Efficient Active Learning Constraints for Improved Semi- Supervised Clustering Performance

Ramkumar Eswaraprasad[1], and Shanmugam Vengidusamy[2]

***Abstract*—** This paper presents a semi supervised clustering technique with incremental and decremental affinity propagation (ID-AP) that structures labeled exemplars into the AP algorithm and a new method for actively selecting informative constraints to make available of improved clustering performance. The clustering and active learning methods are both scalable to large datasets, and can hold very high dimensional data. In this paper, the active learning challenges are examined to choose the must-link and cannot-link constraints for semi-supervised clustering. The proposed active learning approach increases the neighborhoods based on selecting the informative points and querying their relationship between the neighborhoods. At this time, the classic uncertainty-based principle is designed and novel approach is presented for calculating the uncertainty associated with each data point. Further, a selection criterion is introduced that trades off the amount of uncertainty of each data point with the probable number of queries (the cost) essential to determine this uncertainty. This permits us to select queries that have the maximum information rate. Experimental results demonstrate that the proposed ID-AP technique adequately captures and takes full advantage of the intrinsic relationship between the labeled samples and unlabeled data, and produces better performance than the other considered methods Empirically evaluate the proposed method on the eight benchmark data sets against a number of competing methods. The evaluation results indicate that our method achieves consistent and substantial improvements over its competitors.

***Keywords*—** Affinity propagation (AP), decremental learning, incremental learning, clustering, semi-supervised learning..

I. Introduction

MANY semi-supervised clustering methods have been proposed to put into effect top-down construction as clustering [1,2,3]. These methods allow the user to build pairwise constraints, which may be either must-link or cannot-link, on the data as side information. These papers have shown that the use of pairwise constraints can significantly improve the correspondence between clusters and semantic labels when the constraints are selected well. Currently, most work in semi-supervised clustering ignores this problem and simply selects a random constraint set, but some work has now been done on active constraint selection methods [5,6], which allow semi-supervised clustering algorithms to intelligently select constraints based on the structure of the data and/or intermediate clustering results. Active selection methods can be stratified according to whether nodes or node-pairs are the primary element on which the process is based. Node-based methods first select nodes of interest, and then query constraints based on those nodes [8], while those methods that directly seek pair constraints [6,7], define an uncertainty measure on pairs and iteratively seek the most uncertain pairs during constraint selection.

Both of these current approaches have drawbacks, however. Current node-based methods function by selecting all of their constraints in one long selection phase before clustering. Because of this, they cannot incorporate information from actual clustering results into their decisions, and may thus choose many unnecessary constraints (for instance, constraints regarding points that the algorithm is able to cluster correctly even without side information). In contrast, the pair-based methods choose constraints online based on intermediate clustering results, but due to the nature of the pair selection problem (n2) possible constraints to rank and select from) have thus far been limited to either binary or small-scale clustering problems.

In this paper, a critical issue with this approach that it only considers the pairwise uncertainty of the first query and fails to measure the benefit of the ensuing queries that are required to determine the neighbourhood for a point. A semi supervised AP (SAP) clustering approach is proposed which meets the following criteria:

(i) It is able to learn by using prior information, i.e., labelled samples.
(ii) It is able to identify the clusters by selecting the most informative unlabelled data points together with the labelled samples thus avoiding learning bias.
(iii) It does not require any access to the useless labelled samples (which have been fully utilized and do not convey new information) to avoid stability-plasticity dilemma.
(iv) It preserves previously acquired knowledge (i.e., useful labelled samples, which have not been fully utilized) to avoid losing any opportunity to learn new information.

For criterion (i), the labelled samples are used as prior information to adjust the similarity matrix of the AP framework;

For criteria (ii), (iii), and (iv), propose an incremental and a decremental learning principle (ID-LP) for both selecting useful unlabelled data and discarding useless labelled samples.

The proposed algorithm has the above properties and is called incremental and decremental AP (ID-AP). The ID-LP

Ramkumar Eswaraprasad, Senior Lecturer, Botho University, Botswana
Shanmugam Vengidusamy, Lecturer, Botho University, Botswana

iteratively improves the clustering accuracy in the learning process. A novel clustering model is learned at each iteration, and the clustering models learned at each iteration are combined together to form a final clustering model. At each iteration, the set of most informative unlabelled samples is selected based on the incremental learning principle (ILP) to avoid learning bias. In this condition, the fully used labelled samples that do not convey new information are discarded based on the decremental learning principle (DLP) to avoid stability-plasticity dilemma.

