SIMILARITY-DISSIMILARITY PLOT FOR HIGH DIMENSIONAL DATA OF DIFFERENT ATTRIBUTE TYPES IN BIOMEDICAL DATASETS

MUHAMMAD ARIF AND SALEH BASALAMAH
College of Computer and Information Systems
Umm-Alqura University
Makkah, Kingdom of Saudi Arabia
{ mahamid; smbasalamah }@uqu.edu.sa

Received November 2010; revised March 2011

ABSTRACT. In real life biomedical classification applications, feature space may be of high dimension in which visualization of class distribution is impossible. Moreover, attributes of features may be numeric, ordinal, categorical or binary. Most of the time, features may be composed of mixed type of attributes. In this paper, the concept of similarity-dissimilarity is extended to various types of attributes. Similarity-dissimilarity plot projects the high dimensional feature space on two dimensional plot revealing the class separation in the feature space which may be continuous or discrete. Furthermore, effect of various distance measures proposed in the literature for different type of attributes is also studied. An index called percentage of data points above the similarity-dissimilarity line (PAS) is proposed which is the fraction of data points found near to its own class as compared to other classes. Several real life biomedical datasets are used to show the effectiveness of the proposed similarity-dissimilarity plot and the PAS index.

Keywords: Visualization, High dimensional data, Pattern classification, Features quality, Nearest neighbors

1. Introduction. In biomedical applications, data or feature set can be of numeric, nominal, ordinal or categorical type. Moreover, due to high dimensionality of the data, it is not possible to visualize the data and extract important information about the structure of the data in the context of pattern classification. In pattern classification problems [1, 2], raw data are measured from sensors, images or clinical tests and features are extracted so that different classes may be discriminated in the feature space in an appropriate way. Classifier of arbitrary type and settings are applied to these features to get best possible classification accuracy. Classifiers are optimized using classification accuracy as an optimization criterion. However, one question remains unanswered that whether certain classification accuracy is due to the selection of classifier or poor discrimination quality of the feature set. It is easy to visualize different clustering patterns belonging to different classes in low dimensional feature space (up to three dimensions). Hence, quality of features can be assessed easily. However, in the high dimensional feature space, specialized visualization tools are needed that can help in projecting the data on two or three dimensional space in a meaningful and more descriptive manner. Purpose of visualization is to study intra-class and inter-class relationships so that quality of the features in discrimination of different classes can be assessed.

In case of multivariate data sets where some or all variables are correlated to each other in some linear or nonlinear sense, projecting the high dimensional data to lower dimension by using Principle Component Analysis (PCA) [3], projection pursuit [4] or Kohonen’s self organizing map (SOM) [5,6] is very popular and applied to many real life applications. In