



A FAST SEMI-SUPERVISED SINGLE-LINK CLUSTERING METHOD

Y.C.A.PADMANABHA REDDY, P.VISWANATH, AND B.ESWARA REDDY

ABSTRACT. Semi-supervised single-link (SSL) clustering method, can give better results than unsupervised single-link method. This paper proposes constrained leader-based SSL method to achieve almost similar clustering result, but with much reduced time by employing Leaders clustering method working with leader data patterns, instead of working with every individual data pattern in the dataset. The proposed method basically a two step process. First it derives a set of prototypes by using a fast clustering method along with constraints called the Constrained leaders clustering method with threshold parameter α . The second step is by applying the SSL method in the presence of a few labeled data with respecting the constraints (Must-link and Cannot-link) are presented in this paper. The ‘must-link’ constraints (a constraint of this kind says, two particular data patterns must be grouped together), this type of constraints can be merged in the process easily, but “can-not link” constraints (*e.g.*, two particular data patterns must not be grouped together) or “distinct label constraints” having their own difficulties. Both benchmark and synthetic datasets are used in the experimental study. Experimentally, the proposed Constrained leader-based SSL method is shown to work at-least 10 factors faster than SSL method, and achieves almost similar clustering result.

1. INTRODUCTION

The traditional single-link clustering method, can find arbitrary shaped clusters [1] whereas k-means clustering method can’t find. K-means can found only spherical shaped clusters. The single-link method comes under the categorization of hierarchical clustering methods where it follows bottom up approach. The final result is evolved based on the user given selection criteria. The Semi-supervised single-link method is having two demerits, i.e (i) The computational cost of the existing SSL method and (ii) the final result is chosen from user criteria. constrained

1991 *Mathematics Subject Classification.* 68T10, 62H30.

Key words and phrases. Machine Learning, Clustering, semi-supervised, Pattern Recognition.

2 Y.C.A.PADMANABHA REDDY, P.VISWANATH, AND B.ESWARA REDDY

leader-based SSL method can overcome the above two problems, and also produced the almost similar clustering result with slight decrease of rand-index and much reduced time.

Constrained Semi-supervised clustering got more attention in the field of machine learning research, in this type of clustering the constraints are smoothly integrated in the process along with the unlabeled data patterns [2–4]. Pairwise Constraints can be “must-link” constraints in the form of “two specific data patterns must be clustered together” or “cannot-link” constraints, i.e., constraints like “these two specific data patterns can-not not be clustered together” [5,6].

This paper proposes a method called constrained leader-based semi-supervised single-link clustering method which is a two step process where prototypes are generated in the first step with the help of fast clustering method called Leaders method and incorporation of instance level constraints. The method, purely based on Leaders algorithm, first derives a set of leaders set and associated followers comes under specific leader based on Threshold value α and also labels are considered in to account. The efficiency of the proposed method is compared with the Semi-supervised SSL method, constrained complete-link method and also with traditional linkage methods such as single-link and complete-link methods, clearly reported in experimental section [7] (details are given in Sections 2 and 4) using two benchmark datasets and three artificially generated synthetic data sets. Real or benchmark datasets are taken from the UCI machine learning repository [8].

The remaining paper is arranged as follows. Section 2 reviews about the related work related to the semi-supervised clustering methods. Section III reviews about the semi-supervised SSL method [9] Section 4 explains about the proposed Constrained Leader-based SSL method. This method is evaluated with other existing methods such as SSL and CCL methods. Section 5 gives complete information regarding experimental study on proposed method by using the benchmark and synthetic datasets. Section 6 states the conclusions of our research work and gives new direction of research ideas for future researchers.

This paper is an improvement to our conference paper with title “semi-supervised single-link clustering method” [9] where we proved that SSL method performance is better than other existing methods such as single-link, complete-link and constrained complete link methods. The enhanced SSL is named as constrained leader-based ssl, which reduces the computational time by working with constrained leader patterns instead of

working with all the patterns in the dataset and also respecting pairwise constraints.

