IJRSET JANUARY 2017 Volume 4, Issue 1 Pages: 17-26

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

SEMI SUPERVISED CLUSTERING ENSEMBLE FOR HIGH DIMENSIONAL DATA

[1] M. Pavithra, [2] Dr. R. M. S. Parvathi
[1] Assistant Professor, [2] Dean- PG Studies,
[1, 2] Department of C.S.E,
[1] Jansons Institute of Technology, [2] Sri Ramakrishna Institute of Technology,
[1, 2] Coimbatore, India.

ABSTRACT: Cluster analysis methods seek to partition a data set into homogeneous subgroups. It is useful in a wide variety of applications, including document processing and modern genetics. Conventional clustering methods are unsupervised, meaning that there is no outcome variable nor is anything known about the relationship between the observations in the data set. In many situations, however, information about the clusters is available in addition to the values of the features. For example, the cluster labels of some observations may be known, or certain observations may be known to belong to the same cluster. Inother cases, one may wish to identify clusters that are associated with a particular outcome variable. This review describes several clustering algorithms (known as "semi-supervised clustering" methods) that can be applied in these situations. The objective of cluster analysis is to partition a data set into a group of subsets (i.e. "clusters") such that observations within a cluster are more similar to one another than observations in other clusters. For a more detailed discussion, see Hastie et al or Gordon. Traditional clustering methods are unsupervised, meaning that there is no outcome measure and nothing is known about the relationship between the observations in the data set. However, in many situations one may wish to perform cluster analysis even though an outcome variable exists or some preliminary information about the clusters is known.

KEYWORDS: [Semi Supervised clustering, Cluster Ensemble, High Dimensional Data,Similarity-based Methods, Search-based Methods, Partially Labeled Data.]

1. INTRODUCTION

1. a. SEMI-SUPERVISED CLUSTERING METHODS

We will now briefly outline several semi-supervised clustering methods. These methods will be organized according to the nature of the known outcome data. First, we will consider the simplest case, namely the case where the data is partially labeled. In other words, the cluster assignments are known for some subset of the observations [3]. We will then consider the case where some sort of relationship between the features is known, and finally the case where one seeks to identify clusters associated with a particular outcome variable. In addition to the similarity information used by unsupervised clustering,in many cases a small amount of

knowledge is available concerning eitherpairwise (must-link or cannot-link) constraints between data items or classlabels for some items [5]. Instead of simply using this knowledge for the externalvalidation of the results of clustering, one can imagine letting it "guide" or"adjust" the clustering process, i.e. provide a limited form of supervision.The resulting approach is called semi-supervised clustering [6]. We also considerthat the available knowledge is too far from being representative of a targetclassification of the items, so that supervised learning is not possible, evenin a transductive form.Note that class labels can always be translated into pairwise constraintsfor the labeled data items and, reciprocally, by using consistent pairwiseconstraints for some items one can obtain groups of items that should belongto a same cluster [8].

1. b. PARTIALLY LABELED DATA

In some situations, the cluster assignments may be known for some subset of the data. The objective is to classify the unlabeled observations in the data to the appropriate clusters using the known cluster assignments for this subset of the data [9]. This problem is equivalent to a supervised classification problem, where the objective is to develop a model to assign observations in a data set to oneof a finite set of classes based on a training set where the true class labels are known. However, traditional supervised classification methods may be inefficient when only a small subset of the data is labeled [11].

2. RELATED WORK

2 a. KNOWN CONSTRAINTS ON THE OBSERVATIONS

We now consider clustering when more complex relationships among the observationsare known. In particular, we will consider two types of possible constraints amongobservations: "Must-link constraints" require that two observations must be placed inthe same cluster, and "cannot-link constraints" require that two observations must notbe placed in the same cluster [2]. One possible application is when repeated measurementsare collected on some subset of the experimental units. In such a situation, one maywant to assign all of the repeated measurements of the same experimental unit to thesame cluster [6].Note that this is a generalization of the problem considered in the previous section,where the cluster assignments are known for a subset of the features. In that situation,for each feature j that is known to belong to cluster k, one may impose a must-linkconstraint between feature j and all other features known to belong to cluster k and acannot-link constraint between feature j and features known not to belong to cluster k.

