ISSN 2278-3091
Volume 6, No.1, January – February 2017
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse01612017.pdf

A NOVEL APPROACH FOR SEMI SUPERVISED CLUSTERING ALGORITHM

ABDELRAHMAN ELSHARIF KARRAR
COLLEGE OF COMPUTER SCIENCE AND ENGINEERING
TAIBAH UNIVERSITY,SAUDI ARABIA
akarrar@taibahu.edu.sa

ABSTRACT

Semi-supervised clustering (SSC) is an important research problem in machine learning. While it is usually expected that the use of unlabelled data can improve performance, in many cases SSL is outperformed by supervised learning using only labelled data. To this end, the construction of a performance-safe SSL method has become a key issue of SSC study. In this paper classified the effect of fast food on human body by clustering with supervised learning and improve the clustering. This paper also use feature selection and feature extraction. Clustering is the technique used for data reduction. It divides the data into groups based on pattern similarities such that each group is abstracted by one or more representatives. Recently, there is a growing emphasis on exploratory analysis of very large datasets to discover useful patterns. This paper explains extracting the useful knowledge represented by clusters from textual information contained in a large number of emails for text and data mining techniques. E-mail data that are now becoming the dominant form of inter and intra organizational written communication for many companies. The sample texts of two mails are verified for data clustering. The cluster shows the similar emails exchanged between the users and finding the text similarities to cluster the texts. In this paper the use of Pattern similarities i.e., the similar words exchanged between the users by considering the different Threshold values are made for the purpose. The threshold value shows the frequency of the words used. The representation of data is done using a vector space model. .The semi-supervised projected model-based clustering algorithm (SeSProC) also includes a novel model selection approach, using a greedy forward search to estimate the final number of clusters. The quality of SeSProC is assessed using synthetic data, demonstrating its effectiveness, under different data conditions, not only at classifying instances with known labels, but also at discovering completely hidden clusters in different subspaces.

KEYWORDS: Data Mining, Clustering, Semi Supervised Clustering, SesProC, SSL, SSC, Data Clustering

1. INTRODUCTION

The topic of semi-supervised clustering has attracted considerable interests among researchers in the data mining and machine learning community [2, 3, 4, and 8]. The goal of semi supervised clustering is to obtain a better partitioning of the data by incorporating background knowledge. Although current semi-supervised algorithms have shown significant improvements over their unsupervised counterparts, they assume that the background knowledge is specified in the same feature space as the unlabelled data [5]. These algorithms are therefore inapplicable when the background knowledge is provided by another source with a different feature space [9]. The key challenge of semi-supervised clustering with partial background knowledge is to determine how to utilize both the shared and non-shared features while performing clustering on the full feature set of the unlabelled data. The semi-supervised algorithm also has to recognize the possibility that the shared features might be useful for identifying certain clusters but not for others [6]. Clustering is generally an unsupervised criterion. With the help of clusters i can collect same items in one cluster i.e. items which have same properties. So in this way i can make several clusters in which each cluster contain items with similar properties. I can classify each cluster with the help of classification. This classification will be based on clustering information derived from class. The scheme which has derived from this is used to forecast heart disease also it produces the efficient classification mechanism related to multidimensional data [8] .Clustering helps to reduce the dimensions to reduce the error in classification. Clustering is the process of partitioning or dividing a set of patterns (data) into groups. Each cluster is abstracted using one or more representatives. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Clustering is

a type of classification imposed on finite set of objects [3]. The relationship between objects is represented in a proximity matrix in which the rows represent „n" e-mails and columns correspond to the terms given as dimensions. If objects are categorized as patterns, or points in a d-dimensional metric space, the proximity measure can be Euclidean distance between a pair of points. Unless a meaningful measure of distance or proximity, between a pair of objects is established, no meaningful cluster analysis is possible. Clustering is useful in many applications like decision making, data mining, text mining, machine learning, grouping, and pattern classification and intrusion detection. Clustering has to be done as it helps in detecting outliners and to examine small size clusters [1]. The proximity matrix is used in this context and thus serves as a useful input to the clustering algorithm. It represents a cluster of n patterns by m points. Typically, m < n leading to data compression, can use centroid [3]. This would help in prototype selection for efficient classification. The clustering algorithms are applied to the training set belonging to two different classes separately to obtain their correspondent cluster representatives. There are different stages in clustering [2].

