

# Comparison of Signed Distance Function Methods with Support Vector Machines on Binary Classification

Erik M. Boczko, Dept. Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232, erik.m.boczko@vanderbilt.edu

Todd R. Young (Corresponding author), Dept. Mathematics, Ohio University, Athens, OH 45701, young@math.ohiou.edu

Di Wu and Minhui Xie, Dept. Electrical Engineering & Computer Science, Vanderbilt University, Nashville, TN 37235, di.wu, minhui.xie@vanderbilt.edu

Efficient and accurate computational solutions for binary classification problems are currently of interest in many contexts, particularly in biomedical informatics and computational biology where the interesting genomic and proteomic data sets are imbued with dimensional complexity and confounded by noise. Over the past several years it has been effectively demonstrated that binary classification of genomic and proteomic data can be used to connect a molecular snapshot of an individual's internal state with the presence or absence of a disease. This potential promises to revolutionize personalized medicine and is fueling the development and analysis of robust classification algorithms. Among the existing classification algorithms Support Vector Machine (SVM) methods have distinguished themselves as efficient, accurate and robust. Applications of Radial Basis Function Networks (RBFN) to classification have also generated much attention.

We consider only a geometric (rather than statistical) formulation of the binary classification problem. Namely, we suppose that the space of measurements  $X$  is divided into two subsets,  $A$  and its complement  $A^c = X \setminus A$ . We are given data  $\{x_i\}$  for which we know the membership in  $A$  or  $A^c$  of each data point. From this data the binary classification problem is to construct a rule or *classifier* that we can use to predict the class of new, uncharacterized data. As an example, the data may be measurements of genomic activation levels, one class might be measures from individuals known to have a certain type of disease while the other class may be from individuals without the disease.

The linear SVM method was originally designed to be geometric and robust through a constraint that it produce a dividing surface of maximal margin between data of opposite type. However, it has been shown that nonlinear SVM implementations actually are built around reconstruction of the indicator function that ties the location of the data to its class ([2]). The indicator function:

$$i_A(x) = \begin{cases} 1 & \text{if } x \in A \\ -1 & \text{if } x \in A^c, \end{cases}$$

encodes only the most primitive geometric information. In [1] we proposed an alternative tool for classification, the Signed Distance Function (SDF), that measures the signed distance from the data to the boundary between the classes, i.e.

$$b_A(x) = \begin{cases} d(x, A^c) & \text{if } x \in A \\ -d(x, A) & \text{if } x \in A^c, \end{cases} \quad (1)$$

where  $d$  is a distance function. While the SDF has not previously been applied to classification, it has been an important tool in other fields, such as free boundary problems in fluid dynamics, and so has a rich mathematical development that could be exploited. We have tested rudimentary classification algorithms based on the idea of reconstructing the SDF from training data. We note

that this reconstruction could be based on any accepted method of regression, including SVM or RBFN regression. Thus, new SVM or RBFN classification methods could be built on the SDF foundation.

We investigate the performance of a simple SDF based method by direct comparison with standard SVM packages, as well as K-nearest neighbor and RBFN methods. We present experimental results comparing the SDF approach with other classifiers on both synthetic geometric problems and five benchmark clinical microarray data sets. On both geometric problems and microarray data sets, the non-optimized SDF based classifiers perform just as well or slightly better than well-developed, standard SVM methods. These results demonstrate the potential accuracy of SDF-based methods on some types of problems.

[1] E.M. Boczko and T. Young, Signed distance functions: A new tool in binary classification, preprint.

[2] T. Poggio and S. Smale, The mathematics of learning: Dealing with data, *Notices Amer. Math. Soc.*, 50:537–544, 2003.