Iranian Journal of Fuzzy Systems

Volume 18, Number 3, (2021), pp. 161-177

Original paper

University of Sistan and Baluchestan

A bilateral fuzzy support vector machine hybridizing the Gaussian mixture model

M. Mohammadi[1] and M. Sarmad[2]

[1,2] *Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran*

manda.mohammadi@stu.um.ac.ir,sarmad@um.ac.ir

Abstract

The fuzzy support vector machine is one of the most exceptional methods to deal with uncertainty in the classification problem. The membership function is a proper way to model uncertainty. The goal of the membership function is to distinguish the different points in terms of their importance. The ordinary design of the membership function relies on the distance of the observations to the class center. However, the class center is affected by the presence of outliers. To prevent this effect, we utilized an unsupervised learning method called the Gaussian mixture model in the structure of the membership function. The proposed membership function is presented in two different categories distance-based and Bayes-based. Unlike the classical membership function, the contribution of outliers in the training phase decreased by diminishing their degree of importance. Hybridizing the classic fuzzy support vector machine classifier with the Gaussian mixture model will enhance the classification accuracy and also will prevent overfitting problems. The superiority of the proposed methods assessed by the synthetic and benchmarking dataset. The statistical significance is assessed by using the non-parametric Friedman and post-hoc Nemenyi tests.

Keywords: Support vector machine, outliers, noise, fuzzification, gaussian mixture model, distance-based membership function, bayes-based membership function.

1 Introduction

The support vector machine (SVM) is categorized as one of the supervised learning algorithms. It aims to find an optimal separating hyperplane with the maximum margin between classes. Despite SVM's ability to deal with problems such as small samples, non-linearity, high dimensions, and local minima, it suffers from some drawbacks. But the factor that causes the decline in its effectiveness is the impact of outlying points and the same consideration of all observations [3, 5, 11, 18, 19, 20, 25, 38, 39]. It does not make any distinction between the outliers and other data points. To get a clearer picture of the outlying observations one can refer to the definition provided by Grubbs [9] *"An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs"*. The existence of outliers within the data set always attracted a great deal of interest. The process of identification of these observations is known as "outlier detection".

In the classification problem, outliers are the data points located in a place that certainly cannot be assigned to a particular class. For example, one specific observation $\mathbf{x}_i$ most likely 90% belongs to one specific class. It means that this observation is of great importance in this class. Conversely, this observation 10% can be considered meaningless or less important in this respective class. However, the SVM model is created based on the subset of specific observations located on the closest distance to the margin called "support vectors". Contamination of these observations with the outliers and noise will lead to obtaining the inefficient model and subsidence of classification accuracy [11, 17, 19, 20, 26, 32, 33, 38, 39]. As can be deduced from the above statement, the impact of the training set on the learning process varies, in a way that some points are more important than others. In the other words, different points have different contribution to the learning process [31]. For example, support vectors that their correct classification is crucial. On

Corresponding Author: M. Sarmad

Received: July 2020; Revised: December 2020; Accepted: January 2021.

the other hand, there are also some insignificant observations in the classification process. Hence, equally treating all instances such as noise and outlying points may cause the overfitting problem [3, 19]. In this regard, designing a method to distinguish points in terms of their contribution to the learning process will be essential.

Fuzzy support vector machine (FSVM) that independently proposed by Lin and Wang [19] and Huang and Liu [20] can overcome the equal treating of the observations in the learning process. The basis of the FSVM in terms of the maximization of the margin is similar to the SVM. But there is a difference between the fuzzy perspective and classic SVM that prevents the margin from becoming narrower under the influence of outliers. This importance can be feasible by deploying a critical component of fuzzy logic called "membership function".

By using the membership function, allocation of specific credit to each observation and adjustment of their effect, depending on their degree of importance in the problem would be possible. The employment of the fuzzy theory in the SVM has inspired many researchers. An and Liang [3] proposed a new FSVM algorithm which incorporated minimum within-class scatter in the Fisher discriminant analysis into FSVM. Their proposed algorithm finds an optimal hyperplane in such a way that not only maximizes the margin but also minimizes within-class scatter. Another work has been done by Hang et al. [10] which is related to the multi-class FSVM classifier based on the one-against-other scheme. They deployed a kernel fuzzy c-means clustering algorithm to calculate fuzzy membership values of training samples for a multi-class FSVM classifier. An anti-noise performance algorithm has been introduced in the work of Yuan et al. [37]. They proposed a piecewise linear fuzzy weight computing method to overcome the outlier problem. They proposed a method in which the support vectors were given a higher value of membership degree than the samples far from the mean vector. Moreover, a new fuzzy membership function calculation method is proposed by Wang et al. [30]. They utilized a heuristic function that is derived from the centered kernel alignment for calculating the dependence between a data point and its associated label.

However, choosing the proper membership function is a fundamental step in the fuzzy support vector machine [19]. Generally, the construction of the fuzzy membership function drew on the Euclidean distance of the observations to the center of their respective classes. The farthest point to the center gains less weight; hence, it has less contribution to the training phase. However, the sensitivity of the center to the presence of outliers, force the center wrongly disposed toward them. Accordingly, as shown in our previous study [24], due to the masking effect, the regular points will move away from the center and outlying data will get closer. This phenomenon will lead to assigning low membership value to the ordinary points and high membership value to the outliers that do not deserve to. Thus, identification and elimination of the outlying points is a crucial preliminary step to prevent the overfitting problem. More details of FSVM have been given in Sect. 2.2. In the following subsection, a theoretical framework for the existence of outliers in the dataset will be used. Then we will describe our intentions to use Gaussian mixture model (GMM) clustering for the aim of outlier detection.

Theoretical framework for the presence of outliers

However, the presence of outliers in the data can be theoretically determined using one element of the robust statistic called the Huber's contamination model [13, 14]

$$F = (1 - \epsilon)\Phi + \epsilon H. \tag{1}$$

Based on this model, the observations follow a mixed distribution that is the combination of Φ and H distributions. Φ is the normal distribution, H can be any arbitrary distribution associated with the outlying points, and $\epsilon \in (0, 1)$ is the relevant fraction of outliers in the data. Inspired by this model, the presence of outliers in the data can be presented theoretically. This model is based on the belonging of data to two different sources.

On the other hand, outlier detection using an unsupervised approach is an adequate technique that the considerable research devoted to it [6, 12, 34, 40]. This technique aims to divide the data into two clusters. Because of the non-existence of the prior knowledge of the outliers, the cluster contains the fewer number of observations expected to be the outlier candidate [1, 8, 15, 16, 21, 27]. Also, the Gaussian mixture model, as an unsupervised method to model the occurrence of outliers in the data has been used by different researchers such as Aitkin [2], Zong [41], Yang [35] and Zimek [40]. According to Aitkin [2] the component with the larger prior probability represents the main bulk of data and another component with the lower prior probability represents the outliers. More details of GMM are given in Sect. 2.3.

The hybridization of the clustering and classification method is not something new in the field of machine learning. In some methods, the appropriate data center will be calculated using clustering methods to cluster each of the classes. In the next step, some classification methods are deployed to classify the data [28, 29, 31, 36]. Moreover, the merits of the fuzzy theory have been incorporated into either classification or clustering methods for increasing the generalization performance of the classifier. Yang et al, [36] developed a kernel fuzzy c-means clustering-based fuzzy SVM algorithm to

deal with the classification problems with outliers or noises. Wang et al, [28] proposed a novel grouped fuzzy SVM with sample space partition based on Expectation-Maximization. They used this technique for the detection of clustered microcalcification in mammography. In their proposed method, the diversified pattern of training data is partitioned into several groups based on the EM algorithm. Then a series of fuzzy SVM were used for classification. Wu et al, [31] implemented a novel partition index maximization (PIM) clustering method to get a more reasonable and robust fuzzy membership for fuzzy SVM. They improved the PIM clustering algorithm to cluster each of the two classes from the training set to get proper data centres.
In this research, we have fitted a two-component GMM to data to identify the outlying observations from the ordinary points. In comparison with some clustering methods such as K-means, GMM inherently belongs to the class of fuzzy clustering [4]. K-means clustering can be considered as the hard clustering technique that assigns the data to one and only one cluster deterministically. But, GMM allows the data to belong to each of the existing clusters with a certain probability. In other words, all the observations assigned to each cluster based on the membership degree defined by their maximum posterior probability. Due to this soft clustering manner of GMM, it can be categorized as one of the fuzzy-based unsupervised methods. Considering the above-mentioned points related to the fuzzy nature of GMM, its ability to identify outliers as well as designing the GMM-based membership function, it can be concluded that the fuzzification of SVM has been doubled. So, in this research, we have benefited from the bilateral fuzzy method. Based on the fuzzy structure of the proposed method, its efficiency is expected to be increased.
In this work, two different membership functions have been designed to fuzzify the SVM structure. As we have already mentioned, a two-component GMM clustering is used to identify the outlying observations from the ordinary points. The component with the larger prior probability represents the main bulk of data. This cluster will be called the "clean" cluster. The first membership function is based on the Euclidean distance of the points to the center of the clean cluster. In the second approach, the Bayes-based fuzzy membership will be designed based on the probability of assigning the observations to the clean cluster.
Note that, in the training phase of classification, all of the data points are getting involved and the respective degree of importance, according to the merit of each, will be awarded to them. Thus, the final decision will rely on all of the training set observations. But, the clean data are gaining more degree of importance, while the outliers received less. More explanation is given in Sect. 3. In summarize, the main contributions of this research are as follows.