1. Randomly assign each observation to an initial cluster.
2. For each feature j and cluster k, calculate xkj, the mean of feature j in cluster k.
3. Assign each observation i to a new cluster Ci as follows:

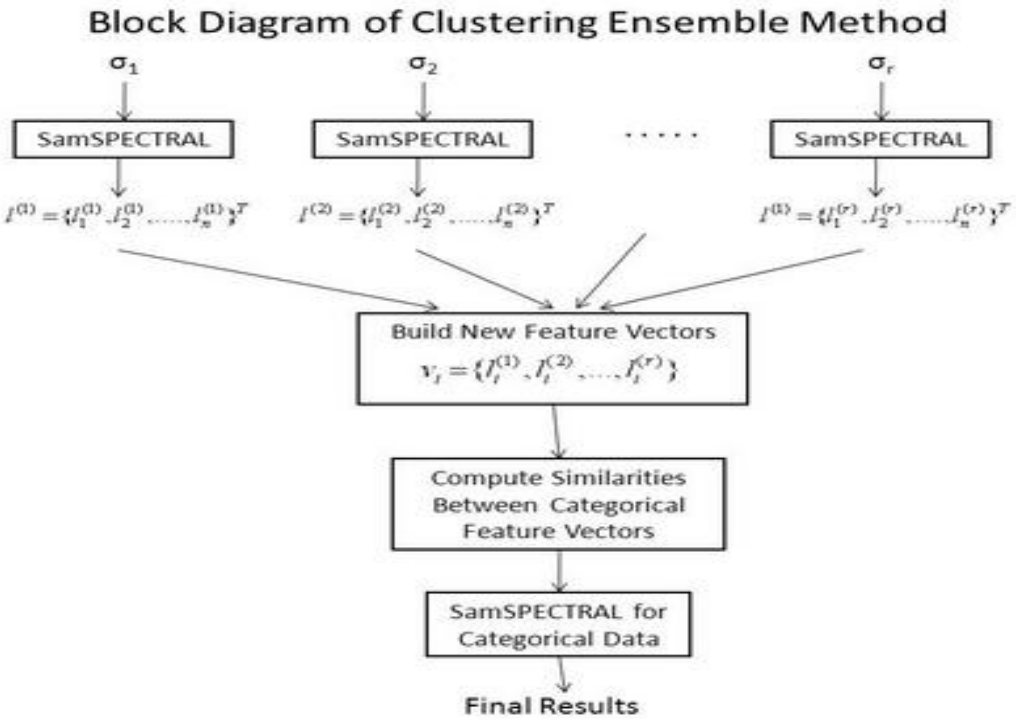
$$C_i = \arg\min_{k \in D_{ik}} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

WhereDik = {k: no constraints are violated when observation i is assigned to cluster k}. 4. Repeat steps 2 and 3 until the algorithm converges. The algorithm fails if Dik =∅for any i at any step of the procedure.

2. b. SEARCH-BASED METHODS

In search-based approaches, the clustering algorithm itself is modified so that user-provided labels or constraints are used to bias the search for an appropriate partitioning [2]. This can be done by several methods, e.g., modifying the clustering objective function so that it includes a term for satisfying specified constraints, enforcing constraints to be satisfied during the cluster assignment in the clustering process, doing clustering using side-information from conditional distributions in an auxiliary space, and initializing clusters and inferring clustering

constraints based on neighborhoods derived from labeled examples [5].



