BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–3

R/dwd: Distance Weighted Discrimination and Second Order Cone Programming

Hanwen Huang [1,2,*], Xiaosun Lu [1,2], Yufeng Liu [1], J. S. Marron [1], Perry Haaland [2,*]

[1]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA.
[2]BD Technologies, 21 Davis Drive, RTP, NC 27709 USA.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: R/dwd is an extensible package for classification and optimization. It provides a new classification tool using a recently developed Distance Weighted Discrimination (DWD) method which is related to, but has been shown to be superior to, Support Vector Machine (SVM) in some high dimensional situations. The key step in the implementation of DWD is to solve a Second Order Cone Programming (SOCP) problem. R/dwd also provides an efficient SOCP solver based on the interior point method originally developed by Toh *et al.*, 1999 for solving semidefinite-quadratic-linear programming. In addition, an efficient Quadratic Programming (QP) solver based on the SOCP is also included in this package. We believe that this package will be very useful for many aspects of bioinformatics research.

Availability: The package is freely available from cran.r-project.org.

Contact: hanwenh@email.unc.edu or Perry_Haaland@bd.com

1 INTRODUCTION

The Support Vector Machine (SVM) is a well known machine learning technique and has achieved great success in bioinformatics applications (see Byvatov *et al.*, 2003 for a review). However, as shown in Marron *et al.*, 2007, SVM suffers from the data piling problem in High Dimensional Low Sample Size (HDLSS) situations. Distance Weighted Discrimination (DWD) is a recently developed classification method which was originally motivated for solving the HDLSS problems (Marron *et al.*, 2007), but can be applied to many other examples as well. One of the big advantages of DWD over SVM is that it can overcome the data piling problem in high dimensional situations as illustrated in Figure 1. The original DWD paper (Marron *et al.*, 2007) only described the implementation of the binary classification method. The multiclass version of DWD has also been developed in Huang *et al.*, 2011. It is suggested that all users refer to these publications in order to understand the DWD terminology and principles in greater detail.

DWD has been widely used in many bioinformatics areas including adjusting batch effects in microarray expression data (Benito *et al.*, 2004) and discovering cancer subtypes (Hu *et al.*, 2006). However, the existing DWD package was written in Matlab which is not a free software package. To make it more convenient for bioinformatics people to use, we have developed an R version of DWD here. R/dwd is based on the existing Matlab version (http://www.unc.edu/m̃arron/marron_software.html), but includes some additional features such as multiclass version. For the convenience of the users who are familiar with using SVM in R, the main classification functions and arguments in `R/dwd` are formatted in a similar way to the one used by the SVM functions in the `kernlab` package. To help the users in using this software, some examples to illustrate the coding are provided.

*to whom correspondence should be addressed

2 OPTIMIZATION ALGORITHM

The implementation of DWD is more challenging than the implementation of SVM because it requires solving an optimization problem called Second Order Cone Programming (SOCP). The DWD Matlab version employed a very efficient SOCP solver from the `SDPT3` (semidefinite-quadratic-linear programming) package developed by Toh *et al.*, 1999. The `SDPT3` implemented an infeasible path-following algorithm for solving conic optimization problems involving semidefinite, second-order and linear cone constraints.

Central to R/dwd is the SOCP solver which was implemented using exactly the same algorithm as the one used by the corresponding Matlab version. The optimization problem that underlies SVM is called Quadratic Programming (QP). An efficient R/QP solver based on the SOCP is also included in this package which provides a useful tool for those users who want to make their own modifications of the SVM method.

3 FEATURES

3.1 Classification

Both SVM and DWD are margin-based classification methods in the sense that they build the classifier through finding a decision boundary to separate the classes (Liu, *et al*, 2011). DWD uses a different criterion from SVM. It seeks to achieve the goal by maximizing the average distance rather than the minimum distance among the classes. Similar to the `ksvm` function from `kernlab`,

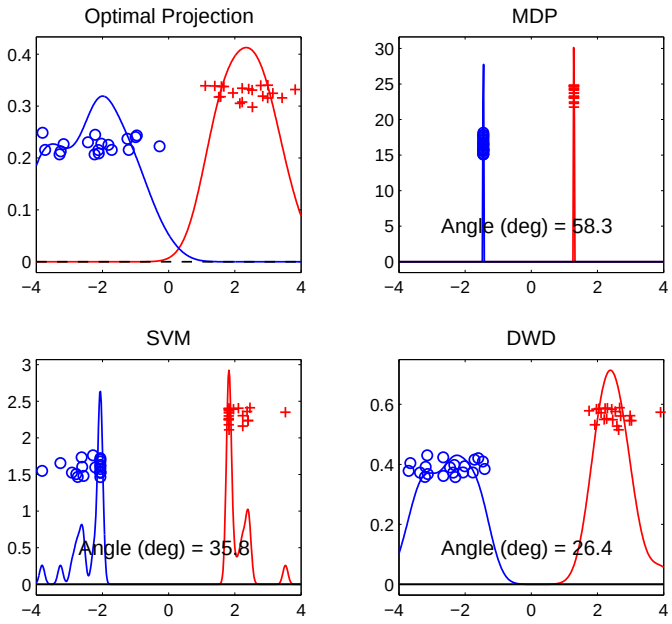

Fig. 1. Superior Performance of DWD over SVM in HDLSS settings. The toy example has dimension $d = 50$, with $n_+ = 20$ data vectors from Class +1 (red plus signs), and $n_- = 20$ data vectors from Class -1 (blue circles). The data were drawn from 2 distributions that are nearly standard normal, except that the mean in the first dimension is shifted to +2.2 (-2.2) for Class +1 (-1). Four important one-dimensional projections are shown here with the horizontal coordinate representing the projection, and with a random vertical coordinate used for visual separation of the points. The top left, top right, bottom left and bottom right are the projections onto the true optimal, Maximum Data Piling (MDP, Ahn *et al.*, 2010), SVM and DWD directions respectively. The DWD direction is closer to the optimal direction than the SVM direction as indicated by both the kernel density estimation (solid curve) and the calculated angle to the optimal direction. The Maximum Data Piling direction in HDLSS is illustrated by the top right plot.

which has been widely used in SVM analysis, we developed a function called `kdwd` in this package for doing DWD analysis.

