Rgtsp: a generalized top scoring pairs package for class prediction

Vlad Popovici 1,2,∗, Eva Budinská 1,3 and Mauro Delorenzi 1

1 Bioinformatics Core Facility, Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland
2 Swiss National Center of Competence in Research Molecular Oncology, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland
3 Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic

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 have 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

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 the expression level of the genes is easily interpretable in the clinical context, making the TSPs attractive for medical tests.

Let \( x = [x_i]_{i=1,...,m} \in \mathbb{R}^m \) be a vector of measurements (e.g. gene expression) representing a sample and let the corresponding class label be \( y \), with two classes denoted by 0 and 1. Then, for all pairs of variables \( i \) and \( j \), a score is computed,

\[
s_{i,j} = P(x_i < x_j|y = 1) - P(x_i < x_j|y = 0), 1 \leq i, j \leq m \tag{1}
\]

where \( P \) are conditional probabilities estimated from the data, and the corresponding decision rule is: if \( \text{sign}(s_{i,j})x_i < \text{sign}(s_{i,j})x_j \) then predict \( y = 1 \), otherwise \( y = 0 \). The pairs are ordered by the absolute values of their scores and the top \( t \) pairs (\( t \geq 1 \)) are then considered for the final model (Xu et al. (2005); Tan et al. (2005); Geman et al. (2004)). Remarkably, training a TSP does not require the optimization of any parameter and does not depend on any threshold. Selecting a suitable value for \( t \) should be done following the usual machine learning paradigm for optimizing meta-parameters (see, for example, Hastie et al. (2001)). Figure 1 shows an example of a TSP predicting the estrogen receptor status. The decision boundary (in grey) is always a line with a slope of 1.

Fig. 1. Predicting estrogen receptor status: if GSTP1 < ESR1, then the sample is considered ER+ (red dots), otherwise ER- (blue dots).

2 IMPLEMENTATION

While the method briefly described above is simple and poses no implementation problems, using it in the context of highly dimensional data requires the evaluation of an extremely large number of pairs 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 package are available from http://lausanne.isb-sib.ch/∼vpopovic/research/. As C++ is the main implementation language, the library can be easily extended and integrated with
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 data set 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.n()` 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.s()`, 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 \( X[,tsp.list[I[p]] < X[,tsp.list[J[p]] \) 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.

For illustration purposes we use only a subset of the full data set. The following function gives more weight to the TSPs with better scores. 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
> yp = predict(h, X, combiner="majority")
> sum(yp != 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.n()` or `tsp.s()` 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)
> print(r)
```

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 (Furnkranz (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 (y/4 is an artificial 4–class label vector):

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

ACKNOWLEDGEMENT

Funding: VP and MD acknowledge the support of the Swiss National Science Foundation NCCR Molecular Oncology. EB acknowledges the support of Fondation Medic.

REFERENCES