On the Use of Spearman's Rho to Measure the Stability of Feature Rankings

Sarah Nogueira, Konstantinos Sechidis and Gavin Brown

School of Computer Science, University of Manchester, Manchester M13 9PL, UK
sarah.nogueira@manchester.ac.uk,
WWW home page: http://www.cs.man.ac.uk/~nogueirs/

Abstract. Producing stable feature rankings is critical in many areas, such as in bioinformatics where the robustness of a list of ranked genes is crucial to interpretation by a domain expert. In this paper, we study Spearman's rho as a measure of stability to training data perturbations - not just as a heuristic, but here *proving* that it is the natural measure of stability when using mean rank aggregation. We provide insights on the properties of this stability measure, allowing a useful interpretation of stability values - e.g. how close a stability value is to that of a purely random feature ranking process, and concepts such as the *expected value* of a stability estimator.

Keywords: Stability, Robustness, Feature Rankings, Ensembles, Spearman's Rho, Mean Rank Aggregation.

1 Introduction

Feature selection is a broad topic that consists in identifying the relevant features for future use in a predictive model or for interpretation by domain experts. The output of a feature selection algorithm might be one of 3 types: a scoring on the features (e.g. the coefficients of a regression model), a ranking on the features (e.g. with any sequential forward selection) or a feature set (e.g. when using hypothesis testing procedures). In this paper we focus on feature rankings.

Stability (or robustness) of a feature ranker (FR) is its *sensitivity to small perturbations in the training set* [12]. In information retrieval, ranking systems on search engines are expected to be robust to spam [8]. In bioinformatics, where by nature the training samples are usually small, the removal of only one example on the training set can cause substantially different rankings making the feature rankings non-interpretable and not reliable for clinical use. For this reason, robust FRs have become a major requirement in the field of gene selection, biomarker identification or molecular profiling [1, 3, 7, 10, 20].

Many measures of stability have been proposed in the literature. Some measures focus on the stability of *partial* feature rankings or in giving more weight to features with higher rankings [11]. In this paper, we focus on a popular measure used to measure the stability of full feature rankings: the Spearman rank-order

correlation coefficient, also commonly called Spearman's ρ. The main contributions of this paper include an understanding of the properties of this measure for useful interpretation of stability values; a proof that unstable FRs yield *better* rankings than individual rankings do on average when aggregated by their mean and an *explanation* of why mean rank aggregation produces more stable FRs.

The paper is structured as follows. Section 2 provides background material to the quantification of stability of FRs. Section 3 provides a statistical interpretation and derives the properties of this measure. Section 4 focuses on the topic of mean rank aggregation and Section 5 illustrates some of our theoretical results on mean rank aggregation.

2 Background

Let us assume there are d features in total. A ranking $\mathbf{r}$ can be modelled as a vector of d **distinct** natural numbers taken from 1 to d (i.e. as a permutation of the numbers from 1 to d). To quantify stability, we measure the variability of the rankings obtained when *small* perturbations are applied to the dataset. In most literature, the general procedure to evaluate the stability of a FR consists in taking M bootstrap samples of the dataset and then to apply the FR to each one of the M samples hence giving M rankings [9, 12].

Let us take an example: assume that we have a dataset with $d = 5$ features and that we apply a FR to $M = 3$ bootstrap samples. Then we can represent the output of the FR as follows:

$$\mathcal{R} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \mathbf{r}_3 \end{bmatrix} = \left.\begin{bmatrix} 5\ 3\ 1\ 4\ 2 \\ 4\ 3\ 1\ 5\ 2 \\ 5\ 3\ 2\ 4\ 1 \end{bmatrix}\right\} M = 3 \text{ feature rankings}$$

where $\mathbf{r}_i$ is the ranking obtained on the i^{th} dataset. In the first ranking $\mathbf{r}_1$, the first feature is ranked 5^{th}, the second feature is ranked 3^{rd} and so on. We can see that there are some variations in between the three rankings $\mathbf{r}_1$, $\mathbf{r}_2$ and $\mathbf{r}_3$. Even though the 2^{nd} feature is always ranked in the 3^{rd} position, the other features present some variations in their ranks. A fully stable FR would have produced identical rankings (i.e. $\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r}_3$) on the different data samples. In general, the M rankings can be represented by a matrix $\mathcal{R}$ as follows:

$$\mathcal{R} = \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_M \end{bmatrix} = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,d} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ r_{M,1} & r_{M,2} & \cdots & r_{M,d} \end{bmatrix}$$

where $\mathbf{r}_i$ is the feature ranking on the i^{th} bootstrap sample. Quantifying the stability of a FR consists in defining a measure $\hat{\Phi}$ taking as an input such a matrix $\mathcal{R}$ to quantify these variations. We can wonder what would be a sensible definition for $\hat{\Phi}$ and which properties should a stability measure have so that the stability values are interpretable and comparable in different contexts.

Let ϕ be a function that takes as an input two feature rankings $\mathbf{r}_i$ and $\mathbf{r}_j$ and returns a *similarity* value between the two rankings. A common approach to measure the stability of a FR is to define the stability as the average pairwise similarities between all possible *unique* pairs of rankings in $\mathcal{R}$ [12], that is:

$$\hat{\Phi}(\mathcal{R}) = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \phi(\mathbf{r}_i, \mathbf{r}_j). \tag{1}$$

Several proposals have been made in the literature for the similarity measure ϕ. Such measures include the Kendall Tau [19], the Canberra Distance [10], the scaled Spearman footrule [17] or the Spearman's ρ [12, 15]. In this paper, we focus on the use of Spearman's ρ which is formally defined as:.

$$\rho(\mathbf{r}_i, \mathbf{r}_j) = 1 - \frac{6 \sum_{f=1}^{d} (r_{i,f} - r_{j,f})^2}{d(d^2-1)}. \tag{2}$$

Hereafter, $\hat{\Phi}$ will denote the stability measure using Spearman's ρ. In the next section, we study the properties of this stability measure and show that $\hat{\Phi}$ should be interpreted as a *random variable*.

3 Pairwise Spearman's Rho as a Stability Measure

3.1 Statistical Interpretation

An important point is that Equation (1) is an *estimator*, based on a random process (bootstrapping) – therefore $\hat{\Phi}$ is a *random variable*, and we can discuss concepts such as the expectation and the convergence of that random variable. Surprisingly, these concepts – the expectation/convergence of stability estimates have not been considered in the literature before. We proceed below by characterising this random variable for the case of Spearman's Rho.

We can see each ranking $\mathbf{r}$ as a draw from a an unknown distribution and therefore $\hat{\Phi}$ is an estimator of a population parameter Φ that depends on the parameters of that distribution. Let X_f be the random variable corresponding to the rank of the f^{th} feature. We can therefore see f^{th} column of $\mathcal{R}$ as a realisation of X_f. The *maximum likelihood estimate* σ_f^2 of the variance of X_f and the unbiased sample variance s_f^2 are by definition:

$$\sigma_f^2 = \left[\left(\frac{1}{M} \sum_{i=1}^{M} r_{i,f}^2 \right) - \left(\frac{1}{M} \sum_{i=1}^{M} r_{i,f} \right)^2 \right] \quad \text{and} \quad s_f^2 = \frac{M}{M-1} \sigma_f^2. \tag{3}$$

We build upon this with a novel theorem, our first contribution, Theorem 1.

Theorem 1. *The stability $\hat{\Phi}$ using Spearman's ρ can be re-written as follows:*

$$\hat{\Phi}(\mathcal{R}) = 1 - \frac{\frac{1}{d} \sum_{f=1}^{d} s_f^2}{V_r}, \tag{4}$$

where $V_r = \frac{d^2-1}{12}$ is a constant only depending on d.

Proof. All proofs of theorems/corollaries are in the Supplementary Material[1].