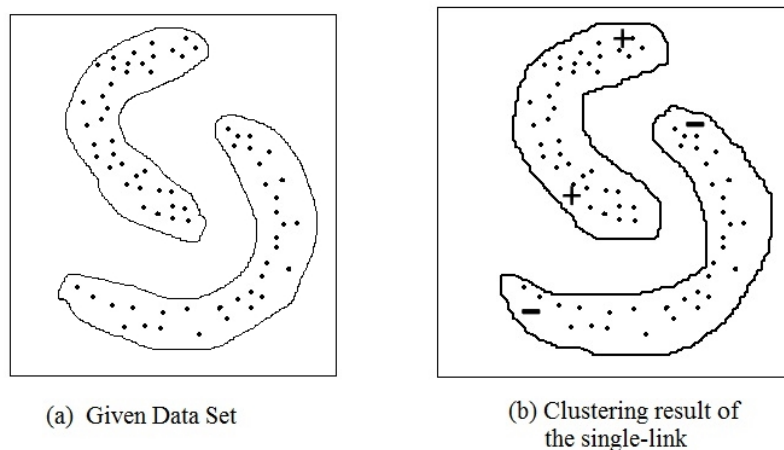


FIGURE 1. Arbitrary shaped clusters from the dataset

2. RELATED WORK

Conventional single-link clustering method can work for large datasets with the help of Leaders Clustering method l-ssl [10] Semi-supervised clustering performs better than conventional clustering process with the help of side information in the form of pairwise constraints (must-link and cannot link). The quality of the semi-supervised clustering is improved by proper incorporating the constraints in to the process. Most of the existing methods must-links are properly handled but can-not link constraint is violated [6].

Hierarchical methods like semi-supervised single-link and density based methods like SS-DBSCAN [11] can find arbitrary shaped clusters by respecting pairwise constraints.

In Fig 1, the dataset is given, where underlying arbitrary shaped clusters are formed which is clearly shown in fig 1(b). If the few data patterns consist of labels then semi-supervised single-link method can work better by respecting the class labels as constraints.

4 Y.C.A.PADMANABHA REDDY, P.VISWANATH, AND B.ESWARA REDDY

Euclidean distance is the metric to find the distance between two data patterns and clusters. Let $distance(a, b)$ is the distance between patterns¹ a and b . From single-link method perspective, Distance between clusters P_i and Q_j is defined as

$$(1) \quad Distance(P_i, Q_j) = Min(\{distance(a, b) | a \in P_i, b \in Q_j\})$$

Distance is the only metric which can distinguish from single-link with other linkage methods. The distance between the clusters is defined as follows

For complete-link,

$$(2) \quad Distance(P_i, Q_j) = Max(\{dist(a, b) | a \in P_i, b \in Q_j\})$$

and for average-link,

$$(3) \quad Distance(P_i, Q_j) = Average(\{dist(x, y) | a \in P_i, b \in Q_j\})$$

Basically, semi-supervised clustering paradigms are categorized in to two types: They are a) Similarity-adapting methods and b) search-based methods. Constrained clustering has got more attention in the recent years [5]. Experimentally The constrained complete link (CCL) method and semi-supervised single-link (SSL) method are clearly explained in [9].

3. SEMI-SUPERVISED SINGLE-LINK CLUSTERING METHOD

Semi-supervised learning (SSL) is a special kind of Machine Learning (ML) technique in which the data patterns are semi-labeled or partially labeled as shown in the fig 1(b). SSL is the mid-way between supervised and unsupervised learning [12]. The main motivation of SSL is to overcome the problems or demerits of both supervised and unsupervised learning. Supervised learning is nothing but classification requires more amount of training data to classify the test data, which is cost effective and time consuming process. On the other hand, example for unsupervised learning is clustering, which doesn't require any labeled data which is cost effective. The clustering process clusters the data which clusters the data based on similarity measures. The main drawback of this approach, it can't cluster the unknown data correctly. To overcome these issues, SSL has been proposed by research community [2], which can learn with small amount of labeled data to label the unlabeled or unknown test data.

The distance matrix in SSL method is created by incorporation of constraints with the metric of euclidean distance is found between the data

¹“Data elements” and “Data pattern” are used interchangeably, and means the same thing.

TABLE 1. Benchmark and Synthetic datasets used for Experimental Study

S.No	Data set	No. of Classes	No. of features	Total Size
1	OCR	10	192	10988
2	PENDIGIT	10	16	9508
3	SYNTH1 (banana)	2	2	14920
4	SYNTH2 (spiral)	2	2	15000
5	SYNTH3 (double circle)	2	2	20000

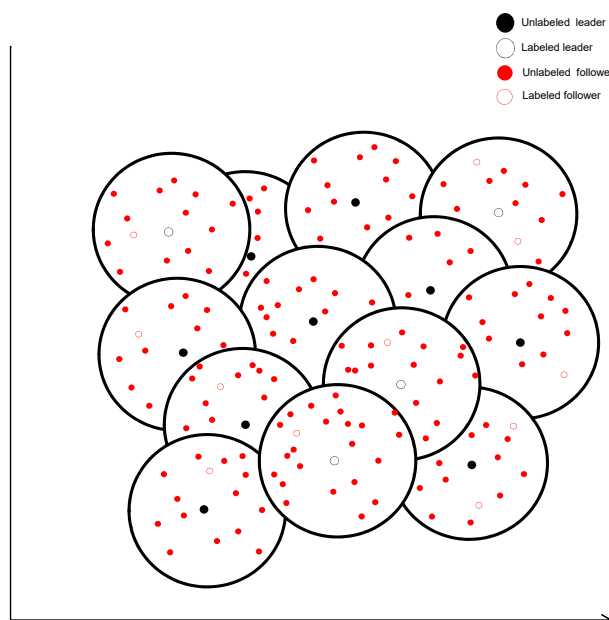


FIGURE 2. Constrained Leaders

patterns and clusters. The pairwise constraints are integrated in the process of clustering by filling the values of "Zero" and "infinity" for must-link and can-not link constraints. the detailed step by step SSI algorithm explained in our conference paper and also related proof are available in [9].

6 Y.C.A.PADMANABHA REDDY, P.VISWANATH, AND B.ESWARA REDDY

```

1 Algorithm Constrained-Leaders( $X, \alpha$ )
2 let us assume the first pattern is always a leader and has label
3 Let  $\mathcal{L}$  be the leaders set
4 Let  $\mathcal{F}$  be the followers set
5 let  $\mathcal{L} \leftarrow x$ 
6 let  $\mathcal{F} \leftarrow \phi$ 
7 for  $x \in X$  do
8   Find a  $l \in L$  such that distance  $\|l - x\| \leq \alpha$  and  $x$  is labeled
9   If there is no such  $l$  then
10  if  $label(l) \neq label(x)$  then
11     $\mathcal{L} \leftarrow \mathcal{L} \cup \{x\}$ ;
12     $count(x) \leftarrow 1$ ;
13     $count(l) \leftarrow count(x) + 1$ ;
14    if  $label(l) = label(x)$  then
15       $\mathcal{F} \leftarrow \mathcal{F} \cup \{x\}$ ;
16    end
17  else
18     $\mathcal{L} \leftarrow \mathcal{L} \cup \{x\}$ ;
19     $count(x) \leftarrow 1$ 
20  end
21 end
22 else
23   If the leader  $l$  is unlabeled such that distance( $x, l$ )  $\leq \alpha$  &  $x$ 
   is labeled
24    $label(l) = label(x)$ ;
25 end
26 end
27 no_of_leaders = len( $\mathcal{L}$ );
28 return  $\langle \mathcal{L}, \mathcal{F}, \# \text{ no.of.leaders} \rangle$ ;

```

Algorithm 1: Selection of Constrained Leaders

3.1. Benchmark Datasets Description. In this paper we employed variety of datasets are shown in 1 viz. two standard datasets: OCR, Pendigit [13] used popularly due to its large size and three synthetic datasets: banana, spiral and double circle. The OCR is created from the printed articles which consists 10988 patterns and it has 192 features are grouped into different labels from zero to nine. Pendigit dataset is created from digital input device from 44 writers. It has 9508 patterns with 16 features and classified into 10 different groups. The other three datasets

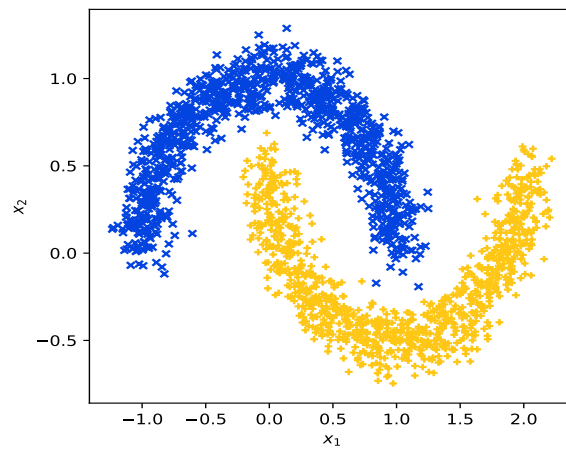


FIGURE 3. banana dataset

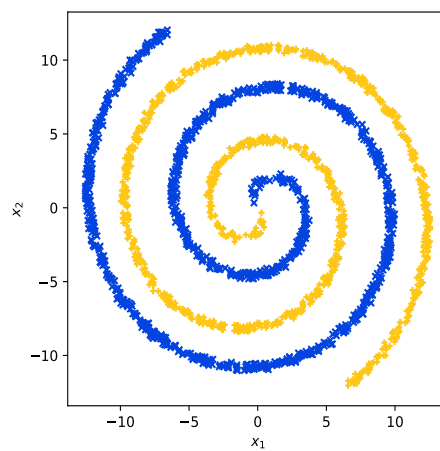


FIGURE 4. spirals dataset

are artificially generated based on norm in different shapes. These artificial datasets has two unique patterns are classified into two groups.

8 Y.C.A.PADMANABHA REDDY, P.VISWANATH, AND B.ESWARA REDDY

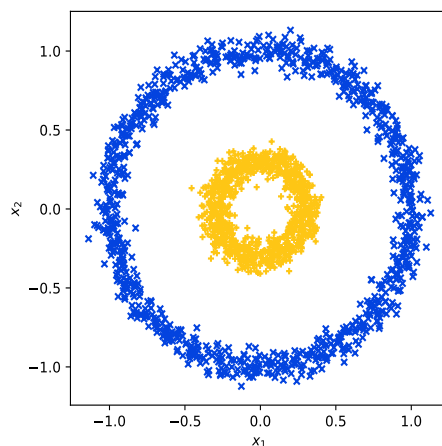


FIGURE 5. Circles dataset

4. THE PROPOSED CONSTRAINED LEADER-BASED SEMI-SUPERVISED SINGLE-LINK CLUSTERING METHOD

In this work, our dataset only few data patterns have labels. Initially the leader set L is with one pattern i.e first pattern having with label. The algorithm finds the leaders and followers until all the patterns are visited once.

First find constrained Leaders set $L = l_1, l_2, \dots, l_n$ through Leaders algorithm then apply Semi-supervised single-link clustering. For reducing the time complexity of semi-supervised single-link clustering (SSL), here we use constrained Leaders algorithm where constraints are taken in to account. In these leaders algorithm we take first pattern as the leader and find out the distance between first pattern to the second pattern. if the distance is less than the threshold value then second pattern becomes as a follower, if the the same distance is greater than the threshold value then that second pattern becomes as the next leader. The process repeat until every pattern in the dataset is scanned at once. Finally, we will get set of leaders with associated followers under each leader.

There are various procedures for selection of threshold value [14] which is significant input parameter. For knowing the threshold value we have generated a random function and pick some randomly generated patterns and see the distance between the patterns and taking the average tau value. After that multiply it with some required percentage we get the average tau value. Base on that threshold value we generate leaders

TABLE 2. Comparison of Clustering result by using Rand-index

Dataset	SL method	CL method	CCL method	SSL method	l-SSL method
OCR	0.88	0.90	0.90	0.94	0.92
PENDIGIT	0.72	0.66	0.78	0.86	0.83
SYNTH1 (banana)	0.88	0.90	0.90	0.94	0.92
SYNTH2 (spiral)	0.72	0.66	0.78	0.86	0.84
SYNTH3 (double circle)	0.98	0.92	0.97	0.96	0.96

and place that leaders in the leader set then we can easily decrease the threshold value. when a leader knows the class label then automatically all the followers get the same label so that time taken for executing is easily decreased. By using this leaders , we can improve the performance of Semi-supervised single-link clustering (SSL).

4.1. Constrained Leader-based SSL method. In this section, we cluster the patterns based on the leader set. Initially, the Algorithm 2 filter the leader patterns returned by Algorithm 1. The algorithm converges till all the leader patterns are covered and returns the different clusters. The clusters are formed based on Euclidean distance between patterns. Constrained clustering applied in many real world applications. Constrained leaders are generated based on the threshold parameter α and also class label considered in to the account.there are labeled leader which has label and assigns its label to follower data patterns.The unlabeled leader pattern will get the label from the follower pattern, if it is labled.the same process is repeated untill all the patterns visited once. the constrained leaders, followers with labeled and without labels are clearly shown in fig 2.

5. EXPERIMENTAL STUDY

Experiments are conducted on PC with Intel i3 processor IV generation with 2GB RAM capacity. Both benchmark and synthetic datasets are used in the experimental study (see in Table-I). OCR dataset is also used in [15–17]. Benchmark datasets are Optical character Recognition (OCR) [18] and pen digit datasets are taken from UCI machine Learning repository [18].

For every pattern, its actual label is available, i.e nothing but its ground truth . Clustering result is evaluated by using the measure Rand-index [19] is used for obtained clustering result with the ground truth clustering result. From table-I, OCR dataset consist of 10,998

10 Y.C.A.PADMANABHA REDDY, P.VISWANATH, AND B.ESWARA REDDY

```

1 Algorithm  $Cl - SSL(X, \mathcal{L})$ 
2 Input:  $N$  training examples with partial labels.
3  $\mathcal{L}$  - has indexes of original dataset
4  $\{X \rightarrow Y\} \{ \langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \dots, \langle x_{n-1}, y_{?} \rangle, \langle x_n, y_{?} \rangle \}$ 
5 Output:  $C$  - clusters with labels
6 Let  $D$  be a distance matrix
7  $\mathcal{K} \in X(\mathcal{L})$ 
8 do
9   for each  $p$  in  $\mathcal{K}$  do
10     //p - pattern (i,j)
11      $l \leftarrow \text{Leaders}(p, \alpha)$ 
12      $\text{dist} = \text{Euclidean distance}(p)$ 
13     if  $\text{dist}_{i,j}$  then
14        $D_{i,j} \leftarrow 0$  //Must-link
15     end
16     else
17        $D_{i,j} \leftarrow \infty$  //Cannot-link
18     end
19      $D = \text{Do distance}(p)$ 
20   end
21 while till cluster similar labels
22 return  $C$ 

```

Algorithm 2: Leaders based Semi-supervised Clustering

TABLE 3. Running Time of SSL and L-SSL in sec

Data set	SSL	L-SSL	% of Running time reduced
OCR	0.931	0.073	92.16
PENDIGIT	1.624	0.089	94.52
SYNTH1 (banana)	0.095	0.018	81.05
SYNTH2 (spiral)	0.223	0.068	69.51
SYNTH2 (double circle)	0.358	0.168	53.07

data patterns with 198 features, 10 classes, where each class label category is associated with equal no of data patterns. On OCR dataset, ssl method performs better with rand-index of 0.94(see in Table-II) when compared with single-link, complete- link, constrained complete-link and constrained l-ssl achieves 0.88, 0.90,0.90, 0.92 respectively reported in Table-II. Our proposed method decreases to 2 percent of accuracy when

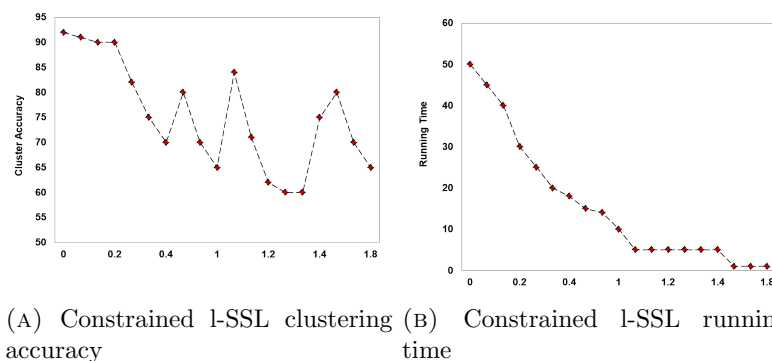


FIGURE 6. Constrained l-SSL accuracy and running time for OCR data set with respect to α

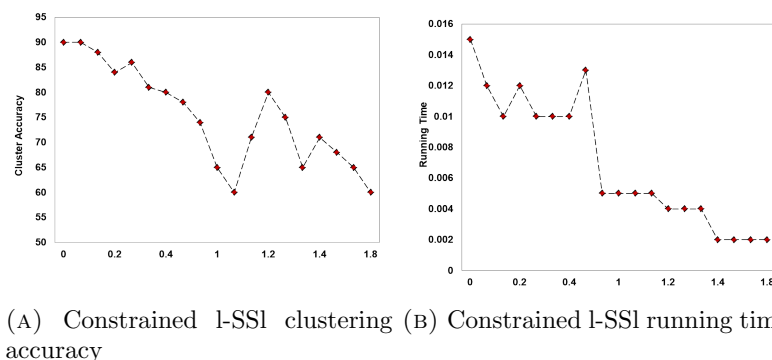
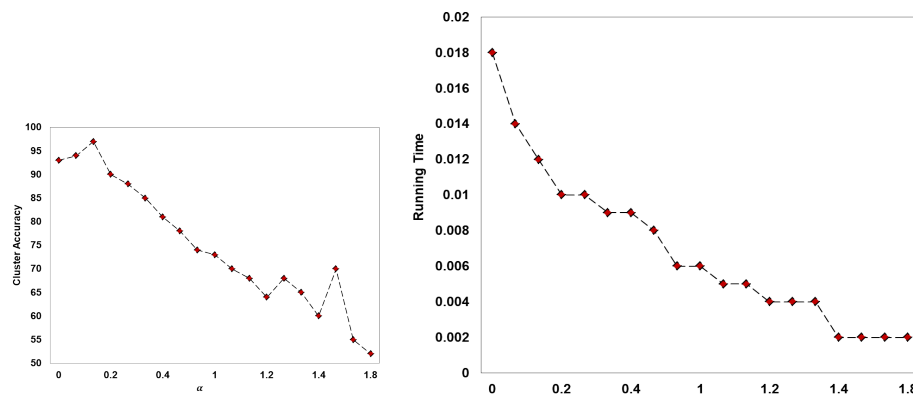


FIGURE 7. Constrained l-SSL accuracy and running time for Pendigit data set with respect to α

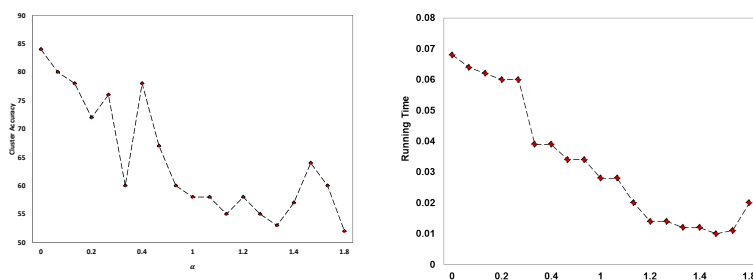
compared with the ssl method but there is good improvement in reducing the running time reported in table-III where ssl running time is 0.931 sec and constrained l-ssl achieves 0.073sec. Our proposed method reduces the running time to 92.16% when compared with ssl method.

From Table-I, Pendigit dataset consist of 9508 data patterns with 16 features and 10 classes. ssl method performs better rand-index 0.86 when compared with single-link, complete-link, constrained complete-link and constrained l-ssl rand-index are 0.72, 0.66, 0.78, 0.83 respectively (see in Table-II). Our proposed method not performed well in terms of accuracy but very good improvement in the running time when compared with

12 Y.C.A.PADMANABHA REDDY, P.VISWANATH, AND B.ESWARA REDDY



(A) Constrained l-SSl clustering accuracy (B) Constrained l-SSl running time

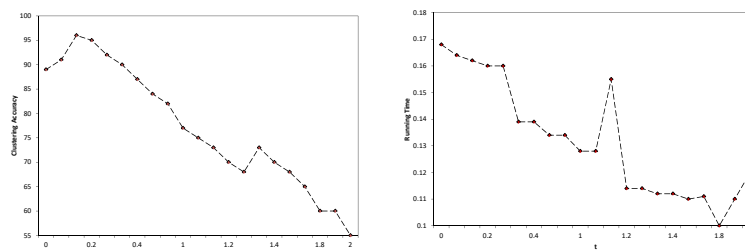
FIGURE 8. Constrained l-SSl accuracy and running time for Banana data set with respect to α 

(A) Constrained l-SSl clustering accuracy (B) Constrained l-SSl running time

FIGURE 9. Constrained l-SSl accuracy and running time for Spiral data set with respect to α

the ssl method. In Table III, we reported that ssl method running time is 1.624 sec and proposed method Constrained l-ssl method is 0.089 sec, which is much reduced time of 94.52%.

In this section we experiments done with both synthetic and standard datasets. Three synthetic datasets are generated and two standard data sets are taken from the UCI repository [8]. and three synthetic data sets. Table 1 describes the details about these five data sets. These data sets



(A) Constrained l-SSL clustering accuracy (B) Constrained l-SSL running time

FIGURE 10. Constrained l-SSL accuracy and running time for double-circle data set with respect to α

are labeled for supervised learning. For Semi-supervised clustering, we have considered only 10% of labels to form as constraints which helps in grouping the patterns during the clustering process. The two data sets named OCR Optical Character Recognition and PENDIGIT from the UCI repository [8] and three synthetic data sets are generated artificially named banana (SYNTH1), spiral (SYNTH2) and double circle (SYNTH3).

The proposed method is compared with traditional linkage based methods such as single-link, complete-link, constrained complete-link and semi-supervised single-link method as shown in the Table-II. The stopping condition for the clustering is based on the number of classes associated in the dataset which is specified in the Table 1 for the five data sets from the UCI repository.

Table 3 demonstrates running time of SSL and Constrained L-SSL methods. Constrained leader based SSL perform better when compared to SSL method. The running time depends on the α . The Figure 6b, 7b, 8b, 9b shows the running time with respect to α . The running time of OCR dataset is reduced to 92% compared to traditional SSL method. Similarly, the other datasets PENDIGIT, Banana and Spiral is reduced to 94.52, 81.05 and 69.51 respectively.

6. CONCLUSIONS AND FUTURE WORK

Constrained leader-based SSL clustering performs with much reduced time when compared with SSL clustering method with decrease of accuracy. The decrease of rand-index value for both synthetic and standard

14 Y.C.A.PADMANABHA REDDY, P.VISWANATH, AND B.ESWARA REDDY

datasets with minimum of 0.1 to maximum of 0.3 clustering result due to the approximation error rate during the clustering process. Stopping condition is based on the no of classes in the dataset.Reducing this error rate and getting similar clustering result for Constrained l-SSL with SSL method is one new future direction of research work.

ACKNOWLEDGEMENTS

The work reported in this paper is partially supported by UGC-SERO Minor Project with Reference No. MRP-4609/14.

REFERENCES

- [1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.
- [2] O. Chapelle, B. Schölkopf, A. Zien *et al.*, "Semi-supervised learning," 2006.
- [3] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 11.
- [4] A. M. Bensaid, L. O. Hall, J. C. Bezdek, and L. P. Clarke, "Partially supervised clustering for image segmentation," *Pattern Recognition*, vol. 29, no. 5, pp. 859–871, 1996.
- [5] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl *et al.*, "Constrained k-means clustering with background knowledge," in *ICML*, vol. 1, 2001, pp. 577–584.
- [6] S. Basu, A. Banerjee, and R. Mooney, "Semi-supervised clustering by seeding," in *In Proceedings of 19th International Conference on Machine Learning (ICML-2002)*. Citeseer, 2002.
- [7] D. Klein, S. D. Kamvar, and C. D. Manning, "From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering," 2002.
- [8] A. Asuncion and D. Newman, "Uci machine learning repository," 2007.
- [9] Y. P. Reddy, P. Viswanath, and B. E. Reddy, "Semi-supervised single-link clustering method," in *Computational Intelligence and Computing Research (IC-CIC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–5.
- [10] B. K. Patra, S. Nandi, and P. Viswanath, "A distance based clustering method for arbitrary shaped clusters in large datasets," *Pattern Recognition*, vol. 44, no. 12, pp. 2862–2870, 2011.
- [11] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.
- [12] X. Zhu, "Semi-supervised learning literature survey," 2005.
- [13] P. Viswanath, K. Rajesh, C. Lavanya, and Y. P. Reddy, "A selective incremental approach for transductive nearest neighbor classification," in *Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE*. IEEE, 2011, pp. 221–226.

- [14] T. H. Sarma, P. Viswanath, and B. E. Reddy, "A hybrid approach to speed-up the k-means clustering method," *International Journal of Machine Learning and Cybernetics*, vol. 4, no. 2, pp. 107–117, 2013.
- [15] V. Ananthanarayana, M. N. Murty, and D. Subramanian, "An incremental data mining algorithm for compact realization of prototypes," *Pattern Recognition*, vol. 34, no. 11, pp. 2249–2251, 2001.
- [16] T. R. Babu and M. N. Murty, "Comparison of genetic algorithm based prototype selection schemes," *Pattern Recognition*, vol. 34, no. 2, pp. 523–525, 2001.
- [17] T. H. Sarma and P. Viswanath, "Speeding-up the k-means clustering method: A prototype based approach," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2009, pp. 56–61.
- [18] C. J. Merz and P. M. Murphy, "{UCI} repository of machine learning databases," 1998.
- [19] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING, RESEARCH SCHOLAR,
JNTUA , ANANTHAPURAMU-515 002 (A.P) , INDIA.

E-mail address: `ananthayc.reddy@gmail.com`

DEPARTMENT OF COMPUTER, SCIENCE & ENGINEERING, IIIT CHITTOOR, SRI
CITY-517541. CHITTOOR(DIST), A.P., INDIA

E-mail address: `Viswanath.p@iiits.in`

DEPARTMENT OF COMPUTER, SCIENCE & ENGINEERING, JNTUA COLLEGE OF
ENGINEERING, KALIKIRI-517 324 CHITTOOR(DIST), A.P., INDIA

E-mail address: `eswar.csejntua@gmail.com`