2. c. SIMILARITY-BASED METHODS

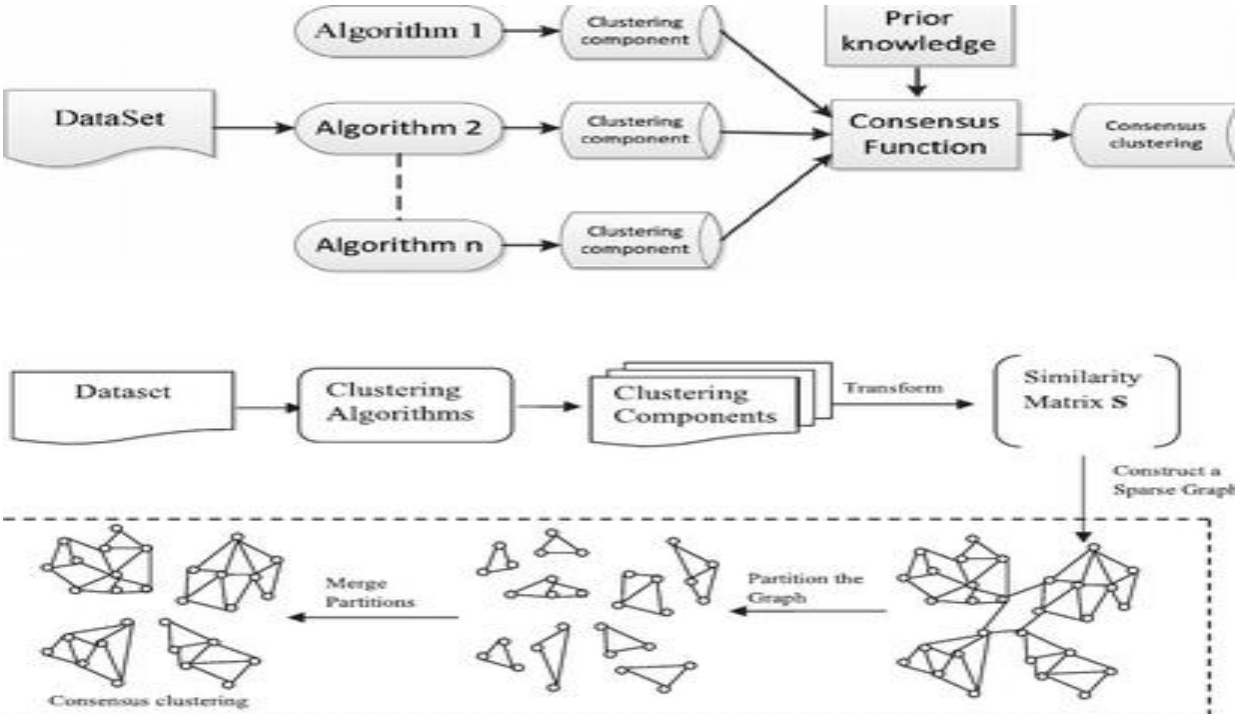
In similarity-based approaches, an existing clustering algorithm that uses a similarity metric is employed; however, thesimilarity metric is first trained to satisfy the labels or constraints in the supervised data. Several similarity metrics have been used for similarity-based semi-supervised clustering, including string-edit distance [7].

2d. GOAL OF PROPOSED SEMI SUPERVISED CLUSTERING

In the proposed thesis, the main goal is to study semi-supervised clustering algorithms, characterize some of theirproperties and apply them to different domains [9]. In our completed work, we have already shown how supervisioncan be provided to clustering in the form of labeled data points or pairwise constraints. We have also developedan active learning framework for selecting informative constraints in the pairwise constrained semi-supervised clusteringmodel, and proposed a method for unifying search-based and similarity-based techniques in semi-supervisedclustering [11]. Investigate the effects of noisy supervision, probabilistic supervision (e.g., soft constraints) or incomplete supervision (e.g., labels not specified for all clusters) in clustering;

- Study model selection issues in semi-supervised clustering, which will help to characterize the difference betweensemi-supervised clustering and classification;
- Study the feasibility of semi-supervising other clustering algorithms, especially in the discriminative clusteringor online clustering framework;
- Create a framework for ensemble semi-supervised clustering;
- Apply the semi-supervised clustering model on other domains apart from text, especially web search engines,astronomy and bioinformatics;
- Study the relation between different evaluation metrics used to evaluate semi-supervised clustering;

- Investigate other forms of semi-supervision, e.g., attribute-level constraints;
- Do more theoretical analysis of certain aspects of semi-supervision, especially semi-supervised clustering withlabeled data and the unified semi-supervised clustering model.



3. PROPOSED WORK

3. a. SEMI-SUPERVISED CLUSTERING USING LABELED DATA

In this section, we give an outline of our initial work where we considered a scenario where supervision is incorporatedinto clustering in the form of labeled data [4]. We used the labeled data to generate seed clusters that initialize a clusteringalgorithm, and used constraints generated from the labeled data to guide the clustering process. The underlyingintuition is that proper seeding biases clustering towards a good region of the search space, thereby reducing thechances of it getting stuck in poor local optima, while simultaneously producing a clustering similar to the userspecifiedlabels.

