Chapter 87
Fuzzification of Euclidean Space Approach in Machine Learning Techniques

Mostafa A. Salama
British University, Egypt

Aboul Ella Hassanien
Cairo University, Egypt & Scientific Research Group in Egypt (SRGE), Egypt

ABSTRACT

Euclidian calculations represent a cornerstone in many machine learning techniques such as the Fuzzy C-Means (FCM) and Support Vector Machine (SVM) techniques. The FCM technique calculates the Euclidian distance between different data points, and the SVM technique calculates the dot product of two points in the Euclidian space. These calculations do not consider the degree of relevance of the selected features to the target class labels. This paper proposed a modification in the Euclidian space calculation for the FCM and SVM techniques based on the ranking of features extracted from evaluating the features. The authors consider the ranking as a membership value of this feature in Fuzzification of Euclidian calculations rather than using the crisp concept of feature selection, which selects some features and ignores others. Experimental results proved that applying the fuzzy value of memberships to Euclidian calculations in the FCM and SVM techniques has better accuracy than the ordinary calculating method and just ignoring the unselected features.

1. INTRODUCTION

Fuzzification algorithms have been applied in most machine learning techniques to provide more human-like behavior and are successful in increasing the performance and accuracy of the classification results. Fuzzy logic introduces to machine learning techniques a framework for dealing quantitatively, mathematically, and logically with semantic and ambiguous concepts (Kyoomarsi et al. 2009). The membership of data points in a set or class label is not crisp but can be specified as a degree of membership. The machine learning techniques under investigation in this paper are C-Means clustering and support vec-

DOI: 10.4018/978-1-5225-1759-7.ch087

tor machine (SVM), where these techniques are fuzzified to generate FCM clustering (FCM) and fuzzy SVM, respectively. In instance-based techniques such as the C-Means clustering technique, fuzzy logic is used to determine the proximity of a given instance to the training set's instances (Kotsiantis 2007). Fuzzy logic allows data instances to belong to two or more clusters where it is based on minimizing an objective or dissimilarity function (Gewenigera et al. 2010). With FCM, the centroid of a cluster is computed as the mean of all points, weighted by their degree of belonging to the cluster according to their proximity in the feature space. The degree of being in a certain cluster is related to the inverse of the distance to the cluster. SVM is a non-linear binary classification algorithm based on the theory of structural risk minimization. SVM solves complex classification tasks without suffering from the overfitting problems that affect other classification algorithms. Computationally speaking, the SVM training problem is a convex quadratic programming problem, meaning local minima are not a problem (Cortes et al. 1995). In fuzzy SVM, the fuzzy membership values are calculated based on the distribution of training vectors where the outliers are given proportionally smaller membership values compared to the other training vectors (Lee et. al. 2006), (Shilton et. al. 2007).

The problem in such fuzzified techniques is that they apply the fuzzy logic concept on the level of objects and ignore the features composing the objects. Each object has a degree of membership in the class labels in the learning problem. For multivariate objects, the features of the objects have different degrees of relevance to the target class labels. When reducing the number of features by selecting the best features or extracting a lower number of features from the higher ones, the reduced features still have different relevance to the classification problem, while irrelevant features are ignored (Janecek et. al. 2008). Consequently, classifiers deal crisply with features without considering the degree of relevance: Either the full data set is used or only the selected features. The degree of relevance is calculated with feature evaluation techniques such as ChiMerge (Abdelwadood et. al. 2007). This technique successfully ranks continuous and discrete features. Using the degree of relevance in classification techniques could enhance their accuracy.

The proposed approach adds the degree of importance in machine learning techniques, especially such techniques as FCM and SVM (Salama et. al 2011). These techniques depend on Euclidean calculations between data points in the space. For high-dimensional data sets, a popular measure, the Minkowski metric (Kim 2010), is used to calculate the distance. Euclidean is a special case of such equation. Most of the existing kernels used in linear-nonlinear SVMs measure the similarity between a pair of data instances based on the Euclidean inner product or the Euclidean distance of corresponding input instances. Calculating the Euclidean distance or product ignores the degree of relevance of each feature to the classification problem and treats all features equally. Then the fuzziness concept must be lowered from the level of data point membership degree to a specific set to the level of the feature membership degree to the classification problem. To apply this, the crisp dot product between data points in the Euclidean distance calculation is transformed into a fuzzy dot product through multiplying the dot product of each feature to the membership function of the corresponding features. The feature rankings are extracted through a feature selection called ChiMerge that calculates the χ^2 value of the features in the input data set. The resulting rankings from the ChiMerge technique are the membership degrees of the corresponding features. This step is considered hybridization between the feature selection technique and these classification techniques.

Seven data sets are used to test the importance of the fuzzification of the Euclidean distance in improving the classification accuracy of the classifiers. The data sets are medical data sets that are ex-