Gaussian Mixture Clustering Based Attribute Weighting for Non-Linearly Separable Datasets

Kemal Polat[1]

[1]Abant İzzet Baysal University, Department of Electrical and Electronics Engineering, Gölköy, Bolu, Turkey
kemal_polat2003@yahoo.com

Abstract

Data transformation or preprocessing methods needs to be used prior to classification or prediction algorithms for improving the classification performance in the classifying of non –linear separable datasets. As data preprocessing method, the Gaussian mixture clustering based attribute weighting (GMCBAW) has been proposed and applied to two artificial datasets including 2-D spiral and chain link datasets that have a non-linear data distribution. After data preprocessing stage, the k-nearest neighbor (k-NN) classifier has been used. In the classification of 2-D spiral dataset, while alone k-NN classifier obtained the success rate of 47.91%, the combination of GMCBAW and k-NN classifier achieved the success rate of 56.20% for the k value of 50 in k-NN classifier. As for chain link dataset, while alone k-NN classifier obtained the success rate of 74.20%, the combination of GMCBAW and k-NN classifier achieved the success rate of 85.80% for the k value of 200 in k-NN classifier. The results demonstrated the feasibility and robustness of GMCBAW on the classification of linear non-separable datasets.

1. Introduction

Data preprocessing methods are commonly used in solutions of many pattern recognition problems. In order to simplify the classification algorithm and to decrease the computational cost of used algorithms, data preprocessing methods are employed to datasets. By means of using data preprocessing techniques, a dataset that has a linearly non-separable distribution is transformed to a linearly separable dataset. Also, data transformation or preprocessing methods are used for noise removing, data imputation, feature reduction, data reduction, data filtering, normalization, data clustering etc.

In literature, many studies have been conducted regarding the data preprocessing or transformation. Among these, Blansché et al. proposed a new process for modular clustering of complex previous data, like remote sensing images. This method carry outs attribute weighting in a wrapper approach [1]. Tahir et al. proposed a novel hybrid approach for simultaneous feature selection and feature weighting of k-NN (k-nearest neighbor) rule based on Tabu Search (TS) heuristic. Gançarski et al. proposed two new feature weighting methods on the basis of coevolutive algorithms. The first one is motivated by the Lamarck theory and uses the distance-based cost function defined in the LKM algorithm as fitness function. The second method uses a fitness function on the basis of a new partitioning quality measure [2]. Polat et al. proposed a new attribute weighting method based on fuzzy logic and called fuzzy weighting method in classification of ECG dataset [3]. Polat et al. have proposed k-NN (k-nearest neighbor) based feature-weighting method to reduce the variance within features in dataset and applied to medical datasets [4, 5, 6]. Polat et al. suggested a novel attribute weighting method based on similarity measure between features and applied to classification of Doppler signals to diagnose Atherosclerosis disease [7]. Dua et al. present a method based on the connected component theory to remove the labels from the image and crop the image to a relevant reduced size as data preprocessing method [8]. Polat et al. proposed a novel attribute weighting called subtractive clustering based attribute weighting for recognizing the traffic accidents and used the support vector machine classifier as classifier algorithm [9].

In this study, a new attribute weighting method based on clustering centers belonging to each class of dataset using Gaussian mixture clustering has been proposed. This weighting method is called Gaussian mixture clustering based attribute weighting (GMCBAW). In this method, first of all, the mean values of each attribute belonging to each class are calculated. Secondly, the probable cluster centers of the whole dataset are obtained using Gaussian mixture clustering. Then, the ratios of mean values to cluster centers belonging to each attribute in each class are computed. Finally, these obtained ratios are multiplied with each attribute and weighted the dataset. To efficiency of proposed weighting method, two artificial datasets have been used. These datasets are 2-D spiral dataset and chain link dataset. The main characteristic of these datasets has a linearly non-separable distribution. To classify the weighted these datasets, k-nearest neighbor (k-NN) classifier has been used since k-NN classifier is a simple and effective algorithm. The classification accuracy, recall, true negative rate (TNR), prediction, g-mean 1, g-mean 2, and f-measure values have been used to evaluate the proposed method.