3. b. ACTIVE LEARNING FOR SEMI-SUPERVISED CLUSTERING

In order to maximize the utility of the limited supervised data available in a semi-supervised setting, supervisedtraining examples should be activelyselected as maximally informative ones rather than chosen at random, if possible[6]. In the PCC framework, this would imply that fewer constraints will be required tosignificantly improve the clustering accuracy. To this end, we developed a new method for actively selecting goodpairwise constraints for semi-supervised clustering in the PCC framework.

3. c. UNIFIED MODEL OF SEMI-SUPERVISED CLUSTERING

In previous work, similarity-based and search-based approaches to semi-supervised clustering have not been adequatelycompared experimentally, so their relative strengths and weaknesses are largely unknown. Also, the two approachesare not incompatible; therefore, applying a search-based approach with a trained similarity metric is clearlyan additional option which may have advantages over both existing approaches [8]. In this work, we presented a newunified semi-

supervised clustering algorithm derived from Means that incorporates bothmetric learning and usinglabeled data as seeds and/or constraints.

3. d. INCOMPLETE SEMI-SUPERVISION AND CLASS DISCOVERY

In semi-supervised classification, all classes are assumed to be known apriori and labeled training data is providedfor all classes. In labeled semi-supervised clustering, when we consider clustering a dataset that has an underlyingclass labeling, we would like to consider incomplete seeding where labeled data are not provided for every underlyingclass [10]. For such incomplete semi-supervision, we would like to see if the labels on some classes can help the clusteringalgorithm discover the unknown classes.

3. e. GENERALIZING THE UNIFIED SEMI-SUPERVISED CLUSTERING MODEL

We proposed a framework for unifying search-based and similarity-based semi-supervised clusteringthat works only with Euclidean Means. I am working with Misha Bilenko, who is formulating an effective metriclearning algorithm in high dimensions, on generalizing our unified PCC framework to work withSPKMeans so that we can apply it to domains like text [11].

3. f. OTHER SEMI-SUPERVISED CLUSTERING ALGORITHMS

We can proceed and run the usual HAC algorithm on the data points using this modified similarity metric, sothat at each cluster-merge step, we consider the similarity between the data points as well as the cost of constraintviolation incurred during the merge operation [12].

A more interesting problem would be when the initial supervision is given in the form of a hierarchy, andthe clustering problem will be to do hierarchical clustering “using” the initial hierarchy. We want to formalize thenotion of using an initial seed hierarchy for hierarchical clustering [13].

Such an approach would be useful for contentmanagement applications, e.g., if the requirement is to hierarchically cluster the documents of a company, and theinitial seed hierarchy is a preliminary directory structure containing a subset of the documents.So far we have mainly focused on clustering algorithms that use a generative model [4]. We also want toapply the pairwise constrained framework to discriminative clustering algorithms (e.g., graph partitioning), for whichpairwise constraints are a natural way for providing constraints. Another interesting research direction would be onlineclustering in the semi-supervised framework.

3. g. ENSEMBLE SEMI-SUPERVISED CLUSTERING

In our work so far, we have assumed constraints to be noise-free. We have also assumed the weights on the constraintsto be uniform (PCKMeans) or changed the weights based on the “difficulty of satisfying the constraints” (unifiedmodel). An interesting problem in the PCC model would be the choice of the constraint weights in the general case ofnoisy constraints [5].

Each PCC clustered can be considered as a weak learner taking pairwise data points as input, and givingan binary output decision of “same-cluster” or “different-cluster”. The must-link and cannot-link constraints can beconsidered as the training data for each weak learner. Given a set of input constraints, the PCC clustered initially setsall constraints to have uniform weight and performs clustering. After clustering is completed, the clustered categorizeseach pair of points as “same-cluster” or “different-cluster”, based on whether the pair ended up in the same cluster or in different clusters [8].

Since the given constraints are noisy, some of them will be violated by the clustering. Theconstraints are reweighted based on the number of errors made by the weak learner,

and a new clustered is created toperform the clustering with the new weights on the constraints. We use boosting for re-weighting of the constraintsand combining the outputs of the clusters in the ensemble [7].

3. h. APPROXIMATION ALGORITHMS FOR SEMI-SUPERVISED CLUSTERING

Another interesting research direction is considering how semi-supervision affects approximation algorithms for someclustering methods, e.g., KMedian. The KMedian problem, which was explained briefly in similarto the facility location problem [3]. In the facility location problem, we are given a set of demand points and a set ofcandidate facility sites with costs of building facilities at each of them. Each demand point is then assigned to itsclosest facility, incurring a service cost equal to the distance to its assigned facility. The goal is to select a subset ofsites where facilities should be built, so that the sum of facility costs and the service costs for the demand points isminimized [6]. The KMedian problem is similar to facility location, but with a few differences in KMedian there areno facility costs and there is bound.We proposea semi-supervised extension to KMedian to handle constraints on the demand points. Theconstrained KMedian problem would be additionally given an input set of must-link and cannot-link constraints on thedemand points (i.e., two demand points should be or should not be assigned to the same facility), and the goal wouldbe to minimize an objective function that is the sum of the service costs for the demand points and the cost of violatingthe constraints [10].

4. EXPERIMENTS

In this section, we empirically demonstrate that our proposed semi-supervised clustering algorithm is both e cient and e ective.

4. a. DATASETS

Four real-world benchmark datasets with varied sizes are used in our experiments, which are:

• Heart Diseases, a dataset containing 20 objects with 1,440 images in total. Each image is represented by a 1024-dimensional vector.

• Diabetics, a widely used handwritten digits dataset including 9,298 handwritten images. Each image is represented by a 256-dimensional vector that belongs to one of 10 classes.

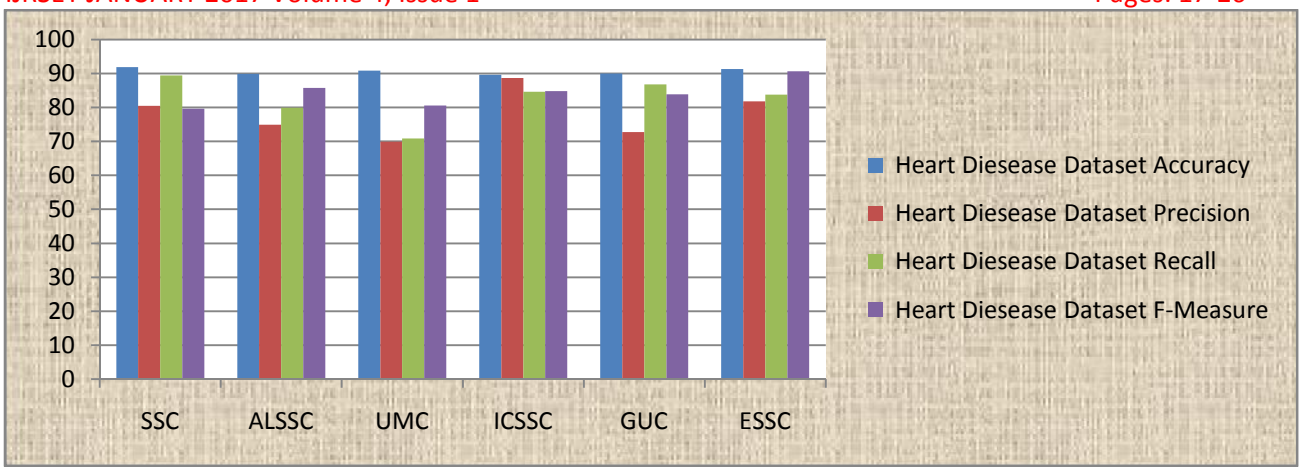
• Cancer, a dataset used to predict forest covers types using cartographic variables. This dataset consists of 581,012 records belonging to seven cover type classes, i.e., spruce/fir, lodge pole pine, ponderosa pine, cottonwood/willow, aspen, Douglas-fir, and krummholz.

• Liver Diseases, a dataset artificially enlarged from the MNIST handwritten digits dataset6. It contains a total of 8,100,000 samples that belong to 10 classes.