One interpretation of the stability of $\hat{\Phi}$, from the form of Equation (1), is the average correlation between the rankings in $\mathcal{R}$. Theorem 1 gives another interpretation. In fact, we can see the value of $\hat{\Phi}$ as an *estimator* of the average variance of X_f over the d features rescaled by a constant depending only on d. First of all, this gives us a natural and novel multivariate extension of Spearman's ρ for a set of M rankings since it reduces to Spearman's ρ for $M = 2$ (which has been a topic of interest in the statistical literature [16]). When the FR is *fully stable*, i.e. when all the rankings in $\mathcal{R}$ are identical, the sample variance of X_f will be equal to 0 and therefore $\hat{\Phi}(\mathcal{R})$ will be equal to 1. Computing the stability using Equation (4) instead of Equation (1) reduces the computational complexity from $\mathcal{O}(M^2 d)$ to $\mathcal{O}(Md)$. Since s_f^2 is an unbiased and consistent estimator of the true variance $\mathrm{Var}(X_f)$, we can derive the result given in Corollary 1. This corollary shows that the estimated stability $\hat{\Phi}$ will converge in probability to the population stability Φ.

Corollary 1. *$\hat{\Phi}(\mathcal{R})$ is an unbiased and consistent estimator of:*

$$\Phi = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} \mathrm{Var}(X_f)}{V_r}. \tag{5}$$

We can wonder what happens if we use the maximum likelihood estimate of the variance σ_f^2 instead of the unbiased estimator s_f^2 to estimate the true variance of X_f in Equation (4). It turns out that that quantity corresponds to the average pairwise Spearman's ρ between all M^2 pairs of rankings (i.e. the $M(M-1)$ pairs we already had plus the M correlations of each ranking with itself). Let us call that latter quantity $\hat{\Phi}^{all}$. We have that:

$$\hat{\Phi}^{all}(\mathcal{R}) = \frac{1}{M^2}\left(M(M-1)\hat{\Phi}(\mathcal{R}) + \sum_{i=1}^{M}\rho(\mathbf{r}_i, \mathbf{r}_i)\right) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} \mathrm{Var}(X_f)}{V_r}. \tag{6}$$

The only difference between $\hat{\Phi}$ and $\hat{\Phi}^{all}$ lies in the way the true variance of X_f is estimated. Even though the maximum likelihood estimator σ_f^2 is biased, it converges to the population parameter $\mathrm{Var}(X_f)$ as M goes to infinity. In other words, when M is large enough, these two quantities can be used interchangeably. This will be critical in introducing the concepts discussed in Section 4.

3.2 Properties

We know from the statistical literature the Spearman's ρ is a *chance-corrected* measure of correlation and that $-1 \leq \rho \leq 1$ [2]. But we can wonder what are the properties of the average pairwise Spearman's ρ? In this section, we prove two properties for $\hat{\Phi}$ that we argue to be useful for interpretation and comparison of stability values in different settings.

[1] Available online at
http://www.cs.man.ac.uk/~nogueirs/files/IbPRIA2017-supplementary-material.pdf

Theorem 2 (Bounds). *$\hat{\Phi}$ is asymptotically bounded ($M \to \infty$) by* 0 *and* 1.

Even though ρ can take negative values, Theorem 2 shows that the resulting stability estimate $\hat{\Phi}$ is asymptotically non-negative. We expect this result since the population parameter Φ we are estimating –Eq. (5)– is in the interval $[0, 1]$.

Theorem 3 (Correction For Chance). *The stability estimate $\hat{\Phi}$ is corrected by chance which means that its expected value is constant and equal to* 0 *when the FR is random (i.e. when all rankings/permutations have equal probability).*

Theorem 3 shows that no matter what is the total number of features d, the stability estimate $\hat{\Phi}$ of a random FR will be 0 in expectation. As pointed out by [2], "chance-corrected measures yield values that are interpreted as a proportion above that expected by chance alone". We can therefore interpret the stability estimate $\hat{\Phi}$ as the proportion of agreement above chance between the rankings in $\mathcal{R}$. Some popular measures of stability used in the literature do not have this property. For instance, the stability of a random FR using the Canberra distance [10] will systematically increase with the number of features d, which means it cannot be used to compare the stability of ranked gene lists of different sizes.