Clustering is a technique to partition a dataset into homogeneous clusters such that the data points in the same cluster are more similar to each other than in different clusters where classification is to label or classify a new unknown data from a collection of labelled, pre-classified, data. Clustering generally known as unsupervised learning where classification known as supervised learning [1]. The term 'learning' states an algorithm that examines a set of points without examining any corresponding class/cluster label [1, 5, and 7]. In various real and practical applications like bioinformatics, medical, pattern recognition etc., a large amount of unknown data is available than the labelled ones. To generate labelled data become a lengthy and slow process using unsupervised method, also is a tedious work to label all data using supervised method [7]. Therefore, one may wish to use large dataset without labelling or generating data should employ semi-supervised learning. Semi-supervised learning is a technique of learning from a combination of labelled and unlabelled data. This can be used for both classification and clustering purpose [4]. Semi-supervised classification uses labelled data along-with some unlabeled data to train the classifier where semi-supervised clustering, involves some labelled class data or pair wise constraints along with the unlabelled data to obtain better clustering [3]. There are several semi-supervised classification algorithms like co-training, transductive support vector machines (SVMs), Expectation maximization etc. for using unlabelled data to improve classification accuracy. The advantage of semi-supervised clustering is that the data categories (clusters) can generate from initial labeled data as well as extend and modify the existing ones to reflect other regularities in the data [1, 7, 9].

2. RELATED WORK

Clustering [2, 11] an unsupervised technique is the process of organizing objects into groups such that similarity within the same cluster is maximized and similarities among different clusters are minimized. In many real world problems, clustering with equal weights for each attribute does not provide the desired results since different attributes have different significance levels [6]. Same weights are assigned to all the attributes in many clustering algorithms irrespective of the fact that all attributes do not have equal importance or weights in most of the real world problems. Weighted k-means clustering is considered as the popular solution to handle such kind of problems [9]. In order to introduce the different weights for different attributes, parametric Minkowski model [3] is used to consider the weightage scheme in weighted kmeans clustering algorithm. In parametric Minkowski model, the distance function is defined by a weighted version of the Minkowski distance measure. The parameters for this model are the weights in different dimensions [10].

PC-Kmeans algorithm is similar to COP-Kmeans, but the main difference is that this algorithm can violate the constraints with some trade off as penalty for doing so [4]. It tries to come up with a good cluster formation while minimizing the penalty that it incurs. A major limitation of this approach is that it assumes a single metric for all clusters, preventing them from having different shapes. Bilenko et. al. [10] has proposed metric pair-wise constraint kmeans (MPCK-Means) algorithm to get rid of this limitation. MPCK-Means is considered as one of the most popular semi-supervised clustering algorithms in the recent past. Therefore, the proposed approach has been compared with MPCK-Means in the paper [7]. The proposed approach based on Hyperlink-Induced Topic Search (HITS) algorithm is introduced to overcome the limitations of earlier work. Using the proposed approach the weights for the attributes are generated automatically from the data, for the weighted k-means using parametric Minkowski's model, some of the preliminaries i.e. parametric Minkowski model and HITS algorithm are described in the next section of the paper [9].

The clustering algorithms are classified into generative and discriminative. Generative is a parametric form of data generation is assumed and the goal in the maximum likelihood formulation is to find the parameters that maximize the probability (likelihood) of generation of the

data given the model [9]. Discriminative tries to cluster the data so as to maximize within-cluster similarity and minimize between-cluster similarity based on a particular similarity metric, where it is not necessary to consider an underlying parametric data generation model. Both can be implemented using Expectation maximization and k-means [7]. In addition, k-means is a flat partitioning type of clustering algorithm that divides the data points into k partitions or clusters by grouping the similar features (usually Euclidean) and assigning each point to the cluster whose mean value on a set of x variables is nearest to it on that set. Furthermore, semi-supervised clustering algorithms categorize into similarity or distance-based or partially labelled data and search-based or pair constraint based methods [5]. The former, used to classify the unlabeled data to the appropriate clusters using the known clusters where the later considers "Must-link constraints" require that two observations must be placed in the same cluster, and "cannot-link constraints" require that two observations must not be placed in the same cluster [1]. Another difference between the two is in distance-based method uses traditional clustering algorithm like k-means that uses a similarity metric where 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 [4].

