ABSTRACT

Summary: A top scoring pair (TSP) classifier consists of a pair of variables whose relative ordering can be used for accurately predicting the class label of a sample. This classification rule has the advantage of being easily interpretable and more robust against technical variations in data, as those due to different microarray platforms. Here we describe a parallel implementation of this classifier which significantly reduces the training time, and a number of extensions, including a multi-class approach, which has the potential of improving the classification performance.

Availability and Implementation: Full C++ source code and R package Rgtsp are freely available from http://lausanne.isb-sib.ch/~vpopovic/research/. The implementation relies on existing OpenMP libraries.

Contact: vlad.popovici@isb-sib.ch

Received on November 12, 2010; revised and accepted on April 8, 2011

1 INTRODUCTION

Top scoring pairs (TSPs; Geman et al., 2004) are simple two-variables binary classifiers, in which the prediction of the class label is based solely on the relative ranking of the expression levels of the two genes. The rank-based approach to classification ensures a higher degree of robustness to technical variations and makes the rule easily portable across platforms. Also, the direct comparison of variables binary classifiers, in which the prediction of the class label is based solely on the relative ranking of the expression levels of variables making its usage impractical, especially in the context of resampling techniques for performance estimation. However, most if not all of the modern desktop computers are multi-core machines, making parallel programs a feasible alternative to classical serial ones.

Our implementation in C++ exploits the multi-core architecture by using the OpenMP libraries of the system (Chapman et al., 2007), and is wrapped in an R package – Rgtsp. The full source code and the R functions are available from http://lausanne.isb-sib.ch/~vpopovic/research/. As C++ is the main implementation language, the library can easily be extended and integrated with other software libraries. Also, the R functions are independent of the domain of application so they could be applied to any kind of data.

*To whom correspondence should be addressed.

Fig. 1. Predicting estrogen receptor status: if GSTP1 < ESR1, then the sample is considered ER+ (circles), otherwise ER− (triangles).
3 USAGE EXAMPLES

We present a typical case of using Rgtsp package. These examples represent solely some code snippets and not the full process of developing and assessing the performance of a classifier.

The data used in these examples consists of 130 samples stage I to III breast cancer (Hess et al., 2006) and the goal is to predict the estrogen receptor status (positive or negative coded with ‘+1’ and ‘0’, respectively). For illustration purposes we use only a subset of full dataset available from GEO repository under accession number GSE16716.

Before starting R, the user has the option of choosing the number of processing units that will be used, by setting the environment variable OMP_NUM_THREADS. If not set, it defaults to the maximum number of processing units available.

The first steps load the library and the data and build a list of TSPs (note that the matrix X contains the variables as columns):

```r
> library(Rgtsp)
> data(mdabr)
> tsp.list = tsp.n(X, y.erpos, 500)
> str(tsp.list)
> print(tsp.list)

The function tsp.nl() returns at most n TSPs as a list with three components: the first two correspond to the indexes of the selected variables and the third one contains the associated scores. A similar function, tsp.n(), returns all the TSPs that have a score larger than a specified value.

For the p-th TSP, the prediction rule can be written as: predict class ‘+1’ if $X[,tsp.list[[p]]] < X[,tsp.list[[p]]]$ and this forms the core of the predict function. The decision function for p=1 in the above example is shown in Figure 1. Given a list of TSPs one has different choices on how to obtain the final predicted labels. Currently, Rgtsp proposes two means of combining the predictions of individual TSPs: either by majority voting or by weighting the votes with the corresponding scores—giving more weight to the TSPs with better scores. This functionality is available through the predict generic function:

```r
> yp = predict(tsp.list, X, combiner="majority")
> sum(yp != y.erpos) # count the errors
1
```

By inspecting the list of TSPs, it becomes clear that there are variables that are selected many times as having always either higher or lower value than all its pairing variables. We call such a structure a TSP hub and we can construct all the hubs larger than a specified size (25 pairs for example) using

```r
> h = tsp.hub(tsp.list, min.hub.size=25)
> print(h)
Hub 1: 194 pairs
   Center: 953
   14 25 42 43 44 45 54 105 140 146 149 150 152 202 ...
```

This corresponds to a TSP hub in which the probeset colnames(X)[953] (205225_at, ESR1) has a higher expression than all other probesets in the list tsp.list. The TSP hubs can also be used in predicting the labels, through the same mechanism as above:

```r
> yph = predict(h, X, combiner="majority")
> sum(yph != y.erpos) # no. of errors: 6
```

We see that in this particular case the prediction by TSP hubs is slightly less accurate than the combined predictions of the individual TSPs.

The generalization performance of the TSPs classifiers can be estimated by various methods. The Rgtsp package provides a function for k-fold cross-validation of the binary TSP classifiers (either tsp.nl() or tsp.n() functions), cv.tsp(), which returns the training and validation performance of the classifier (it defaults to 5-fold cross-validation).

```r
> r = cv.tsp(X, y.erpos)
```

In the case of a multi-class problem, we propose to use classification trees built on top of TSPs predictions. For C>2 classes, one can train TSPs to solve each of the C(C−1)/2 pairwise binary classification problems [called one-versus-one (Hsu and Lin, 2002) or round robin (Fürnkranz, 2002) strategy] and then combine the predictions of the TSPs through a classification tree to predict the original classes. For more details the reader is referred to the package web page. This approach is implemented in the function mtsp() and makes use of the ctree() function in the party R package (y4 is an artificial 4-class label vector):

```r
> m = mtsp(X, y4)
> yp = predict(m, X)
```

**Funding:** Swiss National Science Foundation NCCR Molecular Oncology (to V.P. and M.D.); Fondation Medic (to E.B.).

**Conflict of Interest:** none declared.

**REFERENCES**