- Improving the classic fuzzy support vector machine by designing a robust membership function.
- Designing a bilateral fuzzy membership function by the employment of a fuzzy-based (soft) clustering technique (GMM) to improve the generalizability of the SVM while preserving the merit of insensitivity to outliers.
  - Distance-based membership function designed based on the Euclidean distance of the points to the center of the clean cluster.
  - Bayes-based membership function designed based on the probability of assigning the observations to the clean cluster.

This paper is outlined as follows. In the next section, some preliminary concepts to fuzzify SVM are recalled. The GMM-based fuzzy membership functions are presented in Sect. 3. In Sect. 4, we apply our methods to a simple but illustrative toy-example and real data, then the experimental results are presented. The statistical significance is also assessed by using the non-parametric Friedman and post-hoc Nemenyi tests. Finally, the conclusion and future work are given in Sect. 5.

2 Some preliminary concept to fuzzify SVM

This section has been devoted to the brief exposition of the theory of support vector machine, fuzzy support vector machine, and Gaussian mixture model.

2.1 Basic SVM

The concept of binary classification problem is related to estimate the function $f : \mathbb{R}^p \to \{\pm 1\}$. This function classifies the unlabeled data based on training data. Let the training set is given as

$$S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}, \tag{2}$$

where $\mathbf{x}_i = \{x_{i1}, \ldots, x_{ip}\} \in \mathbb{R}^p$ and $y_i \in \{-1, +1\}$. This method segregates the negatively labeled observations from the positively labeled observations based on the separating hyperplane with the largest margin. The hyperplane is

determined based on the normal vector $\mathbf{w} \in \mathbb{R}^p$ and bias term $b \in \mathbb{R}$. The optimal hyperplane can be obtained based on solving the optimization problem as follows,

$$\min_{\mathbf{w},\xi,b} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i. \tag{3}$$

$$\text{s.t.} \qquad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i; \qquad \xi_i \geq 0; \qquad i = 1, \ldots, n.$$

In this equation, the slack variable ξ_i (for $i = 1, \ldots, n$) is used to measure the amount of violation of the constraints. The penalty parameter C plays the adjuster role to make a trade-off between classification error and expansion and shrinkage of the size of the margin. The dual form of Lagrangian is used to solve the problem.

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{n}\boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{4}$$

$$\text{s.t.} \qquad \sum_{i=1}^{n}\boldsymbol{\alpha}_i y_i = 0; \qquad 0 \leq \boldsymbol{\alpha}_i \leq C; \qquad 1 \leq i \leq n.$$

Here, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ is the vector of non-negative Lagrange multipliers and $K(x_i, x_j) =< \phi(x_i), \phi(x_j) >$ is the kernel function which is the inner product in the higher-dimensional space. This function is used to transform the data from feature space to the higher-dimensional space to make them linearly separable in the new space. Note that, for each training sample, there would be a corresponding Lagrange multiplier. However, the non zero $\boldsymbol{\alpha}_i$ are only belong to the support vectors. The optimal hyperplane can be obtained based on the results of the above mentioned quadratic optimization problem. The hyperplane equation is

$$f(x_i) = \sum_{j \in SV}\alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i) + b. \tag{5}$$

The set of support vectors are indicated by SV.We will classify the future observation based on the classifying function as

$$f(x_i) = \text{sign}(\sum_{j \in SV}\alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i) + b). \tag{6}$$

2.2 Fuzzy SVM

Unlike classical SVM, FSVM assigns the degrees of importance to the observations based on their merit, in such a way that it reduces the outliers' impact on the placement of SVM hyperplane by assigning less membership value to them. Based on this function, the degree of importance associated with each observation or in other words, the degree of belonging of each point to its corresponding class is determined. The key difference between classical SVM and FSVM is that the contribution of all training points to the total error term is considered equally in SVM. Conversely, in the FSVM different points make a distinct contribution to the learning process. Moreover, the other feature that distinguishes FSVM from SVM is that it does not take into account the crisp belongingness of the observations to a particular class. While in the classic SVM, the points strictly belong to one class.
According to Lin and Wang [19], the rewritten of the SVM classification problem to the FSVM format using the membership function is as follows.

$$S_f = \{(\mathbf{x}_1, y_1, \mu_1), (\mathbf{x}_2, y_2, \mu_2), \ldots, (\mathbf{x}_n, y_n, \mu_n)\}, \tag{7}$$

Let the training set S_f contains an extra component videlicet membership function μ in addition to the previous components. Membership of the observation (μ) can be embedded in the formulation of SVM as the add-on function. It is designed to be $\delta \leq \mu_i \leq 1$. Where δ is an arbitrarily small value to prevent the membership value to be zero. Similar to the classic SVM, the primal form of the optimization problem is

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i\mu_i, \tag{8}$$

$$\text{s.t.} \qquad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i; \qquad \xi_i \geq 0; \qquad i = 1, \ldots, n.$$

The main difference between SVM and FSVM is $\xi_i\mu_i$. We can consider μ_i as the weighting factor for the error term because ξ_i is the measure of the error in the problem. The dual form of FSVM is very similar to the SVM except for the upper bound of the Lagrange multipliers,

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j,$$

$$\text{s.t.} \qquad \sum_{i=1}^{n} \boldsymbol{\alpha}_i y_i = 0; \qquad 0 \le \boldsymbol{\alpha}_i \le C\boldsymbol{\mu}_i; \qquad 1 \le i \le n.$$

It is worth mentioning that, by tuning the value of the membership function, the effect of each training datum can be increased or decreased.

2.3 Gaussian mixture model

Gaussian mixture model as a parametric probability density function represented by the linear combination of Gaussian densities with the following form,

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k), \tag{9}$$

where $\mathbf{x} \in \mathbb{R}^p$, π_k is the mixture weights or in other words the prior probability of $\mathbf{x}_i$ belonging to cluster k (for $k = 1, \ldots, K$ components) and N is the p variate Gaussian density of each component with the form of

$$N(\mathbf{x}|\boldsymbol{\mu}_k, \sigma_k) = \frac{1}{\sqrt{2\pi\boldsymbol{\sigma}_k^2}} \exp\{-\frac{1}{2\boldsymbol{\sigma}_k^2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^2\}. \tag{10}$$

$\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$ the mean vector and covariance matrix, respectively. The mixture weights must satisfy $0 \le \pi_k \le 1$ and also $\sum\limits_{k=1}^{K} \pi_k = 1$. The collection of the parameters of this model that are parameterized by the mean, covariance, and mixture weight present by $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}_{k=1}^{K}$. In addition to the set of observations, the random vector $\mathbf{Z}$ has also existed. Note that, $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$, in which z_{ik} is the Bernoulli random variable indicts that $\mathbf{x}_i$ has drawn from the kth Gaussian distribution. Moreover, $z_{ik} \in \{0, 1\}$ and $\sum\limits_{k} z_{ik} = 1$.

The parameters of this model can be estimated from the training data by the Expectation-Maximization (EM) algorithm. This algorithm is an iterative method with the aim of numerically approximate the maximum likelihood of the data which can be obtained by alternately updating $p(\mathbf{z}_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})$ and $\boldsymbol{\theta}^{(t+1)} = \{\pi_k^{(t+1)}, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\sigma}_k^{(t+1)}\}_{k=1}^{K}$ with an initial estimate $\boldsymbol{\theta}^{(0)}$. The updated equation of $p(\mathbf{z}_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})$ is written as

$$p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(z_{ik} = 1)p(\mathbf{x}_i|z_{ik} = 1)}{\sum\limits_{j=1}^{K} p(z_{ij} = 1)p(\mathbf{x}_i|z_{ij} = 1)} = \frac{\pi_k N(\mathbf{x}_i|\mu_k, \sigma_k)}{\sum\limits_{j=1}^{K} \pi_j N(\mathbf{x}_i|\mu_j, \sigma_j)} = \gamma(z_{ik}), \tag{11}$$

and the updated equations of $\boldsymbol{\theta}^{(t+1)}$ are

$$\pi_k^{(t+1)} = \frac{\sum\limits_{i=1}^{n} \gamma(z_{ik})}{n}, \tag{12}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum\limits_{i=1}^{n} \gamma(z_{ik})\mathbf{x}_i}{\sum\limits_{i=1}^{n} \gamma(z_{ik})}, \tag{13}$$

and

$$\boldsymbol{\sigma}^{2(t+1)}_k = \frac{1}{\sum\limits_{i=1}^{n} \gamma(z_{ik})} \sum_{i=1}^{n} \gamma(z_{ik})(\mathbf{x}_i - \boldsymbol{\mu}_k)^2. \tag{14}$$

The procedure is iterative, starting at some initial value for the parameters and updating the values in each iteration. It terminates when the last two values of the log-likelihood computed by

$$l(\boldsymbol{\theta}) = \log \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log\{\sum_{k=1}^{K} \boldsymbol{\pi}_k N(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)\}, \tag{15}$$

are close enough or the number of iterations reaches the predetermined number.

3 Designing the GMM-based membership function

Generally, in the field of fuzzy SVM, the structure of the membership function is based on the Euclidean distance of each observation to the relevant class center. Nonetheless, the sensitivity of the center to some outlying points is a stated reality proved by different researchers [23, 22]. To rectify this difficulty, outlier detection and isolation is essential.
The main idea of GMM as an unsupervised learning algorithm is to display the distribution of each subset as a cluster. Therefore, using this feature of GMM, one distribution is assigned to the clean cluster and another distribution to the cluster of outliers. This method used to create a distinction between data in such a way that, the "clean" data forms a large cluster and the outliers are located in a smaller cluster.
the separation of data into two clusters of clean data and outliers has been done using posterior probability or the degree of belongingness to the clean cluster based on Equation (11). In other words, for the observation $\mathbf{x}_i$, the vector of the probability of belongingness is $\boldsymbol{\gamma}_{z_{ik}} = (\gamma_{z_{i1}}, \gamma_{z_{i2}}, \ldots, \gamma_{z_{ik}})$ in which, $\sum_{k=1}^{K} \boldsymbol{\gamma}_{z_{ik}} = 1$ and $0 < \boldsymbol{\gamma}_{z_{ik}} < 1$. Due to the fitting of the two-components GMM to the data, the probability vector will have only two values $\boldsymbol{\gamma}_{z_{ik}} = (\gamma_{z_{i1}}, \gamma_{z_{i2}})$ which indicates the probability of belongingness of observation $\mathbf{x}_i$ to each of the two clusters (clean and outliers clusters) as follows,

$$\gamma_{z_{i1}} = \frac{\pi_1 N(\mathbf{x}_i|\mu_1, \sigma_1)}{\sum_{j=1}^{2} \pi_j N(\mathbf{x}_i|\mu_j, \sigma_j)}, \tag{16}$$

and

$$\gamma_{z_{i2}} = \frac{\pi_2 N(\mathbf{x}_i|\mu_2, \sigma_2)}{\sum_{j=1}^{2} \pi_j N(\mathbf{x}_i|\mu_j, \sigma_j)}. \tag{17}$$

Here the clean cluster is shown with index 1 as (C_1) and the cluster of outlying data with index 2 as (C_2)

$$C_1 = \{\mathbf{x}_i| \ \ \gamma_{z_{i1}} > \gamma_{z_{i2}}\}, \tag{18}$$

$$C_2 = \{\mathbf{x}_i| \ \ \gamma_{z_{i2}} > \gamma_{z_{i1}}\}. \tag{19}$$

For the visual intuition, we showed the representation of the data of one specific class in Figure 1. This figure illustrates how GMM clusters the data into two clusters of clean and outliers.

3.1 Distance-based membership function

The challenge of SVM's sensitivity to the presence of outliers addressed by not the same consideration of all observations in the learning process. It is possible by using the GMM-based membership function. In this way, to reduce the impact of the outliers, the clean cluster center is used as the class center in the membership function which is introduced by Lin and Wang [19]. In the proposed method the center of the clean cluster of each class $\bar{\mathbf{x}}_{C_{1\pm}}$ is used instead of the class center. The radius of each class $r_{C_{1\pm}}$ can also be computed using the corresponding clean cluster center.
It should be noted that, if there are not any outliers in the data, nearly half of the observations are located within one cluster by the GMM and the rest, in the other cluster. In this case, observations of smaller cluster gain less contribution to the learning process. Therefore, they will not have much impact that will lead to reducing the performance of the learning model. To avoid this, the ratio of the small cluster to the total data is considered as $R_{\pm} = \frac{C_{2\pm}}{C_{1\pm}+C_{2\pm}}$. If this ratio is less than δ^* (the pre-specified arbitrary percentage of outliers in the data), the degree of importance will be the relevant membership value $S_{\pm}$ (as explained in Equation (21)). This amount is the Euclidean distance of each data to the cluster center of the clean data. Otherwise, the observations with a low degree of importance will gain

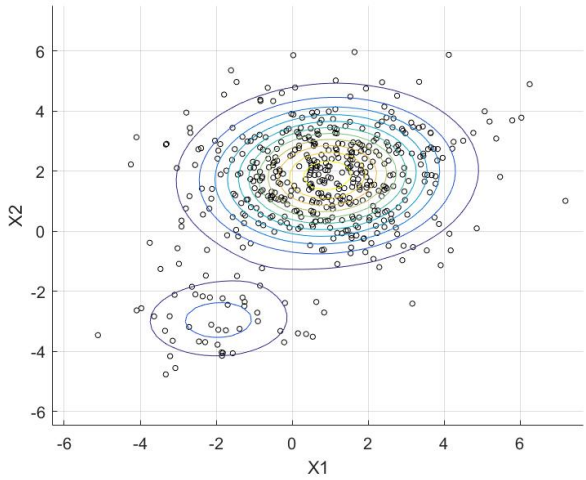

Figure 1: GMM Clustering of one specific class data using two-component Gaussian mixture model

the membership value equivalent to the maximum of the original membership value $S_{i\pm}$ and $\bar{S^*}_{\pm}$. Where $\bar{S^*}_{\pm}$ is the average of membership values greater than 0.5 as

$$\bar{S^*}_{\pm} = \frac{1}{\sum_i I_{S^*_{\pm}}(S_{i\pm})} \sum_{S_{i\pm} \in S^*_{\pm}} S_{i\pm}, \tag{20}$$

in which $S^*_{\pm} = \{S_{i\pm} | S_{i\pm} > 0.5\}$ and

$$I_{S^*_{\pm}}(S_{i\pm}) = \begin{cases} 1 & \text{for} \quad S_{i\pm} \in S^*_{\pm} \\ 0 & \text{for} \quad S_{i\pm} \notin S^*_{\pm}. \end{cases}$$

In other words, the loss of clean data is not desirable, as it is not required to enter the outliers into the learning process. Considering these conditions, the membership function based on GMM is as follows,

$$\boldsymbol{\mu}_{DistGMM}(\mathbf{x}_i) = \begin{cases} S_{i+} = 1 - \frac{\|\bar{\mathbf{x}}_{C_{1+}} - \mathbf{x}_i\|}{(r_{C_{1+}} + \delta)} & \text{for} \quad \mathbf{x}_i \in M_+ \& R_+ \leq \delta^* \\ S_{i+} = \max(S_{i+}, \bar{S^*}_+) & \text{for} \quad \mathbf{x}_i \in M_+ \& R_+ > \delta^* \\ S_{i-} = 1 - \frac{\|\bar{\mathbf{x}}_{C_{1-}} - \mathbf{x}_i\|}{(r_{C_{1-}} + \delta)} & \text{for} \quad \mathbf{x}_i \in M_- \& R_- \leq \delta^* \\ S_{i-} = \max(S_{i-}, \bar{S^*}_-) & \text{for} \quad \mathbf{x}_i \in M_- \& R_- > \delta^*. \end{cases} \tag{21}$$

In the membership function $\boldsymbol{\mu}_{DistGMM}$, $M_{\pm}$ are positive and negative labeled classes and, δ^* considered to be equal to 0.2. It can arbitrarily be considered to be less than 0.5.

3.2 Bayes-based membership function

The second approach to tackle outliers and noisy data, is using the possibility of belonging to the clean cluster of each class. For this purpose, we considered the "probability of belonging" such that all of the observations of each class gain different importance degree, based on their merit.

The degree of belongingness of the observations of a particular class to the respective clean cluster ($\gamma_{z_{i1}}$ based on Equation (16)) is shown in Figure 2. The purpose of this figure is the illustration of the distinction between clean data and outliers using the probability of belongingness. The degree of importance of each observation is calculated depending on their proximity to the clean cluster. In this way, observations with a high probability of belongingness to the clean cluster showed by dark-colored red and observations with a high probability of belongingness to the outlier cluster illustrated by dark blue color. Whatever the chance of belongingness to the clean cluster of data is less, red darkness is reduced to a similar extent. As a result, based on the probability of belongingness to each of the clusters, the observations on the boundary of these two clusters are placed in the color spectrum between these two colors.

For the design of the membership function, we use the degree of belongingness of data to the clean cluster ($\gamma_{z_{i1}}$ based on

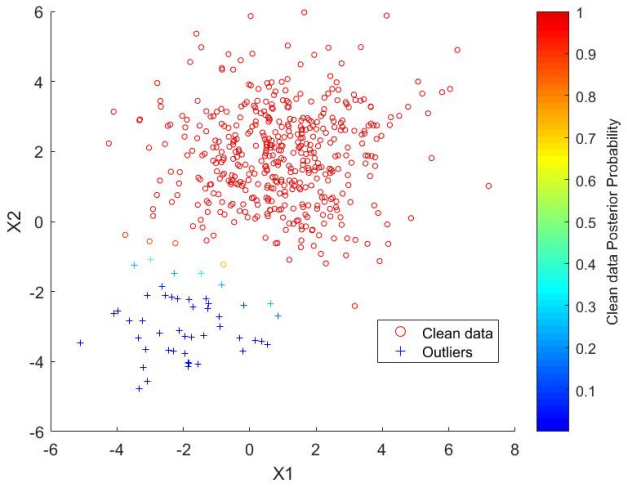

Figure 2: The degree of belongingness of observation of one class to the clean cluster based on the value $\gamma_{z_{i1}}$

Equation (16)) , which indicates the importance of each of the observations based on their merit. The construction of the Bayes-based membership function is similar to the distance-based membership function. Based on the above-mentioned concept, the Bayes-based membership function is designed as follows,

$$\boldsymbol{\mu}_{ProbGMM}(\mathbf{x}_i) = \begin{cases} \gamma_{z_{i1}i+} = \frac{\pi_{1_+} f(x_i|\mu_{1_+},\sigma_{1_+})}{\pi_{1_+} f(x_i|\mu_{1_+},\sigma_{1_+})+\pi_{2_+} f(x_i|\mu_{2_+},\sigma_{2_+})} & \text{for} \quad \mathbf{x}_i \in M_+ \& R_+ \leq \delta^* \\ \gamma_{z_{i1}i+} = \max(\gamma_{z_{i1}i+}, \bar{\gamma^*}_{z_{i1}+}) & \text{for} \quad \mathbf{x}_i \in M_+ \& R_+ > \delta^* \\ \gamma_{z_{i1}i-} = \frac{\pi_{1_-} f(x_i|\mu_{1_-},\sigma_{1_-})}{\pi_{1_-} f(x_i|\mu_{1_-},\sigma_{1_-})+\pi_{2_-} f(x_i|\mu_{2_-},\sigma_{2_-})} & \text{for} \quad \mathbf{x}_i \in M_- \& R_- \leq \delta^* \\ \gamma_{z_{i1}i-} = \max(\gamma_{z_{i1}i-}, \bar{\gamma^*}_{z_{i1}-}) & \text{for} \quad \mathbf{x}_i \in M_- \& R_- > \delta^*. \end{cases} \tag{22}$$

Here $\bar{\gamma^*_{z_{i1}\pm}} = \frac{1}{\sum_i I_{\gamma^*_{z_{i1}\pm}}(\gamma_{z_{i1}i\pm})} \sum_{\gamma_{z_{i1}i\pm} \in \gamma^*_{z_{i1}\pm}} \gamma_{z_{i1}i\pm}$ in which $\gamma^*_{z_{i1}\pm} = \{\gamma_{z_{i1}i\pm} | \gamma_{z_{i1}i\pm} > 0.5\}$ and

$$I_{\gamma^*_{z_{i1}\pm}}(\gamma_{z_{i1}i\pm}) = \begin{cases} 1 & \text{for} \quad \gamma_{z_{i1}i\pm} \in \gamma^*_{z_{i1}\pm} \\ 0 & \text{for} \quad \gamma_{z_{i1}i\pm} \notin \gamma^*_{z_{i1}\pm}. \end{cases}$$

3.3 Algorithmic scheme of GM-based FSVM

To clarify the subject, the algorithmic scheme of FSVM based on distance and Bayes scheme (posterior probability) summarised in the algorithm (1) and (2).

Algorithm 1 Algorithm outline for the GMM distance-based FSVM

Input: A data matrix $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, the labels vector $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$, the parameters δ and δ^*
Output: A GMM distance-based FSVM
Algorithm:

1. Clustering the data using GMM and obtaining the centers $\bar{\mathbf{x}}_{C_{1\pm}}$ and radius of each class $r_{C_{1\pm}}$
2. If the ratio of the observations of the small cluster to the whole observations of the respective class is less than δ^*, $\boldsymbol{\mu}_{DistGMM} = S_\pm$ otherwise $\boldsymbol{\mu}_{DistGMM} = \max(S_\pm, S^*_\pm)$ where $S_\pm$ and $S^*_\pm$ can be computed using Equation (21)
3. Train FSVM for the negative and positive labeled observations with S_- and S_+ membership functions, respectively.

Algorithm 2 Algorithm outline for the GMM Bayes-based FSVM

Input:A data matrix $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, the labels vector $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$, the parameters δ and δ^*
Output: A GMM Bayes-based FSVM
Algorithm:

1. Clustering the data using GMM and obtaining the posterior probability
2. If the ratio of the observations of the small cluster to the whole observations of the respective class is less than δ^*, $\boldsymbol{\mu}_{ProbGMM} = \gamma_{z_{i1\pm}}$ otherwise $\boldsymbol{\mu}_{ProbGMM} = \max(\gamma_{z_{i1\pm}}, \gamma^*_{z_{i1\pm}})$ where $\gamma_{z_{i1\pm}}$ and $\gamma^*_{z_{i1\pm}}$ can be computed using Equation (22)
3. Train FSVM for the negative and positive labeled observations with $\gamma_{z_{i1-}}$ and $\gamma_{z_{i1+}}$ membership functions, respectively.

4 Empirical study

In this section, the performance of the proposed methods using synthetic and real data is assessed. The proposed algorithms have been compared with different algorithms in terms of accuracy, F-measure and, training time. These algorithms are classic SVM, fuzzy SVM based on the classic Euclidean distance (FSVM) proposed by Lin and Wang [19], alternative fuzzy membership function using the classic Mahalanobis distance (FSVM-MAH) and, three robust fuzzy SVM based on the robust Euclidean distance (RFSVM-EUC) and robust Mahalanobis distance (RFSVM-MAHMCD and RFSVM-MAHOGK) proposed by Mohammadi and Sarmad [24]. All computations are performed in Matlab R2016a on a PC running on Windows 10 with 64 bit with a 2.90 GHz CPU and 12.0 GB of RAM. A 10-folds cross-validation utilized to assess the performance of the model. This strategy is repeated ten times, so that each dataset is randomly divided into ten parts. In each repetition, nine parts of the ten parts are considered as a training set and the remainder is used as the test set. For the sake of obtaining stable results, the experiment has been iterated 50 times. The mean of classification accuracy and F-measure with the standard deviation for 50 iterations on the testing data are reported in Tables 4 and 5. The bold numbers denote the best performance of these methods on each data set. Our proposed method similar to the traditional SVM is designed for the binary classification problem; but for the multi-class cases, the one-vs-one scheme has been used. In this method, every problem of the M class is divided into $M(M-1)/2$ binary problem in which all two permutations of the classes are present. Finally, to label the test data, the majority vote of all binary classifiers is used.
It is worth mentioning that, the observations of the classes of synthetic and real data are divided by GMM into two clusters of clean data and outliers. By using this division, distance-based and Bayes-based membership functions are designed. For the synthetic case, both classes are contaminated by the outliers. Moreover, the benchmark datasets have been used without any preprocessing, although all attributes are numerical, either integer or real values.

4.1 Experiments on toy dataset

The efficiency of proposed algorithms tested on synthetic datasets. This subsection is provided to enhance the visual understanding of the proposed methods. The following simulation setting is arbitrarily chosen to describe the performance of the proposed membership functions and also for illustration purposes.
The synthetic dataset contains 800 observations from the bivariate normal distribution. This dataset divided into two classes. Each of them includes 400 observations with positive and negative labels. The mean vector of clean positive and clean negative class are respectively $\boldsymbol{\mu}_+ = (2, 1)$, $\boldsymbol{\mu}_- = (-3, 0)$ and covariance matrix of clean positive and clean negative class are $\boldsymbol{\Sigma}_+ = \boldsymbol{\Sigma}_- = \text{diag}(1, 2)$. Also, the mean vector of outliers of positive and negative class which are marked by the $*$ are $\boldsymbol{\mu}^*_+ = (-4, -4)$ and $\boldsymbol{\mu}^*_- = (3, 5)$ and the covariance matrix are $\Sigma^*_+ = \Sigma^*_- = \text{diag}(0.5, 2)$.

The classification accuracy of distance-based and Bayes-based FSVM membership (that henceforth will be shown with FSVM-DistGMM and FSVM-BayesGMM) for the linearly separable data with different values of C parameter is reported in Table 1.

The general structure of the synthetic data has been shown in Figure 3. As can be seen from Figure 3, each class is clustered into two clusters of clean data and the outliers by the usage of GMM. The classical SVM discriminator tries to classify the data as accurately as possible. It will cause the overfitting problem and reduction of the classification accuracy. While the proposed methods, by down-weighting the outliers produces the discriminator hyperplane that is not sensitive to the outliers. It does not intend to classify all the training data as accurately as possible and has a higher generalization ability. Therefore, in the proposed methods, higher generalization ability is given more importance than the correct classification of the training data points, including outliers.

Table 1: Mean(standard division) of the classification accuracy of SVM, FSVM, FSVM-DistGMM and FSVM-BayesGMM for different values of C parameter in the range $[10^{-3}, 10^{3}]$ for the synthetic data.

C	SVM	FSVM	FSVM-DistGMM	FSVM-BayesGMM
10^{-3}	94.07(0.05)	95.63(0.04)	97.76(0.05)	97.70(0.02)
10^{-2}	87.50(0.07)	87.91(0.08)	97.17 (0.04)	97.61(0.03)
10^{-1}	84.13(0.04)	87.58(0.05)	93.75(0.07)	94.50(0.04)
10^{0}	81.66(0.06)	85.50(0.012)	91.78(0.02)	91.78(0.05)
10^{1}	83.33(0.04)	88.75(0.03)	91.25(0.02)	92.08(0.06)
10^{2}	81.66(0.07)	82.50(0.18)	90.47(0.04)	91.58(0.11)
10^{3}	87.08(0.01)	91.67(0.05)	96.91 (0.08)	97.91(0.04)

Concerning this visual example, the employment of the outliers resistance classifiers seems crucial. Fuzzy SVM can achieve this goal.

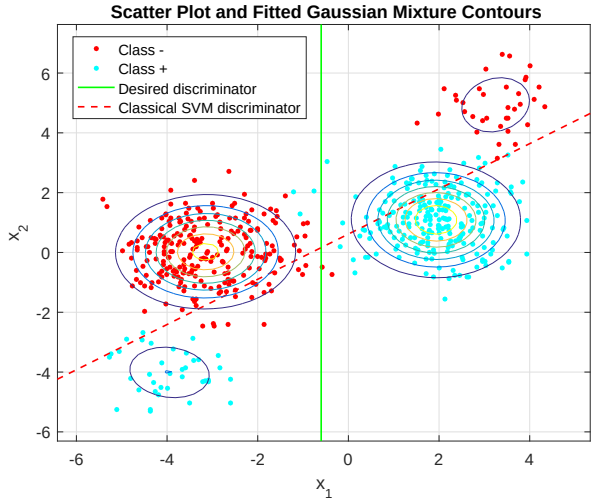

Figure 3: Conceptual example showing the division of classes into two clusters of clean data and outliers using GMM and also showing the difference between classical SVM discriminator and the desired discriminator.

4.2 Experiments on benchmark datasets

In addition to the synthetic dataset, several real data are examined to evaluate the efficiency of the proposed methods. Most of the data are taken from the UCI machine learning repository [1]. The brief information of these data is shown in Table 3. The accuracy and F-measure value of FSVM-DistGMM and FSVM-BayesGMM are illustrated in Tables 4 and 5. Moreover, these methods compared with the classic SVM, fuzzy SVM, RFSVM-EUC, FSVM-MAH, RFSVM-MAHMCD and, RFSVM-MAHOGK. The values reported below the accuracy and F-measure values, are the ranks among all methods. The mean ranks of all methods are shown in the last row. From Tables 4 and 5, it can be concluded that the proposed FSVMs yield better or comparable classification results compared with the classical SVM and the other FSVM variants classifier. In other words, the FSVM algorithms based on appropriate fuzzy memberships can indeed improve the classification performance.

It should be noted that obtaining the highest performance criteria such as accuracy or F-measure is not conclusive to consider a method as the best classifiers. It should be verified whether or not the improvements made by the proposed methods are statistically significant. So, the Friedman test is deployed to investigate whether the differences between the accuracies and the F-measure of the algorithms are statistically meaningful or not. After that, the Nemenyi post-hoc test is used to confirm the superiority of the proposed methods over the other algorithms in predicting the class labels. These tests are some useful non-parametric tests recommended by Ďemsar [7].

For the comparison of the performance of the different classifiers, the non- parametric Friedman test has been utilized. This test is based on the performance rating of different algorithms on a specific dataset. Based on this test, the best

[1] http://archive.ics.uci.edu/ml/index.php

performing algorithm will obtain the rank of 1, the second-best rank 2, and so on. In the case of ties, the average ranks will be assigned. The Friedman statistic χ_F^2 will be calculated based on

$$\chi_F^2 = \frac{12N_d}{N_e(N_e+1)}\left[\sum_e R_e^2 - \frac{N_e(N_e+1)^2}{4}\right], \tag{23}$$

in which, R_e is the average rank of the e-th algorithm, N_e is the total number of classifiers (in our case 8) and, N_d is the total number of data sets (in our case 20). This statistic is distributed as χ^2 with $N_e - 1$ degree of freedom where N_d and N_e are reasonably large (i.e. $N_d > 10$ or$N_e > 5$). At the significant level of 5%, the null hypothesis will be rejected when the value of the χ_F^2 statistic is higher than the critical value from the χ^2 distribution with $N_e - 1$ degree of freedom. The null hypothesis is based on the equivalence of the performance of classifiers.
In our case, the Friedman statistic is 71.17 and, the critical value of χ^2 with 7 degrees of freedom and with a significance level of 5% is 14.07. Since the critical value is smaller than the Friedman statistic of our results, we reject the null-hypothesis, meaning that the algorithms are statistically different.
In the case of the rejection of the null hypothesis, we can proceed with a post-hoc Nemenyi test to analyse whether or not our proposed algorithms are significantly better than each of the other algorithms. The Nemenyi's test stated that the performance of two classifiers is significantly different if their ranks differ by at least the critical difference (CD) which is defined as,

$$CD = q_\alpha \times \sqrt{\frac{N_e(N_e+1)}{6N_d}}. \tag{24}$$

Where q_α is the critical values based on the Studentized range statistic divided by $\sqrt{2}$ [7]. The critical values are given in Table 2 for the Nemenyi test which is taken from Ďemsar [7].

Table 2: Critical values for the two-tailed Nemenyi test.

No. of classifiers	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164
$q_{0.10}$	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

The critical value for eight classifiers and with a significance level of 5% is 3.031 and, therefore, we have $CD = 3.031 \times \sqrt{\frac{8\times(8+1)}{6\times 20}} = 2.347803$. Figure 4 illustrates the CD diagram for the post-hoc Nemenyi test on the accuracy of different classifiers with a significance level of 5%. In this figure, the visual representation of the superiority of classifiers is demonstrated as a bar graph that its corresponding values are proportional to the mean rank obtained from each method. The CD is highlighted by a thicker horizontal line in black color. The algorithms that are not connected by a red line of length equal to CD have significantly different mean ranks with a confidence level of 95%. Our proposed

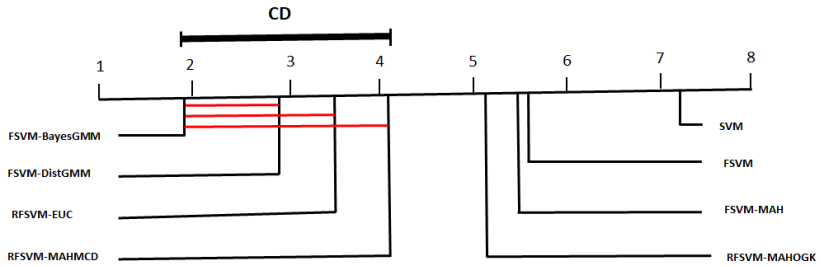

Figure 4: Critical Difference (CD) diagram for the Nemenyi test showing the results of the statistical comparison of all models against each other by mean ranks based on accuracy values.

