Sequential rank agreement methods for comparison of ranked lists

April 4, 2015

Abstract

Key words: sequential rank agreement, rank agreement, order statistic, methods comparison, variable selection

1 Introduction

Ranked lists occur in many applications of statistics. Regression methods rank predictor variables according to magnitude of association with outcome, prediction models rank subjects according to their risk of an event, and genetic studies rank genes according to their difference in expression across samples. A common research question is where to stop, i.e., to decide the maximal significant rank. In the three examples this would correspond to the number of significantly associated predictor variables, the number of patients at high risk of an event, and the number of genes that are worth to pursue in further experiments, respectively.

In this article we describe some new breakthrough tools for measuring agreement across a set of lists which soon will enter the state-of-the art. The methods should be useful whenever there are multiple rankings of the same list. The idea is to define agreement based on the ranks of the first k elements in each list.

2 Methods

Consider a set of P different items $X = \{X_1, \dots, X_P\}$. An ordered list is a permutation function, $R : \{1, \dots, P\} \to \{1, \dots, P\}$, such that $R(X_p)$ is the

Table 1: Example set of ranked lists. (a) shows the ranked list of items for each of three lists, (b) presents the ranks obtained by each item in each of the three lists and (c) shows the cumulative set of items up to a given depth in the three lists.

(a)

Rank	π_1	π_2	π_3
1	A	A	B
2	B	C	A
3	C	D	E
4	D	B	C
5	E	E	D

(b)

Item	R_1	R_2	R_3
A	1	1	2
B	2	4	1
C	3	2	4
D	4	3	5
E	5	5	3

(c)

Depth	S_d
1	{A, B}
2	{A, B, C}
3	{A, B, C, D, E}
4	{A, B, C, D, E}
5	{A, B, C, D, E}

rank of item X_p in the list. The inverse mapping $\pi = R^{-1}$ assigns to rank $r \in \{1, \ldots, P\}$ the item $\pi(r)$ found at that rank. The methods described below work for a set of L lists $R_1, \ldots, R_L$, $L \geq 2$. We denote $\pi_l = R_l^{-1}$ for the corresponding inverse mappings. Panels (a) and (b) of Table 1 show a schematic example of these mappings.

The agreement of the lists regarding the rank given to an item can be measured by

$$A(p) = f(R_1(p), \ldots, R_L(p)), \tag{1}$$

for a distance function f. Throughout this paper we will use the sample standard error as our function f and hence use

$$A(p) = \sqrt{\frac{\sum_{i=1}^{L}(R_i(p) - \bar{R}(p))^2}{L-1}},$$

but other choices could be made (see the discussion). The sample standard error has an interpretation as the average distance of the individual rankings of the lists from the average ranking.

We now describe what is exemplified in Panel (c) of Table 1 and how it can be used to define *sequential rank agreement*. For an integer $1 \leq d \leq P$ we define the unique set of items found in the L top d parts of the lists, i.e., the set of items ranked less than or equal to d in any of the lists:

$$S_d = \{\pi_l(r); r \leq d, l = 1, \ldots, L\}. \tag{2}$$

The *sequential rank agreement* is the pooled standard deviation of the items found in the set S_d:

$$\text{SRA}(d) = \sqrt{\frac{\sum_{\{p \in S_d\}}(L-1)A(p)^2}{(L-1)|S_d|}}, \tag{3}$$

and small values close to zero suggests that the lists agree on the ordering while larger values suggests disagreement. If the ranked lists are identical then the value of SRA will be zero for all depths d. The sequential rank agreement can be interpreted as the average distance of the individual rankings of the lists from the average ranking for each of the items we have seen until depth d.

2.1 All lists fully observed

We shall start by the simplest case where all L lists are fully observed, *i.e.*,, we have the rank of all P items for all of the L lists. This situation occurs is common when we have the original dataset available and when we wish to, say, compare the results from different analysis methods.

For the fully obseved list situation we can plot the sequantial rank agreement (3) as a function of depth d. If there is a ... An example is seen in Figure **??**

2.2 Analysis of top k lists

Not uncommon for lists to be

censored

Let $\Lambda_l, l = 1, \ldots, L$ be the set of items found in list l so Λ_l is the top k_l list of items from list l where $k_l = |\Lambda_l|$. Note that we observe the top k items for each of the L lists if $k_1 = \cdots = k_L = k$. For censored lists the rank function becomes

$$\tilde{R}_l(p) = \begin{cases} \{\pi_l^{-1}(p)\} & \text{for } p \in \Lambda_l \\ \{k_l + 1, \ldots, P\} & \text{for } p \notin \Lambda_l \end{cases} \tag{4}$$

where we only know that the rank for the unobserved items in list l must be larger than the largest rank observed in that list.

In the case of censored lists it is sufficient (FIXME: requires argument) to look at depths where we have corresponding observations so the largest rank we should consider will be

$$d \leq \max(k_1, \ldots, k_L). \tag{5}$$

We cannot directly compute $A(p)$ for all predictors because we only observe a censored version of $\tilde{R}$ for some of the lists. Instead we assume that the rank assigned to predictor p in list l is uniformly distributed among the ranks that have been unassigned for list l. The ranks are clearly not independent since each of the lists essentially contains full set of ranks

$$\tilde{A}(p) = \frac{\sum_{r_1; r_1 \in \tilde{R}_1(p)} \cdots \sum_{r_L; r_L \in \tilde{R}_L(p)} A(p)}{\prod_l |\tilde{R}_l(p)|} \tag{6}$$

FIXME: If instead of running through all elements of $\tilde{R}_1(p)$ one would use the average rank in $\tilde{R}_1(p)$ we would end up with a too small variance.

Noget med

3 Benchmarks

Three approaches

3.1 Independent lists

3.2 Permuted outcomes + analyses

Do the following a large number of times

1. Permute outcome vector
2. Redo analyses for all L methods
3. Compute sequential rank agreement for

Lidt spøjs ting man matcher dem op imod

3.3 Changepoint analysis

4 Applications

4.1 Comparing results across different method

In a classical paper by Golub (1999) a dataset of 3051 gene expression values were measured on 38 tumor mRNA samples in order to improve the classification of acute leukemias between two types: acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). Preprocessing of the gene expression data was done as described in (Dudoit et al., 2002).

4.2 Stability of results

4.3 Evaluating results from top-k lists

Bootstrap across a single method and compare results. Discuss collinearity

5 Discussion

Mention/discuss different measures.

References

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97(457):77–87.

Golub, T. R. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537.