DOI 10.29042/2019-4802-4807

Helix Vol. 9 (1): 4802- 4807

Classifying Unbalanced Datasets Using Iterative Fuzzy Support Vector Machine

[*1]P. Aruna Kumari, [2]Dr. G. Jaya Suma

[1]JNTUK Research Scholar, Assistant Professor, Department of CSE, JNTUK-UCEV, Vizianagaram, AP, India.

[2]HOD-IT & Professor, Department of IT, JNTUK-UCEV, Vizianagaram, AP, India.

Email: arunakumarip.cse@jntukucev.ac.in, gjsuma.it@jntukucev.ac.in

Received: 26th October 2018, Accepted: 29th January 2019, Published: 28th February 2019

Abstract

In real world applications, training the classifier using unbalanced dataset is the major problem, as it decreases the performance of Machine Learning algorithms. Unbalanced dataset can be prominently classified based on Support Vector Machine (SVM) which uses Kernel technique to find decision boundary. High Dimensionality and uneven distribution of data has a significant impact on the decision boundary. By employing Feature selection (FS) high dimensionality of data can be solved by selecting prominent features. It is usually applied as a pre-processing step in both soft computing and machine learning tasks. FS is employed in different applications with a variety of purposes: to overcome the curse of dimensionality, to speed up the classification model construction, to help unravel and interpret the innate structure of data sets, to streamline data collection when the measurement cost of attributes are considered and to remove irrelevant and redundant features thus improving classification performance. Hence, in this paper, two different FS approaches has been proposed namely Fuzzy Rough set based FS and Fuzzy Soft set based FS. After FS the reduced dataset has been given to the proposed Iterative Fuzzy Support Vector Machine (IFSVM) for classification which has considered two different membership functions. The Experiments has been carried out on four different data sets namely Thyroid, Breast Cancer, Thoracic surgery, and Heart Disease. The results shown that the classification accuracy is better for Fuzzy Rough set based FS when compared other.

Keywords

Support Vector Machine, Fuzzy Logic, Rough Sets, Soft Sets, Feature Selection

Introduction

One of the most well notorious supervised machine learning algorithms for classification or prediction is SVM [1]. By calculating a decision boundary called hyper plane, SVM performs the classification of the data points. One of the vital benefits of the SVM is that it can, with relative ease, conquer the high dimensionality problem [2]. It can handle a dataset with a few numbers of instances and with large feature space because of its data driven nature, discriminative power for classifying data points. It has been successfully employed in health care analytics which has improved prediction accuracy [3, 4]. Moreover, in the field of bioinformatics the classification performance has been greatly improved by using SVM [5, 6].

In many real world applications, the training data obtained has been often contaminated by noises. Moreover, some of the data points in the given dataset may be located away from original data space or may be migrated towards incorrect class. One of the main downside of the standard SVM is that the outliers and noisy points present in the dataset can use overfitting while training SVM. In many real world applications of classification due to presence of noise and outliers, a training point may not belongs to one of the class exactly or cannot be considered as noise point. The data point nearer to decision boundary may belong to one of the class or it may be a noisy point. However, this type of uncertainty data points plays a vital role in decision making process because which drives the training process towards overfitting problem. This uncertainty problem can be best handled by fuzzy approaches which diminishes the role of less significant data [7]. In this approach a fuzzy membership value has been calculated for each data point which will be considered as a weight for that point. Basing on these weights the significance of the data points will be evaluated. So many fuzzy approaches are developed and proposed in literature to reduce the effect of outliers. In [8] the fuzzy memberships have been calculated by adopting a similarity measure. Nevertheless, this approach has made an assumption that the outliers must exhibit some variations compares to normal data points. A triangular membership function has been employed for all the data points to eliminate the effect of the parameter C while modeling traditional binary class SVM. On the other hand, based on some assumptions this approach can be adopted [9]. Above two problems are solved by Fuzzy SVM.

The method proposed in [10] is based on the supposition that outliers in the training vector set are less trustworthy, and hence of less significant over other training vectors. As outliers are detected based solely on their relative distance from their class mean, this method may be expected to produce good results if the distributions of training vectors x_i of each class are spherical with central means (in the space used to calculate the memberships). In general, however, this assumption may not hold, which motivates us to seek a more

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

universally applicable method. Hence, calculating fuzzy membership values has been still became a major challenge. This problem can be solved by IFSVM.