The rest of this paper is arranged as follows. The section 2 presents the material and method. In the section 3, the experimental results and discussion is given. Section 4 concludes the remarking findings.

2. Material and Method

2.1. Data

In this study, two artificial datasets including 2-D spiral dataset and chain link dataset have been used.

Two (2D) spiral dataset is a linearly non-separable classification problem. Since this dataset is a nonlinear separable dataset, the classification of two spiral dataset is a challenge task. In the dataset, there are 194 samples including each class of 94 samples. There are two attributes in the 2D spiral dataset. Figure 1 presents the class distribution of two spiral dataset [10].

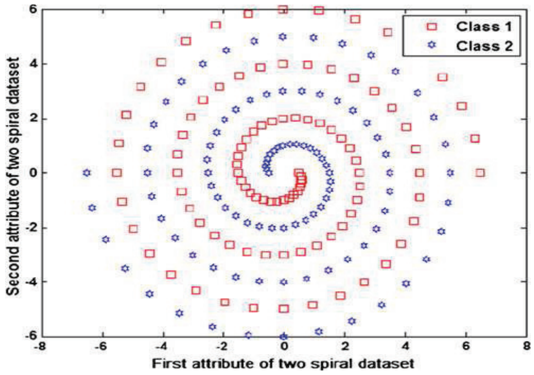


Fig. 1. The class distribution of two spiral dataset

Chain link dataset also called intertwined rings is a classic example of a data set, which provokes topology preservation violations. The data set includes two rings, each two-dimensional, which is intertwined in a three-dimensional space. In this dataset, there are 500 data points for each class. There are totally 1000 data points. There are two attributes in chain link dataset. Figure 2 demonstrates the class distribution of chain link dataset [11].

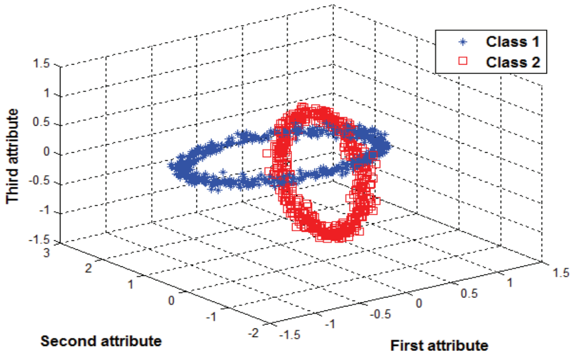


Fig. 2. The class distribution of chain link dataset

2.2. Method

A hybrid system with two stages has been proposed. In first stage, to weight the datasets or to transform from non-linearly separable dataset to linearly separable dataset, Gaussian mixture clustering based attribute weighting method has been proposed and used to scale the datasets. In the second stage, after data preprocessing stage, k-NN classifier has been used. The used stages have been explained in the following subsections.

2.2.1. Gaussian mixture clustering based attribute weighting (GMCBAW)

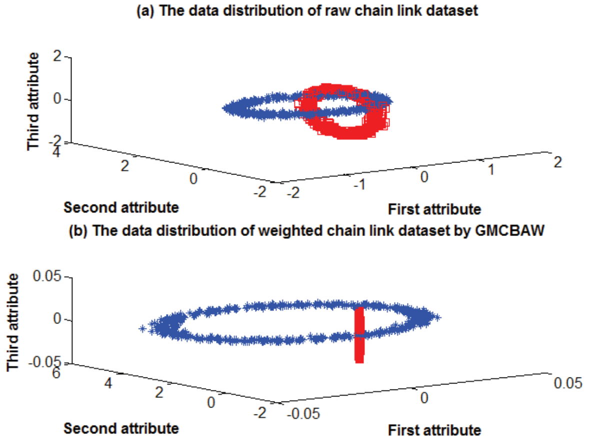
Gaussian mixture models are formed by combining multivariate normal density components. They are often used for data clustering. Clusters are assigned by selecting the component that maximizes the posterior probability. Like k-means clustering, Gaussian mixture modeling uses an iterative algorithm that converges to a local optimum. Gaussian mixture modeling may be more appropriate than k-means clustering when clusters have different sizes and correlation within them. Clustering using Gaussian mixture models is sometimes considered a soft clustering method. The posterior probabilities for each point indicate that each data point has some probability of belonging to each cluster [12, 13].