II. Methodology

This paper describes about the methodology to effectively choose pairwise queries to produce an accurate clustering assignment. Through active learning, the number of queries is reduced to achieve a good clustering performance. Consider this as an iterative process such that the decision for selecting queries should depend on what has been learned from all the previously formulated queries.

2.1. Neighborhood - Based Framework based on proposed semi supervised learning approach

Definition: A neighborhood contains a set of data instances that are known to belong to the same class (i.e., connected by must-link constraints). Furthermore, different neighbourhoods are connected by cannot-link constraints and, thus, are known thus, are known to belong to different classes.

Given a set of constraints denoted by K, can identify a set of m neighborhoods $H = \{H_1, \ldots, H_m\}$, such that m<K and K is the total number of classes. Denote each data instance as y and the label as "LB". One way to the neighborhoods is to view them as the "labeled examples" of the underlying classes because instances belonging to different neighborhoods are guaranteed to have different class labels [10], and instances of the same neighborhood must belong to the same class. A key advantage of using the neighborhood concepts is that by leveraging the knowledge of the neighborhoods, can acquire a large number of constraints via a small number of queries. In particular, if the neighborhood of an instance x is identified, immediately infer its pairwise relationship with all other points that are currently confirmed to belong to any of the existing neighborhoods. After selecting the most informative data point and querying Incremental Learning Principle (ILP) and Decremental Learning Principle (DLP) is used as a iterative procedure.

The goal of ILP is to iteratively identify the most informative unlabelled data. It implies that the labels assigned to the unlabelled data should be correct; thus expanding the training set with the most informative unlabelled samples should improve the clustering performance. The selection of the unlabeled data is based on their potential contribution to the training of subsequent models. Therefore, at each iteration, select the unlabeled data that are the most similar to the labeled samples. This requires measuring the similarity between the labelled samples and unlabeled data. However, at each step of the incremental process, ILP requires access to previously learned labeled samples. This may be critical in terms of stability-plasticity. If the previously learned labelled samples are preserved, the learning may not accommodate any information from the new acquired labeled samples. Therefore, have to discard these reiterative redundant/useless labelled samples, while keeping unused labelled samples to avoid losing acquired information. The overall algorithm is as follows

Algorithm 1. The Neighborhood-based Framework

Input: A set of data points M; the total number of classes K; the maximum number of pairwise queries P.

Output: a clustering of S into K clusters.

1: Initializations: K= 0; H1= {x}, where x is a random point in M; H= H1; m= 1; t= 0;

2: repeat

3: π =Semi-supervised-Clustering (M, K);

4: y^* =MostInformative (M, π ,H);

5: for each $H_i \in H$ in decreasing order of $p(y^* \in H_i)$ do

6: Query y^* against any data point $y_i \in H_i$;

7: s++;

8: Update *K* based on returned answer;

9: if (y^*, y_i,LB) then $H_i = H_i U\{y^*\}$; break;

10: end for

11: if no must-link is achieved

12: then m++; Hl= {y^*}; $H = H \cup Hl$;

13: until s > P

14: return Semi-supervised-clustering (M, *K*)

Briefly, the algorithms begin by initializing the neighborhoods by selecting a random point to be the initial neighborhood (line 1). In each iteration, given the current set of constraints K, it performs semisupervised clustering on M to produce a clustering solution (line 3). A selection criterion is then applied to select the "most informative" data point y* based on the current set of neighborhoods and the clustering solution (line 4). The selected point y* is then queried against each existing neighborhood N_i to identify where y* belongs, during which the constraint set K is updated (lines 5-12). In line 5, go through the neighborhoods in decreasing order based on $p(y \in H_i), i \in \{1, \ldots m\}$i.e., the probability of y* belonging to each neighborhood, which is assumed to be known. This query order will allow us to determine the neighborhood of y* with the smallest number of queries. This process is repeated until reach the maximum number of queries allowed (line 13).