Generally fuzzy approach based machine learning techniques faces two main difficulties that are how to set fuzzy memberships and how to decrease computational complexity. In literature, it was known clearly that the performance of any fuzzy SVM has been significantly affected by the calculation of fuzzy membership values. Therefore, here a new approach has been proposed in this paper for the calculation these membership values. Instead of calculating membership values for all data points, only for misclassified data points this membership values will be calculated in training process. Two approaches have been proposed for calculating fuzzy membership values for misclassified points. In the first approach, based on the positions of training vectors with respect to decision boundary iteratively the membership values has been generated. In second approach, which is fuzzy clustering approach, a clustering method has been employed on the training data to obtain the clusters. Then for the clusters in mixed regions the membership values has been set to 1 and for the remaining points the membership values has been calculated based on nearest clusters respectively. After calculation these membership values then data set has been fed to IFSVM for training process. In this paper, proposed a new approach as depicted in fig1 along with membership values for reducing the misclassification rate that is set the threshold value based on membership values, after training Fuzzy approaches still the point is misclassified but membership value is greater than the threshold it is predicated as actual class.

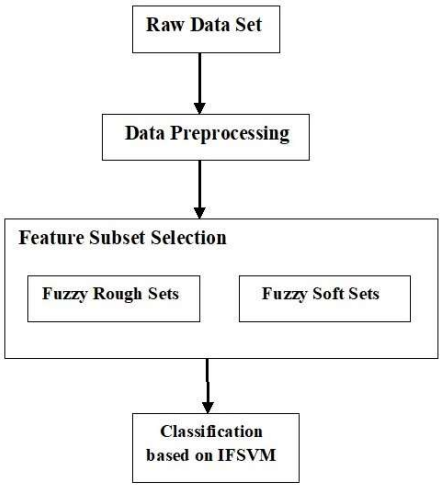

Fig 1: Proposed Methodology

The paper is organized as follows. The preprocessing of raw data has been presented in Section2 and the proposed FS approaches have been discussed in Section 3. The classification using IFSVM is presented in Section4. The Experimental results and analysis on selected four datasets has been discussed in setcion5 and the work has been concluded in Section6.

Preprocessing

In this work, the data sets from UCI machine Repository have been considered. Before developing a model the data has to be analyzed and it should be understood to know the structure and relevance of features. The data includes a countable number of missing values for number of features and some of feature values are continuous and other discrete. An even more noise can be present in the data, which demands the cleaning of data in preparing the dataset for classification analysis [15]. In this hybrid approach, as a part of cleaning, missing values of attributes have been replaced by mean of all the values of the attribute and based on equal area method the continuous values have been discretized in this paper.

Feature Selection

After preprocessing feature subset selection has been employed. FS refers to the problem of selecting those input attributes that are most predictive of a given outcome. Unlike other dimensionality reduction methods, feature selectors preserve the original meaning of the features after reduction. FS has been mainly employed where huge number of features or presence of large feature space has found in the given dataset. Because the noisy features, irrelevant and redundant feature present in the given feature space may leads to poor classification performance. FS techniques have also been applied to small and medium-sized datasets in order to locate the most informative features for later use. The importance of feature selection is to reduce the problem size by improving the quality and speed of classification. In this work FS has been performed based on fuzzy rough sets and fuzzy soft sets.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

1. Fuzzy Rough

The most important and widely used concept of fuzzy rough was originated by Dubois and prade [11]. This demonstrates the power of fuzzy-rough set theory in handling the vagueness and uncertainty often present in data. Because of fuzzy nature present in most of real world problems, expansion of rough approximation into fuzzy space helps in solving real world problems. In that first fuzzification is done after calculate the dependency of each attribute is calculated after select highest dependency attribute continue this procedure until dependency does not change. At last attribute set is obtained with reduced number of features.

2. Fuzzy Soft

Fuzzy Soft sets are the combination of both fuzzy sets and soft sets. Fuzzification is a process of converting continuous data into categorical data by assigning a membership value ranging from [0, 1]. Fuzzification is done by applying membership functions. There are many membership functions like triangular, sigmoid, trapezoidal member ship functions. Triangular membership function has been selected for our work because of simple formula and computational efficiency. In Fuzzy Soft sets first Fuzzification will be done and next normal parameter reduction will be done.

Definition

Let $\Psi(U)$ denote the set of all fuzzy sets of U. Let $A_i \subset E$. A pair (F_i, A_i) is called a fuzzy soft set over U, where F_i is a mapping given by $F_i : A_i \rightarrow \Psi(U)$[12].

Classification Based on IFSVM

After selecting the relevant features from the dataset then train with IFSVM classification technique. In this work fuzzy approach based machine learning technique has been implemented. One of the main drawbacks of the standard SVM is that the training process of the SVM is sensitive to the outliers or noise in the training dataset due to over fitting. In many real world applications community, due to over-fitting problem in SVMs, the training process is particularly sensitive to those sample points which are far away from their own class in the training dataset. While performing classification, each data point receives equal importance by SVM. But Different input points can make different contributions to the learning of decision surface. But these kinds of uncertainty points may be more important than others for making decision, which leads to the problem of over fitting. As fuzzy approaches are effective in solving uncertain problems, this problem for SVM can be handled with IFSVM. It is very important to assign each data point in the training dataset with a membership in order to decrease the effect of those outliers or noises.

Data points with large membership value can be treated as more envoy point of that class where as the points with small membership value should be considered as less significant point, then the contribution of abnormal data points with minimum membership towards error will get reduced. In fact, this fuzzy at present, Fuzzy based ML techniques faces two main difficulties: How to set fuzzy memberships and how to decrease computational complexity. But computing fuzzy memberships is still a challenge. This paper majorly focused on only two methods for calculation of membership values namely iterative and fuzzy clustering approaches. This membership value significantly gives the relative importance of each data point for accurate classification.

1. Determining Membership Values

1.1 Iterative Approach

The membership values for only misclassified points by using two membership values have been calculated as follows:

Calculated the $\xi_i = \max\{0, 1 - d_i g(x_i)\}$ i.e slack variable

For Membership Function1 (MF1) ξ_i is given as

$$l_{cnt}(\xi_i) = \begin{cases} 0 & \text{if } \xi_i \leq 0 \\ \xi_i & \text{if } 0 \leq \xi_i \leq 0 \\ 1 + \ln(\xi_i) & \text{if } \xi_i \geq 1 \end{cases}$$

For Membership Function2 (MF2) slack variable is $l_{sig}(\xi i) = \tanh(\xi_i)$

The MF1 applied is given blow:

$$h_{cnt}(\xi) = h(\xi) = 1 \text{ if } \xi<1$$
$$= 1/\xi \text{ otherwise}$$

Where, the membership is inversely proportional to the distance from the hyper plane.

The MF2employed is presented below:

$$h_{sig}(\xi) = \text{sech2}(\xi)$$

Where, $\xi_i = \max\{0, 1 - d_i g(x_i)\}$, d_i is class label, and $g(x_i) = w^T \phi(x_i) + b$.

1.2 Fuzzy Clustering

The steps applied in fuzzy clustering are as follows:

1) Clustering approach has been selected

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

2) On the given dataset, the selected clustering approach has been applied
3) From the set of clusters formed by the given approach, identified the clusters which contain normal data points and abnormal data points. And then mark the set of cluster as MIXEDCLUS.
4) For each data point x in MIXEDCLUS, fuzzy membership value has been set to1
5) For each data point x not in MIXEDCLUS, Identified the center of closest cluster to x and then calculated fuzzy membership of x with that cluster.

2. IFSVM

In IFSVM, membership values are generated iteratively based on the positions of training vectors relative to the SVM decision surface itself. The calculation of the membership values using an IFSVM that makes no a-priori assumptions about the shape of the distribution of the training vectors. This method makes use of the result of the SVM training process and information about incorrectly classified training vectors (error vectors) to tune the membership values. The FSVM is then retrained with these new values, and the process repeated either for a fixed number of iterations or until the membership values converge [14].

Algorithm

Step 1:

SVM has been applied on the given dataset. For applying or Training SVM on given dataset we need to calculate α, w, b values. $\alpha_1 \ldots \alpha_N$ values have been calculated such that

$$\mathbf{Q(\alpha)} = \Sigma\alpha_i - \tfrac{1}{2}\Sigma\Sigma\alpha_i\alpha_j y_i y_j \mathbf{x_i}^T\mathbf{x_j} \text{ is maximized and } \Sigma\alpha_i y_i = 0,\ 0 \le \alpha_i \le C \text{ for all } \alpha$$

After calculating α value w value has been calculated such that

$$\text{Weight vector} = \sum\alpha_i y_i x_i (0<=\alpha_i<=C)$$

Then b value is obtained using equation.

$$\text{Bias} = 1/n\sum_{sv}(Y_{sv} - \sum\alpha_i y_i \mathbf{x_i}.\mathbf{x_{sv}})$$

Classification:

If $w^T\phi(x_i) + b > 0$ is 1
else $w^T\phi(x_i) + b < 0$ is -1

Misclassified points have been determined by comparing original data with predicted classes. Calculated MF values for each misclassified point using MF1 and MF2.

Step 2:

1) Set s = 1.
2) Solved the FSVM dual training problem. $\alpha_1 \ldots \alpha_N$ values haven calculated such that
   $Q(\alpha) = \Sigma\alpha_i - \tfrac{1}{2}\Sigma\Sigma\alpha_i\alpha_j y_i y_j x_i^T x_j$ is maximized and $\Sigma\alpha_i y_i = 0,\ 0 \le \alpha_i \le S_i * C$ for all α
3) For all $i \in Z_N$ set (where $0 < \mu < 1$):
   $s_i = s_{previousi} + \mu(h(\xi_i) - s_{previousi})$
4) Stop if termination condition has been met.
5) Otherwise repeat from step 2

The termination conditions employed are:

- The first (and simpler) termination condition is to stop after a fixed number (n) of iterations
- The second termination condition is to continue until the rate of change of s becomes sufficiently small; indicating that the class membership vector s has converged to some value. $s_i - s_{previous\ i} \le 10{-}3\ \forall_i \in Z_N$.

Step 3: The threshold value has been fixed based on MF.

Step 4: Prediction:

If predicate=actual Then Final prediction=predicted classes
Else if predicated =! Actual and Threshold value<=MF Then Final prediction = Predicted class
Otherwise Final prediction= Actual class

Step 5: Calculated Accuracy.

Experimental Results and Analysis

The experiments are carried out on various datasets namely Thyroid, Thoracic Surgery, Heart Disease, and Breast Cancer which have been taken from UCI Machine Learning Repository. The number of instances and features of each dataset has been presented in Table1. The reduced number of features after two FS approaches has been presented in table1. Three different kernel functions linear, RBF, and poly have been used in IFSVM. The performance of IFSVM with respect to these three kernel functions before FS and after two FS approaches have been presented in table2. The results shown that the classification accuracy has been improved after FS for each dataset. And the accuracy is high in fuzzy soft FS approach when compared to fuzzy rough FS approach for all the given datasets. But the time taken to select the prominent features is more in fuzzy soft approach when compared to fuzzy rough approach.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Name of the Data Set	Number Of Instances	Number of Attributes		
		Before Feature Selection	Fuzzy Soft Feature Selection	Fuzzy Rough Feature Selection
Thyroid	3772	30	21	16
Thoracic Surgery	470	16	7	14
Heart Disease	270	13	10	7
Breast Cancer	699	9	8	8

Table 1: Comparisons of the Different Feature Selection Approaches

Data Set	Classifier	Kernel Function	Membership Function	Prediction Accuracy		
				Before Feature Selection	Fuzzy Soft Feature Selection	Fuzzy Rough Feature Selection
Thyroid	IFSVM	Linear	MF1	96.78	93.67	97.13
			MF2	94.78	98.53	98.41
		RBF	MF1	90.61	90.62	98.89
			MF2	87.67	89.56	97.56
		Poly	MF1	93.45	97.53	97.53
			MF2	93.5	97.53	97.93
Breast Cancer	IFSVM	Linear	MF1	99.14	99.8	98.29
			MF2	99.14	99	97.43
		RBF	MF1	99.14	99.87	99.14
			MF2	99.8	99.14	96.58
		Poly	MF1	99.25	99.5	98.29
			MF2	99.14	99.5	68.29
Thoracic Surgery	IFSVM	Linear	MF1	96	99.5	99.5
			MF2	96.2	96	99.5
		RBF	MF1	99.56	98	99.5
			MF2	96	97	99.5
		Poly	MF1	96.2	98.93	97.46
			MF2	95	95.23	96
Heart Disease	IFSVM	Linear	MF1	99.5	95.5	99.56
			MF2	97.7	95.5	99.5
		RBF	MF1	99.25	97.7	98.52
			MF2	97.7	95.5	98.52
		Poly	MF1	97.7	95.5	97.7
			MF2	95.5	93.3	97.2

Table 2: Comparisons of Performance of the Classifier

Conclusion

In this work, the proposed Fuzzy Approach based machine learning technique has been generated great results on four datasets. The fuzzy rough and fuzzy soft approaches for FS have been employed which has greatly improved the classification performance. Among these FS approaches, Fuzzy soft has given better results with high time complexity. The proposed IFSVM has given promising results on different medical data sets. This can be extended for multi-classification problems where unclassifiable regions may exist if a data point belongs to more than one class or does not belong to any class. When no of classes increases time complexity also increases. This problem can be handled novel model based on support vector domain combined with kernel-based fuzzy clustering.

References

[1] V. N. Vapnik, The nature of statistical learning theory, Springer-Verlag, New York, 1955.
[2] X.G. Zhang, Using class-center vectors to build support vector machines, in proceedings of IEEE signal processing society workshop, Madison, USA, 1999, pp. 3-11.
[3] S. Shilaskar, A. Ghatol , Feature selection for medical diagnosis : Evaluation for cardiovascular diseases, Expert Systems with Applications 40 (2013) 4146–4153.
[4] West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnosis decision support system: A breast cancer diagnosis application. European Journal of Operational Research(162), 532–551.
[5] K. Rajeswari, V.Vaithiyanathan, Fuzzy based modeling for diabetic decision support using Artificial Neural Netwotk, International Journal of Computer Science and Network Security, Vol. 11 No.4, April 2011.
[6] Z. Pawlak, rough sets, International Journal of Computer & Information Sciences, vol. 11 (5), 1982.
[7] Z. Pawlak, Hard and soft sets, ICS research report, Institute of Computer Science, Poland, 1994.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

[8] R.S. Huan Liu, H. Motoda, Z.Zhao, Feature selection: an ever evolving frontier in data mining, JMLR: workshop and Conference proceedings 10: the fourth workshop on feature selection in data mining, 2010, pp.4-13.
[9] A. R. Roy & P.K. Maji, A fuzzy Soft set- theoretic approach to decision making problems, Journal of computational and applied mathematics, vol.203, pp. 412-418, 2007.
[10] D. Dubois, H. Prade, rough fuzzy sets and fuzzy rough sets, International Journal of General Systems, 17(23), 1990.
[11] C. F. Lin, S. D. Wang, Fuzzy support vector machines, IEEE transactions on neural networks, 13(2002) 268-276.
[12] X. F. Jiang, Z. Yi, J. C. Lv, fuzzy SVM with a new membership function, Neural Computer Applications, 15(2006) 268-276.
[13] P.K.Maji, R. Biswas, A. R. Roy, Fuzzy soft sets, Journal of fuzzy Mathematics, 9(3), 2001, 589-602.
[14] A.Shilton, D. T. Lai, Iterative fuzzy support vector machine classification, in fuzzy systems conference, 2007, FUZZ-IEEE 2007,pp. 1-6.
[15] G. H. Lee, J.S. Taur, C. W. Tao, A robust fuzzy support vector machine for two-class pattern classification, international journal of fuzzy systems, 8(2), 2006, 76-87.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495