To solve multiclass classification problem, two different methods are used. The first one is to use the "one-versus-one" approach, in which classifiers are trained on each pair of classes and the class label is predicted by a voting scheme. The second one is to build a single classifier including all classes simultaneously and solve a big optimization problem.

Every DWD analysis requires two elements from a dataset: (1) a matrix of predictors, which should be in the form of an $n \times d$ matrix, where n represents the sample size and d represents the dimension. (2) a response vector of length n with each element corresponding to one sample. The appropriate scaling steps can be taken by using the `scaled` option. It should be noted that in the current version of `DWD`, missing values are not allowed, and must be imputed prior to analysis.

The basic output from `kdwd` is an object of class *kdwd*. Showing objects of class *kdwd* will print details on the results for all classifiers included in the model. For each classifier, the optimal solution of the parameters are displayed along with the final primal and dual objective values. The cross validation error rate can also be returned with the argument `cross = k`, where `k` is the number of folds used in the cross-validation.

3.2 Optimization

The application of SOCP is not limited to DWD, it can be used to solve many other problems as well. The package `R/dwd` provides a stand-alone SOCP solver called `sqlp`. The algorithm implemented in `sqlp` is an infeasible primal-dual path-following algorithm, described in detail in Toh *et al.*, 1999. The basic idea is that, at each iteration, we first compute a predictor search direction aimed at decreasing the gap between the primal and dual objective values. After that, the algorithm generates a corrector step with the goal of keeping the iterations close to the central path. The most crucial part is to solve a linear system which is especially challenging in situations when a big sparse matrix is included. To increase the speed and efficiency, the sparse matrix package `Matrix` is incorporated to deal with high dimensional large datasets.

QP is a special case of SOCP. SVM is only one of the important applications of QP, but QP can be applied to many other areas as well. The QP solver in this package is formatted in a similar way but has been shown to be more numerically stable than the existing QP solver `solve.QP` in `quadprob` package.

4 DISCUSSION

The main purpose of R/dwd is to do classification and optimization, both of which are very important in bioinformatics fields. In classification, the DWD method is implemented. In optimization, the efficient SOCP and QP solvers are provided which can be integrated into a variety of applications. R/dwd is under continual development. Only the linear discrimination method is considered in the current version. Future plans include incorporating some kernel tricks into the DWD method such that it can be used to solve more general nonlinear problems. In addition to prediction of class labels, we are also investigating ways to predict the probability of a data point belonging to a certain class based on the given measurements. Moreover, we also intend to develop some imputation methods to handle missing values in the input file.

REFERENCES

[1]Ahn, J. and Marron, J. S. (2010), *The Maximal Data Piling Direction for Discrimination*, Biometrika, 97(1):254-259

[2]Byvatov E, Schneider G. (2003), *Support vector machine applications in bioinformatics*, Appl Bioinformatics. 2003;2(2):67-77.

[3]Monica Benito, Joel Parker, Quan Du, Junyuan Wu, Dong Xiang, Charles M. Perou, and J. S. Marron (2004), *Adjustment of systematic microarray data biases*, Bioinformatics (2004) 20 (1): 105-114.

[4]Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, Nobel A, Parker J, Ewend MG, Sawyer LR, Wu J, Liu Y, Nanda R, Tretiakova M, Ruiz Orrico A, Dreher D, Palazzo JP, Perreard L, Nelson E, Mone M, Hansen H, Mullins M, Quackenbush JF, Ellis MJ, Olopade OI, Bernard PS, Perou CM. *The molecular portraits of breast tumors are conserved across microarray platforms*. BMC Genomics. 2006 Apr 27;7:96.

[5]H. Huang, Y. Liu, Y. Du, C.M. Perou, D.N. Hayes, M.J. Todd, and J.S. Marron, (2011), *Multiclass distance weighted discrimination with applications to batch adjustment*, submitted.

[Karatzoglou]Alexandros Karatzoglou, Alex Smola, Kurt Hornik, *Kernel-based Machine Learning Lab*, CRAN - Package kernlab.

[6]Liu, Y., Zhang, H. H., and Wu, Y. (2011). *Soft or hard classification? Large margin unified machines*. Journal of the American Statistical Association, 106, 166-177.

[7]Marron, J. S., Todd, M. J. and Ahn,J., (2007), *Distance-Weighted Discrimination*. Journal of the American Statistical Association, Vol. 102, No. 480, pp. 1267-1271.

[8]K.C. Toh, M.J. Todd, and R.H. Tutuncu, *SDPT3 — a Matlab software package for semidefinite programming*, Optimization Methods and Software, 11 (1999), pp. 545–581.

[Turlach]Berwin A. Turlach, *quadprog: Functions to solve Quadratic Programming Problems*, CRAN - Package quadprog