seen that the components of the head segment correspond to the axis of the ellipsoid. The fluctuation is relatively insignificant when the centre of the ellipsoid is at zero. When we look at those hubs and their head segments, we overlook a small quantity of data. Calculating the axis of the ellipsoid involves first finding the mean of each variable in the dataset, then resolving the information around that mean, and finally finding the axis.

After being computed, the covariance data matrix is then put to use in an evaluation of the Eigen values and vectors of co-change framework. After that, the Eigen vector arrangement is given an orthogonal zed and standardized appearance by having it transformed into unit vectors. This procedure is carried out on a regular basis in order to generate unit Eigenvectors that pivot the ellipsoid to accommodate the information. To determine the extent to which this process can be carried out, a partitioning of the Eigen values and a comparison of those values with the Eigenvector are used. This method allows for a high level of sensitivity to be attained. The principal component analysis can be represented numerically as a symmetrical straight change.

Take into consideration an information matrix (X) that has a mean of zero. In this context, n-lines represent an alternative for investigating redundancy, and p-segments stand in for a particular kind of highlights. The formula for calculating the dimensional vector weights is as follows: w (vk) = (wv_1, wv_2,...., wv_p)(v_k). Every row of the vector x_1 that represents X is mapped onto a new vector called t(i) = (t_1,...., t_m)(i), which can be expressed as tk(i) minus x(i). w(vk) fori=1,….,n k=1,….,m

The fluctuations in the value of 'v' can be determined by looking at the 't' factors individually. The stacking vector w must become a unit vector in order to continue.

3.2. Optimal feature selection using PSO

Highlighting the choices is an important strategy and significant pre-processing venture in example acknowledgment and AI. The fundamental point of the element to choose includes subsets from the first list of capabilities and expand the presentation of learning calculations. The feature selection process in the graph pattern mining process is shown in figure 1.

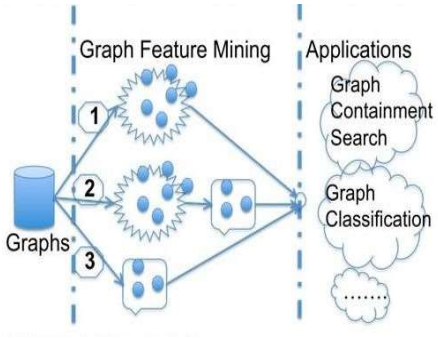

Figure 1: Graph feature selection process.

The algorithm of the proposed method works in three stages. During the initial phase, a graph indicates the full feature set. In the subsequent development, the highlights are divided into numerous groups utilizing implies calculation. Lastly, a novel pursuit procedure dependent on the PSO enhancement strategy is created to choose the last subset of highlighted features. Here separability index is used for the fitness evaluation to ensure the accurate feature selection result.

Behaviours of Bird flocking is simulated by using PSO method. Fowl gathering in a zone is looked randomly. At least there must be one sustenance in a zone. The nourishment location is not known among the winged creatures. The feathered creatures are known for their emphasis on sustenance. Powerful creatures are used to pursue the winged creature. Development-related issues are calculated using PSO gained from the circumstance. Inside the inquiry territory each and every arrangement is a winged creature.

This is called as ‘Molecule’. Wellness esteems of the particle are used to evaluate the wellness capacity. Speed of the particles is advanced based on the esteems. The particles follow the ideal particles. PSO is used to gather the irregular particles. Ages of the particles are refreshed to find the optima value. Best values are used to refresh the molecule. Best arrangement is accomplished up to this point. This worth is classified as "pbest".

The agent carrying out the particle swarm streamlining is used to find the excellent esteem. The worldwide best is called as "gbest". A particle shares its place in the population and also

in the role of its geographical neighbours; and the best worth is nearby best and it is designated "lbest". The velocity and position associated with the particles are refreshed to find best values as,

vel[ ] = vel[ ] + lf1 * r and N( ) * (pbest[ ] - present[ ]) + lf2 * r and N( ) * (gbest[ ] – present[ ])

Present [ ] = present[ ] + vel[ ] vel[ ] represents the particle speed and present [ ] the present particle (i.e. arrangement). pbest[ ] and gbest[ ] are characterized as expressed previously. randN( ) is an irregular number which ranges from 0 and 1. lf1, lf2 refer to the learning factors. Typically, $c_1=c_2=2$. The pseudo code associated with PSO is given as per the following:

