iJRASET

International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 Issue: III Month of publication: March 2018

DOI: http://doi.org/10.22214/ijraset.2018.3657

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

Ensembling Solutions for Semi – Supervised Clusters

Viveka Priya[1], Meera Vijayan[2]
[1]Department of Computer Science and Engineering, [2]Department of Mobile and Pervasive Computing, Anna University(BIT Campus), Trichy, Tamilnadu

Abstract: Regular semi-regulated bunching approaches have a few deficiencies, for example, not completely using all helpful must-interface and can't connect requirements, not considering how to manage high dimensional information with commotion, and not completely tending to the need to utilize a versatile procedure to additionally enhance the execution of the calculation, We initially propose the transitive conclusion based imperative engendering approach, which makes utilization of the transitive conclusion administrator and the proclivity proliferation to address the primary restriction. At that point, the arbitrary subspace based semi-regulated bunching troupe structure with an arrangement of proposed certainty factors is intended to address the second constraint and give more steady, robust, and exact outcomes to perform the clustering process. The proposed approaches function admirably on the greater part of this present reality datasets.

Keywords: Semi - Supervised, Clustering, Transitive Closure, Random Subspace

I. INTRODUCTION

Semi-supervised clustering is an important sub-field of clustering and it is broadly applied in different areas, such as image processing, multimedia, pattern recognition and bioinformatics and in many different fields. Here, we proposed a semi-supervised clustering ensemble approach, which is called as the knowledge based clustering ensemble framework. When compared with conventional clustering approaches, the semi-supervised clustering approaches makes use of the prior knowledge, which is represented by a small number of labelled data or pairwise constraints, to improve the performance of the clustering process.

In this paper, we focus on constrained clustering, which belongs to the class of semi-supervised clustering approaches. Constrained clustering integrates a set of must-link constraints and cannot-link constraints into the clustering process. The must-link constraint means that two datasets should belong to the same cluster, while the cannot-link constraint means that two datasets cannot be assigned to the same cluster. Traditional constrained clustering approaches have two limitations: (1) They do not consider how to make full use of must-link constraints and cannot-link constraints; (2) Some methods do not take into account how to deal with high dimensional data with noise. In order to address the limitations of traditional constrained clustering approaches, we first propose a transitive closure based constraint propagation approach (tccpa), which not only expands the constraint set using transitive closure, but also adopts the label propagation approach to disseminate the pairwise constraints. The TCCPA, which makes use of the transitive closure operator and the affinity propagation to explore and decomposes the pair wise constraint propagation procedure into a set of independent semi-supervised binary classification problems, which can be solved by label propagation in parallel. Streaming K-Means approach is used to find, if the volume of data is too large to be stored in the main memory available, the K-Means algorithm is not suitable, as it is a batch processing mechanism iterates over all the data points. The labelled examples which has number of clusters and it is to known something about the characteristic of the data where it belongs to the same class or in different class. There are two types of clustering algorithm they are Balanced clustering and Data Stream Clustering this information can be expressed by means of constraints among examples: must links and cannot links. Then, a Random subspace based SEMI-supervised Clustering Ensemble framework (RSEMICE) explores the underlying structure of the dataset in the feature subspace, which will reduce the effect of noise features. It is able to combine multiple clustering solutions into a unified one, which will provide more accurate, stable and robust results. Finally, a set of nonparametric tests are used to compare different approaches over multiple datasets.

II. RELATED DATA

Semi-supervised clustering is one of the important research directions in the area of data mining, which is able to make use of prior knowledge, such as pairwise constraints or a small amount of labelled data, to guide the search process and improve the quality of clustering. A number of semi-supervised clustering approaches have been previously proposed, which can be divided into five categories.

The approaches belonging to the first category focus on designing new kinds of semi-supervised clustering algorithms, such as semi-supervised hierarchical clustering, semi-supervised kernel mean shift clustering , semi-supervised maximum margin clustering , semi-supervised clustering corresponding to spherical K-means and feature projection , semi-supervised linear discriminant clustering , semi-supervised information-maximization clustering , semi-supervised clustering based on seeding , active semi-supervised fuzzy clustering , semi-supervised kernel fuzzy c-means , semi-supervised clustering framework based on discriminative random fields , semi-supervised subspace clustering, semi-supervised matrix decomposition, semi-supervised fuzzy clustering based on competitive agglomeration , constrained clustering based on Minkowski weighted K-means, semi-supervised non-negative matrix factorization based on constraint propagation , semi-supervised kernel mean shift clustering , and so on. In general, most of the new semi-supervised clustering approaches are extensions of traditional clustering algorithms by taking into account label information or pairwise constraints.

III.TCCPA

The transitive closure based constraint propagation Clustering approach (TCCPA) first applies the transitive Closure operator to fully explore the constraint set. Then, it decomposes the pair wise constraint propagation procedure in U into a set of independent semi-supervised binary classification problems, which can be solved by label propagation in parallel. The propagated constraints are used to adjust the similarity matrix for constrained spectral clustering. TCCPA first applies the transitive closure operator. Since the must-link constraint is an equivalence relation; it satisfies the properties of reflexivity, symmetry and transitivity**.**

Then, the propagation procedure can be viewed as a semi-supervised binary classification sub-problem. Data samples which are must-linked can be considered as positive samples, while data samples which are cannot-linked will be labeled as negative samples. Our goal is to propagate supervised information from labeled data samples to unlabeled ones, which can be solved by label propagation based on the algorithm. The semi-supervised classification can be formulated as an optimization problem with the following objective function.

$$)\min_{U_{il}} \frac{1}{2}\{\sum_{i=1}^{n}(U_{il} - R_{il})^2 + \frac{\mu}{2}\sum_{i,j}^{n} w_{ij}(\frac{U_{il}}{\sqrt{d_{ii}}} - \frac{U_{jl}}{\sqrt{d_{jj}}})^2\}$$

Transitive Closure based Constraint Propagation approach is used to allocate the space for the given dataset that is Random subspace. The text file will choose their space and find matches with each other this process has been done by TCCPA. Each file compare with each other and their matches will be shown in Percentage. The fig.1 shows the tccpa data

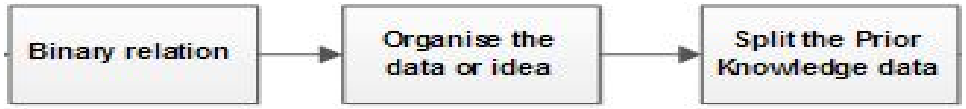


Fig. 1 Transitive Closure Constraint Propagation Approach

The TCCPA is calculated by certain java code there is no formula have been used. If the text file has the same content 75% matches with the same file means the duplication was removed. We able can view the text files there was no duplication of files will be there. The same process has been used for the all datasets.

IV.RANDOM SUBSPACE ENSEMBLE FRAMEWORK

RSEMICE are characterized with two properties. It explores the underlying structure of the dataset in the feature subspace, which will reduce the effect of noise features. It is able to combine multiple clustering solutions into a unified one, which will provide more accurate, stable and robust results, it will send the data to the local environment and co-operate with that different environment, and then it finally random cooperation and accurate results in the adaptive process for the Random Subspace System. The fig.2 shows the random subspace framework with the distributed region of attribute bagging and the data sets are splitted and distributed

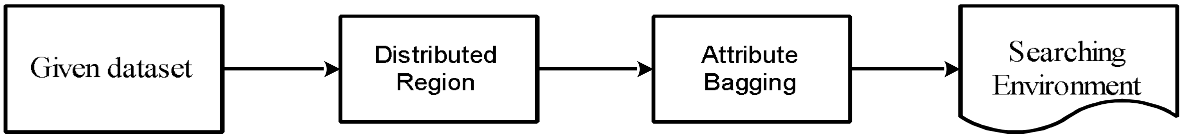


Fig. 2 Random Subspace Ensemble Framework

The random subspace technique is used to split the data and it is then distributed into the certain regions with the prior knowledge