Therefore, need to discard these reiterative redundant/useless labelled samples, while keeping unused labeled samples to avoid losing acquired information. According to the above analysis, combine ILP and DLP into a single entity, i.e., ID-LP. The proposed algorithm is based on the ID-LP, and is called ID-AP. Algorithm 2 illustrates the proposed ID-AP technique. It should be noted that the purpose

of the DLP is to reduce/remove the effects of useless labeled exemplars by considering the unlabeled data points in the process.

Algorithm 2: Proposed Incremental and Decremental Affinity Propagation

X: data set obtained from neighbohood framework, L: labeled set U: unlabeled set

Initialization:

Consider the labeled samples set and unlabeled data set to be L_t and U_t, respectively. Let $X = L_t \cup U_t$ and the total number of the data points $N = l_t + u_t$, where t is the iterative times.

Steps

Step 1: Calculate the similarity matrix using the (1) [11] for the data set, and then adjust the similarity matrix according to L_t.

Let *X* be the data set and *s* be the similarity set. Let $s(x_i,x_j)$ be the similarity between data points x_i and x_j , i.e., the suitability of data point $x_i$to serve as the exemplar for data point x_j. In conventional AP, a common choice for similarity is the negative Euclidean distance

$$s = (x_i,x_j) = -\|x_i - x_j\|^2 \quad (1)$$

Step 2: Update availability and responsibility according to (2)–(5);

$$a(x_i,x_j) = \begin{cases} \sum_{k\neq j}\max\{0,r(x_k,x_j)\} & i = j \\ \{\min\{0,r(x_i,x_j) + \sum_{k\neq j,i}\max\{0,r(x_k,x_j)\}\} & i \neq j \end{cases} \quad (2)$$

$$r(x_i,x_j) = s(x_i,x_j) - \max_{k\neq j}\{s(x_i,x_k) + a(x_i,x_k)\} \quad (3)$$

The two kinds of messages have an intuitive interpretation. The responsibility indicates how appropriate that candidate exemplar would be as a cluster exemplar. The availability indicates how well-suited the data point would be as a member of the cluster of candidate exemplars. When AP converges, the exemplars are obtained by calculating the set of positive $a(x_i, x_i) + r(x_i, x_i)$ messages for each x_i. Let C be the set of exemplars, non-exemplars are assigned to their respective exemplars according to the following rule:

$$\max_{x_j \in C}\{a(x_i,x_j) + r(x_i,x_j)\} \quad (4)$$

Therefore, the two kinds of messages could be damped according to the following equations:

$$R^{t+1} = \alpha R^{t-1} + (1-\alpha)R^t \quad (5)$$

$$A^{t+1} = \alpha A^{t-1} + (1-\alpha)A^t$$

where R and A represent responsibility and availability vectors, respectively; α is the factor of damping, which should satisfy $0.5 \le \alpha < 1$; t is the number of iterations. Higher values of α will lead to slower convergence.

Step 3: Identify cluster exemplars by the maximum value of the availabilities and responsibilities;

Step 4: Repeat Steps 1–3 until the decisions for cluster exemplars are unchanged for some number of iterations. Then record the temporary cluster exemplars.

Incremental Learning for Unlabeled Data Selection

Step 5: Calculate labeling function using (6).

- ✓ Select the new labeled samples set $V_t = \{v_1, v_2, \ldots, v_m\}$ from U_t according to the labeling function, where *m* is the number of the selected new labeled samples.

$$Lab(V) \leftarrow \begin{cases} if\ x_i = \arg\max_{x_j}\{a(x_j,x_k) + r(x_j,x_k)\}1 \le k \le N \ \ and\ x_i \in C \\ if\ x_i = \arg\max\{a(x_i,x_k) + r(x_i,x_k)\}1 \le k \le N \ \ and\ x_i \notin C \end{cases} \quad (6)$$

where V is the new labeled sample set picked from U according to the labeling function.

Step 6: $L_{t+1} = L_t \cup V_t, l_{t+1} = l_t + m$

Step 7: $U_{t+1} = U_t \backslash V_t, u_{t+1} = u_t - m$

Decremental Learning for Discard Labeled Samples:

Step 8: Discard the useless labeled samples.