For every particle
    Introduce the First particle
END
Do
    For every particle
        Compute the Fitness value
          In the event that the wellness value (current worth) is superior to anything in the best wellness value (pBest) in past history
    End
    particle with best fitness value is then picked amongst all the particles as then assumed as being gBest
    For every particle
        Compute the particle velocity by using equation (a)
         Update the position of the particle by using equation (b)
    End

while the maximum number of iterations or the minimum of error condition has not been ascertained. In each measurement Particles' speeds are assured into a most extreme speed (Vmax). The average speed value surpasses Vmax. It is a parameter predetermined by the client. The speed of this measurement is restricted to Vmax.

3.3. Graph-based representation

At last, the objective list of capabilities is mapped into the identical diagram G=(F, E, wF), where F={F1, F2,F3, ... ,Fn} signify the first highlighted features, $E=\{(F_i,F_j) : F_i,F_j \in F\}$ the edges of the diagram and w_{ij} the similarity between the highlighted features F_i and F_j and associated by the edge (F_i, F_j). The strategies for estimating the comparability values (i.e., edge loads) fundamentally decide the exhibition of ensuing chart-based component-choice technique. Diverse likeness estimates that can be utilized to decide the edge loads and various techniques may prompt various outcomes. Therefore, it is imperative to give preference to the most suitable measure with caution. In General, the Pearson connection coefficient and Euclidean separation are both, for the most part, utilized for likeness measures. In the proposed work, the Pearson connection coefficient system is accustomed to estimating the comparability

esteems between various highlights of a given information set. The sample graph representation of the gene features is illustrated in figure 2.

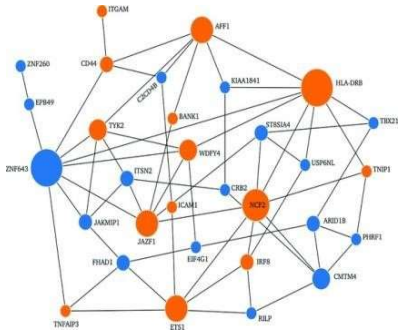

Figure 2: SLE exome information

Chart hub represents the top 31 variations. They are dictated by Encore pipeline (Evaporative Cooling highlight choice strategy, reGAIN, and SNPrank) and mapped into specific qualities. Known SLE-related qualities are named in orange speak. Disclosure qualities are named in blue speak. Edges in the blue have an average association score of 20. Hub range is measured in degree (i.e., number of edges).

3.4. Graph clustering using NMF method

NMF also known as Non negative matrix factorization technique gives the lower rank estimation of a nonnegative lattice, and it has been effectively utilized as a grouping system. Although NMF has been widely applied in the clustering process and often proven to achieve a remarkable clustering quality in comparison with classical techniques such as K-means, it is not a generic clustering approach, whose performance is excellent in all kinds of circumstances. The grouping capacity of a calculation and its impediments can be credited to its supposition on the bunching structure. The general graph clustering technique outcome is shown in Figure 3.

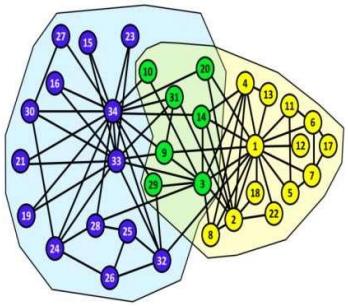

Figure 3: The Graph Clustering Process

NMF accepts that each bunch can be speaking to a solitary premise vector, and various groups ought to relate distinctive premise vectors. It tends to be seen as the best approach to execute and utilize GT (referring to the arrangement of X in a lower-dimensional space) to induce the grouping allocations. The objective of NMF is to surmise the first information network originating from the reality (utilizing the straight blend of premise vectors). The basic of 'k' bunches has non-direct structure and lies on a non-straight complex. NMF can't discover any

sort of 'k' premise vectors, which speak to the 'k' groups individually. Then again, one more prominent set of clustering methodologies exists, which gives due importance to the data points present in a plotted graph and minimizes some kind of trims in the graph. Although a graph model is usually designed over the coordinates of data points in their original Euclidean space, the information on the coordinate has no role to play after the graph is designed based on the similarity present between each pair of data points. Therefore, clustering techniques that belong to this class are just associated with an n×n similarity matrix A, where n specifies the number of data points or nodes.