3. PROPOSED WORK

3.1 SEMI HARD CLUSTERING (SHC)

In Semi Hard clustering, [tij] takes the value from {0, 1}. The objective function for hard clustering can be re-written as follows:

$$Q = \sum_{i=1}^{n} A_i^t T_i + b_i$$

During the E-step, i should minimize the contribution of each point to the objective function Q. Clearly, minimizing Qsubject to the constraints is a linear programming problem. From [6], the minimum can be achieved by setting tij as follows:

$$a_{ij} = \begin{cases} d_{ij}^2 - Rw_j & \text{if } f_{ij} = 1 \\ d_{ij}^2 + Rw_j & \text{otherwise} \end{cases}$$

$$t_{ij} = \begin{cases} 1 & \text{if } j = \min_l a_{il} \\ 0 & \text{otherwise} \end{cases}$$

During the M-step, since the configuration matrix is fixed, the second term of the objective function is unchanged. Minimizing Q is therefore equivalent to minimizing the first term [9]. The centroid update formula is:

$$c_j = \frac{\sum_{i=1}^{n} t_{ij} u_i}{\sum_{i=1}^{n} t_{ij}}$$

3.2 CLUSTER WEIGHTING (CW)

The proposed objective function should also take into account how informative is the shared feature set. I use the weight wj to reflect the importance of the shared feature set in terms of discriminating cluster j from other clusters [9]. The concept of feature weighting has been used in clustering by Frigui et. al. [7]. Let D (d,j) denote the discriminating ability of the

$$D(d,j) = \frac{\sum_{u_i \in c_j} \sum_{l \neq j} d(x_i, c_l, \Re^d)}{\sum_{u_i \in c_j} d(x_i, c_j, \Re^d)}$$

Where d (ui, cj, <d) is the distance between ui and centroid cj based on the feature set <d. Intuitively, D (d, j) is the ratio of the between-cluster distance and the within-cluster distance [7]. The higher the ratio, the more informative is the feature set in terms of discriminating cluster j from other clusters.

The weight of cluster j is then determined as follows:

$$w_j = D(p,j)/D(d,j)$$

Where p is the number of shared features and d is the total number of features in U.

3.3 SESPROC FOR SEMI-SUPERVISED SUBSPACE MODEL BASED CLUSTERING USING THE EM ALGORITHM (SESPROC)

I apply the above mixture model theory to a clustering problem with two specific characteristics:

1. The groups of instances can be hidden in different feature subspaces. Therefore, an LFSS is required in each mixture component. This way i can identify data structures that would remain undiscovered using all features or GFSS [2].
2. Theclassinformationofsomeinstancesisavailable.Thisk nowledgeisusedduring the EM process to improve the final clustering; therefore, this is a semi-supervised clustering task [5].

$$p(\mathbf{x}_i \mid \Theta) = \sum_{m=1}^{K} \pi_m \prod_{j=1}^{F} \Big(\rho_{mj}\, p(x_{ij} \mid \theta_{mj}) + (1 - \rho_{mj}) p(x_{ij} \mid$$

SeSProC uses the available instance label information to guide the clustering of the unlabeled instances [4]. Based on this information, the model learning process can be divided in to two learning parts: the labelled instances are correctly classified in to known classes {1... C} (classification term) and the unlabeled instances can be grouped either in those known or in other unknown components {C +1... K} (clustering term) [9].

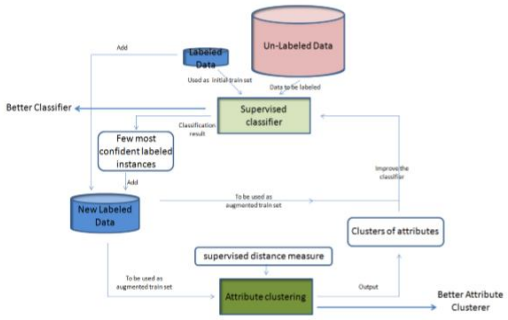
$$\mathbb{E}_{v_{mj},|x_{ij},\theta_{mj}}[v_{mj}] = \gamma(v_{mj}) = \frac{\rho_{mj}\, p(x_{ij} \mid \theta_{mj})}{\rho_{mj}\, p(x_{ij} \mid \theta_{mj}) + (1-\rho_{mj}) p(x_{ij} \mid \lambda_{mj})} = p(v_{mj} = 1 \mid x_{ij}, \theta_{mj}),$$

3.4 SEMI-SUPERVISED ATTRIBUTE CLUSTERING APPROACH (SSAC)

The general algorithm for this semi-supervised attribute clustering framework is defined in algorithm 1. The Input includes labelled training set L with M instances and A attributes, Unlabeled training set U with N instances and A attributes and also Number of clusters of attributes, P. P is optional based on attribute clustering approach used in step 5 of the algorithm. Output of this framework includes both an improved attribute clustering and an improved classifier [6].

3.4.1 ALGORITHM 1 PROPOSED SEMI-SUPERVISED ATTRIBUTE CLUSTERING APPROACH

1: Input Labelled training set L with M instances and A attributes, Unlabeled training set U with N instances and A attributes, Number of Clusters of attributes P (optional)
2: repeat
3: if First Iteration then
4: Initialize the framework: If a multi- view Classifier would be used define the initial views of attributes, perform initial attribute selection, perform initial missing attribute value handling
5: else
6: Use new clusters of attributes F to improve the classifier used in step 2 by Updating views of multi-view classifier or performing attribute selection or performing missing attribute value handling
7: end if
8: Train a classifier C using the Labelled training set L and the result of previous step.
9: Classify unlabeled instances U
10: Add K most confident new labelled instances to the labelled set (M = M +K,L = L∪{K most confident new labelled instance})
11: Use new labelled set with an attribute clustering method to provide clusters of attributes F 12: until No new instances added
13: Output:
The set of clusters of attributes and the classifier trained in the final iteration



Figure(1): PROPOSED SEMI-SUPERVISED ATTRIBUTE CLUSTERING APPROACH

3.5 SEMI-SUPERVISED ENRICHMENT (SSE)

Most times in realistic settings, users provide only a few rating histories of movies. The lack of rating data leads to the cold start problem in recommendation. I enrich the user profiles with a semi-supervised way when the number of rating histories is lower than a threshold [7]. An iterative co-training technique of multi view learning is used here. First, the prediction scores of unrated movies from each view are computed respectively [4]. Then the items rated with the same score by all the view predictors are added to the training set. This process repeats until there is no new item to add. The algorithm of this process is shown in Algorithm 1 and denoted as MVE (Z, Q, and k). The enriched training set is used to predict the remaining items in the testing set in multi-view recommendation. Note that the enrichment will be executed only when the rated history of users is less than a threshold σ [1].

3.5.1 ALGORITHM 1. SEMI-SUPERVISED ENRICHMENT PROCEDURE

Input: rated movies Z: (u, m, and r), movies to be rated
Q: (u, m), views: {v, t, a}
Output: Z0
1 for each q ∈Q, view k ∈ {v, t, a}
Pk (q) = Recom SV (Z, Q, k)
End for

	Balance Dataset			
Algorithm	**Accuracy**	**Precision**	**Recall**	**F-Measure**
SHC	89.45	87.91	92.77	90.89
CW	79.91	76.08	74.78	86.56
SESPROC	70.92	79.67	79.89	85.78
SSAC	84.67	90.67	86.78	77.67
SSE	90.07	83.66	82.33	72.88

2 for each q ∈Q
If ∀Pk (q) = r
Z0 ←−Z ∪ {(q, r)}
End if
End for

4. EXPERIMENTS

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

4.1 DATASETS

The data sets used in my experiments include six UCI data sets1. Here is some basic information of those data sets. Table(1) summarizes the basic information of those data sets.

• Balance. This data set was generated to model psychological experimental results. There are totally 625 examples that can be classified as having the balance scale tip to the right, tip to the left, or be balanced.

• Iris. This data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.

• Ionosphere. It is a collection of the radar signals belonging to two classes. The data set contains 351 objects in total, which are all 34-dimensional.

• Soybean. It is collected from the Michalski's famous soybean disease databases, which contains 562 instances from 19 classes.

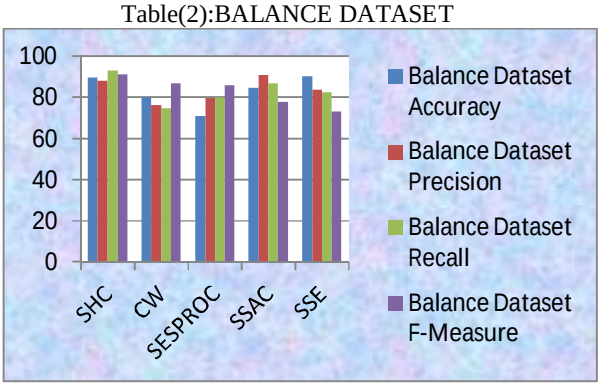
Table(1) : SIX UCI DATA SETS

Datasets	**Size**	**Classes**	**Dimensions**
Balance	625	3	4
Iris	150	3	4
Ionosphere	351	2	34
Soybean	562	19	35

5. EXPERIMENTAL RESULTS

5.1 BALANCE DATASET RESULTS

Table(2):BALANCE DATASET



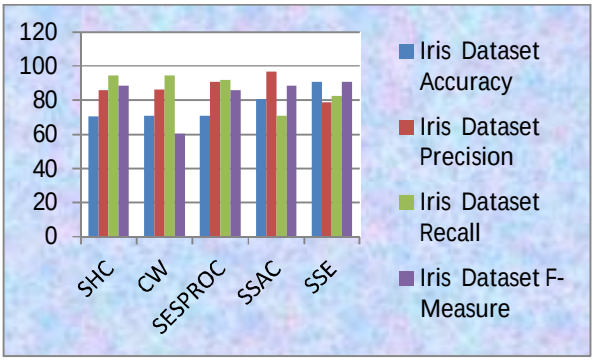
Figure(2): PERFORMANCE OF BALANCE DATASET

The above graph shows that performance of Balance dataset. The Accuracy of SSE algorithm is 90.07 which is higher when compare to other four (SHC, CW, SESPROC, SSAC) algorithms. The Precision of SSAC algorithm is 90.67 which is higher when compare to other four (SHC, CW, SESPROC, SSE) algorithms. The Recall of SHC algorithm is 92.77 which is higher when compare to other four (SSE, CW, SESPROC, SSAC) algorithms. The F-Measure of SHC algorithm is 90.89 which is higher when compare to other four (SSE, CW, SESPROC, SSAC) algorithms.

5.2 IRIS DATASET RESULTS

Table(3) : IRIS DATASETS

	Iris Dataset			
Algorithm	**Accuracy**	**Precision**	**Recall**	**F-Measure**
SHC	70.45	85.91	94.77	88.89
CW	70.91	86.08	94.78	60.56
SESPROC	70.92	90.67	91.89	85.78
SSAC	80.67	96.67	70.78	88.67
SSE	90.78	78.76	82.54	90.89



Figure(3): PERFORMANCE OF IRIS DATASET

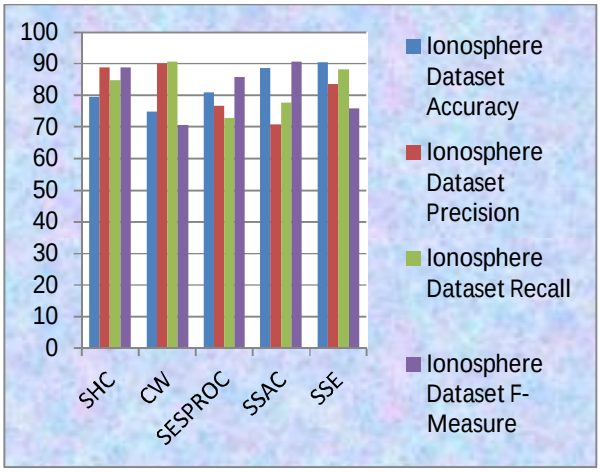
	Soybean Dataset			
Algorithm	**Accuracy**	**Precision**	**Recall**	**F-Measure**
SHC	79.89	88.65	84.23	88.34
CW	74.03	90.89	90.67	71.23
SESPROC	81.08	76.32	72.45	85.9
SSAC	88.54	71.32	77.89	90.56
SSE	90.08	83.78	88.78	75.9

The above graph shows that performance of Iris dataset. The Accuracy of SSE algorithm is 90.78 which is higher when compare to other four (SHC, CW, SESPROC, SSAC) algorithms. The Precision of SSAC algorithm is 96.67 which is higher when compare to other four (SHC, CW, SESPROC, SSE) algorithms. The Recall of CW algorithm is 94.78 which is higher when compare to other four (SSE, SHC, SESPROC, SSAC) algorithms. The F-Measure of SSE algorithm is 90.89 which is higher when compare to other four (SHC, CW, SESPROC, SSAC) algorithms.

	Ionosphere Dataset			
Algorithm	**Accuracy**	**Precision**	**Recall**	**F-Measure**
SHC	79.45	88.91	84.77	88.89
CW	74.91	90.08	90.78	70.56
SESPROC	80.98	76.67	72.89	85.78
SSAC	88.67	70.67	77.78	90.67
SSE	90.56	83.45	88.34	75.89

5.3 IONOSPHERE DATASET RESULTS

Table(4):IONOSPHERE DATASET

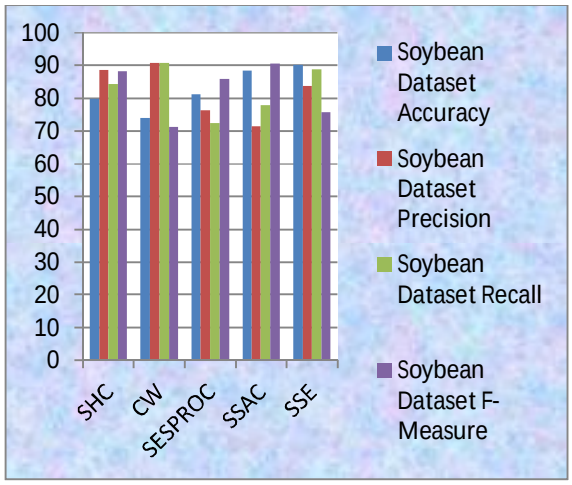


Figure(4): PERFORMANCE OF IONOSPHERE DATASET

The above graph shows that performance of Ionosphere dataset. The Accuracy of SSE algorithm is 90.56 which is higher when compare to other four (SHC, CW, SESPROC, SSAC) algorithms. The Precision of CW algorithm is 90.08 which is higher when compare to other four (SHC, SSAC, SESPROC, SSE) algorithms. The Recall of CW algorithm is 90.78 which is higher when compare to other four (SSE, SHC, SESPROC, SSAC) algorithms. The F-Measure of SSAC algorithm is 90.67 which is higher when compare to other four (SSE, CW, SESPROC, SHC) algorithms.

5.4 SOYBEAN DATASET RESULTS

Table(5): SOYBEAN DATASET



Figure(5): PERFORMANCE OF SOYBEAN DATASET

The above graph shows that performance of Soybean dataset. The Accuracy of SSE algorithm is 90.08 which is higher when compare to other four (SHC, CW, SESPROC, SSAC) algorithms. The Precision of CW algorithm is 90.89 which is higher when compare to other four (SHC, SSAC, SESPROC, SSE) algorithms. The Recall of CW algorithm is 90.67 which is higher when compare to other four (SSE, SHC, SESPROC, SSAC) algorithms. The F-Measure of SSAC algorithm is 90.56 which is higher when compare to other four (SSE, CW, SESPROC, SHC) algorithms.

6. CONCLUSION

The results that i have here are with small cluster number and are just the start in order to address problem with larger magnitude. A novel approach for clustering with closeness is put forth. It is not just the threshold value but the dynamic change in closeness value that generated the clusters accurately [2]. Extension of the work include investigating CW approach for some more datasets as well as application of the approach as a baseline method

for incremental update of clusters that can be applied in semi-supervised way. The effectiveness of the proposed approach over the existing clustering algorithm has been illustrated using UCI machine learning repository datasets and compared with the popular clustering algorithms such as SSE and SSAC. The proposed clustering approach produces better results even with the widely used semi-supervised clustering algorithm like SESPROC [4]. It can be applied to large scale of practical problems as most of the real world problems do not have equal weights for each of the attributes and weights are either unknown or hard to obtain. The approach can play an important role for wider variety of clustering problems especially where the attributes of a dataset do not have equal weights [3]. I have proposed a semi-supervised method, called SeSProC, capable of discovering unknown clusters, based on EM algorithm, and including a LFSS. This algorithm includes available information in the search for subspaces and clusters. Besides, SeSProC has two major advantages over related algorithms. The first one is that my proposal has only one, easily adjustable input parameter [7]. Whereas other algorithms are unable to find a final solution without proper parameter tuning, SeSProC always obtains a clustering solution regardless of the value of the input parameter. The second advantage is related to the known labels. SeSProC is able to find hidden clusters that are not represented by the labelled instances. It uses a novel greedy process to find these clusters, assuming that instances that fit the known clusters worst are candidates for initializing new clusters [9].

REFERENCES

[1] A. Demiriz, K. P. Bennett, and M. J. Embrechts, “Semi-supervised clustering using genetic algorithms,” Artificial Neural Networks in Engineering (ANNIE), 809–814, 2010.

[2] A. K. Jain, M. N. Murty, P. J. Flynn, “Data clustering: a review, ACM Computing Surveys”, 31(3), 264–323, 2011.

[3] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, “UCI Repository of machine learning databases,” [http://www.ics.uci.edu/mlearn/MLRepository .html]. Irvine, CA: University of California, 2011.

[4] M. Bilenko, S. Basu and R. Mooney, “Integrating constraints and metric learning in semi-supervised clustering,” in Proc. ICML, 2014, pp. 81–88.

[5] S. Basu, M. Bilenko, and R. J. Mooney,” A Probabilistic Framework for Semi-Supervised Clustering”, In Proc. of SIGKDD, 59-68, 2012.

[6] T. Li, C. Ding and M. I. Jordan, “Solving Consensus and Semi-supervised Clustering Problems Using Nonnegative Matrix Factorization”, In Proc. of ICDM. 2009.

[7] Miller D, Chu-Fang L, Kesidis G, Collins,” Semi supervised mixture modelling with fine-grained component-conditional class labelling and transductive inference”, In: IEEE international workshop on machine learning for signal processing, pp. 1–6,2010.

[8] X. Zhu,” Semi-supervised clustering literature survey”, Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2011.

[9] X. Zhu and A. B. Goldberg,” Introduction to Semi-supervised Clustering “, Morgan & Claypool Publishers, 2009.

[10] X. Zhu, Z. Ghahramani, and J. Lafferty,” Semi-supervised learning using Gaussian fields and harmonic functions”, In Proceedings of International Conference on Machine Learning, pages 912– 919, 2013.

[11] Kulis, B., Basu, S., Dhillon, I. S. & Mooney R. J,” Semi-supervised Graph Clustering: a Kernel Approach”, In: Proc. of International Conference on Machine Learning (ICML) (pp. 457464), 2011.

[12] S. Basu, M. Bilenko, and R. J. Mooney,” A probabilistic framework for semi-supervised clustering”, In KDD ’04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining, pages 59–68. ACM Press, 2014.