iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** IV **Month of publication:** April 2017

DOI: http://doi.org/10.22214/ijraset.2017.4210

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Ensembled Semi Supervised Clustering Approach for High Dimensional Data

M. LeelaReddy[1], Naveen Sai[2], G. Keerthi[3], M. Druga Kalyani[4]
[1,2,3,4] *Department of IT, LBRCE*

Abstract*: *we first propose an incremental semi-supervised clustering ensemble framework (ISSCE) that makes use of the advantage of the random subspace technique, the constraint propagation approach, the incremental ensemble member selection process, and the normalized cut algorithm to perform high dimensional data clustering. The random subspace technique is effective for handling high dimensional data, while the constraint propagation approach is useful for incorporating prior knowledge. The incremental ensemble member selection process is newly designed to remove redundant ensemble members based on a newly proposed local cost function and a global cost function, and the normalized cut algorithm is adopted to serve as the consensus function for providing more stable, robust, and accurate results.
Keywords*: *Random subspace, Constraint, normalized, ensemble, incorporate

I. INTRODUCTION

Cluster ensemble techniques are gaining more attention, due to its enormous helpful applications in the areas of pattern recognition, data mining, bioinformatics, and so on. When compared with traditional single clustering algorithms, these can integrate multiple clustering solutions obtained from different data sources into a unified solution, and provide a more robust, stable and accurate result. However, conventional cluster ensemble approaches have several limitations.

They do not consider how to make use of prior knowledge given by experts, which is represented by pair wise constraints. Pair wise constraints are often defined as the must-link constraints and the cannot-link constraints. The must-link constraint means that two feature vectors should be assigned to the same cluster, while the cannot-link constraints means that two feature vectors cannot be assigned to the same cluster. Most of the cluster ensemble methods cannot achieve satisfactory results on high dimensional datasets. Not all the ensemble members contribute to the result.

to, address the first and second limitations, we first propose the random subspace based semi supervised clustering ensemble framework (RSSCE), which combines the random subspace technique, the constraint propagation approach, and the normalized cut algorithm into the cluster ensemble framework to perform high dimensional data clustering. Then, the incremental semi supervised clustering ensemble framework (ISSCE) is designed to remove the redundant ensemble members. When compared with traditional semi-supervised clustering algorithm, ISSCE is characterized by the incremental ensemble member selection (IEMS) process based on a newly proposed global objective function and a local objective function, which selects ensemble members progressively. The local objective function is calculated based on a newly designed similarity function which determines how similar two sets of attributes are in the subspaces. Next, the computational cost and the space consumption of ISSCE are analysed theoretically. Finally, we adopt several nonparametric tests to match multiple semi supervised clustering ensemble techniques among different datasets. The experiment results show the improvement of ISSCE over existing semi-supervised clustering ensemble techniques or conventional cluster ensemble methods on six real-world datasets from UCI machine learning repository and 12 real-world datasets of cancer gene expression profiles.

The contributions of the paper are fourfold. First, we propose an incremental ensemble framework for semi-supervised clustering in high dimensional distinct attribute spaces. Secondly, a local cost function and a global cost function are calculated to incrementally select the ensemble members. Third, the newly designed similarity function is adopted to measure the extent to which two sets of attributes are similar in the subspaces. Fourth, we use non-parametric tests to compare multiple semi-supervised clustering ensemble approaches over different datasets**.**

II. BACK GROUND

It presents a novel pair wise constraint propagation approach problem into a set of independent semis supervised classification sub problems which can be solved in quadratic time using label propagation based on k-nearest neighbour graphs it provides an efficient solution for exhausting propagating pair wise constraints throughout the entire data set [1]. There exists, two types of pair wise

things they are- must link and cannot link constraints. A pair of data points with same label denotes must-link constraints. Otherwise cannot-link constraints. Here, two sets of initial pair wise constraints are directly used to adjust the similarities between data points.

The ensemble clustering technique combines multiple clustering's into a probably better and more robust clustering. The three limitations to most of the existing ensemble clustering methods [2]. Initially they generally consider the cluster number of the final clustering as input and thereafter lacks the ability to estimate the cluster number. Second, most of them treat each base clustering with equally and next overlooks the different reliability of the base clustering's are considered. Third, mostly the existing methods work at the object-level and does not increase or consider well for large ensembles. Here in this paper, we are proposing a specialized ensemble clustering approach termed as ensemble clustering using factor graph (ECFG). The advantages of this method are as follows: the cluster number that is obtained automatically are not need to be specified in before or in advance. Later the reliability of each base clustering can be estimated in an unsupervised manner and finally exploited in the consensus process. Finally, our approach is also efficient for processing ensembles with large data sizes, large ensemble sizes and large dimensional data. We use the concept of super-object as a compact and adaptive representation for ensemble data. Instead of using the original data objects.

The three effective and efficient techniques are proposed for obtaining high quality combiners (consensus functions) [3]. The first combiner induces a similarity measure from the partitioning's and then re clusters the objects. The second combiner is based on hyper graph partitioning. The third one collapses groups of clusters into meta-clusters which then compete for each object to firmly decide the combined clustering. Because of the low computational costs of our techniques, it is quite feasible to use a supra-consensus function that evaluates all three approaches against the objective function and picks the best solution for a given situation. Unlike classification or regression settings, there have been very few approaches proposed for combining multiple clustering's.

A random double clustering based cluster ensemble approach (RDCCE) to perform tumour clustering based on gene expression data. [4] Particularly, RDCCE generates a set of representative features using a randomly chosen clustering algorithm in the ensemble, and then assigns samples to their corresponding clusters based on the grouping results. In addition to that, we also introduce the random double clustering based fuzzy cluster ensemble technique (RDCFCE), which is designed to improve the performance of RDCCE by combining the newly proposed fuzzy extension model into the ensemble framework.

The expert's knowledge as constraints in the process of clustering, and propose a feature selection based semi-supervised cluster ensemble approach (FSSSCE) for tumour clustering from bio-molecular data. [5] Compared with existing tumour clustering approaches, the proposed framework FS-SSCE is featured by two properties: The choice of feature selection techniques to make disappear the effect of noisy genes. The utilization of the binate (dual/binary) constraint based K-means algorithm to consider the effect of experts' knowledge. Then, a double selection based semi-supervised cluster ensemble framework (DS-SSCE) that not only uses the feature selection technique to perform gene selection on the gene dimension, but also chooses an optimal subset of representative clustering solutions in the ensemble and betters the performance of tumour clustering using the normalized cut algorithm. DS-SSCE also introduces a confidence factor into the process of constructing the consensus matrix by using the knowledge of the data set. Finally, they designed a modified double selection based semi-supervised cluster ensemble framework (MDS-SSCE) which uses several clustering solution selection techniques and an aggregated solution selection function to choose an optimal subset of clustering solutions

III. EXISTING SYSTEM

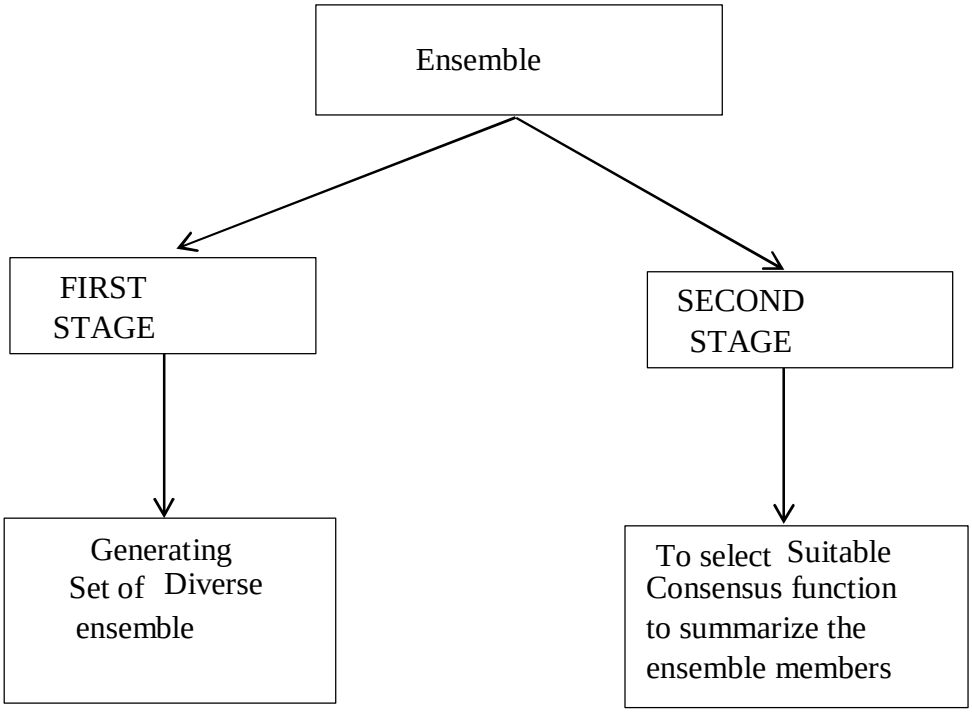
Conventional cluster ensemble approaches have several limitations: They do not consider how to make use of prior knowledge given by experts, which is represented by pair wise constraints. Pair wise constraints are often defined as the must-link constraints and the cannot-link constraints. The must-link constraint means that two feature vectors should be assigned to the same cluster, while the cannot-link constraints means that two feature vectors cannot be assigned to the same cluster. Most of the cluster ensemble methods cannot achieve satisfactory results on high dimensional datasets. Not all the ensemble members contribute to the result.

IV. PROPOSED SYSTEM

Cluster ensemble, is referred to as consensus clustering, one of the important research directions in ensemble learning, that can be divided into two stages: the first stage aims at generating a set of various ensemble members, while the objective of the second stage is to choose a suitable consensus function to summarize the ensemble members and search for an optimal unified clustering solution. To attain these objectives, we use a knowledge reuse framework which integrates multiple clustering solutions into a unified one. While there are various kinds of cluster ensemble techniques, some of them consider how to handle high dimensional data clustering,

Volume 5 Issue IV, April 2017
ISSN: 2321-9653

and how to make use of prior knowledge of the given data set. High dimensional datasets have too many attributes relative to the number of samples, which will lead to the over fitting problem. Most of the conventional cluster ensemble methods do not consider how to handle the over fitting problem, and cannot obtain satisfactory results when handling high dimensional data. Our method uses the random subspace technique to generate the new datasets in a low dimensional space, which will evaluate this problem.



In summary, most of the cluster ensemble approaches only consider using a similarity score or feature selection technique to remove the redundant ensemble members, and few of them study how to apply an optimization method to search for a suitable subset of ensemble members.

A. Constraint Propagation Approach

The constraint propagation approach (E2CP), that propagate constraints in a more exhaustive and efficient way, as the basic clustering algorithm in ISSCE. This approach has two advantages: The time complexity of E2CP is proportional to the total number of all possible pairwise constraints, which is O (Kn2)(where K is the number of neighbours in the K-NN graph, and n is the number of feature vectors in the given dataset). It is much smaller than that of traditional constraint clustering techniques, which is O(n4). E 2CP achieves better results on various real-world datasets, such as image datasets, UCI datasets, cross-modal multimedia retrieval, and so on. When compared with other approaches, our proposed incremental semi-supervised clustering ensemble framework uses the most effective constraint propagation technique to convey supervised data from the labelled data samples to the unlabelled samples, and solve the label propagation problem in parallel.

B. Incremental Semi-Supervised Clustering Ensemble Framework

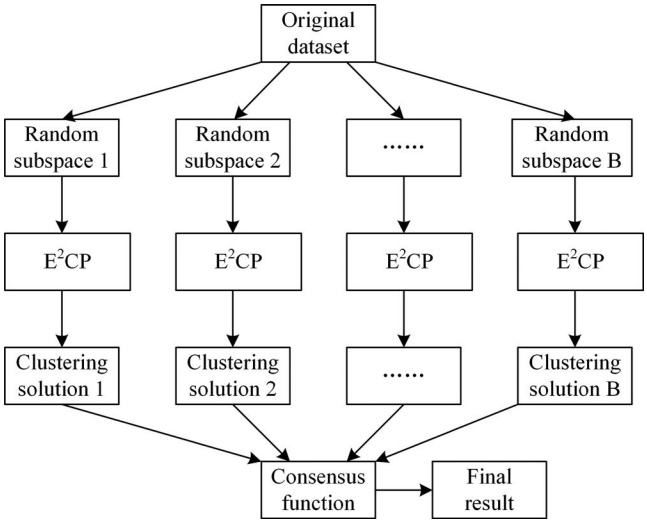
The incremental semi-supervised clustering ensemble framework (ISSCE) is designed to remove the redundant ensemble members. When compared with traditional semi-supervised clustering algorithm, ISSCE is characterized by the incremental ensemble member selection (IEMS) process based on a newly proposed global objective function and a local objective function, which selects ensemble members progressively. The local objective function is calculated based on a newly designed similarity function which determines how similar two sets of attributes are in the subspaces. Next, the computational cost and the space consumption of ISSCE are analysed theoretically. Finally, we adopt several nonparametric tests to compare multiple semi supervised clustering ensemble approaches over different datasets.

We focus on semi-supervised clustering ensemble approaches that have been successfully practiced to different areas, such as data mining, bioinformatics, and so on. The constraint neighbourhood projection based semi-supervised clustering ensemble technique, and achieved better performance on UCI machine learning datasets. We represented prior knowledge provided by experts as pair

Volume 5 Issue IV, April 2017
ISSN: 2321-9653

wise constraints, and proposed the knowledge based cluster ensemble method and the double selection based semi-supervised clustering ensemble approach. Both are successfully used for clustering gene expression data. Though, some of them consider how to handle high dimensional datasets. To address this limitation, we first propose the random subspace based semi-supervised clustering ensemble approach, RSSCE first adopts the random subspace technique to generate B random subspaces A1, A2, ..., AB. Then, the constraint propagation approach (E2CP) is applied to perform clustering in the subspaces, and generate a set of clustering solutions I1, I2, ..., IB. Next, a consensus matrix is constructed based on the set of clustering solutions. Finally, the normalized cut algorithm (N cut) is used to serve as the consensus function, partition the consensus matrix, and obtain the result. Specifically, given a very high dimensional dataset P with each feature vector pi (i in{1,2….n}) containing m attributes, RSSCE uses the random subspace technique to generate a set of random subspaces in the first step. Specifically, a sampling rate t= {tmin, tmax} of the number of attributes in the subspace over the total number of attributes in the original set. The new subspace is constructed by these selected attributes. Finally, RSSCE will generate a set of random subspaces A1, A2, ..., AB by repeating the process B times. The advantage of the random subspace technique is to provide multiple ways to explore the underlying structure of the data in a low dimensional space.

1) *Input:* *A* high dimensional dataset P.
2) *Ensure:*
a) Generate B random subspacesfA1, A2, AB.
b) Generate the semi-supervised clustering models x1, x2, xB.
c) Call incremental member selection process. New Ensemble generation.
d) Generate B0 random subspaces fA1, A2, AB (B0 <B).
e) Generate semi-supervised clustering models x1, x2, xB.
f) Obtain consensus matrix O by summarizing the clustering solutions.
g) Consensus function for the final results using the normalized cut approach.
3) *Output:* The labels of the samples in P.



C. The Incremental Ensemble Member Selection Process

The proposed ISSCE framework uses a newly designed incremental ensemble member selection technique to generate a maximum set of members. In addition, conventional cluster ensemble methods do not consider how to make use of prior knowledge, which is usually represented in the form of pairwise constraints or a very small set of labelled data. Single semi supervised clustering algorithms can handle prior knowledge, and use them to guide the search in the process of clustering. It is natural to adopt a suitable single semi-supervised clustering method as the basic clustering algorithm in the cluster ensemble. This algorithm provides an overview of the incremental ensemble member selection process. The input is the original ensemble, while the output is a newly generated ensemble with smaller size. Specifically, IEMS considers the ensemble members one by one, and calculates the objective function Delta Ib for each clustering solution Ib generated by E2CP with respect to the subspace Ab in the first step. In the second step, it sorts all the ensemble members in b G in ascending order according to the corresponding D values.

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

1) *Input:* The sample set P = (p1, p2, . . ., pn).
The must-link set M.

The cannot link set N.

a set of random subspaces A = {A1, . . .AB}.

2) *Ensure:*

a) Calculate the objective function for each clustering solution Ib generated by E2CP.

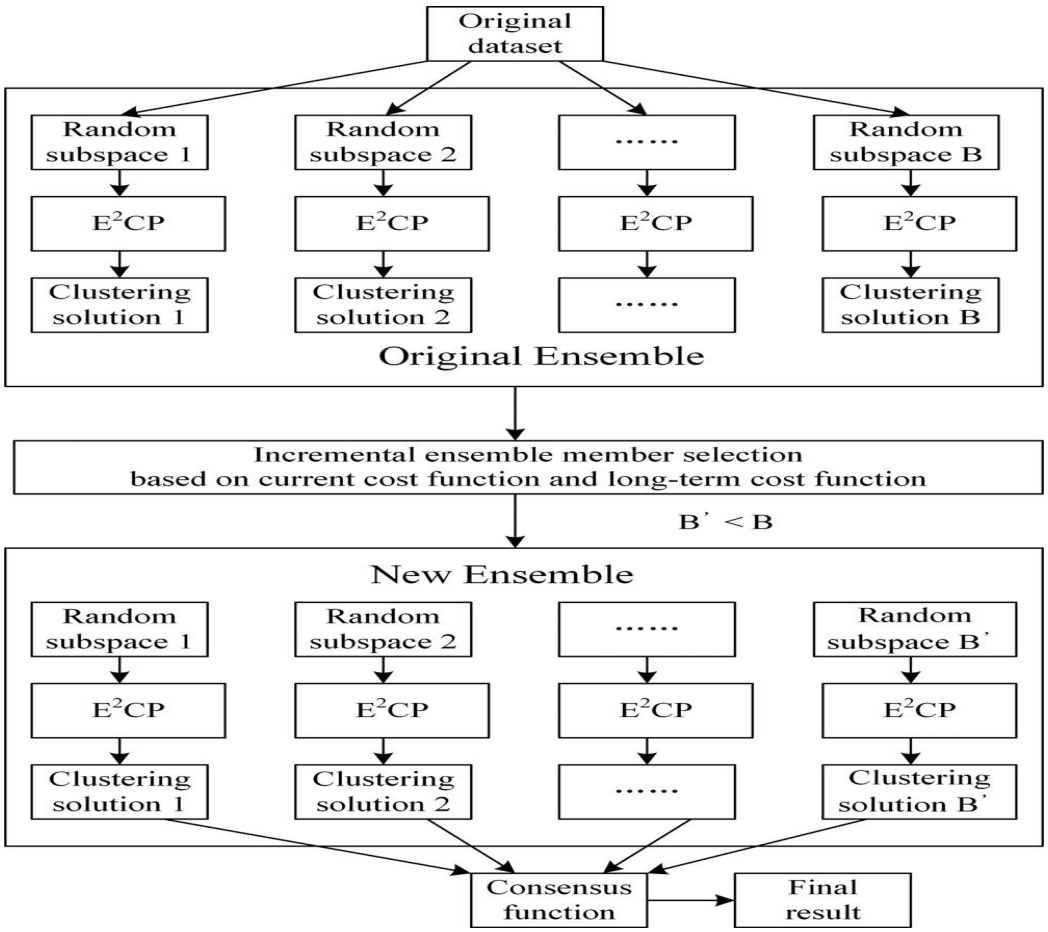
b) Sort ensemble members in ascending order and pick up the first ensemble member.

c) Calculate the local objective function.

d) Sort ensemble members in ascending order as per the corresponding local objective function.

e) Calculate the global objective function for the clustering solutions.

3) *Output:* The new ensemble G.



D. Similarity Function

Given the subspaces Ab and At, the set of attributes in these subspaces can be represented by Gaussian mixture models respectively (where wb h1, mb h1 and Sb h1, (h1 2f 1; k1g) denote the weight value, the mean vector and the covariance matrix for the h1-th component Fb h1 of Vb, respectively. wt h2, mt h2 and St h2, (h2 2 f1; k2g) denote the weight value, the mean vector and the covariance matrix for the h2-th component Ft h2 of Vt, respectively). The expectation maximization approach (EM) is used to perform clustering on the set of attributes in the subspace, and resolve the optimal parameter values of GMMs. This algorithm provides a flowchart of the similarity function (SF) for S (Ab to At). The input of SF is two Gaussian mixture models Vb and Vt, while the output is the similarity value S (Ab, At) between two subspaces Ab and At. Specifically, the similarity function first considers the similarity of all the pairs of components in Vb and Vt. The Bhattacharyya distance is used to calculate the similarity between two components Fb h1 in Vb and Ft h2 in Vt. Then, SF sorts all the component pairs in ascending order as per the

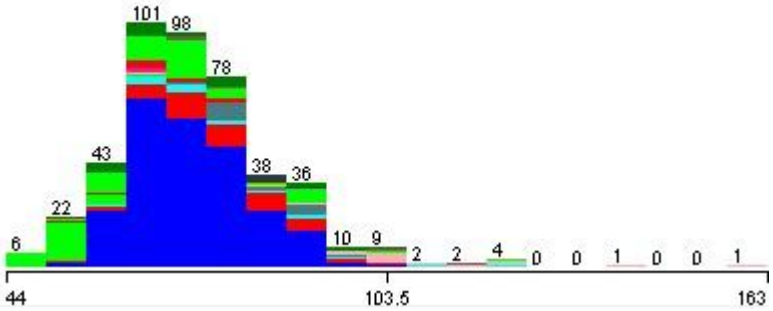
Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

corresponding Bhattacharyya distance values, and inserts them into a queue. Next, it sets S (Ab, At)=0, and performs a de-queue operation, and considers the component pair Fb h1; Ft h2 one by one. If wt. h1 > 0 and wt. h2 > 0, the minimum weight w between the two is selected in the first step. In the second step, the similarity value S (Ab, At) is assigned a new value. And the calculated weights are updated in the third step. Finally, the similarity value S (Ab, At) will be obtained by considering all the component in the queue.

1) *Input:* The Gaussian mixture model.
2) *Ensure:*
a) Calculate the Bhattacharyya distance.
b) Sort all the component pairs in ascending order according to the corresponding Bhattacharyya distance values, and insert them into a queue.
c) Perform a de-queue operation and obtain the component pair.
3) *Output:* The similarity value.

V. EXPERMENTAL RESULTS

These are the results obtained for the ARRYTHMIA dataset based on the heart rate column



This graph shows the heart rate column of the dataset

A. *Confusion Matrix*

	Predicted +ve	Predicted -ve
Actual +ve Cases	21	10
Actual -ve Cases	4	10

B. *Accuracy*

$$A = \frac{TN+TP}{TN+FP+FN+TP} = \frac{10+21}{10+10+4+21} = 0.78$$

1) *TN – True Negative:* It represents the value which must be present in the output but it is not in the output. These refer to the negative tuples that were correctly labelled by the classifier.
2) *TP – True Positive:* It represents the value which should be present in the output and it is in the output. These refer to the positive tuples that were correctly labelled by the classifier.
3) *FN – False Negative:* It represents the value which must not be present in the output and it is not present in the output. These refer to the positive tuples that were mislabelled as negative.
4) *FP – False Positive:* It represents the value which should not be present in the output and it is present in the output. These refer to the negative tuples that were mislabelled as positive.

VI. CONCLUSION AND FUTURE WORK

We propose a new semi-supervised clustering ensemble approach, which is referred to as the incremental semi-supervised clustering

ensemble approach. Our major contribution is the development of an incremental ensemble member selection process based on a global objective function and a local objective function. To design a good local objective function, we also propose a new similarity function to quantify the extent to which two sets of attributes in the subspaces are similar to each other. We conduct experiments on six real-world datasets from the UCI machine learning repository and 12 real world datasets of cancer gene expression profiles, and obtain the following: The incremental ensemble member selection process is a general technique which can be used in different semi-supervised clustering ensemble approaches. The prior knowledge represented by the pair wise constraints is useful for improving the performance of ISSCE. ISSCE outperforms most conventional semi-supervised clustering ensemble approaches on many datasets, especially on high dimensional datasets. In the future, we shall perform theoretical analysis to further study the effectiveness of ISSCE, and consider how to combine the incremental ensemble member selection process with other semi supervised clustering ensemble approaches. We shall also investigate how to select parameter values depending on the structure/complexity of the datasets.

In the future, we shall perform theoretical analysis to further study the effectiveness of ISSCE, and consider how to combine the incremental ensemble member selection process with other semi supervised clustering ensemble approaches. We shall also investigate how to select parameter values depending on the structure/complexity of the datasets.

REFERENCES

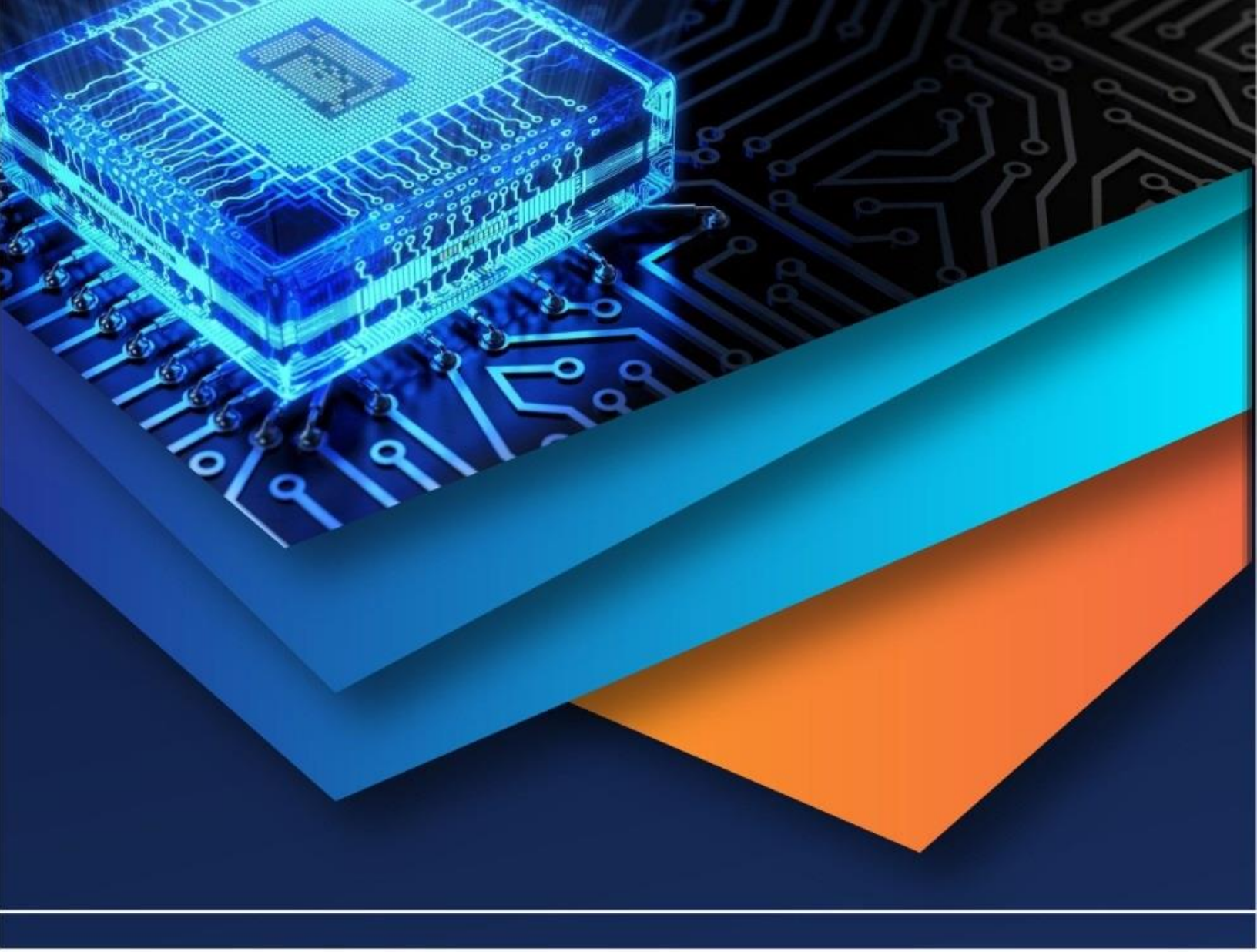
[1] Z. Lu and Y. Peng, "Exhaustive and efficient constraint propagation: A graph-based learning approach and its applications," Int. J. Comput. Vis., vol. 103, no. 3, pp. 306–325, 2013.

[2] Dong Huanga, d, Jianhuang Laia, n, Chang-Dong Wang, "Ensemble clustering using factor graph" S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.

[3] A. Strehl and J. Ghosh, "Cluster ensembles — A knowledge reuse framework for combining multiple partitions," J. Machine Learning Res., vol. 3, pp. 583–617, 2002**.** R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[4] Z. Yu, H. Chen, J. You, H.-S. Wong, J. Liu, G. Han, and L. Li, "Adaptive fuzzy consensus Clustering framework for clustering analysis of cancer data," IEEE/ACM Trans. Comput. Biol. Bioinformat., vol. 12, no. 4, pp. 887–901, Jul./Aug. 2015.

[5] Z. Yu, H. Chen, J. You, H.-S. Wong, J.Liu, L.Li, and G.Han, "Double selection based semi- Supervised clustering ensemble for tumor clustering from gene expression profiles," IEEE/ACM Trans. Comput. Biol.Bioinformat., vol.11, no.4, pp.1–14, Jul./Aug.2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY