COLLABORATIVE CLUSTERING: AN ALGORITHM FOR SEMI-SUPERVISED LEARNING

P.Padmaja$^{\#1}$,V.R.V.Vamsi Krishna.N^{*2}

DEPARTMENT OF INFORMATION TEHNOLOGY,

GIT, GITAM UNIVERSITY

***Abstract*— Supervised learning is the process of disposition of a set of consanguine data items which have known labels. The apportion of an unlabeled dataset into a conglomeration of analogous items(clusters) by the optimization of an objective function to attenuate the inter-class similarity and augment the intra-class similarity is called unsupervised learning. But when multi-modal data is used, there ensues a predicament with algorithms of either type. Hence a new breed of clustering known as Semi-Supervised clustering has been popularized. This algorithm partitions an unlabelled data set into a congregation of data items by taking only the limited available information from the user.**

When contemporary clustering algorithms are applied on a single dataset, different result sets are obtained. Hence an algorithm is needed to reveal the underlying structure of the dataset. In this paper an algorithm for semi-supervised learning is endowed, quartered on the principle of collaboration of clusters. This analytical study can be justified by carrying out various experiments.

***Keywords* - semi-supervised learning, collaboration of clusters, multi-modal data, unsupervised learning**

I.INTRODUCTION

Data mining is termed as the extraction of data from multiple data sources. The obtained data is known as information and has numerous applications. For this reason data mining has been the epicentre of research for the past few decades. Learning is one of the most intricate concepts of data mining where extensive research is being conducted. Two kinds of classification algorithms have emerged: supervised classification algorithms and unsupervised classification algorithms [1]. When a fixed set of classes with known class labels are used to build a classification function based on which the class patterns are formed, it is known as supervised classification [2]. In unsupervised classification an unlabeled dataset is partitioned into groups of similar items based on the object optimization function built on the dataset. Both these methods minimize the inter-class similarity and maximize the intra-class similarity.

The two methods of learning have their own disadvantages. In the case of supervised learning, it is impossible to find a label for each and every sample in the dataset. The problem of over-fitting can occur. In case of unsupervised classification, the determination of the ideal number of clusters has always been a problem and is still under extensive research.

To eliminate the problems in the two approaches, a new breed of algorithm known as semi-supervised learning has been proposed. The main objective of semi-supervised learning [3] is to obtain a better partition of the dataset by incorporating background knowledge. The algorithm partitions an unlabeled dataset by making use of some known information provided by the user. This approach is mainly used when the data is multi-modal where the dataset contains both labelled and unlabeled data. The background information is incorporated into the dataset based on constraints which are known as must-link constraints [4] and cannot-link constraints. Must-link constraints indicate that the two selected objects in data should be placed in the same cluster. Similarly, cannot-link constraints indicate that the two selected objects should not be placed in the same cluster. Clustering based on this approach generally leads to a more generalizable function.

This paper concentrates mainly on proposing an algorithm for semi-supervised based on the collaborative clustering approach which utilizes the labelled samples. This algorithm is very useful in case of a multi-modal dataset where the user does not know the coherence of the samples in the dataset to the information provided.

II.SEMI-SUPERVISED CLUSTERING

Semi-Supervised clustering [5] is designed as an improvement for the existing clustering algorithms. The essence of semi-supervised clustering is to use the background knowledge provided by the user. The datasets used in this case are multi-modal where the samples are both labelled and unlabeled.

Basu et al. [6] proposed a probabilistic model for semi-supervised learning based on Hidden Markov Random Fields(HMRF), that provides a principled framework for incorporating supervision into prototype-based clustering. Experimental results on several text data sets demonstrate the advantages of the proposed framework.

Wagstaff et al. [4] presented a constrained version of the k-means algorithm to bias the partition of samples to clusters. Not only do the constraints define the algorithm, they also enable the user to optimize the distance function based on which the clustering process is performed.

Basu et al. [7] articulated that a subset of data objects is used to seed the clusters of the k-means algorithm. The paper also proposes two algorithms, seeded k-means and constrained algorithms. The results are analyzed with different datasets. Both these algorithms work efficiently independent of the size of the dataset. However, the performance degrades as the noise level rises.

W.Pedrycz et al. [8] introduced various types of background knowledge and discusses the way to exploit and incorporate these background knowledge.

J.Gao et al [3] have exclusively stated that incorporating background knowledge into unsupervised clustering has been the subject of extensive research in recent years. This paper emphasizes on incorporation of only partial background knowledge which is available from the user in the form of constraints. This is useful when the labelled samples have moderate overlapping features with unlabeled data. This is also known as semi-supervised clustering.

Germain Forestier et al [9] proposed a new algorithm for semi-supervised clustering. This paper defines semi-supervised clustering as process of grouping an unlabeled dataset into groups of similar items by taking partial available information from the user. It also emphasizes that the importance of the information given by the user is voluminous in nature.

This paper also states that collaborative clustering has the following steps: collaboration of clusters and cluster labelling. In collaborative clustering, initially different clustering algorithms are applied on an unlabeled dataset and the corresponding result sets are refined to find a general agreement about the classification of the data. This is used to discover the structure of the dataset. Next, a cluster labelling process is used for final class allocation. The problem arises only when there is no available sample is present in the cluster and hence it cannot be labelled. This process helps in discovering the unknown relevant patterns. This is useful when the data is multi-modal. It also maintains the degree of freedom to discover relevant unknown patterns.

III. SEMI-SUPERVISED COLLABORATIVE CLUSTERING

The essence of collaborative clustering is to apply the background knowledge provided by the user. There are two main processes in the proposed algorithm. Initially, an unlabelled dataset partitioned by applying different existing clustering algorithms. The corresponding result sets are refined to obtain a generalization among the samples. Then an algorithm is used to label the datasets with the information provided by the user. Initially, different clustering algorithms are applied on the same multi-modal dataset and the corresponding result sets obtained. These result sets are then refined using a systematic refinement approach to obtain result sets with similar structure and the same number of clusters. Then, a voting algorithm is used to unify the clusters into one single result set. Finally, a cluster labelling process is employed to label the samples.

IV. PROPOSED SYSTEM

The proposed system consists of three major steps where the unlabeled dataset is initially clustered using the different clustering algorithms. Then the results are refined using the concept of distribution of local resolution of conflicts which is performed iteratively. Finally, the unification process is done by using an appropriate algorithm

Collaborative Process:

Initial Clustering: Different clustering algorithms are applied on the same unlabeled dataset. Each algorithm is initialized with its own parameters. Hence, the corresponding results consist of different number of clusters and dissimilar structures.

Results Refinement: There are four main phases which are iterated to eliminate the conflicts between the clusters and arrive at the result such that both the datasets consist of similar structures and possibly the same number of clusters. The entire concept is based on the distributed local resolution of conflicts [9].

Conflict Detection: This process involves in finding all the couples (C_k^i , R^j), i ≠j, such as $C_k^i \neq CC(C_k^i , R^j)$ [9]
Each conflict is associated with the conflict importance function:

$$CI(K_k^{i,j}) = 1- S(C_k^i, CC(C_k^i , R^j)) \quad (1)$$

where $S(C_k^i, CC(C_k^i , R^j))$ defines the similarity of the clusters and $K_k^{i,j}$ is the conflict chosen

Local Resolution of Conflicts: The most important conflict is initially chooses based on the conflict importance function define in equation(1) and one operator of the following operators is applied on it.

(1) Merging of Clusters: clusters whose correspondence value is less than the threshold value are merged [10].

(2) Splitting of Clusters: clusters whose correspondence value is greater than the threshold value are split.

The values of correspondence are calculated based on the corresponding clusters [9] function.