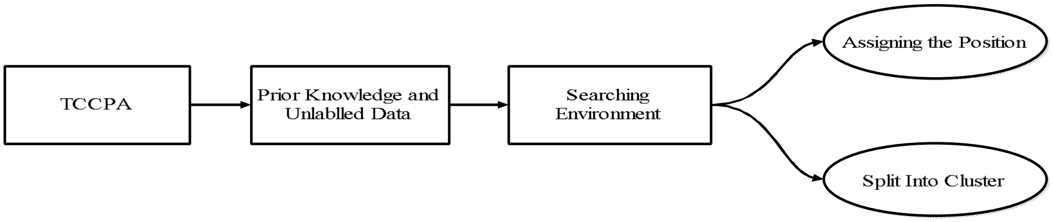


Fig. 3 Split into cluster with assigned position

V. CLUSTERING FRAMEWORK

Clustering framework is the process of analysing the sequence of clustering deviations with the help of Ensembling process of accessing their groups, it will assign in the separate orders, and clustering path is always the sequence and the analysis of each gateway. The Semi-supervised cluster ensemble approach, which is called the knowledge based clustering ensemble framework, and applied it to bio-molecular pattern mining. When compared with conventional clustering approaches, the semi-supervised clustering approaches make use of prior knowledge, which is represented by a small number of labelled data or pair wise constraints, to improve the performance of the clustering process.

Here, we propose two algorithms known as Balanced clustering and Data streaming algorithm.

A. *Balanced Clustering Algorithm*

1) Balanced Clustering is the Special case of clustering where the strictest cluster Sizes are constraint to n/k, when n is the number of points and k is the number, a typical algorithm is balanced k means, which minimizes the square mean error.
2) The number of cluster is divided by two numbers that is will balance the cluster.
3) It will balance the data and load according to their balancing position.

B. *Data Stream Algorithm*

1) Data stream clustering is defined as the clustering process of data that arrives continuously such that telephone records, financial transactions etc can be verified using clustering.
2) Data stream clustering is usually studied as a streaming algorithm and the main objective is when given a sequence of points, to construct a good clustering of the stream
3) It uses a small amount of memory and time.

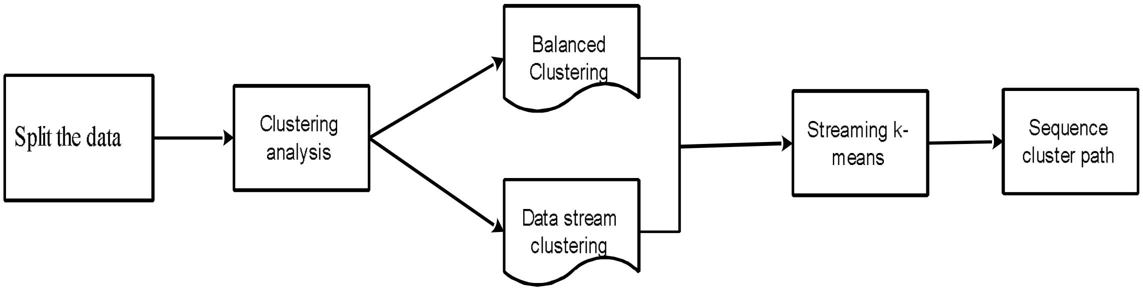


Fig. 4 Clustering Framework

VI. ALGORITHM

A. *Semi-Supervised Cluster Ensemble Framework*

1) *Input:*

the sample set P = (p1, p2, ..., pl);
the must-link set M;
the cannot-link set N;
a set of random subspaces A = {A1, ...AB};
a set of semi-supervised clustering models $\chi = \{\chi 1, ..., \chi B\}$;
a set of ensemble members $\Gamma = \{(A1, \chi 1),(A2, \chi 2), ...,(AB, \chi B)\}$;
the empty ensemble Γ;

```
Repeat
t=t+1;
For each (Ab, χb) in Γ
Calculate the local objective function ζb;
Sort ensemble members in Γ in ascending order according to the corresponding local objective function ζb;
Set b = 0;
Repeat
Set b = b + 1;
Generate new ensemble Γ = Γ + {(Ab, χb)} (where (Ab, χb) ∈ Γb );
Calculate the global objective function Δ(I’ ) and Δ(I)for the clustering solutions I ‘ and I generated by Γ ‘ and Γ respectively;
Until Δ(I ‘ ) ≤ Δ(I);
Add to new ensemble: Γ = Γ + {(Ab, χb)}, Γ = b Γ − {(Ab, χb)};
Until t ≥ B’ or Γ =b Ø;
```