NMF technique deploys non-negative matrix which is given by $X \in \mathbf{R}_{+}^{m \times n}$ and $k \leq \min(m,n)$. X is a product of two non-negative matrices $W \in \mathbf{R}_{+}^{m \times k}$ and $H \in \mathbf{R}_{+}^{m \times n}$.

$$X = WxH$$

The collection of non-negative real numbers is given by R+.

X is a data matrix. Rows of the matrix represent features and their columns represent n' non-negative data points in the m-dimensional space.

Many types of data have such representation as high-dimensional vectors. For example, a document in the Bag-of-Words Model is represented as distribution of all the words on the vocabulary; and a raw image (without feature extraction) as a vectorized array of pixels. In high dimensional data analysis, rather than training or making prediction relying on these high-dimensional data directly, it is often desirable to discover a small set comprising latent factors using a dimension-reduction method. In fact, high-dimensional data such as documents and images are usually embedded in a space with much lower dimensions.

EXPERIMENTAL RESULTS

In this section, the general model used to generate appropriate instances for the experimental evaluation is described, and the results of the evaluation are discussed.

4.1. Real genomic data

Real genomic data graphs are formed and it consists of localized clusters, which exploits the FBA technique, as the PMMG and PMSG are permitted much in advance to characterize the localized clusters. It means, the two probabilistic parameters $\mathbf{r_{ik}}$ and $\boldsymbol{\eta}_{\mathbf{mk}}$ are used to generate every real genomic graph representation. To formulate a graph, the two nodes $\mathbf{v_i}$ and $\mathbf{v_j}$(the edge$\mathbf{e_{ij}}$) are randomly generated according to $\mathbf{r_{ik}}$and $\boldsymbol{\eta}_{\mathbf{mk}}$ and to allow the value of the respective weight in a graph (or a matrix) to be equivalent to one, i.e., $\mathbf{W_{ij}} = \mathbf{W_{ji}} = \mathbf{1}$.

In this experimental test, there are five gene networks utilized. The cut-off values for $\mathbf{W}^{(SS)}$ and $\mathbf{W}^{(CC)}$have been adjusted so as to attain the number of edges existing in these two graphs to be approximately equal to those of the other three groups of networks (in which the number of edges cannot be controlled). We concentrated on 1,207 digestion associated qualities that were observed in the maximum associated segment (MCC) of the association of the five systems.

4.2. Evaluation measure

Standardized Mutual Information (NMI) strategy is a benchmark technique used for estimating the bunching outcomes [15]. NMI supposes that the actual clusters can be provided as the input. NMI demonstrates the cover between anticipated groups and genuine bunches, which implies that the exhibition is better just as NMI technique is bigger.

$$\mathbf{NMI}\ (\mathbf{Y}, \mathbf{C}) = [\ \mathbf{2} \times \mathbf{I}\ (\mathbf{Y}; \mathbf{C})]/[\mathbf{H}\ (\mathbf{Y}) + \mathbf{H}\ (\mathbf{C})]$$

where,

In the above equation class labels are represented by Y, cluster labels by Y, entropy by H (.) and Mutual Information between Y and C by I (Y;C). For example, assume class label=3 and cluster label=2.

P (Y=1) = 5/20 = 0.25

P (Y=2) = 5/20 = 0.25

P (Y=3) = 10/20 = 0.5

H (Y) = − 0.25log (0.25) – 0.25log (0.25) – 0.5log (0.5) = 1.5

H (C) represents the entropy of Cluster Labels

P (C=1) = 10/20 = 0.5

P (C=2) = 10/20 = 0.5

H (Y) = − 0.5 log (0.5) – 0.5 log (0.5)=1

For every cluster change, this calculation is done.

Mutual information I (Y;C)is expressed as:

I (Y; C) = H (Y) – H (Y | C)

H (Y). H (Y | C) represents entropy of class marks inside every bunch. Shared Information reveals to us the decrease in the entropy of class marks that we get on the off chance that we realize the group names.

H (Y|C) represents the conditional entropy of class labels.

Assume Cluster-1:

P (Y=1|C=1)=3/10 (three triangles in cluster-1)

P (Y=2|C=1)=3/10 (three rectangles in cluster-1)

P (Y=3|C=1)=4/10 (four stars in cluster-1)

Conditional entropy is calculated as:

$$\mathbf{H}\ (\mathbf{Y}\ |\mathbf{C} = \mathbf{1}) = -\mathbf{P}\ (\mathbf{C} = \mathbf{1})\sum(\mathbf{P}\ (\mathbf{Y} = \mathbf{y}\ |\mathbf{C} = \mathbf{1})\mathbf{log}\ (\mathbf{P}\ (\mathbf{Y} = \mathbf{y}\ |\mathbf{C} = \mathbf{1}\ )))$$

$\mathbf{y} \in \{\ \mathbf{1},\mathbf{2},\mathbf{3}\ \}$

= - ½ ×[3/10log(3/10) + 3/10log(3/10) + 4/10log(4/10)] = 0.7855

Now, assume Cluster-2:

P (Y=1|C=2)=2/10 (two triangles in cluster-1)

P (Y=2|C=2)=7/10 (seven rectangles in cluster-1)

P (Y=3|C=2)=1/10 (one star in cluster-1)

Compute conditional entropy as:

$$\mathbf{H}\ (\mathbf{Y}\ |\mathbf{C} = \mathbf{2}\ ) = -\mathbf{P}\ (\ \mathbf{C} = \mathbf{2}\ )\sum(\mathbf{P}\ (\mathbf{Y} = \mathbf{y}\ |\mathbf{C} = \mathbf{2}\ )\ \mathbf{log}(\ \mathbf{P}\ (\mathbf{Y} = \mathbf{y}\ |\mathbf{C} = \mathbf{2}\ )))$$

$\mathbf{y} \in \{\mathbf{1},\mathbf{2},\mathbf{3}\}$

= -½ × [2/10log(2/10) + 7/10log(7/10) + 1/10log(1/10)] = 0.5784

At last the Mutual Information is given by:

$I(Y;C) = H(Y) - H(Y|C)$
$= 1.5 - [0.7855+0.5784] = 0.1361$
Hence the NMI is,

$$\mathbf{NMI}(\mathbf{Y}, \mathbf{C}) = [\mathbf{2} \times \mathbf{I}(\mathbf{Y}; \mathbf{C})]/[\mathbf{H}(\mathbf{Y}) + \mathbf{H}(\mathbf{C})]$$
$$\mathbf{NMI}(\mathbf{Y}, \mathbf{C}) = \mathbf{2} \times \mathbf{0.1361}/[1.5+1] = \mathbf{0.1089}$$

For the entire dataset computation has been done and is possible before the commencement of clustering, because this aspect does not alter based on clustering output that is engendered. The results of the comparison analysis are clearly depicted in table 1.

Table 1: Simulation values of performance metrics

Methods	Metrics								
	NMI	Purity				Entropy			
		Clusters Count				Clusters Count			
		2	3	4	5	2	3	4	5
IF-NMFCM	**0.15**	**0.35**	**0.45**	**0.51**	**0.55**	**0.27**	**0.33**	**0.4**	**0.6**
FBA	**0.2**	**0.3**	**0.4**	**0.48**	**0.5**	**0.32**	**0.4**	**0.55**	**0.8**
PMMG	**0.7**	**0.23**	**0.3**	**0.4**	**0.47**	**0.6**	**0.6**	**1.1**	**1.5**
PMSG	**0.8**	**0.15**	**0.25**	**0.35**	**0.45**	**0.63**	**0.81**	**0.9**	**1.5**

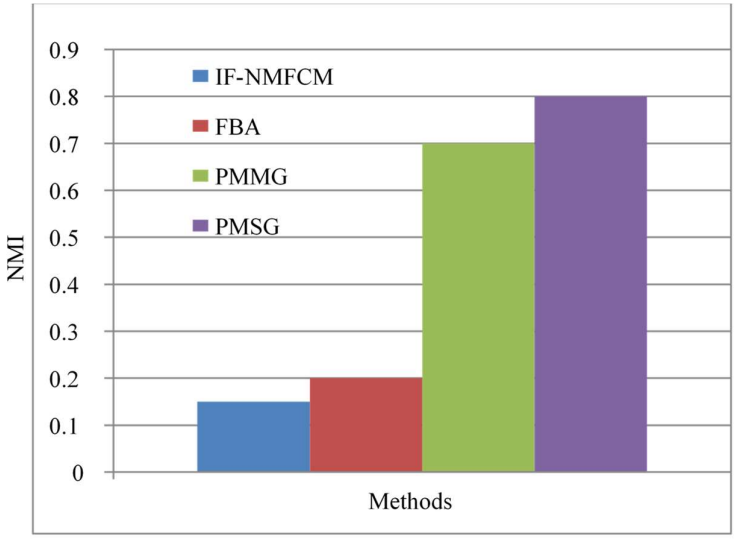

Figure 4: the NMI of three methods

The performance and advantage of IF-NMFCM are compared with other two existing methods, i.e., FBA, PMMG, PMSG. On an interesting note, this feature was observed quite clearly. Therefore, the graphical comparison can be proved for the proposed research method. IF-NMFCM shows 25% better outcome than FBA, 79% better result than PMMG and 81% better result than PMSG.

Purity and Entropy: The usage of Purity and Entropy is extensive to facilitate external clustering evaluation criteria. The purity aspect with respect to a cluster is defined as the ratio of the cluster size indicated by the largest set of objects designated to that cluster. Entropy of a cluster j has been deployed for the measurement of objects that have been mixed as part of the cluster group. The weighted average has been taken into consideration for the entropy overall as well as the individual cluster entropies:

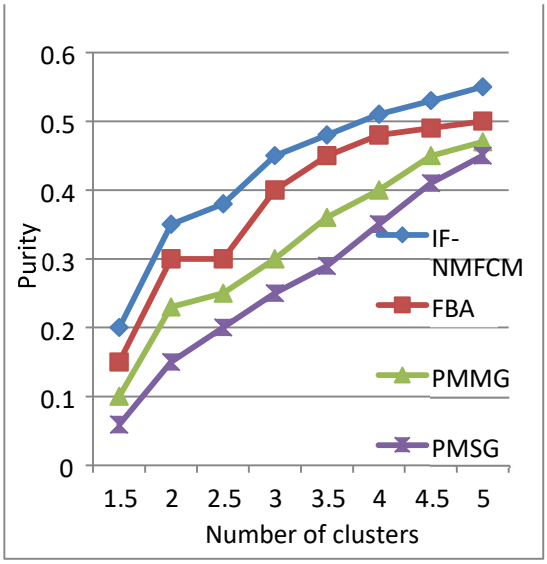

Figure 5: purity vs. no of clusters

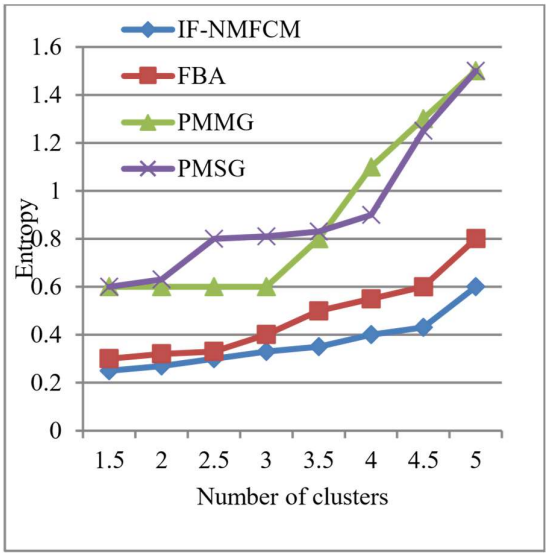

Figure 6: entropy vs no of clusters

Fig. 5 and 6 show good clustering. Therefore, the clustering that is good depicts an HP also known as High Purity and LE also referred to as Low Entropy. From figure 5, the proposed method IF-NMFCM shows 12% better result than FBA, 35% better result than PMMG and 60% better result than PMSG in terms of purity. From figure 6, it is proved that the proposed method IF-NMFCM shows 23% better result than FBA, 59% better result than PMMG and 60% better result than PMSG.

CONCLUSION

Mining the graphs to extract the interaction flow would be a challenging task. By aggregating the nodes that make up the graphs' deep communication, clustering is the most widely used data mining technique for determining the flow of graphical interaction. The proposed research strategy makes use of the NMF clustering technique known as Irrelevant Feature aware NMF clustering method (IF-NMFCM). This technical work finds that the clustering accuracy of the approaches is improved after irrelevant features are removed from the dataset, which reduces the complexity of the research work. Principal component analysis is used to begin the process of data pre-processing. Then, Particle Swarm Optimization (PSO) is applied to feature selection to filter out the superfluous details. As a result of employing this methodology, a superior set of features can be generated, which in turn can improve clustering precision. Finally, the pre-processed data would be clustered using the hierarchical NMF clustering method after being represented in a graph-based format. The Matlab simulation environment is used to analyze and evaluate the proposed research's overall performance. Therefore, the proposed research methods yield superior outcomes and results to the status quo methods due to their higher accuracy rate.

REFERENCES

[1] M. Deshpande, M. Kuramochi, Nikil Wale, and G. Karypis, Frequent Substructure-Based Approaches for Classifying Chemical Compounds, IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 8, pp. 1036-1050, Aug., 2005.

[2] N. Wale and G. Karypis. Acyclic Subgraph-based Descriptor Spaces for Chemical Compound Retrieval and Classification. In Proc of IEEE International Conference on Data Mining (ICDM), 2006.

[3] Moonesinghe, H. D. K., et al. "A probabilistic substructure-based approach for graph classification." Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on. Vol. 1. IEEE, 2007.

[4] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis, “Frequent substructure-based approaches for classifying chemical compounds,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 8, pp. 1036–1050, Aug. 2005.

[5] H. Saigo, S. Nowozin, T. Kadowaki, T. Kudo, and K. Tsuda, “gboost: A mathematical programming approach to graph classification and regression,” Mach. Learning, vol. 75, pp. 69–89, 2009.

[6] J. Zhou, J. Chen, and J. Ye, MALSAR: Multi-Task Learning via Structural Regularization. Tempe, AZ, USA: Arizona State Univ., 2012.

[7] Chen X, Pan W, Kwok JT, Carbonell JG (2009) Accelerated gradient method for multi-task sparse learning problem. In: IEEE international conference on data mining (ICDM09), pp 746–751, ISSN 1550– 4786

[8] Structured feature selection and task relationship inference for multi-task learning

[9] Zhang Y, Yeung D-Y, Xu Q (2010) Probabilistic multi-task feature selection. In: Proceedings of the advances in neural information processing systems (NIPS 2010), pp 2559–2567

[10] S. Pan, J. Wu, and X. Zhu, "Cogboost: Boosting for fast cost-sensitive graph classification," IEEE Trans. Knowl. Data Eng., vol. 27, no. 11, pp. 2933–2946, Nov. 2015.

[11] Vogelstein, Joshua T., et al. "Graph classification using signal-subgraphs: Applications in statistical connectomics." Pattern Analysis and Machine Intelligence, IEEE Transactions on 35.7 (2013): 1539-1551.

[12] Pan, Shirui, and Xingquan Zhu. "Graph classification with imbalanced class distributions and noise." Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press, 2013.

[13] Thoma, Marisa, et al. "Near-optimal Supervised Feature Selection among Frequent Subgraphs." SDM.

2009.

[14] X. Kong and P. Yu, "Semi-supervised feature selection for graph classification," in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2010, pp. 793–802.

[15] Godbole, Shantanu, and SunitaSarawagi. "Discriminative methods for multi-labeled classification." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2004. 22-30.

[16] Liu, Jun, ShuiwangJi, and Jieping Ye. "Multi-task feature learning via efficient l 2, 1-norm minimization." Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. AUAI Press, 2009

[17] Fei, Hongliang, and Jun Huan. "Structure feature selection for graph classification." Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008.

[18] Shi, Xiaoxiao, Xiangnan Kong, and S. Yu Philip. "Transfer Significant Subgraphs across Graph Databases." SDM. 2012.

[19] J. Zhou, J. Chen, and J. Ye, MALSAR: Multi-Task Learning via Structural Regularization. Tempe, AZ, USA: Arizona State Univ., 2012.