ALGORITHM FOR OPERATOR APPLICATION

```
let n= $|CCS(C_k^i , R^j)|$
let $R_1^i$(or $R_1^j$) be the result of application of an operator $R^i$(or $R^j$)
let t be the threshold fixed
if n>t then
    $R_1^i = R^i \backslash \{C_k^i\} \cup \{ split(C_k^i,n)\}$
else
    $R_1^i = R^i \backslash \{C_k^i\} \cup \{ merge(C_k^i,n)\}$
End
```

Global Resolution of Conflicts: Once the conflicts are resolved locally then global resolution of conflicts is performed to ensure the compatibility during the unification process. This is performed based on the global agreement coefficient [9].

Unification: In the final step the result sets are unified using a voting algorithm [11]. This is possible only when the result sets have similar structures with the same number of clusters. For each data sample in the cluster C_k^i the most similar data sample in C_k^j is identified and they are grouped. This is the essence of the voting algorithm.

The algorithm for the proposed system:

STEP 1: begin

STEP 2: collaborative clustering process

Step 2.1: Initial Clustering

Step 2.2: Results Refinement

Step 2.2.1: conflict detection

Step 2.2.2: local resolution of conflicts

Step 2.2.3 : Global resolution of conflicts

Step 2.3 : unification

STEP 3 : cluster labeling

STEP 4 : end

V. CLUSTER LABELLING

Once the unification process is complete, the result set R^u consisting of n_u clusters $\{C_k^u\}_{1<k<n.}$ is obtained.

Let $S = \{s_i\}$ be the set of samples available for the dataset. The resultant cluster sets R^u are labelled on applying a voting method. This enables the user to discover the unknown patterns.

VI. CONCLUSION AND FUTURE SCOPE

In this paper, an algorithm for semi-supervised learning was proposed using the concept of collaboration of clusters and cluster labelling. This has numerous applications especially when the datasets are multi-modal. The algorithm also discovers relevant patterns that are unknown to the user.

The limitation for the proposed algorithm is that it cannot label the data samples if no information is provided by the user. The only solution being provided is to label the samples with the corresponding label of the nearest known sample.

REFERENCES

[1] Jain, A.K., & Dubes, R.C. Algorithms for clustering data. Prentice hall. 1998.

[2] Z. Fan-Zi and Q. Zheng-Ding. A survey of classification learning algorithm. International Conference on Signal Processing, 2:1500–1504, 2004.

[3] J. Gao, P.-N. Tan, , and H. Cheng. Semi-supervised clustering with partial background information. In *SIAM InternationalConference on Data Mining*, 2006.

[4] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In

[5] *International Conference on Machine Learning*, pages 557–584, 2001.

[6] I. Davidson, K.L Wagstaff, and S.Basu. Measuring constraint-set utility for partitional clustering algorithms. In European conference on machine learning, pages 115-126, 2006.

[7] S. Basu, A. Banerjee, and R.J.Mooney. Semi-supervised clustering by seeding. In *International Conference on Machine Learning*, pages 19–26, 2002.

[8] S. Basu, A. Banerjee, and R. J. Mooney. Active semisupervision for pairwise constrained clustering. In *SIAM International Conference on Data Mining*, 2004.

[9] W. Pedrycz. Fuzzy clustering with a knowledge-based guidance.*Pattern Recognition Letters*, 25(4):469–480, 2004.

[10] Germain Forstier, Cedric Wemmert and Pierre Gancarski, Semi-supervised collaborative clustering with partial background knowledge. IEEE International Conference on Data Mining Workshops Pg 211-217. 2008

[11] C.Wemmert and Gancarski. A multi-view voting method to combine unsupervised classifications. *Artificial Intelligence and Applications*, pages 362–324, 2002.

[12] Mingwei Leng, Haitao Tang , Xiaoyun Chen Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing,pp-815-820,2007

[13] C.Wemmert and Gancarski. A multi-view voting method to combine unsupervised classifications. *Artificial Intelligenceand Applications*, pages 362–324, 2002.