5. EXPERIMENTAL RESULTS

5. a. HEART DISEASE DATASET

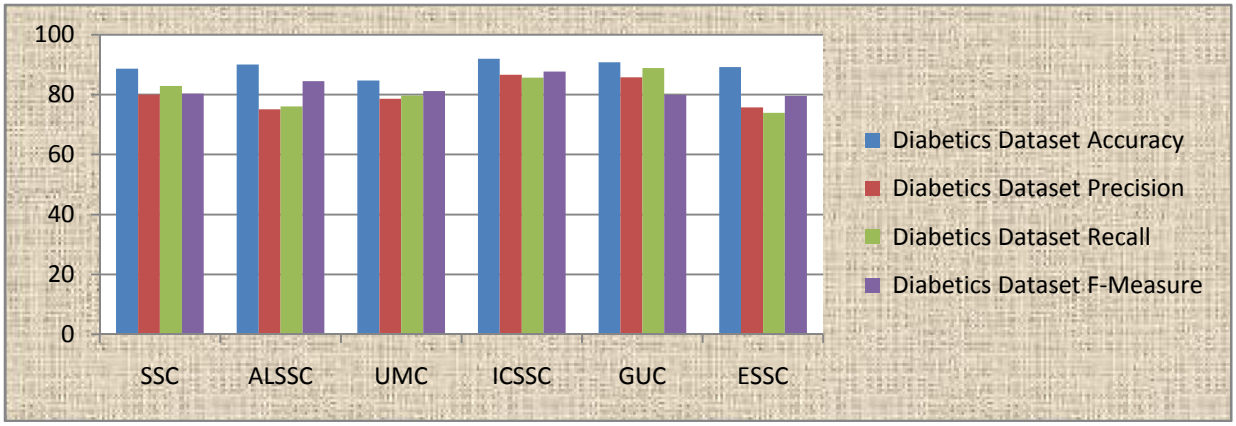
	Heart Disease Dataset			
Algorithm	Accuracy	Precision	Recall	F-Measure
SSC	91.91	80.45	89.45	79.68
ALSSC	89.93	74.90	79.90	85.78
UMC	90.92	69.90	70.90	80.56
ICSSC	89.67	88.67	84.67	84.78
GUC	90.00	72.78	86.78	83.90
ESSC	91.33	81.78	83.78	90.67



The above graph shows that performance of Heart Disease dataset. The Accuracy of SSC algorithm is 91.91 which is higher when compare to other five (ALSSC, UMC, ICSSC, GUC, ESSC) algorithms. The Precision of ICSSC algorithm is 88.67 which is higher when compare to other five (ALSSC, UMC, SSC, GUC, ESSC) algorithms. The Recall of SSC algorithm is 89.45 which is higher when compare to other five (ALSSC, UMC, ICSSC, GUC, ESSC) algorithms. The F-Measure of ESSC algorithm is 90.67 which is higher when compare to other five (ALSSC, UMC, ICSSC, GUC, SSC) algorithms.

5. b. DIABETICS DATASET

	Diabetics Dataset			
Algorithm	Accuracy	Precision	Recall	F-Measure
SSC	88.67	79.98	82.9	80.34
ALSSC	90.11	75.08	76.08	84.56
UMC	84.67	78.67	79.67	81.23
ICSSC	92.01	86.67	85.67	87.67
GUC	90.78	85.78	88.88	79.89
ESSC	89.21	75.78	73.98	79.56

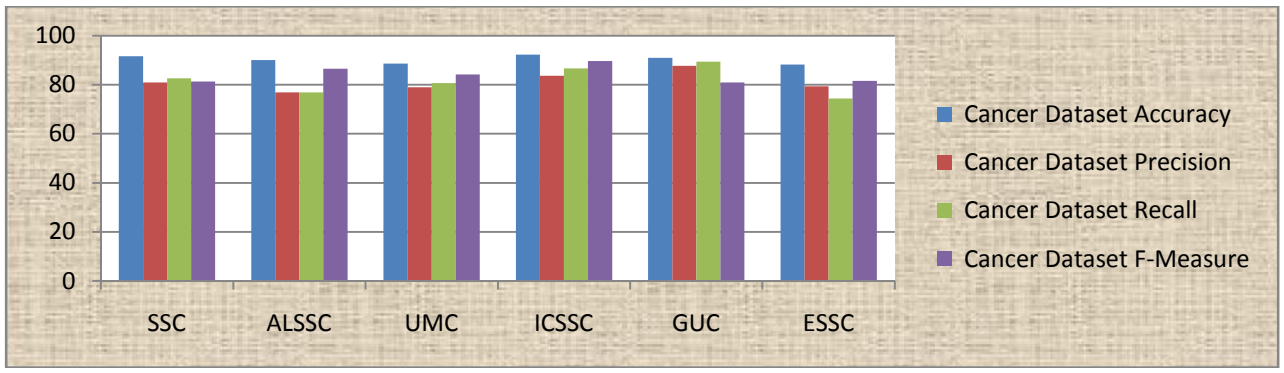


The above graph shows that performance of Diabetics dataset. The Accuracy of ICSSC algorithm is 92.01 which is higher when compare to other five (ALSSC, UMC, GUC, SSC, ESSC) algorithms. The Precision of ICSSC algorithm is 86.67 which is higher

when compare to other five (ALSSC, UMC, SSC, GUC, ESSC) algorithms. The Recall of GUC algorithm is 88.88 which is higher when compare to other five (ALSSC, UMC, ICSSC, SSC, ESSC) algorithms. The F-Measure of ICSSC algorithm is 87.67 which is higher when compare to other five (ESSC, UMC, ALSSC, GUC, SSC) algorithms.

5. c. CANCER DATASET

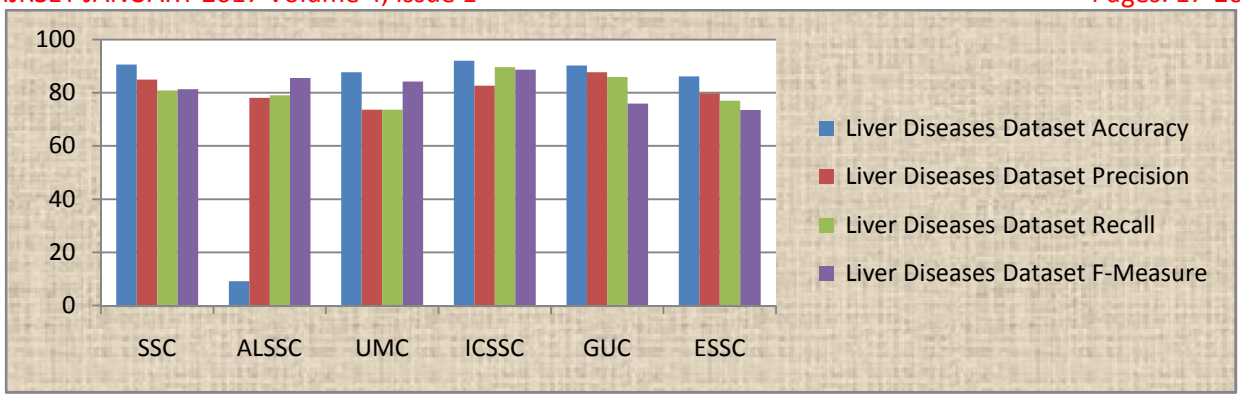
	Cancer Dataset			
Algorithm	Accuracy	Precision	Recall	F-Measure
SSC	91.67	80.98	82.67	81.34
ALSSC	90.11	76.88	76.88	86.56
UMC	88.67	78.97	80.67	84.23
ICSSC	92.33	83.67	86.67	89.67
GUC	90.98	87.78	89.48	80.89
ESSC	88.21	79.38	74.38	81.56



The above graph shows that performance of Cancer dataset. The Accuracy of ICSSC algorithm is 92.33 which is higher when compare to other five (ALSSC, UMC, GUC, SSC, ESSC) algorithms. The Precision of GUC algorithm is 87.78 which is higher when compare to other five (ALSSC, UMC, SSC, ICSSC, ESSC) algorithms. The Recall of GUC algorithm is 89.48 which is higher when compare to other five (ALSSC, UMC, ICSSC, SSC, ESSC) algorithms. The F-Measure of ICSSC algorithm is 89.67 which is higher when compare to other five (ESSC, UMC, ALSSC, GUC, SSC) algorithms.

5. d. LIVER DISEASES DATASET

	Liver Diseases Dataset			
Algorithm	Accuracy	Precision	Recall	F-Measure
SSC	90.67	84.98	80.91	81.34
ALSSC	90.11	78.08	79.08	85.56
UMC	87.67	73.67	73.67	84.23
ICSSC	92.11	82.67	89.67	88.67
GUC	90.28	88.48	85.88	75.89
ESSC	86.21	79.78	76.98	73.56



The above graph shows that performance of Liver Diseases dataset. The Accuracy of ICSSC algorithm is 92.11 which is higher when compare to other five (ALSSC, UMC, GUC, SSC, ESSC) algorithms. The Precision of GUC algorithm is 88.48 which is higher when compare to other five (ALSSC, UMC, SSC, ICSSC, ESSC) algorithms. The Recall of ICSSC algorithm is 89.67 which is higher when compare to other five (ALSSC, UMC, GUC, SSC, ESSC) algorithms. The F-Measure of ICSSC algorithm is 88.67 which is higher when compare to other five (ESSC, UMC, ALSSC, GUC, SSC) algorithms.

CONCLUSION

Our main goal in the proposed thesis is to study search-based semi-supervised clustering algorithms and apply them to different domains. As explained in Chapter 3, our initial work has shown: (1) how supervision can be provided to clustering in the form of labeled data points or pairwise constraints; (2) how informative constraints can be selected in an active learning framework for the pairwise constrained semi-supervised clustering model; and (3) how searchbased and similarity-based techniques can be unified in semi-supervised clustering. In our work so far, we have mainly focused on generative clustering models, e.g. KMeans and EM, and ran experiments on clustering low-dimensional UCI datasets or high-dimensional text datasets. In this thesis, we want to study other aspects of semi-supervised clustering, like: (1) the effect of noisy, probabilistic or incomplete supervision in clustering; (2) model selection techniques for automatic selection of number of clusters in semi-supervised clustering; (3) ensemble semi-supervised clustering. In future, we want to study the effect of semi-supervision on other clustering algorithms, especially in the discriminative clustering and online clustering framework. We also want to study the effectiveness of our semi-supervised clustering algorithms on other domains, e.g., web search engines (clustering of search results), astronomy (clustering of Mars spectral images) and bioinformatics (clustering of gene microarray data).

REFERENCES

[1] X. Fern and C. E. Bradley, “Random projection for high dimensional data clustering: A cluster ensemble approach”, In ICML, pages 186–193, 2013.

[2] X. Fern and C. E. Bradley, “Solving cluster ensemble problems by bipartite graph partitioning”, In ICML, 2010.

[3] Lucas Franek and Xiaoyi Jiang,”Ensemble clustering by means of clustering embedding in vector spaces”, Pattern Recognition, 47(2):833–842, 2014.

[4] W. Liu, J. Wang, and S. Chang,”Robust and scalable graph-based semi supervised learning”, Proceedings of the IEEE, 100(9):2624–2638, 2012.

[5] Z. Lu and T. K. Leen, “Semi-supervised learning with penalized probabilistic clustering”, In NIPS, 2009.

[6] Alexander P. Topchy, Anil K. Jain, and William F. Punch, “A mixture model for clustering ensembles”, In SDM, 2009.
[7] C. Blake and C. Merz, “Uci repository of machine learning databases,” 2011.
[8] A. Strehl and J. Ghosh, ”Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions”, Journal of Machine Learning Research 3, pp. 583-617, 2012.
[9] S. Basu, M. Bilenko, and R. J. Mooney, ”A probabilistic framework for semi-supervised clustering”, In Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-04), pages 59–68, 2014.
[10] Cohn, D., Caruana, R., & McCallum, A, “Semi-supervised clustering with user feedback”, 2010.
[11] Jain, A. K., & Dubes, R. C,”Algorithms for Clustering Data”, Prentice Hall, New Jersey,
Jain, 2009.
[12] A. K., Myrthy, M. N., & Flynn, P. J,” Data clustering: A survey”, ACM Computing Survey, 31(3), 264–323, 2011.
[13] Kaufman, L., & Rousseau, P,”Finding Groups in Data: An Introduction to Cluster Analysis”, John Wiley and Sons, New York, 2011.