Figure 3 gives the pseudo code of GMCBAW. In this method, first of all, the mean values of each attribute belonging to each class are calculated. Secondly, the probable cluster centers of the whole dataset are obtained using Gaussian mixture clustering. Then, the ratios of mean values to cluster centers belonging to each attribute in each class are computed. Finally, these obtained ratios are multiplied with each attribute and weighted the dataset.

```
Initialize total values (inputs, outputs, and distances) to zero
Reading the datasets
For each attribute in dataset
        Step 1.Finding the clustering centers using Gaussian mixture clustering algorithm
        Step 2. Computing the mean values of each attribute belonging to each class
        Step 3.Calculating the ratios of mean values of attributes to cluster centers for each attribute belonging to each class
        Step 4.Multiplying these ratios with each data point in attributes (Weighting process)
End For
```

Fig. 3. The pseudo code of GMCBAW

After GMCBAW method applied to 2-D spiral dataset, the class distributions of weighted and raw two spiral datasets have been given in the Figure 4. Figure 5 presents the class distributions of weighted and raw chain link datasets. As can be seen from both figures, the power of distinguishing of two artificial datasets has increased by means of Gaussian mixture clustering based attribute weighting method. In this way, both non-linear artificial datasets could be classified using classifier algorithms.


(a) The data distribution of raw chain link dataset

(b) The data distribution of weighted chain link dataset by GMCBAW

Fig. 4. The class distributions of raw (a) and weighted (b) 2-D spiral datasets

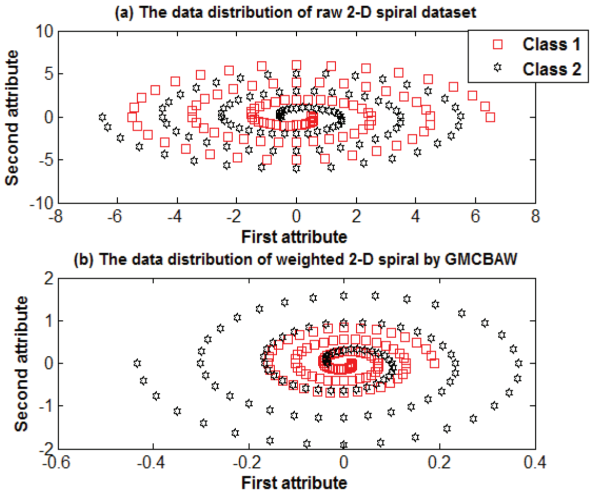


Fig. 5. The class distributions of raw (a) and weighted (b) chain link datasets

2.2.2. k-Nearest Neighbor (k-NN) Classifier

To classify the weighted datasets, k-NN (nearest neighbor) classifier algorithm has been used. The k-nearest neighbor algorithm is one of the simplest algorithms among all machine learning algorithms. In k-nearest-neighbor classification, the training dataset is used to classify each data of a "target" dataset. The structure of the data is that there is a classification variable of interest and a number of additional predictor variables [14, 15].

Robust models can be obtained by locating k, where k >1, neighbors and enabling the majority vote decide the outcome of the class labeling. A higher value of k causes in a smoother function. The disadvantage of increasing the value of k is that as k approaches n, where n is the size of the instance base, the performance of the classifier will approach that of the most straightforward statistical baseline, the assumption that all unknown instances belong to the class most frequently represented in the training data [14, 15].

The several k values in k-NN classifier have been used and compared with each other. By this way, the best k values have been found for classification of two artificial datasets.

3. Results and Discussion

In this study, a new data preprocessing method called Gaussian mixture clustering based attribute weighting (GMCBAW) has been proposed to transform from non-linearly separable datasets to linearly separable datasets. Also, the other aims of this weighting method are to increase the classification performance and to decrease the computational cost of used algorithms. As clustering algorithm, Gaussian mixture clustering algorithm has been used since this method is an effective and robust in the clustering of large datasets. As classifier algorithm, k-NN classifier has been used to classify the weighted datasets. Two artificial datasets have been used. These datasets are two (2-D) spiral and chain link datasets.

To evaluate the proposed method, the classification accuracy, recall, true negative rate (TNR), prediction, g-mean 1, g-mean 2, and f-measure values have been used. The readers can refer to [16] for more information about these measures. Table 1 shows the obtained results in the classification of raw 2-D spiral dataset using k-NN classifier for various k values. Table 2 gives the obtained results in the classification of weighted 2-D spiral dataset using k-NN classifier for various k values. Table 3 presents the obtained results in the classification of raw and weighted 2-D spiral dataset using k-NN classifier for k value of 50. Table 4 depicts the obtained results in the classification of raw chain link dataset using k-NN classifier for various k values. Table 5 demonstrates the obtained results in the classification of weighted chain link dataset using k-NN classifier for various k values. Table 6 exhibits the obtained results in the classification of raw and weighted chain link dataset using k-NN classifier for k value of 200. In the classification of 2-D spiral dataset, the effect of GMCBAW could be seen to classification accuracy. When k value was increased, the classification performance was also increased. As for chain link dataset, when k value was increased, the classification performance was also decreased for raw chain link dataset. In the classification of weighted chain link dataset, when k value was increased, the classification performance was increased.

As can be seen from obtained results, Gaussian mixture clustering based attribute weighting could be safely used as data preprocessing in the classification of non-linearly separable datasets.

Table 1. The obtained results in the classification *o*f raw 2-D spiral dataset using k-NN classifier for various k values

k value in k-NN classier	**Classification Accuracy (%)**
1	39.58
2	39.58
3	37.50
4	36.46
5	36.46
6	36.46
7	35.42
8	35.42
9	35.42
10	38.54
50	47.92

Table 2. The obtained results in the classification of weighted 2-D spiral dataset using k-NN classifier for various k values

k value in k-NN classier	**Classification Accuracy (%)**
1	50.00
2	50.00
3	50.00
4	50.00
5	50.00
6	50.00
7	50.00
8	50.00
9	50.00
10	50.00
50	56.25

Table 3. The obtained results in the classification of raw and weighted 2-D spiral dataset using k-NN classifier for k value of 50

Performance Measures	**Raw 2-D Spiral dataset**	**Weighted 2-D Spiral dataset by GMCBAW**
Classification Accuracy (%)	47.91	56.20
Recall	0.478	1
TNR	0.483	0.533
Precision	0.458	0.125
G-mean 1	0.468	0.265
G-mean 2	0.479	0.547
F-measure	0.468	0.204

Table 4. The obtained results in the classification of raw chain link dataset using k-NN classifier for various k values

k value in k-NN classier	**Classification Accuracy (%)**
1	100
100	100
150	99.40
200	74.20
250	68.80

Table 5. The obtained results in the classification of weighted chain link dataset using k-NN classifier for various k values

k value in k-NN classier	**Classification Accuracy (%)**
1	99.80
100	92.00
150	89.60
200	85.80
250	82.00

Table 6. The obtained results in the classification of raw and weighted chain link dataset using k-NN classifier for k value of 200

Performance Measures	**Raw chain link dataset**	**Weighted chain link dataset by GMCBAW**
Classification Accuracy (%)	74.20	85.80
Recall	0,75	0,778
TNR	0,73	1
Precision	0,72	1
G-mean 1	0,73	0,926
G-mean 2	0,73	0,9262
F-measure	0,73	0,9235

6. Conclusions

Data preprocessing methods are commonly used in solutions of many pattern recognition problems. In this paper, a novel data weighting called Gaussian mixture clustering based attribute weighting has been proposed and applied to transform from two artificial non-linear datasets to linear datasets. The used datasets were the 2-D spiral and chain link datasets. The results demonstrated that the non-linear separable datasets have been converted to linearly separable datasets by means of GMCBAW.

In future, using Gaussian mixture clustering, a new data weighting method could be formed by inter-cluster measure.

7. References

[1] A. Blansché, P. Gançarski, J.J. Korczak, “MACLAW: A modular approach for clustering with local attribute weighting”, *Pattern Recognition Letters*, 27(11), 2006, 1299-1306.

[2] Tahir, M.A., Bouridane, A., Kurugollu, F., “Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier”, Pattern Recognition Letters, 28(4), 2007, 438-446.

[3] Polat, K., Şahan, S., Güneş, S., A new method to medical diagnosis: Artificial immune recognition system (AIRS) with fuzzy weighted pre-processing and application to ECG arrhythmia, Expert Systems with Applications, 31(2), 2006, 264-269.

[4] Polat, K., Şahan, S., Güneş, S., “Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing”, Expert Systems with Applications, 32(2), 2007, 625-631.

[5] Polat, K., Güneş, S., “A hybrid medical decision making system based on principles component analysis, k-NN based weighted pre-processing and adaptive neuro-fuzzy inference system”, Digital Signal Processing, 16(6), 2006, 913-921.

[6] Polat, K., Güneş, S., “The effect to diagnostic accuracy of decision tree classifier of fuzzy and k-NN based weighted pre-processing methods to diagnosis of erythemato-squamous diseases”, Digital Signal Processing, 16(6), 2006, 922-930.

[7] Polat K, Latifoğlu F, Kara S, Güneş S, “Usage of Novel Similarity Based Weighting Method to Diagnose the Atherosclerosis from Carotid Artery Doppler Signals”, Medical & Biological Eng. & Computing, 46, 2008, 353-362.

[8] Dua, S., Singh, H., Thompson, H.W., “Associative classification of mammograms using weighted rules”, Expert Systems with Applications, 36(5), 2009, 9250-9259.

[9] Polat, K., Durduran, S.S., “Subtractive clustering attribute weighting (SCAW) to discriminate the traffic accidents on Konya–Afyonkarahisar highway in Turkey with the help of GIS: A case study”, Advances in Engineering Software, 42(7), 2011, 491-500.

[10] Estévez, P.A., Figueroa, C. J., “Online data visualization using the neural gas network”, Neural Network, 19(6-7), 2006, 923-934.

[11] Ultsch, A., “Clustering with SOM: U*C”, In Proceedings of the workshop on self-organizing maps, 2005, pp. 75–82.

[12] Zeng, H., Cheung, Y.M., “A new feature selection method for Gaussian mixture clustering”, Pattern Recognition, 42(2), 2009, 243-250.

[13] Yuzhong Wang, Jie Yang, Yue Zhou, “Color-texture segmentation using JSEG based on Gaussian mixture

modeling", Journal of Systems Engineering and Electronics, 17(1), 2006, 24-29.

[14] Dasarathy, B. V., editor (1991) Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, ISBN 0-8186-8930-7

[15]http://www.fon.hum.uva.nl/praat/manual/kNN_classifiers_1__What_is_a_kNN_classifier.html , (last accessed: 2011)

[16] Polat, K., "Application of Attribute Weighting Method based on Clustering Centers to Discrimination of Linearly Non-separable Medical Datasets", *Journal of Medical Systems*, Article in Press, 2011.