

# [In of attributes are considered and to](https://assignbuster.com/in-of-attributes-are-considered-and-to/)

[](https://assignbuster.com/)[Design](https://assignbuster.com/essay-subjects/design/)

In real world applications, training the classifierusing unbalanced dataset is the major problem, as it decreases the performanceof Machine Learning algorithms. Unbalanced dataset can be prominently classifiedbased on Support Vector Machine (SVM) which uses Kernel technique to finddecision boundary. High Dimensionality and uneven distribution of data has asignificant impact on the decision boundary. By employing Feature selection (FS) high dimensionalityof data can be solved by selecting prominent features. It is usually applied asa pre-processing step in both soft computing and machine learning tasks. FS isemployed in different applications with a variety of purposes: to overcome thecurse of dimensionality, to speed up the classification model construction, tohelp unravel and interpret the innate structure of data sets, to streamlinedata collection when the measurement cost of attributes are considered and toremove irrelevant and redundant features thus improving classificationperformance. Hence, in this paper, two different FS approaches has beenproposed namely Fuzzy Rough set based FS and Fuzzy Soft set based FS.

After FSthe reduced dataset has been given to the proposed Iterative Fuzzy SupportVector Machine (IFSVM) for classification which has considered two differentmembership functions. The Experiments has been carried out on four differentdata sets namely Thyroid, Breast Cancer, Thoracic surgery, and Heart Disease. The results shown that the classification accuracy is better for Fuzzy Roughset based FS when compared other.

Keywords: Support Vector Machine, Fuzzy logic, RoughSets, Soft Sets, Feature selection.————————————————————————————————————————————— 1. Introduction:      SVM is one of the mostwell­ known supervised machine learning algorithms for classification orprediction method developed by Cortes and Vapnik 1 in the 1990s as a resultof the collaboration between the statistical and the machine learning researchcommunity. SVM tries to classify cases by finding a separating boundary calledhyper plane. The main advantage of the SVM is that it can, with relative ease, overcome ‘ the high dimensionality problem’, i. e.

, the problem that arises whenthere is a large number of input variables relative to the number of availableobservations 2. Also, because the SVM approach is data­  driven and possible without theoreticalframework, it may have important discriminative power for classification, especially in the cases where sample sizes are small and a large number offeatures (variables) are involved (i. e., high dimensional space).

Thistechnique has recently been used to improve methods for detecting diseases inclinical settings 3, 4. Moreover, SVM has demonstrated high performance insolving classification problems in bioinformatics 5, 6.       In many practicalengineering applications, the obtained training data is often contaminated bynoises. Furthermore, some points in the training data set are misplaced faraway from main body or even on the wrong side in feature space. One of the maindrawbacks of the standard SVM is that the training process of the SVM issensitive to the outliers or noise in the training dataset due to over fitting. A training data point may neither exactly belong to any of the two classes whenthe outliers or noises exist in many real-world classification problems. The datapoint nearer to decision boundary may belong to one of the class or it may be anoisy point.

But these kinds of uncertainty points may be more important thanothers for making decision, which leads to the problem of over fitting. Fuzzyapproaches are effective in solving uncertain problems, which reduces thesensitivity of less important data 7. This approach assigns a fuzzymembership value as a weight to each training data point and uses this weightto control the importance of the corresponding data point. So many fuzzyapproaches are developed and proposed in literature to reduce the effect ofoutliers.

A similarity measure function to compute fuzzy memberships wereintroduced in 8. However, they had to assume that outliers should be somewhatseparate from the normal data. The effect of the trade-off parameter C to the model of conventionaltwo-class SVM and introduced a triangular membership function to set highergrades to the data points in regions containing data of both classes.

Howeverthis method could be applied with some assumptions involved 9. Above twoproblems are solved by Fuzzy SVM.     The method proposed in 10is 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 aredetected based solely on their relative distance from their class mean, thismethod may be expected to produce good results if the distributions of trainingvectors xi of each class are spherical with central means (in thespace used to calculate the memberships).

In general, however, this assumptionmay not hold, which motivates us to seek a more universally applicable method. Hence computing fuzzy memberships is still a challenge. This problem can besolved by IFSVM.

Generally fuzzy approachbased machine learning techniques faces two main difficulties that are how toset fuzzy memberships and how to decrease computational complexity. It has beenfound that the performance of fuzzy SVM highly depends on the determination offuzzy memberships therefore in this paper; we proposed a new method to computefuzzy memberships that calculates membership values for only misclassifiedpoints and calculates membership values for all training data points. Forcalculating the membership values for misclassified points an iterative methodhas been employed where membership values are generated iteratively based onthe positions of training vectors relative to the SVM decision surface itself. For calculating the membership values for misclassified points a fuzzyclustering based technique has been adopted where  clustering method has been applied on the dataand determines  the clusters in mixedregions and set Fuzzy membership value as 1 and fuzzy memberships of other datapoints are determined by their closest cluster accordingly.