2) *Output*: the new ensemble Γ

B. Data Stream Clustering

1) *Input:* K (number of clusters), Training set x(1), x(2),...,x(m)

```
Randomly initialize K cluster μ1, μ2,...,μK
 repeat
 // cluster assignment step: find closest
for i = 1 to m data points do
 c(i) := index of cluster closest to x(i)
end for
// update  step: compute means based on assignment
for k = 1 to K do μk := average(mean) of points assigned to cluster k
end for
until N iterations
```

2) *Output:* assignments c(1), c(2),...,c(m) , and learned μ1, μ2,...,μK

VII. EXPERIMENT RESULT

The performance of the proposed approaches are evaluated using datasets (where n denotes the number of data samples, m denotes the number of attributes, and k denotes the number of classes). Conventional semi-supervised clustering approaches cannot obtain satisfactory results on the dataset. In this case, traditional constraint clustering methods cannot be effectively applied to this dataset due to the large number of classes. In summary, these datasets can be used to more thoroughly explore the performance bounds of the semi-supervised clustering ensemble approaches. The pre-processing procedure for the datasets are processed.

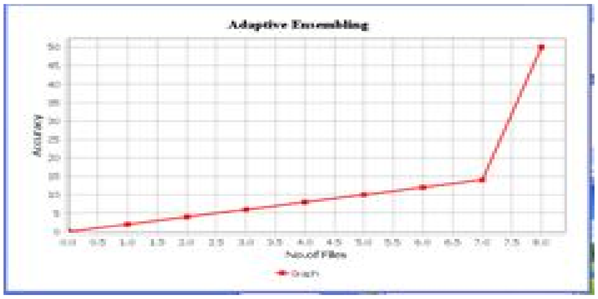


Fig. 5 Clustering Performance with high accuracy is enabled

VIII. CONCLUSION

A newly proposed a transitive closure based constraint propagation approach (TCCPA) is adopted to make use of the transitive closure operator and the constraint propagation to fully explore how to use all useful must-link and cannot-link constraints. The transitive closure operator and the confidence factor each plays an important Role in attaining good performance for the Ensembling solution. The clustering performance is improved and the noisy disturbance is reduced from the high dimensional data. The redundant copies of datasets had been removed and the accuracy is shown. The incremental ensemble member selection process is a general technique which can be used in different semi-supervised clustering ensemble approaches.

IX.FUTURE WORK

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. 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.

REFERENCES

[1] Biswas and D. Jacobs, “Active image clustering: Seeking constraints from humans to complement algorithms,” in Proc. IEEEConf. Comput. Vis. Pattern Recog., 2012, pp. 2152–2159.

[2] X. Zhu, “Semi-supervised learning literature survey,” Dept. Comput.Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech.Rep. 1530, 2008.

[3] X. Zhu and A. B. Goldberg, Introduction to Semi-Supervised Learning,San Rafael, CA, USA: Morgan & Claypool, 2009.

[4] M. S. Baghshah, F. Afsari, S. B. Shouraki, and E. Eslami, “Scalable semi-supervised clustering by spectral kernel learning,” Pattern Recog. Lett., 2014..

[5] Z. Lu and M. A. Carreira-Perpinan, “Constrained spectral clustering through affinity propagation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2008.

[6] Z. Yu, L. Li, J. Liu, and G. Han, “Hybrid adaptive classifier ensemble,” IEEE Trans. Cybernetics, 2015.

[7] S. C. H. Hoi, W. Liu, M. R. Lyu, and W. Y. Ma, “Learning distance metrics with contextual constraints for image retrieval,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog., 2006

[8] J. V. Davis, B. Kulis P. Jain, S. Sra, and I. S. Dhillon, “Information Theoretic Metric Learning,” in Proc. 24th Int. Conf. Mach. Learn., 2007