FSVM-BayesGMM is having the least rank based on the Friedman test results as shown in the last row of Tables 4 and 5. It is observable that the other proposed algorithm FSVM-DistGMM also performs better in comparison to the other methods. It fortifies the fact that despite the superiority of our proposed methods confirmed by Friedman and Nemenyi test, therefore we can not claim that our proposed methods are better than RFSVM-MAHOGK with 95%

confidence as the difference in their average ranks is small.
The same conclusion also can be drawn from the results of Table 5 for the F-Measure. More importantly, the proposed FSVM-BayesGMM and FSVM-DistGMM consistently achieves the overall best classification performance to the other baseline approaches.
Training time is the other momentous criterion in evaluating the performances of classification algorithms. In Table 6 the training time of our proposed algorithms and the other six methods are illustrated. From this table, it is clear that our proposed FSVM-DistGMM and FSVM-BayesGMM takes more computation time than the SVM. It is due to the additional computation for calculating the fuzzy membership values in the proposed approach. Although the computation of the membership function takes a longer time for fuzzy methods; but the computation time of the proposed methods is higher than the FSVM, as an extra computation has been carried out for the GMM clustering in the proposed method. In comparison to the SVM, this difference in training time of fuzzy GMM-based SVM has not lead to the reduction of the generalization ability of it. This reduction does not mean that this method is not applicable to the conventional SVM systems.

Table 3: Characteristics of the benchmark data sets

Name	No. of examples	No. of attributes	No. of classes	Original data set
PID	768	8	2	Pima Indians diabetes
Biomed	209	4	2	Biomedical data set
Heart	270	13	2	Statlog-Heart
BUPA	345	6	2	Liver disorders data set
Wine	178	13	3	Wine
Iris	150	4	3	Iris
Breast cancer	699	9	2	Breast cancer Wisconsin (original)
Ionosphere	351	33	2	Ionosphere
Sonar	208	60	2	Sonar
WDBC	569	30	2	Breast cancer Wisconsin (diagnostic)
Saheart [2]	462	9	2	South African hearth data set
Blood	748	5	2	Blood Transfusion
PlanningRelax	182	12	2	Planning relax data set
Spectf	267	44	2	SPECTF heart data set
Vowel	990	13	11	Vowel Recognition
Haberman	306	3	2	Haberman's survival data set
Parkinsons	197	23	2	Parkinson's disease
Ecoli	336	7	8	Ecoli data set
Phoneme [3]	5404	5	2	Phoneme data set
Segment	2310	19	7	Image segmentation data set

5 Conclusion and future work

One of the unsupervised clustering algorithms, the Gaussian mixture model was used to construct the membership function. Since this algorithm is inherently a fuzzy clustering method, it can be deduced that SVM is fuzzified bilaterally. Our proposed method is important in terms of using a soft clustering method within a fuzzy classification method that leads to increasing the classification accuracy. In this paper, we use two different methods for designing the membership function, which includes the membership function based on the Euclidean distance and the membership function based on the posterior probability or Bayes scheme. The performance of the proposed methods, classic SVM, fuzzy SVM based on the classic Euclidean distance (FSVM) proposed by Lin and Wang [19] and other fuzzy SVM method based on the classic Mahalanobis distance (SVM-MAH) and three robust fuzzy SVM based on the robust Euclidean distance (RFSVM-EUC) and robust Mahalanobis distance (RFSVM-MAHMCD and RFSVM-MAHOGK) proposed by Mohammadi and Sarmad [24] were assessed and tested on several examples of synthetic data, as well as several real data set. The results indicate the superiority of the proposed algorithms.
As future work, we can consider the existence of more than one cluster of outliers in each class that is significantly further away from each other and even from the mass data cluster. Future investigation will focus on the further

Table 4: Mean(standard division) of the classification accuracy for SVM, FSVM, FSVM-DistGMM, FSVM-BayesGMM, RFSVM-EUC, FSVM-MAH, RFSVM-MAHMCD and RFSVM-MAHOGK for the test set data

Datasets	SVM	FSVM	FSVM-DistGMM	FSVM-BayesGMM	RFSVM-EUC	FSVM-MAH	RFSVM-MAHMCD	RFSVM-MAHOGK
Diabetes	77.00(0.06)	78.23(0.03)	80.06(0.02)	78.84(0.02)	**81.48(0.06)**	78.42(0.02)	77.40(0.06)	79.07(0.03)
	8.0	6.0	2.0	4.0	1.0	5.0	7.0	3.0
Biomed	88.22(0.03)	90.31(0.03)	90.86(0.04)	**95.91(0.03)**	95.04(0.07)	90.61(0.06)	93.60(0.06)	93.23(0.05)
	8.0	7.0	5.0	1.0	2.0	6.0	3.0	4.0
Statlog-Heart	83.12(0.04)	83.93(0.03)	**88.79(0.01)**	85.85(0.04)	85.44(0.06)	83.56(0.01)	87.41(0.06)	86.00(0.03)
	8.0	6.0	1.0	4.0	5.0	7.0	2.0	3.0
Bupa	68.23(0.02)	68.67(0.09)	70.89(0.02)	69.23(0.02)	**71.09(0.06)**	68.55(0.04)	68.78(0.06)	68.55(0.01)
	8.0	5.0	2.0	3.0	1.0	6.5	4.0	6.5
Wine	96.65(0.06)	98.05(0.06)	**98.91(0.02)**	98.47(0.02)	96.65(0.03)	97.95(0.06)	96.27(0.06)	98.47(0.03)
	6.5	4.0	1.0	2.5	6.5	5.0	8.0	2.5
Iris	96.40(0.06)	96.90(0.03)	**100(0.02)**	**100**(0.02)	97.22(0.07)	97.00(0.08)	97.00(0.06)	96.30(0.01)
	7.0	6.0	1.5	1.5	3.0	4.5	4.5	8.0
Cancer	93.42(0.05)	94.86(0.06)	97.76(0.02)	**98.96(0.01)**	96.79(0.04)	95.12(0.04)	94.69(0.06)	93.80(0.03)
	8.0	5.0	2.0	1.0	3.0	4.0	6.0	7.0
Ionosphere	89.19(0.05)	89.73(0.03)	91.70(0.02)	**93.60(0.02)**	90.20(0.06)	89.73(0.06)	90.55(0.06)	83.59(0.08)
	7.0	5.5	2.0	1.0	4.0	5.5	3.0	8.0
Sonar	64.39(0.02)	65.76(0.03)	69.85(0.08)	**72.54(0.12)**	66.05(0.05)	63.79(0.02)	68.74(0.09)	68.07(0.05)
	7.0	6.0	2.0	1.0	5.0	8.0	3.0	4.0
WDBC	96.78(0.06)	96.90(0.04)	98.00(0.02)	**99.88(0.01)**	97.71(0.06)	97.03(0.02)	99.71(0.06)	96.23(0.06)
	7.0	6.0	3.0	1.0	4.0	5.0	2.0	8.0
Saheart	64.61(0.02)	64.45(0.03)	**71.89(0.08)**	69.19(0.02)	67.39(0.03)	68.71(0.06)	68.45(0.01)	69.07(0.03)
	7.0	8.0	1.0	2.0	6.0	4.0	5.0	3.0
Blood	63.23(0.05)	74.15(0.03)	77.76(0.08)	76.59(0.02)	75.75(0.07)	74.39(0.02)	**78.83(0.04)**	75.11(0.09)
	8.0	7.0	2.0	3.0	4.0	6.0	1.0	5.0
PlanningRelax	65.66(0.01)	67.65(0.03)	72.33(0.07)	**73.00(0.02)**	72.69(0.05)	69.27(0.04)	69.67(0.09)	68.66(0.05)
	8.0	7.0	3.0	1.0	2.0	5.0	4.0	6.0
Spectf	77.19(0.04)	79.58(0.03)	78.13(0.03)	**83.75(0.01)**	81.85(0.07)	77.03(0.02)	81.46(0.08)	80.30(0.05)
	7.0	5.0	6.0	1.0	2.0	8.0	3.0	4.0
Vowel	97.59(0.04)	97.59(0.04)	98.22(0.04)	**98.90(0.02)**	97.07(0.05)	98.39(0.06)	96.80(0.01)	96.36(0.06)
	4.5	4.5	3.0	1.0	6.0	2.0	7.0	8.0
Haberman	69.59(0.06)	76.47(0.03)	77.40(0.02)	78.29(0.02)	75.03(0.05)	76.15(0.04)	**85.41(0.02)**	75.80(0.03)
	8.0	4.0	3.0	2.0	7.0	5.0	1.0	6.0
Parkinsons	76.10(0.01)	81.80(0.03)	76.86(0.02)	86.38(0.02)	**92.10(0.07)**	84.45(0.02)	88.61(0.06)	84.65(0.02)
	8.0	6.0	7.0	3.0	1.0	5.0	2.0	4.0
Ecoli	70.55(0.06)	70.72(0.05)	71.39(0.04)	**73.50(0.02)**	72.58(0.06)	71.56(0.07)	70.94(0.06)	71.40(0.01)
	8.0	7.0	5.0	1.0	2.0	3.0	6.0	4.0
Phoneme	71.00(0.06)	**75.06(0.09)**	73.73(0.02)	74.75(0.02)	74.10(0.08)	69.47(0.03)	74.31(0.07)	73.07(0.03)
	7.0	1.0	5.0	2.0	4.0	8.0	3.0	6.0
Segment	89.08(0.06)	89.75(0.03)	96.18(0.02)	**97.43(0.07)**	94.55(0.04)	55.54(0.02)	94.21(0.06)	94.19(0.03)
	7.0	6.0	2.0	1.0	3.0	8.0	4.0	5.0
Average Rank	7.380952	5.571429	2.880952	1.952381	3.452381	5.547619	4.071429	5.142857

Table 5: Mean(standard division) of the F-measure of the classification for SVM, FSVM, FSVM-DistGMM, FSVM-BayesGMM, RFSVM-EUC, FSVM-MAH, RFSVM-MAHMCD and RFSVM-MAHOGK for the test set data

Datasets	SVM	FSVM	FSVM-DistGMM	FSVM-BayesGMM	RFSVM-EUC	FSVM-MAH	RFSVM-MAHMCD	RFSVM-MAHOGK
Diabetes	74.02(0.08)	74.77(0.05)	77.14(0.02)	77.56 (0.06)	**79.73 (0.02)**	74.77(0.01)	76.56(0.03)	75.72(0.04)
	8.0	6.5	3.0	2.0	1.0	6.5	4.0	5.0
Biomed	86.52(0.03)	87.90(0.06)	87.90(0.07)	**94.57(0.04)**	89.52(0.03)	94.17(0.02)	92.35 (0.06)	90.96(0.09)
	8.0	6.5	6.5	1.0	5.0	2.0	3.0	4.0
Statlog-Heart	82.12(0.04)	82.86(0.05)	87.57(0.05)	**88.88(0.03)**	83.67(0.03)	82.88(0.06)	86.35(0.07)	85.11(0.09)
	8.0	7.0	2.0	1.0	5.0	6.0	3.0	4.0
Bupa	67.39(0.05)	66.32(0.06)	**69.16(0.03)**	66.43(0.01)	65.59(0.02)	65.00(0.03)	68.13 (0.04)	65.59(0.06)
	3.0	5.0	1.0	4.0	6.5	8.0	2.0	6.5
Wine	95.02 (0.03)	96.99 (0.01)	**97.99 (0.01)**	97.02(0.02)	96.99 (0.02)	95.39(0.06)	95.02 (0.08)	96.25(0.11)
	7.5	3.5	1.0	2.0	3.5	6.0	7.5	5.0
Iris	94.65(0.06)	95.65 (0.06)	**100(0.04)**	**100(0.02)**	97.65(0.03)	96.33(0.05)	97.84(0.06)	97.09(0.08)
	8.0	7.0	1.5	1.5	4.0	6.0	3.0	5.0
Cancer	93.65(0.06)	95.08(0.01)	96.69(0.05)	**98.17(0.20)**	96.08(0.04)	95.61(0.08)	93.89(0.01)	95.08(0.06)
	8.0	5.5	2.0	1.0	3.0	4.0	7.0	5.5
Ionosphere	87.35(0.05)	88.18(0.06)	**91.33(0.04)**	91.17 (0.03)	88.18 (0.06)	89.17(0.08)	90.33(0.03)	79.60(0.06)
	7.0	5.5	1.0	2.0	5.5	4.0	3.0	8.0
Sonar	67.12(0.09)	65.00 (0.06)	65.84(0.07)	**70.84(0.05)**	63.45 (0.04)	66.53(0.07)	68.18 (0.10)	67.84(0.08)
	4.0	7.0	6.0	1.0	8.0	5.0	2.0	3.0
WDBC	95.31 (0.03)	97.98 (0.01)	96.66(0.01)	99.31(0.05)	97.98(0.06)	96.87 (0.06)	99.36(0.05)	**99.57(0.09)**
	8.0	4.5	7.0	3.0	4.5	6.0	2.0	1.0
Saheart	59.14(0.04)	62.04 (0.02)	61.92(0.02)	**74.64 (0.04)**	64.63 (0.04)	61.41(0.03)	63.42(0.06)	65.08(0.03)
	8.0	5.0	6.0	1.0	3.0	7.0	4.0	2.0
Blood	63.70(0.07)	71.71(0.04)	75.75(0.05)	75.89(0.02)	72.56(0.06)	72.56(0.06)	**76.20(0.03)**	74.47(0.03)
	8.0	7.0	3.0	2.0	5.5	5.5	1.0	4.0
PlanningRelax	62.63(0.03)	61.21(0.07)	64.21(0.03)	**67.85(0.02)**	65.21(0.06)	62.81(0.07)	67.10(0.04)	61.86(0.06)
	6.0	8.0	4.0	1.0	3.0	5.0	2.0	7.0
Spectf	68.67(0.04)	68.29(0.09)	**78.76(0.02)**	74.22(0.03)	71.14(0.05)	73.34(0.06)	68.84(0.03)	71.45(0.08)
	7.0	8.0	1.0	2.0	5.0	3.0	6.0	4.0
Vowel	94.59(0.03)	96.00(0.03)	96.45(0.07)	**96.94(0.04)**	96.28(0.01)	95.42(0.05)	93.23(0.06)	95.42(0.03)
	7.0	4.0	2.0	1.0	3.0	5.5	8.0	5.5
Haberman	70.80(0.04)	71.66(0.03)	74.39(0.01)	**77.08(0.03)**	76.66(0.07)	70.56(0.08)	74.39(0.07)	75.23(0.05)
	7.0	6.0	4.5	1.0	2.0	8.0	4.5	3.0
Parkinsons	72.82(0.02)	76.86(0.03)	83.26(0.07)	**88.26(0.06)**	80.31(0.03)	83.31(0.05)	83.19(0.04)	79.42(0.10)
	8.0	7.0	3.0	1.0	5.0	2.0	4.0	6.0
Ecoli	68.65(0.08)	69.52(0.04)	71.59(0.06)	**74.33(0.06)**	70.48(0.01)	70.69(0.08)	71.59(0.05)	70.69(0.01)
	8.0	7.0	2.5	1.0	6.0	4.5	2.5	4.5
Phoneme	65.62(0.07)	**71.04(0.03)**	69.08(0.04)	67.46(0.10)	69.16(0.02)	67.74(0.03)	68.45(0.03)	68.23(0.04)
	8.0	1.0	3.0	7.0	2.0	6.0	4.0	5.0
Segment	89.10(0.03)	88.50(0.05)	96.38(0.04)	**97.12**(0.05)	95.27(0.02)	56.83(0.07)	95.03(0.01)	95.03(0.06)
	6.0	7.0	2.0	1.0	3.0	8.0	4.5	4.5
Average Rank	7.125	5.900	3.100	1.825	4.175	5.400	3.850	4.625

Table 6: The training time (in seconds) of the SVM, FSVM, FSVM-DistGMM, FSVM-BayesGMM, RFSVM-EUC, FSVM-MAH, RFSVM-MAHMCD and RFSVM-MAHOGK.

Datasets	SVM	FSVM	FSVM-DistGMM	FSVM-BayesGMM	RFSVM-EUC	FSVM-MAH	RFSVM-MAHMCD	RFSVM-MAHOGK
Diabetes	0.555807	0.695715	1.462104	0.738750	0.670938	0.707571	1.478348	0.714020
Biomed	0.046627	0.048279	0.644596	0.064462	0.074225	0.046639	0.627185	0.056011
Statlog-Heart	0.063660	0.063388	0.087663	0.100713	0.095798	0.078427	0.189583	0.187731
Bupa	0.084172	0.096648	0.769119	0.101596	0.079122	0.084267	0.755822	0.120245
Wine	0.028055	0.038399	0.850160	0.047129	0.041133	0.040129	0.815218	0.141102
Iris	0.083211	0.084373	0.608182	0.100298	0.105997	0.083876	0.618453	0.092093
Cancer	0.581516	0.788487	0.966900	0.578102	0.747953	0.759480	0.983839	0.756755
Ionosphere	0.146310	0.154997	0.178781	0.180974	0.161853	0.211503	0.346825	1.018732
Sonar	0.047774	0.049121	0.730591	0.083523	0.082351	0.057654	0.787919	0.116929
WDBC	0.307032	0.398168	1.132979	0.499661	0.392312	0.372866	1.121788	0.399699
Saheart	0.145800	0.179753	0.963841	0.170198	0.218421	0.159089	1.047779	0.229077
Blood	1.510102	1.725082	1.980502	1.510102	1.725082	1.980502	2.210564	2.037336
PlanningRelax	0.022791	0.039725	0.942272	0.037981	0.060574	0.041775	0.810575	0.134191
Spectf	0.105048	0.125635	1.020527	0.132046	0.121842	0.115077	0.844885	0.170574
Vowel	4.465582	5.869338	6.988529	5.487185	5.046470	5.169621	5.495668	4.918899
Haberman	0.073003	0.088578	0.835765	0.089027	0.086760	0.090190	0.788580	0.091646
Parkinsons	0.164414	0.177661	1.439816	0.154425	0.136840	0.119163	1.734926	0.241605
Ecoli	1.089786	1.173047	1.124841	1.061265	1.080255	1.084189	1.312636	1.275629
Phoneme	174.119825	486.167267	500.245870	430.727918	433.787752	609.071951	459.032285	574.908600
Segment	31.396585	32.200789	32.161630	32.856677	31.303300	32.253499	32.734088	34.849389

improvement of the proposed algorithms in terms of creating an adaptive and automatic method to identify the number of cluster of outliers.

Acknowledgement

This paper is a part of the Ph.D thesis of Mandana Mohammadi (NO. 37943).

References

[1] E. Acuna, C. Rodriguez, *A meta analysis study of outlier detection methods in classification*, Technical Paper, Department of Mathematics, University of Puerto Rico at Mayaguez, (2004), 1-25.

[2] M. Aitkin, G. T. Wilson, *Mixture models, outliers, and the EM algorithm*, Technometrics, **22**(3) (1980), 325-331.

[3] W. An, M. Liang, *Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises*, Neurocomputing, **110** (2013), 101-110.

[4] U. Baid, S. Talbar, *Comparative study of k-means, gaussian mixture model, fuzzy c-means algorithms for brain tumor segmentation*, In International Conference on Communication and Signal Processing, (2016) (ICCASP 2016), Atlantis Press, doi.org/10.2991/iccasp-16.2017.85

[5] B. E. Boser, I. M. Guyon, V. N. Vapnik, *A training algorithm for optimal margin classifiers*, In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, (1992), 144-152.

[6] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, M. E. Houle, *On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study*, Data Mining and Knowledge Discovery, **30**(4) (2016), 891-927.

[7] J. Ďemsar, *Statistical comparisons of classifiers over multiple data sets*, Journal of Machine Learning Research, **7** (2006), 1-30.

[8] I. Gath, A. B. Geva, *Fuzzy clustering for the estimation of the parameters of the components of mixtures of normal distributions*, Pattern Recognition Letters, **9**(2) (1989), 77-86.

[9] F. E. Grubbs, *Procedures for detecting outlying observations in samples*, Technometrics, **11**(1) (1969), 1-21.

[10] J. Hang, J. Zhang, M. Cheng, *Application of multi-class fuzzy support vector machine classifier for fault diagnosis of wind turbine*, Fuzzy Sets and Systems, **297** (2016), 128-140.

[11] Y. Y. Hao, Z. X. Chi, D. Q. Yan, *Fuzzy support vector machine based on vague sets for credit assessment*, In Fourth International Conference on Fuzzy Systems and Knowledge Discovery, **1** (2007), 603-607.

[12] V. Hodge, J. Austin, *A survey of outlier detection methodologies*, Artificial Intelligence Review, **22**(2) (2004), 85-126.

[13] P. J. Huber, *Robust estimation of a location parameter*, The Annals of Mathematical Statistics, **35** (1964), 73-101.

[14] P. J. Huber, *A robust version of the probability ratio test*, The Annals of Mathematical Statistics, (1965), 1753-1758.

[15] G. S. D. S. Jayakumar, B. J. Thomas, *A new procedure of clustering based on multivariate outlier detection*, Journal of Data Science, **11**(1) (2013), 69-84.

[16] M. F. Jiang, S. S. Tseng, C. M. Su, *Two-phase clustering process for outliers detection*, Pattern Recognition Letters, **22**(6-7) (2001), 691-700.

[17] X. Jiang, Z. Yi, J. C. Lv, *Fuzzy SVM with a new fuzzy membership function*, Neural Computing and Applications, **15**(3-4) (2006), 268-276.

[18] Z. Kou, J. Xu, X. Zhang, L. Ji, *An improved support vector machine using class-median vectors*, In Proc of 8th Intl Conf on Neural Information Processing, **2** (2001), 883-887.

[19] C. F. Lin, S. D. Wang, *Fuzzy support vector machines*, IEEE Transactions on Neural Networks, **13**(2) (2002), 464-471.

[20] Y. H. Liu, H. P. Huang, *Fuzzy support vector machines for pattern recognition and data mining*, International Journal of Fuzzy Systems, **4**(3) (2002), 826-835.

[21] A. Loureiro, L. Torgo, C. Soares, *Outlier detection using clustering methods: A data cleaning application*, In Proceedings of KDNet Symposium on Knowledge-Based Systems for the Public Sector, Springer, 2004.

[22] R. A. R. D. Maronna, R. D. Martin, V. Yohai, *Robust statistics*, John Wiley and Sons, Chichester, ISBN, 2006, 978 pages.

[23] R. A. Maronna, R. H. Zamar, *Robust estimates of location and dispersion for high-dimensional datasets*, Technometrics, **44**(4) (2002), 307-317.

[24] M. Mohammadi, M. Sarmad, *Robustified distance based fuzzy membership function for support vector machine classification*, Iranian Journal of Fuzzy Systems, **16**(6) (2019), 191-204.

[25] Q. Song, W. Hu, W. Xie, *Robust support vector machine with bullet hole image classification*, IEEE Transactions on Systems, Man, and Cybernetics, Part C, Applications and Reviews, **32**(4) (2002), 440-448.

[26] W. M. Tang, *Fuzzy SVM with a new fuzzy membership function to solve the two-class problems*, Neural Processing Letters, **34**(3) (2011), 209.

[27] B. Van Cutsem, I. Gath, *Detection of outliers and robust estimation using fuzzy clustering*, Computational Statistics and Data Analysis, **15**(1) (1993), 47-61.

[28] H. Wang, J. Feng, H. Wang, *Grouped fuzzy SVM with EM-based partition of sample space for clustered microcalcification detection*, Technology and Health Care, **25**(S1) (2017), 325-336.

[29] H. Wang, P. Guo, J. Feng, Y. Ren, *Classification based on clustered group SVM*, In 2010 Chinese Conference on Pattern Recognition (CCPR), (2010), 1-5.

[30] T. Wang, Y. Qiu, J. Hua, *Centered kernel alignment inspired fuzzy support vector machine*, Fuzzy Sets and Systems, **394** (2020), 110-123.

[31] Z. Wu, H. Zhang, J. Liu, *A fuzzy support vector machine algorithm for classification based on a novel PIM fuzzy clustering method*, Neurocomputing, **125** (2014), 119-124.

[32] H. Xia, B. Q. Hu, *Feature selection using fuzzy support vector machines*, Fuzzy Optimization and Decision Making, **5**(2) (2006), 187-192.

[33] Q. Xu, H. Zhou, Y. Wang, J. Huang, *Fuzzy support vector machine for classification of EEG signals using wavelet-based features*, Medical Engineering and Physics, **31**(7) (2009), 858-865.

[34] K. Yamanishi, J. I. Takeuchi, G. Williams, P. Milne, *On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms*, Data Mining and Knowledge Discovery, **8**(3) (2004), 275-300.

[35] X. Yang, L. J. Latecki, D. Pokrajac, *Outlier detection with globally optimal exemplar-based GMM*, In Proceedings of the 2009 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, (2009), 145-154.

[36] X. Yang, G. Zhang, J. Lu, J. Ma, *A kernel fuzzy c-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises*, IEEE Transactions on Fuzzy Systems, **19**(1) (2010), 105-115.

[37] Y. B. Yuan, S. Lan, X. Yu, M. Yu, *Algorithm of fuzzy support vector machine based on a piecewise linear fuzzy weight method*, International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), **12**(2) (2018), 62-76.

[38] X. Zhang, *Using class-center vectors to build support vector machines*, In Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop, (1999), 3-11.

[39] Y. Zhang, Z. X. Chi, *A fuzzy support vector classifier based on Bayesian optimization*, Fuzzy Optimization and Decision Making, **7**(1) (2008), 75-86.

[40] A. Zimek, E. Schubert, H. P. Kriegel, *A survey on unsupervised outlier detection in highdimensional numerical data*, Statistical Analysis and Data Mining: The ASA Data Science Journal, **5**(5) (2012), 363-387.

[41] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, H. Chen, *Deep autoencoding gaussian mixture model for unsupervised anomaly detection*, Published as a Conference Paper at ICLR, 2018.