- ✓ Set n = 0 as the number of the useless labeled samples. For each labeled sample $x_i(1 \le i \le l_{t+1}), if\ x_i \in C \&\& x_i \neq L_o, then\ L_{t+1} = L_{t+1}\backslash x_i, nn + 1.$

Step 9: Reset the data set

$$X = L_{t+1} \cup U_{t+1}, N = N - n.$$

Step 10: t = t + 1.

Identify the Cluster Exemplar of Each Data Point

Repeat Steps 1–10 until no unlabeled data points left in U_t, and then record the final cluster exemplars.

Hence compared with the conventional semisupervised clustering methods, the proposed algorithm makes it possible to learn the most important information to avoid the learning bias, while discarding the useless knowledge to avoid stability-plasticity dilemma (i.e., it forgets the previously acquired information which has been fully used to avoid losing any opportunity to learn new information).

2.2. Selection of Most Informative Instance

Given a set of existing neighborhoods, would like to select an instance such that knowing its neighborhood will allow us to gain maximal information about the underlying clustering structure of the data. Our method is based on the following key observation. It can be able to predict with high certainty to which neighborhood an instance belongs based on our current understanding of the clustering structure, querying about that instance will not lead to any gain of information. Similar observations have been used to motivate the widely used uncertainty-based sampling principle for active learning of classifiers [12].

2.2.1. Measuring Uncertainty

In uncertainty-based sampling for supervised learning, an active learner queries the instance about which the label uncertainty is maximized. Numerous studies have investigated different approaches for measuring uncertainty given probabilistic predictions of the class labels [10]. In our context, one can take a similar approach and measure the uncertainty of each data instance belonging to different clusters. Instead, our approach estimates the probability of each instance belonging to each neighborhood using a similarity based approach, where the similarity measure is learned under the supervision of the current clustering solution. This learning-based approach allows us to transfer the knowledge that have learned from the constraints to the similarity measures.

Random forest [13] is an ensemble learning algorithm that learns a collection of decision trees. Each decision tree is trained using a randomly bootstrapped sample of the training set and the test for each node of the tree is selected from a random subset of the features. Given the learned random forest classifier, compute the similarity between a pair of instances by sending them down the decision trees in the random forest and count the number of times they reach the same leaf, normalized by the total number of trees. This will result in a value between 0 and 1, with 0 for no similarity and 1 for maximum similarity. Note that random forest has previously been successfully applied to estimating similarities between unsupervised objects [14]. In that work, a random forest classifier is built to distinguish the observed data from synthetically generated data, whereas our work builds the random forest classifier to distinguish the different clusters.

Estimation of Neighborhood Probability

Let S denotes the similarity matrix generated by previous steps, let S(yi,yj) denotes the similarity between instance yi and instance yj. For any unconstrained data point y, assume that its Probability of belonging to a neighborhood Hi to be proportional to the average similarity between y and the instances in Hi. More formally, estimate the probability of an instance y belonging to neighborhood Hi,

$$p(y \in H_i) = \frac{\frac{1}{|H_i|}\sum_{y_j \in H_i} S(y,y_j)}{\sum_{p=1}^{m}\frac{1}{|H_p|}\sum_{y_j \in H_p} S(y,y_j)} \quad (7)$$

where $|H_i|$ indicates the number of instances in neighborhood H_i, and m is the total number of existing neighborhoods. Note that in the early stages of our algorithm, when all neighborhoods are small, it is possible for an unconstrained data point y to have zero average similarity with every neighborhood. In such cases, assign equal probabilities to all neighborhoods for y. This will essentially treat instance x as highly uncertain, making it a good candidate to be selected by our algorithm. This behaviour is reasonable because it will encourage the discovery of more neighborhoods in early stages. Finally, measure the uncertainty of an instance by the entropy of its neighborhood membership, which denote as $E(H|y)$

$$E(H|y) = -\sum_{i=1}^{m} p(y \in H_i)\log_2 p(y \in H_i)$$

Where m is the total number of existing neighborhoods.

To demonstrate the effectiveness of the proposed method, first compare its performance to a set of competing methods, including a random policy, the Min-Max approach introduced by Mallapragada et al., and a variant of Huang's method [11] to make it applicable to non-document data types.