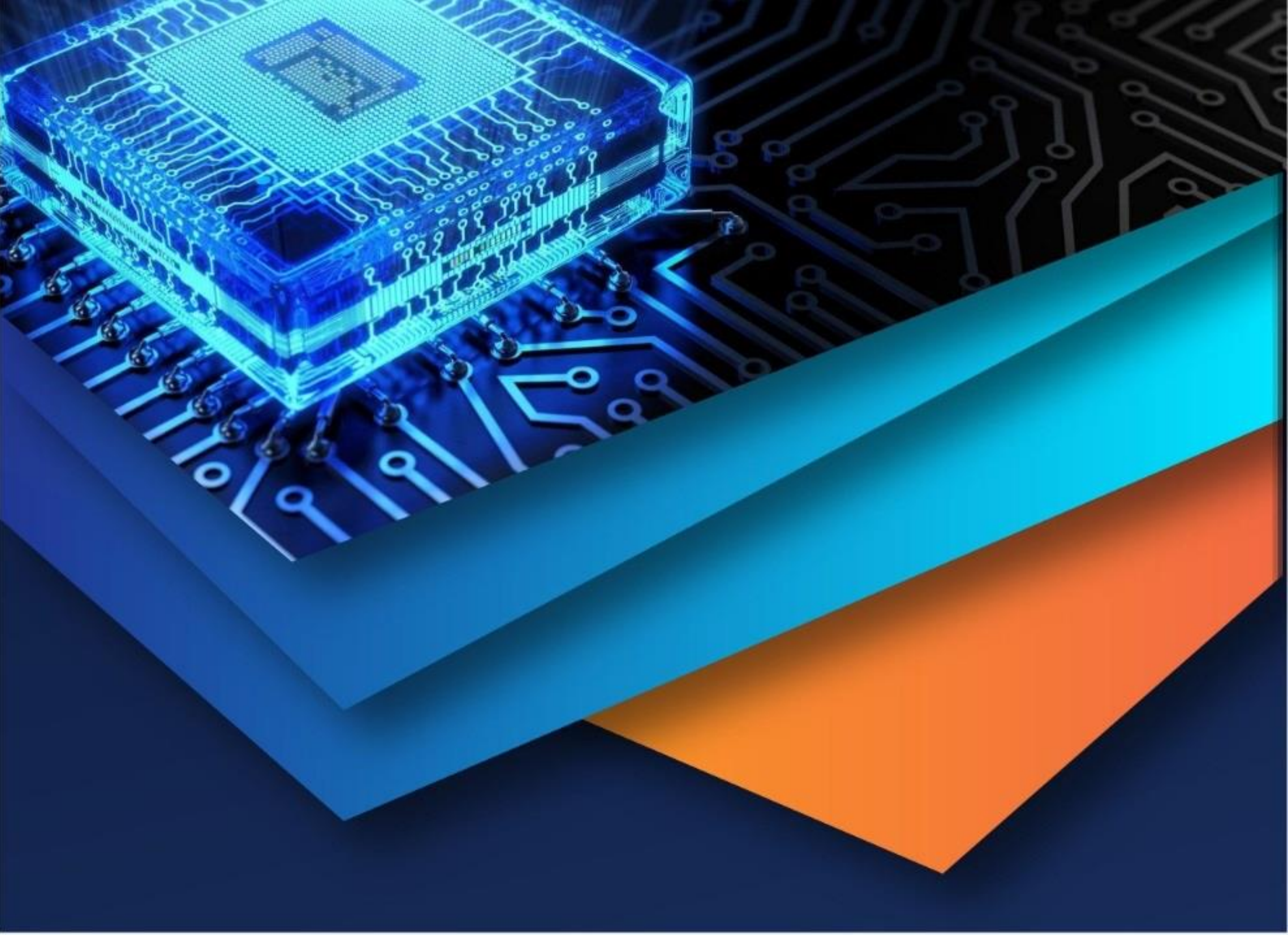
[9] B. Kulis, S. Basu, I. Dhillon, and R. Mooney, “Semi-supervised graph clustering: A kernel approach,” in Proc. 22nd Int. Conf. Mach. Learn., 2005

[10] Z. Yu, et al., “Progressive subspace ensemble learning,” Pattern Recog. 2016.

[11] I. A. Maraziotis, “A semi-supervised fuzzy clustering algorithm applied to gene expression data,” Pattern Recog., vol. 45, no. 1,pp. 637–648, 2012.

[12] Z. Yu, H.-S. Wong, J. You, Q. Yang, and H. Liao, “Knowledge basedcluster ensemble for cancer discovery from biomolecular data,”IEEE Trans. NanoBioScience, vol. 10, no. 2, pp. 76–85, Jun. 2011.

[13] L. Zheng and T. Li, “Semi-supervised hierarchical clustering,” inProc. IEEE 11th Int. Conf. Data Mining, 2011, pp. 982–991.



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY