EAI Endorsed Transactions
on Scalable Information Systems Research Article **EAI.EU**

Improving Semi-Supervised Classification using Clustering

J. Arora[1], M. Tushir[2,*] and R. Kashyap[3]

[1]Assistant Professor, Dept. of Information Technology, MSIT, Affiliated to GGSIPU, Delhi, India
[2]Professor, Dept. of Electrical & Electronics Engineering, MSIT, Affiliated to GGSIPU, Delhi, India
[3]Graduate, Dept. of Information Technology, MSIT, Affiliated to GGSIPU, Delhi, India

Abstract

Supervised classification techniques, broadly depend on the availability of labeled data. However, collecting this labeled data is always a tedious and costly process. To reduce these efforts and improve the performance of classification process, this paper proposes a new framework, which combines a most basic classification technique with the semi-supervised process of clustering. Semi-supervised clustering algorithms, aim to increase the accuracy of clustering process by effectively exploring available supervision from a limited amount of labeled data and help to label the unlabeled data. In our paper, a semi-supervised clustering is integrated with naive bayes classification technique which helps to better train the classifier. To evaluate the performance of the proposed technique, we conduct experiments on several real world benchmark datasets. The experimental results show that the proposed approach surpasses the competing approaches in both accuracy and efficiency.

Keywords: Semi-Supervised Clustering, Naive Bayes Classification, Probability, Fuzzy C- means.

Received on 10 May 2019, accepted on 18 July 2019, published on 05 August 2019

doi: 10.4108/eai.29-7-2019.159793

*Corresponding author. Email:meenatushir@yahoo.com

1. Introduction

Supervised Classification is a process, in which a sample data is used as the representative structure of a class, termed as a training set. This training set is used to train the model for further classification process. Training sets are particularly selected based on the availability of knowledge or labeled data set [1-2]. Now days, supervised classification techniques have been extensively used in the field of pattern matching, including medical diagnosis, face recognition, document classification, banking sector and many other application areas [3-5]. The performance of these techniques mostly depends on the availability of the labeled data. However, labeled data is very limited and is difficult to obtain. It is very time consuming process and requires expert involvement, often resulting in very expensive process [6]. To overcome the problems of supervised classification, semi-supervised learning is used in the process of classification. A lot of research has been carried out in this area. Semi-supervised learning is applicable for both clustering as well as classification process. There are a variety of semi-supervised clustering techniques that have been given and proved to perform better as compared to unsupervised approach. Basically semi-supervised clustering deals with the methods to incorporate additional information gathered by different sources into unsupervised clustering process. Basu et al. (2002) [7] proposed semi-supervised clustering algorithm based on center initialization mechanism. In this algorithm, seeds are used to initialize the centers of clusters using labeled data and then updated using clustering process. Demiriz et al. (1999) [8] have used genetic algorithm along with supervised and unsupervised clustering to design semi-supervised clustering algorithm. Blum et al (2001) [9] used graph based method to provide information regarding labeling in the process of unsupervised clustering. Gan et. al. (2013) [10] have given the concept of self-training, involving semi-supervised clustering along with classification technique. Further Zhang et al. (2004)[11] have defined the semi-supervised clustering along with kernel based approach. They have defined the objective function by reducing the classification errors of both the labeled and unlabeled data. But kernel based approach is more complex and time consuming. Piekari et. al. (2018)[12] have used cluster analysis to improve the results of semi-supervised learning and further used supervised classification approach to classify pathological images of breast. Dunlop et. al. (2019) [13] defined a graph-based semi-supervised learning approach and introduced large data limits of the probit and Bayesian level set problem formulations. Wang et. al. (2019) [14] have used unsupervised approach integrated with classification process for the prediction of day ahead electricity price.

There are several such algorithms available in the literature which are quite effective [15-16]. However question always remain whether we can still improve the results. It is a well-known fact that there is no single algorithm which is capable of giving good results on all kinds of datasets. Hence there is always a scope of improvement over the existing algorithm and thus we have proposed a new model for classification where clustering is integrated with the process of classification. In our framework, semi-supervised fuzzy c-means clustering (SSFCM) is used to train the most basic classifier, resulting in the process of self-training. The main advantage of the model is the efficient use of available labeled and unlabeled data. The classification process helps to reveal the internal structure of the data, which is used by the clustering process along with unlabeled data to further improve the process of classification. Basically, this paper works on the following factors:-

1. A semi-supervised fuzzy c-means clustering is used to label the unlabeled data and further helps to improve the training of the classifier.
2. The naive bayes classification is used for the process of classification and further it helps to provide better internal structure of the data in the process of semi-supervised clustering.
3. Both labeled and unlabeled data are used by the model and help to better classify the unlabeled data with better results.
4. The simplicity of classification process helps to give better results and is more efficient.

Further, the paper is organized in the following section. Section 2 gives the details of the related work including Semi-Supervised Fuzzy C-means Clustering and Naive Bayes classification clustering technique. Section 3 describes the framework of the proposed algorithm. Section 4 includes the experimental results carried out on various real datasets along with the discussions. The paper is concluded in section 5.

2. Related Work

The methodologies of different machine learning techniques widely depend on the availability of labeled and unlabeled data. Supervised learning requires a close attention towards development of training data. If the training data is poor or not representative the classification results will also be poor [17]. On the other hand, unsupervised learning suffers from the problem of local traps due to random initialization of the process [18]. To avoid the above problems, the semi-supervised approaches basically involve an intermediate link between supervised and unsupervised learning techniques [19].

2.1. Semi-supervised Fuzzy C-means (SSFCM)

A salient feature of partial supervision in the clustering algorithms is the availability of labeled data in the given data or some other constraints that can provide some supervision to the process of unsupervised clustering. Pedcryz et. al. (1997) [20] have proposed semi-supervised fuzzy c-means (SSFCM) and updated objective function of conventional FCM by adding the concept of partial supervision. Further, Tong et. al. (2009) [21] extended the definition of SSFCM to cover the hidden information in a better way with the availability of some amount of labeled information. E. Bair (2013) [15] presented the review on the effectiveness of semi-supervised clustering methods.

An algorithm for SSFCM can be defined as:-

Here data is represented as X with n number of vectors in feature space S^p such that

$X = \{x_1, x_2, x_3, \ldots\ldots\ldots . x_n\}$

Here p denotes the dimension of dataset.

The most basic goal of clustering algorithms is to provide label to each data point implying its belongingness to some class [22]. Let c denotes the number of classes, $1 \le c \le n$. In semi-supervised clustering, a small set of labeled data is provided along with large set of unlabeled data. So X in case of semi-supervised clustering can be represented as

$X = \{x_1^l, x_2^l, x_3^l \ldots.. x_k^l . | x_1^u, x_2^u, x_3^u \ldots\ldots x_m^u\}$ where data with l superscript is labeled and with u superscript is unlabeled.

Consider number of data points in a labeled set as n^l and in unlabeled set as n^u. The total number of data points are denoted as $n = n^l + n^u$. Here membership matrix of labeled data will be set as crisp matrix of 0 and1, where

μ_{ik}^{l} is set to 1, if x_k is a member of class i and 0 otherwise and μ_{im}^{u} will be initialized randomly. Further initial seed (cluster center) values will be calculated from labeled data and labeled membership matrix as

$$s_i^0 = \frac{\sum_{k=1}^{n^l} \left(\mu_{ik}^l\right)^t x_k^l}{\sum_{k=1}^{n^l} \left(\mu_{ik}^l\right)}, 1 \le i \le c, 1 \le k \le n^l \quad (1)$$

Further $\left(\mu_{im}^{u}\right)$ membership matrix of unlabeled data is updated using calculated seed (cluster center) values from eq. (1) as

$$\left(\mu_{im}^{u}\right) = \frac{d(x_m^u, s_i)^{1/t-1}}{\sum_{j=1}^{c} d\left(x_m^u, s_j\right)^{1/t-1}}, 1 \le i \le c, 1 \le m \le n^u \quad (2)$$

where $d(x_m^u, s_i)$ is euclidean distance between data point and seed value. Finally the seed values are updated for the complete data set

$$s_i = \frac{\sum_{k=1}^{n_l} \left(\mu_{ik}^l\right)^t x_k^l + \sum_{m=1}^{n_u} \left(\mu_{im}^u\right)^t x_m^u}{\sum_{k=1}^{n_l} \left(\mu_{ik}^l\right)^t + \sum_{m=1}^{n_u} \left(\mu_{im}^u\right)^t} \quad (3)$$

2.2. Naive Bayes Classification (NB)

In probability, bayes theorem finds the probability of an event based on prior condition that might have a relation to the event. So similarly, naive bayes classifier is a probabilistic classification model which is completely based on Bayes theorem. It is used in supervised learning where we already have some information to train the classifier [23]. It is an efficient algorithm which does not need much time to train and quickly classifies the data. The most notable advantages of using a naive bayes classifier is that it does not require much training data and it is not easily affected by outlier values. However the efficiency of classification process increases with the increase in the training information.

Bayes Theorem is represented as follows:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \quad (4)$$

Similarly, Naïve Bayes is represented as follows:

$$P(C_k/x) = \frac{P(x/C_k)P(C_k)}{P(x)} \quad (5)$$

Where, C_k represents the class for k possible outcomes

$P(C_k/x)$ – Posterior probability is the end result that we want to obtain for a given sample for all k classes

$P(x/C_k)$ – Likelihood i.e. the probability of observing x given that it belongs to class C_k

$P(x)$- Marginal Likelihood i.e. the probability of the sample data for having the given attributes

$P(C_k)$ – Prior Probability is the initial probability of the class C_k

3. Proposed Semi-supervised Fuzzy baysian classification algorithm (SSFBC)

In our implementation, we have used SSFCM and naive bayes classification, respectively for semi-supervised clustering and classification. Thus Semi-supervised fuzzy baysian classification (SSFBC) is an extended classification procedure defined with the integration of clustering with partial supervision. In partial supervision, clustering is provided with data having some percentage of labeled data along with a large amount of unlabeled data. In the complete process, the classifier is trained using labeled data and semi-supervised fuzzy c-means is used for labeling unlabeled data as shown in Fig. 1. The percentage of labeled data increases with every iteration and is used for better training of the classifier. Thus, semi-supervised clustering provides classification to unlabeled data and the classifier is retrained. Here Naive Bayes classifier is used for the process of classification. This classifier is known for its simplicity and accuracy, as it does not require any training for classification [23]. So, the simplicity of n**ai**ve bayes classification process reduces the time-complexity of the whole process and is use to obtain accurate results quickly. The complete process is repeated until all unlabeled data are labeled.

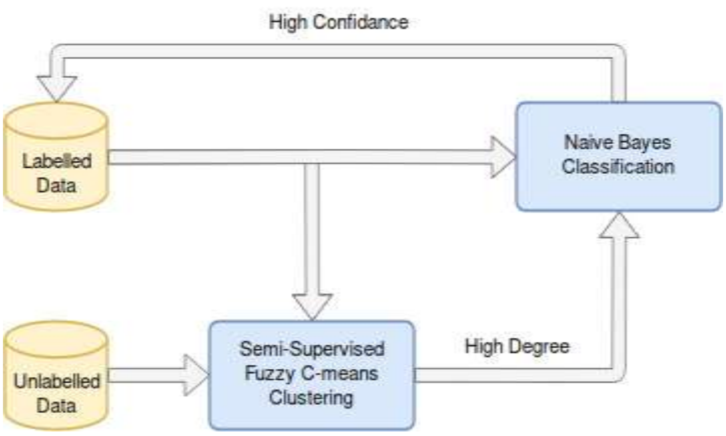


Figure 1. Process of Semi Supervised Fuzzy Bayesian Classification

On comparing the proposed technique with other commonly used techniques like SVM or Naive Bayes, we noticed that these techniques do not adapt to the unlabeled data. In SVM, the classifier trains itself on the labeled data, which then creates a straight line, and classifies the unlabeled data whereas in Naive Bayes, no such type of training is required but when the unlabeled data is being classified, it just calculates the probability of the most probable classification using the labeled data as its reference. In both these cases, only the labeled data is used for the process of learning. These approaches fail in the cases where the unlabeled data is continuously expanding as in such scenarios, it is possible for the pattern to change, which causes these algorithms to fail. In our proposed algorithm, the data is labeled continuously with high degree and high confidence. By doing so, our algorithm is able to adapt to the changes in the pattern of data and outperforms the usual algorithms. In Fig. 2, a synthetic data is generated with partial labeled points. This synthetic data is based on the situation where the data is continually expanding and the pattern is changing as well. The data is trained with the help of partially labeled data using SVM, Naive Bayes and proposed SSFBC technique. Fig. 2(a) shows the result of classification done using SVM where lot of data is misclassified; Fig. 2(b) shows the result of Naïve Bayes classification where the data is learned inaccurately. Fig. 2(c) shows the result of proposed technique and shows that the proposed technique gives high quality result with much better accuracy. The algorithm properly adapts the changes in the structure and results in better classification process. It does not get affected by the continuously changing trend in data and is able to give results with high accuracy. From Fig. 2(a) and Fig. 2(b), we observed that the results continuously degrades as the classifier does not adapts well with the increasing data. The algorithm of the proposed technique, Semi-Supervised Fuzzy Bayesian classification (SSFBC) is given below. We are denoting degree threshold and confidence threshold by $e1$ and $e2$, respectively.

Algorithm – SSFCM+NB (SSFBC)

Step 1: Define the dataset with labeled and unlabeled data.

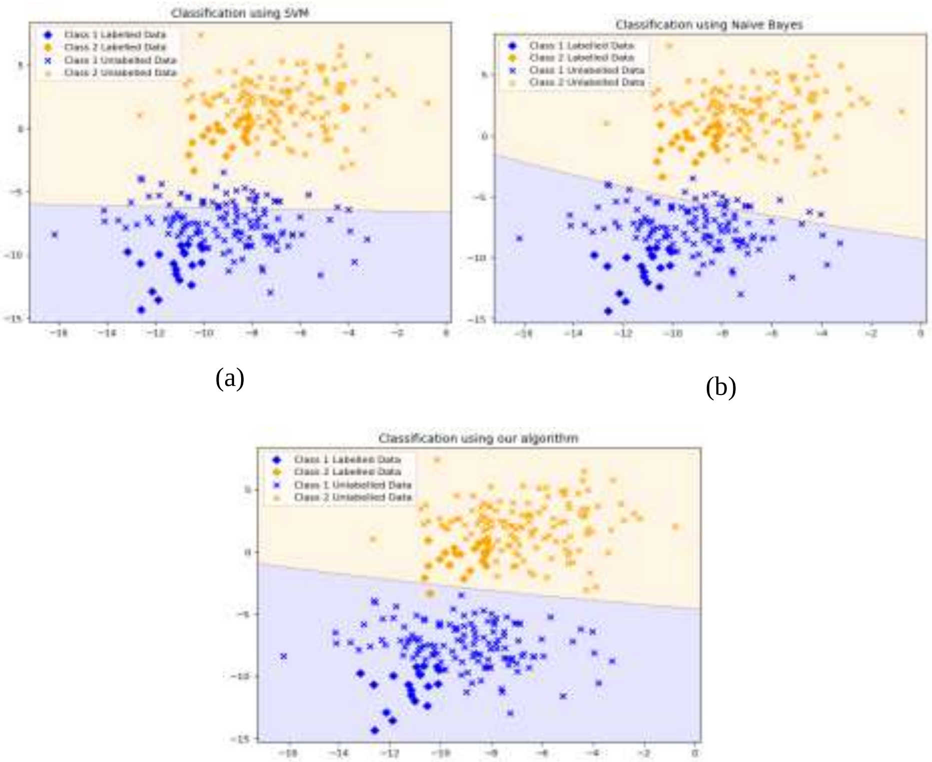
Step 2: Initialize $e1$ and $e2$ as 0.95 and 0.60 respectively

Step 3: Repeat until flag == true

(a) Calculate membership using SSFCM clustering for unlabeled data

(b) Select samples with high degree i.e. samples whose degree is more than $e1$ -

$$x_i > e1$$



(a) (b)

(c)

Figure 2. Classification of synthetic data using (a) SVM (b) Naive Bayes (c) SSFBC (Our Proposed algorithm)

- if there is no such sample, reduce the value of $e1$ by 0.05 and start (Step 3) again

- if $e1$ becomes less than 0.60, set flag as false and break the iteration

(c) Train the Naive Bayes classifier using labeled dataset

(d) Using the trained model, compute the confidence of high degree sample

(e) Select the samples having confidence more than $e2$ - $x_i > e2$

- if there is no such sample, reduce the value of $e1$ by 0.05 and start (Step 3) again

- if $e1$ becomes less than 0.60, set flag as false and break the iteration

(f) Finally, add these samples to labeled data and remove them from unlabeled data.

(g) If size (unlabel) == 0, set flag as false

Step 4: Train the Naive Bayes classifier using labeled data

Step 5: Using the trained model, classify the remaining samples in unlabeled data.

4. Experiment Results and Discussions

In our experiments, we have tested our algorithm on three real benchmark datasets from UCI dataset repository: Iris, Wine and Wireless Indoor Localization datasets. All of these are multi-class datasets where Iris and Wine datasets have three classes whereas, Wireless Indoor Localization dataset has four classes. We have divided our datasets in the ratio of 1:4 i.e. we have used 20% as labeled and the rest 80% as unlabeled. Table 1 shows the details about the datasets.

Table 1. Details of real datasets taken from UCI repository

Dataset	Attributes	Class	Labeled Size	Unlabeled Size
Iris	4	3	30	120
Wine	13	3	35	143
Wireless Indoor Localization	7	4	400	1600

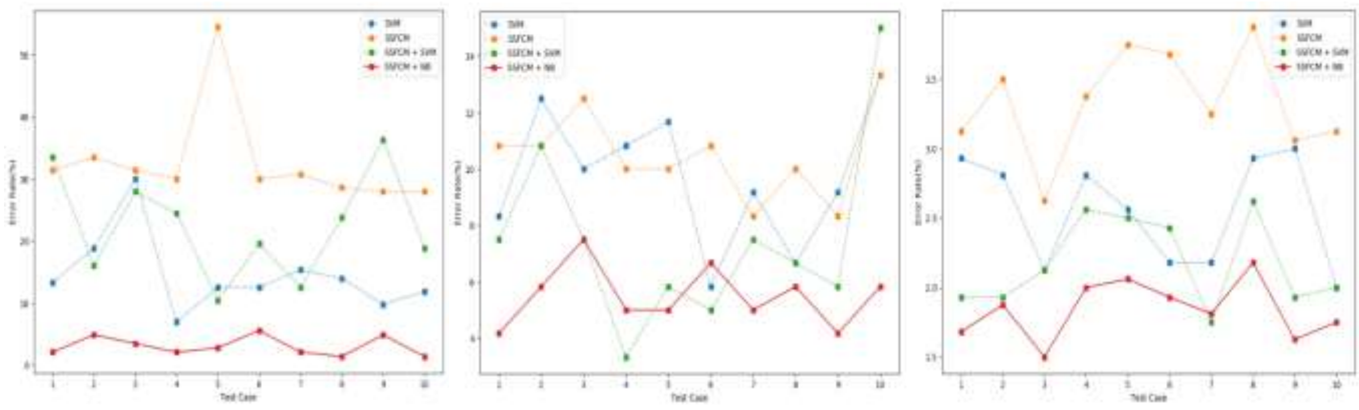
We have compared our proposed SSFBC method with a range of successful supervised and semi-supervised classification methods available in the literature. We have done investigations on standard supervised SVM technique, semi-supervised SSFCM clustering based classifier and semi-supervised SSFCM+SVM[10] technique. SSFCM+SVM works on the same line as our proposed method. It first finds the underlying structure in the data space by applying the semi-supervised clustering on both labeled and unlabeled data and then SVM classifier is trained on the labeled data.

4.1 Error rate of labeling unlabeled data

In this section, we compared the results of our proposed method with other stated algorithms on the basis of error rate in classifying the unlabeled data. The results of experiments are assessed using Huang's accuracy measure [24]:

$$r = \frac{\sum_{i=1}^{k} n_i}{n} \tag{6}$$

where n_i represents the true positive of data occurring in both the i^{th} cluster and its corresponding true cluster and n is the total number of data points in the data set. The higher value of accuracy measure r proves superior clustering results with perfect clustering generating a value $r = 1$. The labeled data will allow the clustering process to specify the number of clusters c on the basis of class information available. The constraint for labeling is that the data should be labeled from every class in order to provide training patterns that could capture a training set from every class. As the labeling of data is done randomly, so on each dataset, we have taken 10 observations and then finally calculated the mean error and standard deviation. Fig. 3 shows the



(a) (b) (c)

Figure 3. Error Rate of Misclassification in Real data (a) Iris (b) Wine (c) Wireless Indoor Localization

comparative behaviour of iris, wine and wireless data sets with SVM, SSFCM, SSFCM+SVM and proposed SSFBC. Here the graph shows percentage error rate of misclassification of data points with their respective clusters for all ten test cases. This gives us an overview of the performance of the above algorithms for different cases. Fig. 3(a) shows the result of iris data set where the proposed technique shows minimum error rate as compared to other techniques in each and every run. Fig. 3(b) shows the result of wine dataset. It reveals that SSFBC performs comparative well in most of the cases. Fig. 3(c) shows the result of wireless localization dataset. In all the cases, proposed technique SSFBC reveals its superiority and performs well as compared to other techniques.

Table 2 shows the mean and standard deviation of the error rate calculated after every run. Here in each pass, labeling is done randomly, and with same set of labeled data, classification process is carried out for every technique. Finally after 10 runs, mean and standard deviation is calculated to obtain the final results. Table 2 shows that the proposed technique SSFBC shows better results with minimum mean and standard deviation of error rate on all the three datasets.

Table 2. Mean and Standard Deviation of Error Rate for SVM, SSFCM, SSFCM integrated with SVM and proposed technique SSFBC

		Supervised SVM Classifier	**SSFCM based classifier**	**Semi-supervised SSFCM + SVM classifier**	**Our proposed method SSFBC**
Iris	**Mean**	9.75	10.498	7.499	**5.5**
	Std D	2.423	1.581	3.287	**1.053**
Wine	**Mean**	14.54	32.651	22.373	**3.07**
	Std D	6.295	7.884	8.534	**1.555**
Wireless Indoor Localization	**Mean**	2.553	3.337	2.178	**1.841**
	Std D	0.392	0.378	0.318	**0.21**

4.2. Time consumption for labelling unlabelled data

In this section, we will compare our algorithm with the

Table 3. Computational time spent for labelling Iris Dataset

Selection	**Supervised SVM Classifier**	**SSFCM**	**Semi-supervised SSFCM + SVM classifier**	**Our proposed method SSFBC**
1	5.32	0.0113	12.06	1.05
2	8.11	0.0115	12.99	1.07
3	5.53	0.0114	11.94	1.16
4	6.85	0.0109	10.58	1.02
5	4.68	0.0107	10.49	1.15
6	7.24	0.0111	10.15	1.07
7	4.55	0.0113	8.48	0.99
8	6.82	0.0113	12.75	1.02
9	6.21	0.0108	9.96	1.05
10	8.13	0.0118	9.43	1.05
Mean	6.344	0.011	10.883	**1.063**
Std D	1.304	0	1.489	**0.054**

previously used algorithms based on the time taken to label the unlabeled data. Time consumption can be basically defined as time taken by the particular technique for the completion of the process. Here the process of clustering is carried out on data with different size and dimension. As the size of data increases, more time will be required for the computational process. Table 3 gives the detail of time spent in the classification of iris data set for each and every run with different techniques. The value illustrates that SSFCM takes minimum time for every run as compared to other techniques. But SSFCM fails to classify the data completely into different classes and gives maximum error rate as compared to other techniques. Table 4 gives the detail of time spent in the classification of wine dataset and Table 5 gives the value for wireless indoor localization dataset. In each case the proposed technique SSFBC outperforms the other techniques of SVM, SSFCM+SVM in the terms of error rate and computational time.

4.3. Impact of threshold parameter e1

In this section, we have discussed how the proposed technique responds to the change in the value of parameter $e1$ i.e. how does and how much influence parameter $e1$ actually has on SSFBC. The parameter $e1$ is the threshold value for the membership of each point that is obtained using SSFCM for the unlabeled data prior to train the classifier. To observe the impact, we have calculated the error rate at different values of $e1$ on all the three datasets. It reflects that when the value of $e1$ is kept close to 0.6, it shows minimum deviation of error rate. Fig. 4 shows the variation in the error rate with varying values of $e1$. It depicts that when we accept only high value for the threshold, we observe an inconsistency in the change of error rate. But after a certain value, further decreasing the value of $e1$ i.e. degree of threshold, does not have any

Table 4. Computational time spent for labelling Wine Dataset

Selection	**Supervised SVM Classifier**	**SSFCM**	**Semi-supervised SSFCM + SVM classifier**	**Our proposed method SSFBC**
1	3.81	0.014	19.7	1.32
2	6.31	0.015	13.52	1.35
3	2.46	0.016	15.44	1.33
4	2.76	0.014	17.23	1.34
5	5.33	0.014	22.26	1.34
6	3.27	0.014	17.6	1.36
7	3.8	0.017	27.31	1.26
8	3.06	0.013	12.14	1.2
9	5.93	0.013	15.15	1.28
10	2.45	0.013	14.46	1.36
Mean	3.918	0.014	17.481	**1.314**
Std D	1.436	0.001	4.563	**0.052**

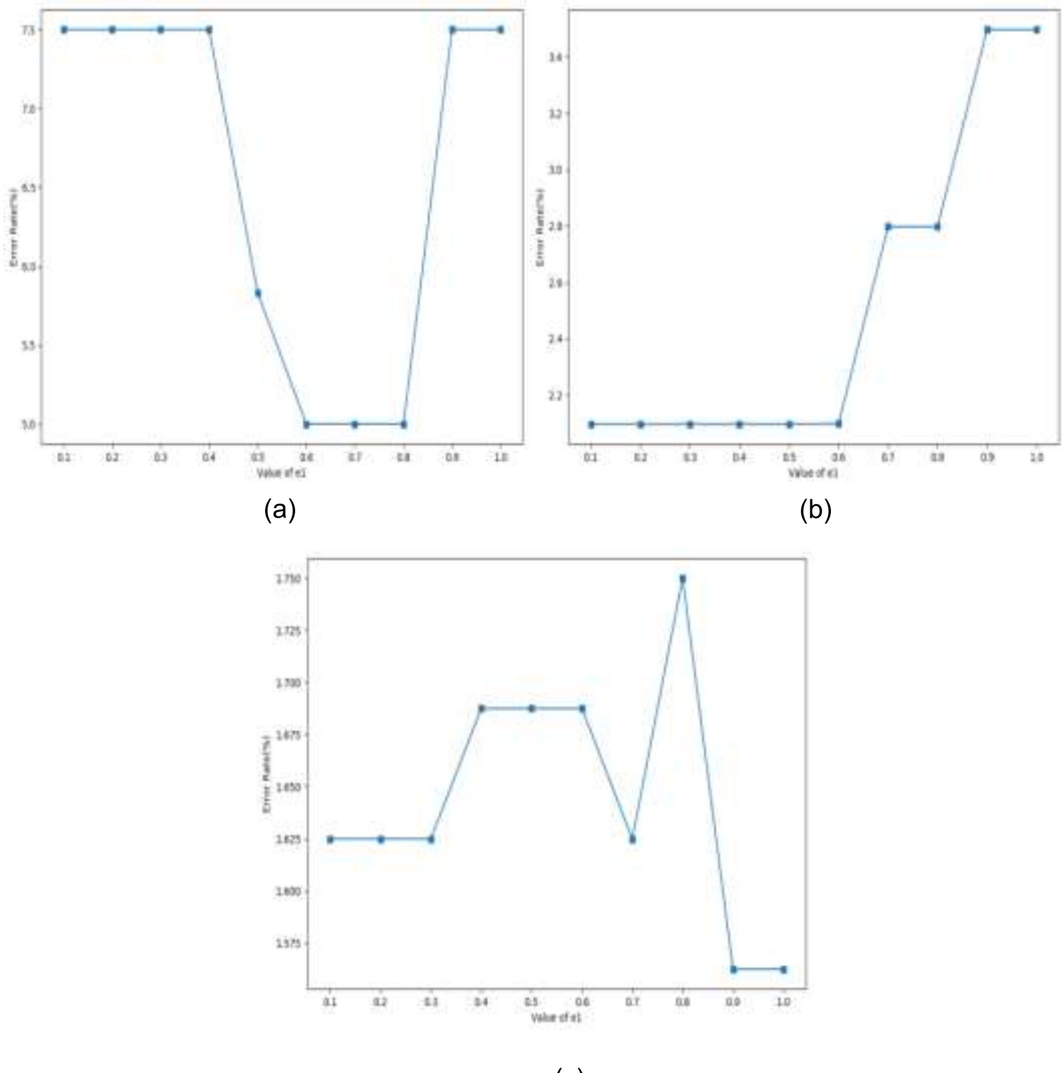
Table 5. Computational time spent for labelling Wireless Indoor Localization Dataset

Selection	**Supervised SVM Classifier**	**SSFCM**	**Semi-supervised SSFCM + SVM classifier**	**Our proposed method SSFBC**
1	99.69	0.1548	136.38	1.43
2	127.63	0.024	158.47	1.48
3	127.54	0.023	149.13	1.45
4	122.66	0.024	148	1.59
5	114.24	0.023	134.38	1.58
6	130.04	0.023	146.64	1.45
7	110.52	0.023	137.32	1.6
8	112.63	0.02	148.15	1.5
9	122.74	0.023	147.37	1.53
10	124.15	0.023	158.37	1.47
Mean	119.184	0.036	146.421	**1.508**
Std D	9.608	0.042	8.369	**0.063**

impact on the error rate of the final result. It reflects that when the value of $e1$ is kept close to 0.6, it shows minimum deviation of error rate.

5. Conclusion

In this paper, we discussed about integration of two different algorithms to produce a synergic effect as clustering provided us with better insights and help us train a better classifier. This was done to tackle the problem where we have lack of labeled data and our data is continuously growing. To do this we used SSFCM as clustering algorithm while Naïve Bayes to train the classifier as it provides accurate results in short amount of time. We carried out multiple experiments on different datasets and compared the results with different supervised/semi-supervised algorithms. In all the experiments, we observed that the error rate of the proposed algorithm SSFBC was exceptionally low as compared to other techniques. In case of time consumption, we observed that regardless of the different sizes in the datasets, the time consumed by the proposed algorithm was very less as compared to SVM and SSFCM+SVM classifier. Only SSFCM was able to give instant results but if we also include the error rate, our algorithm has clearly performed better. Hence the proposed technique proved to perform better in the terms of accuracy and effectiveness.



(a) (b)

(c)

Figure 4. Impact of e1 in terms of error rate on Real data (a) Iris (b) Wine (c) Wireless Indoor Localization

References

[1] Cao, Y., He, H. and Huang, H. (2011) A new framework of learning from testing data for face recognition, Neurocomputing 74: 916-929.

[2] Rish, I. (2001) An empirical study of the naive Bayes classifier, *Proceedings of IJCAI- workshop on Empirical Methods in AI* : 41—46.

[3] Duda, R. O., Hart, P.E. and Stork D. G. (2000) *Pattern Classification*, 2nd ed(Wiley)

[4] Kaur, P. and Sharma, M. (2019) Diagnosis of Human Psychological Disorders using Supervised Learning and Nature-Inspired Computing Techniques: *A Meta-Analysis. Journal of Medical Systems*: 43: 204.

[5] Sharma, M., Sharma, S. and Singh, G. (2018) Performance Analysis of Statitical and Supervised Learning techniques in Stock Data Mining. *MDPI Journal* 3(54): 1-16.

[6] Chapelle, O., Scholkopt, B. and Zien, A. (2006) *Semi-Supervised Learning* (MIT Press, Cambridge, MA)

[7] Basu, S., Banerjee, A. and Mooney, R. (2002) Semi-supervised clustering by seeding, *Proceedings of the Int. Conference on Machine Learning*:19–26.

[8] Demiriz, A., Bennett, K. and Embrechts, M. (1999) Semi-supervised clustering using genetic algorithms, *Intelligent Engineering Systems*: 809–814.

[9] Blum, A. and Chawla, S. (2001) Learning from labeled and unlabeled data using graph mincuts, *Proceedings in 18th International Conference in Machine Learning*.

[10] Gan, H., Sang, N., Huang, R., Tong, X. and Dan, Z. (2013) Using Clustering analysis to improve semi-supervised classification, *Neurocomputing:*290-298.

[11] Zhang, D., Tan, K. and Chen, S. (2004) Semi supervised Kernel Based Fuzzy C- Means, N.R. Pal et al. (Eds.): *ICONIP LNCS* 3316:1229-1234.

[12] Piekari, M., Salama, S., Mozes, S. N. and Martel, A. L. (2018) A Cluster-then-label Semi- supervise Learning Approach for Pathology Image Classification, *Scientific Reports* 8(7193): 1-13.

[13] Dunlop, M. M., Slepcev, D., Stuart, A. M. and Thorpe, M. (2019) Large data and zero noise limits of graph-based semi-supervised learning algorithms. *Applied and Computational Harmonic Analysis.* 1-43.

[14] Wang, F., Li, K., Zhou, L., Ren, H., Contreras, J., Shafie-Khan, M. and Catalao, J. P. S. (2019) Daily pattern prediction based classification modeling approach for day- ahead electricity price forecasting. International Journal of Electrical Power & Energy Systems. 105: 529-540.

[15] Bair, E. (2013) Semi-supervised clustering methods, *Wiley Interdiscip Rev Comput Stat.* 5(5): 349–361.

[16] Huanga, D. V., Morillo, P. and Ferri, C. (2017), Semi-Supervised Clustering Algorithms for Grouping Scientific Articles *Procedia Computer Science, Elsevier.* 108: 325-334.

[17] Sharma, M., Singh, G. and Singh, R. (2017) Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques. *Elsevier Manon IRBM* 38 : 305-324.

[18] Abney, S. (2007) Semisupervised Learning for Computational Linguistic. *Chapman & Hall/ CRC, Computer Science & Data Analysis.*

[19] Ruspini, E. H., Bezdek, J. C. and Keller, J. M. (2019) Fuzzy Clustering: A Historical Perspective. *IEEE Computational Intelligence Magazine* 14(1): 45-55.

[20] Pedrycz, W. and Waletzky, J. (1997) Fuzzy Clustering with partial supervision. *IEEE Trans. on Systems, Man, and Cybernetics, Part B-Cybernetics* 27 : 787-795

[21] Tong, X., Jiang, Q., Sang, N., Gan, H. and Zeng, S. (2009) The Feature weighted FCM algorithm with semi-supervised. *Proceedings of Eighth International Symposiumon Distributed Computingand Applications to Business, Engineering and Science* : 22-26.

[22] Bensaid, A. M., Hall, L. O. and Bezdek, J. C. (1996) Partial Supervised Clustering for Image Segmentation. *Pattern Recognition*, 29: 859-871.

[23] Zhang, H. *(*2004) The Optimality of Naive Bayes. *FLAIRS2004 conference.*

[24] Arora, J. and Tushir, M. (2017) A new kernel-based possibilistic intuitionistic fuzzy c-means clustering, *International Journal of Artificial Intelligence and Soft Computing*, 6(4): 306-325.