III. Experimental Results

This section presents the experimental result, which compares our proposed method to the baseline methods. In the experimentation, eight benchmark UCI data sets are taken, that have been used in previous studies on constraint based clustering. Out data sets include breast, pen-based recognition of handwritten digits (3, 8, 9), ecoli, glass identification, statlog-heart, parkinsons, statlog image segmentation, and wine. For the ecoli data set, removed the smallest three classes, which only contain 2, 2, and, 5 instances, respectively. The characteristics of the eight data sets are shown in Table 1.

TABLE I
CHARACTERISTICS OF THE DATA SETS

Datasets	# of classes	# of features	# of Examples	Datasets
Breast	2	9	683	Breast
Digits-389	3	16	3165	Digits-389
Ecoli	5	7	327	Ecoli
Glass	6	9	214	Glass
Heart	2	13	270	Heart
Parkinsons	2	22	195	Parkinsons
Segment	7	19	2310	Segment
Wine	3	13	178	Wine

IV. Evaluation Criteria

Two evaluation criteria are used in our experiments. First, use normalized mutual information (NMI) to evaluate the clustering assignments against the ground-truth class labels. NMI considers both the class label and clustering assignment as random variables, and measures the mutual information between the two random variables, and normalizes it to a zero-to-one range. In general, let C be the random variable representing the cluster assignments of instances, and K be the random variable representing the class labels of the instances, the NMI is computed by the following equation:

$$NMI = \frac{2I(C;K)}{H(C) + H(K)}$$

Where $I(X;Y) = H(X) - H(X|Y)$ is the mutual information between random variables X and Y . $H(X)$ is the entropy of X, and $H(X|Y)$ is the conditional entropy X given Y .

Second, consider F-measure as another criterion to evaluate how well the pairwise relationship between each pair of instances is predicted which is compared based on the relationship defined by the ground-truth class labels. F-measure is defined as the harmonic mean of precision and recall, which are computed by the following equations:

$$Precision = \frac{\#\ Pairs\ correctly\ predicted\ in\ same\ cluster}{\#\ Total\ pairs\ predicted\ in\ same\ cluster}$$

$$Recall = \frac{\#\ Pairs\ correctly\ predicted\ in\ same\ cluster}{\#\ Total\ pairs\ actually\ in\ same\ cluster}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

From Fig. 1, it can see that the constraints selected by NPU generally leads to clustering results that are more consistent with the underlying class labels, as can be seen by the dominating curve of NPU compared to other baseline curves. It is interesting to note that random actually degrades the performance in some data sets as we include more constraints, namely the breast, heart, and wine data sets. Previous studies on semi-supervised clustering have reported similar results, where randomly selected constraints actually degrade the clustering performance for some data sets.

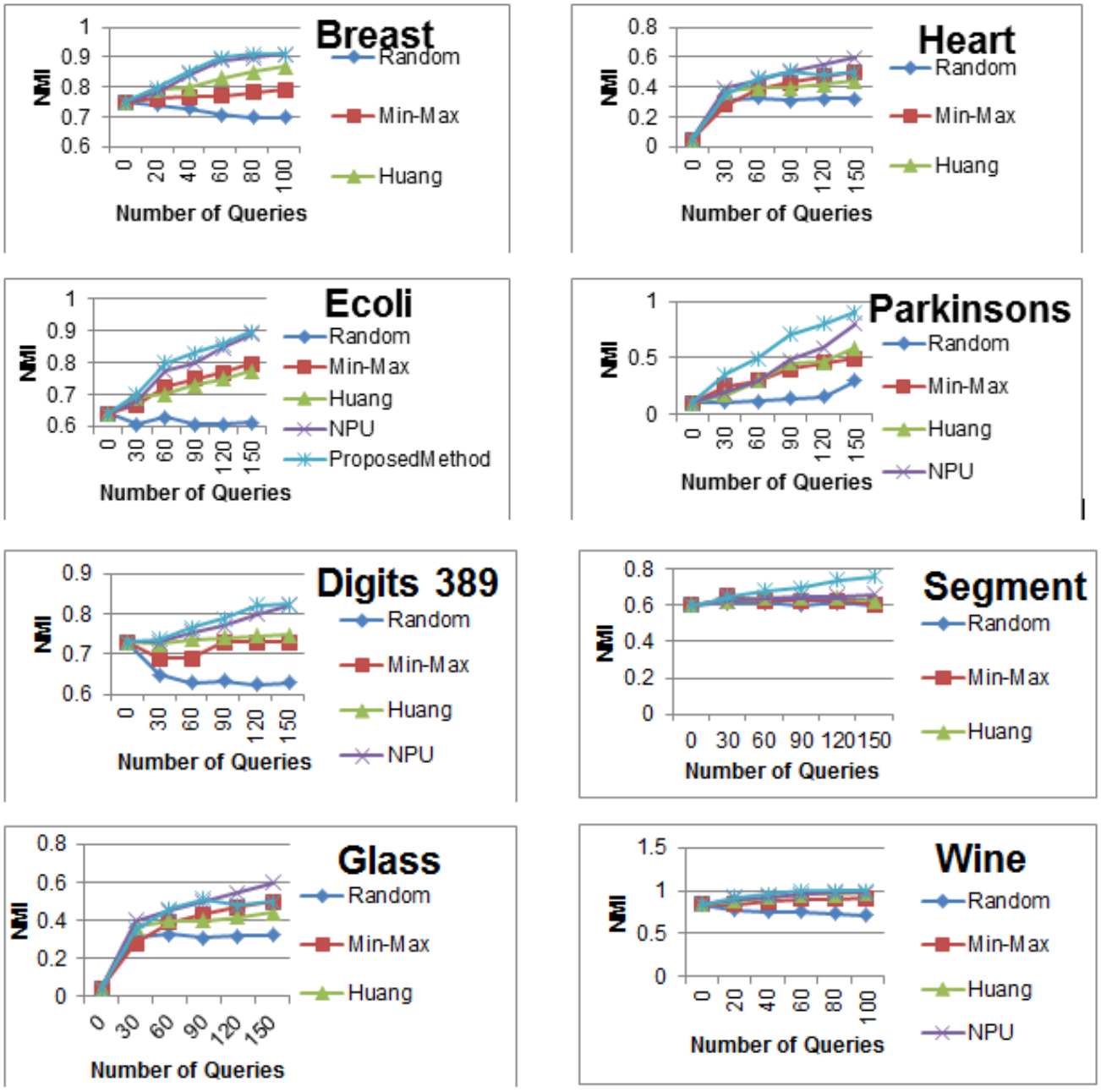

Fig. 1: The NMI values of different methods on eight data sets as a function of the number of pairwise queries (mean and the confidence interval of t-test at 95 percent significance level).

TABLE II
COMPARISON ON F-MEASURE

Data	Algorithm	Number of Queries				
		20	40	60	80	100
Breast	Random	0.927	0.924	0.921	0.916	0.918
	Min max	0.934	0.937	0.939	0.941	0.946
	Huang	0.941	0.951	0.957	0.963	0.967
	NPU	0.943	0.959	0.972	0.976	0.978
	Proposed approach	0.946	0.963	0.976	0.982	0.983
Digits-389	Random	0.762	0.774	0.749	0.752	0.752
	Min max	0.805	0.788	0.797	0.842	0.842
	Huang	0.814	0.826	0.842	0.851	0.853
	NPU	0.808	0.848	0.857	0.870	0.883
	Proposed approach	0.810	0.852	0.860	0.874	0.886
Ecoli	Random	0.642	0.628	0.700	0.653	0.659
	Min max	0.648	0.779	0.836	0.851	0.858
	Huang	0.687	0.762	0.801	0.833	0.829
	NPU	0.673	0.798	0.858	0.879	0.900
	Proposed approach	0.679	0.804	0.865	0.885	0.904
Glass	Random	0.440	0.403	0.410	0.410	0.413
	Min max	0.432	0.418	0.463	0.484	0.493
	Huang	0.480	0.481	0.476	0.474	0.473
	NPU	0.493	0.492	0.481	0.496	0.495
	Proposed approach	0.498	0.499	0.486	0.499	0.499
Heart	Random	0.659	0.636	0.612	0.598	0.587
	Min max	0.700	0.726	0.743	0.760	0.790
	Huang	0.680	0.682	0.709	0.744	0.789
	NPU	0.682	0.725	0.766	0.812	0.845
	Proposed approach	0.686	0.730	0.771	0.816	0.850
Parkinsons	Random	0.594	0.607	0.633	0.637	0.682
	Min max	0.593	0.615	0.666	0.705	0.747
	Huang	0.593	0.605	0.652	0.694	0.736
	NPU	0.597	0.637	0.695	0.759	0.814
	Proposed approach	0.601	0.641	0.698	0.764	0.819
Segment	Random	0.546	0.553	0.552	0.548	0.549
	Min max	0.571	0.582	0.582	0.566	0.569
	Huang	0.567	0.576	0.576	0.573	0.575
	NPU	0.565	0.579	0.579	0.585	0.587
	Proposed approach	0.569	0.585	0.582	0.589	0.592
Wine	Random	0.871	0.853	0.836	0.843	0.827
	Min max	0.909	0.935	0.945	0.953	0.959
	Huang	0.931	0.964	0.982	0.988	0.994
	NPU	0.945	0.992	1.00	1.00	1.000
	Proposed approach	0.950	0.996	1.00	1.00	1.00

V. CONCLUSION

In this paper, a semisupervised clustering technique with incremental and decremental affinity propagation (ID-AP) that build with labeled exemplars into the AP algorithm and a novel method for actively selecting informative constraints to offer improved clustering performance. The iterative framework requires repeats reclustering of the data with an incrementally growing constraint set. This can be computationally demanding for large data sets. To address this problem, it would be interesting to consider an incremental semi-supervised clustering method that updates the existing clustering solution based on the neighborhood assignment for the new point. An alternative way to lower the computational cost is to reduce the number of iterations by applying a batch approach that selects a set of points to query in each iteration. A active learning approach would be to select the top k points that have the highest normalized uncertainty to query their neighborhoods. However, such a strategy will typically select highly redundant points. The experimental results are evaluated to predict the performance of the proposed method based on eight benchmark data sets against a number of competing methods. The evaluation results indicate that the proposed method achieves consistent and substantial improvements over its competitors.

REFERENCES

[1] S. Basu, M. Bilenko, and R.J. Mooney, 'A probabilistic framework for semi-supervised clustering', *in SIGKDD*, pp. 59–68. ACM, (2004).

[2] Ling Chen and Chengqi Zhang, 'Semi-supervised variable weighting for clustering', *in SDM*, pp. 862–871, (2011)

[3] S.C.H. Hoi, R. Jin, and M.R. Lyu, 'Learning nonparametric kernel matrices from pairwise constraints', *in ICML*, pp. 361–368. ACM, (2007)
[4] Davidson, K. Wagstaff, and S. Basu, 'Measuring constraint-set utility for partitional clustering algorithms', *PKDD*, 115–126, (2006)
[5] X. Wang and I. Davidson, 'Active Spectral Clustering', *in ICDM*, (2010).
[6] Q. Xu, M. Desjardins, and K. Wagstaff, *Active constrained clustering by examining spectral eigenvectors*, in Discovery Science, pp. 294–307. Springer, (2005)
[7] S.C.H. Hoi and R. Jin, 'Active kernel learning', *in ICML*, pp. 400–407. ACM, (2008)
[8] Y Fu, B LI, X Zhu, and C Zhang, 'Do they belong to the same class: active learning by querying pairwise label homogeneity', *in CIKM*, pp. 2161–2164. ACM, (2011)
[9] S.J. Huang, R. Jin, and Z.H. Zhou, 'Active learning by querying informative and representative examples'. NIPS, (2010).
[10] B. Settles, "Active Learning Literature Survey," technical report, 2010.
[11] Sicheng Xiong, Javad Azimi, and Xiaoli Z. Fern, "Active Learning of Constraints for Semi-Supervised Clustering", *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, January 2014
[12] R. Huang and W. Lam, "Semi-Supervised Document Clustering via Active Learning with Pairwise Constraints," *Proc. Int'l Conf. Date Mining*, pp. 517-522, 2007.
[13] L. Breiman, "Random Forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.
[14] T. Shi and S. Horvath, "Unsupervised Learning with Random Forest Predictors," *J. Computational and Graphical Statistics*, vol. 15, pp. 118-138, 2006