3.3 Relationship to Other Stability Measures

Finally, we can point out that the use of Spearman's ρ is in line with the stability measures used for different types of feature selection outputs. Since the sample Pearson correlation coefficient reduces to Spearman's ρ in the cases of untied ranks, it is strongly related to the literature that makes use of the average pairwise sample Pearson's correlation coefficient in the case of feature weights [12] and in the case of feature sets [14], which suggests that the use of Spearman's ρ goes towards a unification of the stability literature. We can also point out that the use of the average pairwise Pearson's correlation has been shown to hold a set of desirable properties and to reduce to the very popular Kuncheva's measure [13] in the case of feature sets [14].

4 Ensemble Feature Ranking

In the ensemble learning literature, it is known that a set of *diverse* regression models can be aggregated together to form a more robust model [5]. Inspired by that field, *ensemble feature selection* [15] aims at building more robust feature selectors by using a set of unstable individual FRs and aggregating them together to increase the stability. Nevertheless, there has been no theoretical work that guarantees that the error and the stability of a FR will be improved by the ensemble. Moreover, some works question the use of ensembles since it has empirically been shown not to always increase the stability [6]. We focus on the case of mean rank aggregation. We first show that the *error* of the aggregated ranking will be guaranteed to be lower than the one of an individual ranking on average. Then, we give a theoretical argument showing why the stability of the aggregated rank should improve as the number of ensemble members increases.

4.1 Mean Rank Aggregation

The general procedure to build feature ranking ensembles is to take K bootstrap samples of the data, to apply the FR to each one of the samples and then to combine the K resulting feature rankings using a given *rank aggregation technique*. The reason why we denote by K the number of FRs in ensemble (and not by M) is because we want to distinguish it from the number of bootstrap samples M used to *estimate* $\hat{\Phi}$. In this paper, we focus on the popular mean rank aggregation that consists in taking the mean ranking $\bar{\mathbf{r}} = (\bar{r}_1, ..., \bar{r}_d)$ of each feature over the K rankings (i.e. $\bar{r}_f = \frac{1}{K}\sum_{i=1}^{K} r_{i,f}$). For full rankings, the mean rank aggregation has been proved to be equivalent to the Borda Count aggregation technique [20].

4.2 The "Ambiguity" Decomposition

Let us assume there exists a *true* ranking $\mathbf{r}^* = (r_1^*, ..., r_d^*)$, where r_f^* is the true rank of the f^{th} feature, and that the FR is trying to estimate that true ranking $\mathbf{r}^*$ when producing a ranking $\mathbf{r}_i$. One way to measure the quality of the rank of the f^{th} feature is to measure the *squared error* (SE) of $r_{i,f}$ compared to the true ranking r_f^* as follows: $(r_{i,f} - r_f^*)^2$. Now, assuming we have K ranks $(r_{1,f}, .., r_{K,f})$ for the f^{th} feature, we can define the mean squared error (MSE) as the mean of the squared errors of each one of the K rankings: $\frac{1}{K}\sum_{i=1}^{K}(r_{i,f} - r_f^*)^2$. Similarly to the *ambiguity decomposition* that exists for ensembles of regression predictors [4], we provide an ambiguity decomposition for mean rank aggregation.

Theorem 4. *The average squared error of the mean rank over the d features can be decomposed into two* ***positive*** *terms as follows:*

$$\overbrace{\frac{1}{d}\sum_{f=1}^{d}(\bar{r}_f - r_f^*)^2}^{\text{av. SE of the mean ranker}} = \overbrace{\frac{1}{d}\sum_{f=1}^{d}\left(\frac{1}{K}\sum_{i=1}^{K}(r_{i,f} - r_f^*)^2\right)}^{\text{av. MSE of the K rankers}} - \overbrace{(1 - \hat{\Phi}^{all})V_r}^{\text{ambiguity term}}, \quad (7)$$

where the ambiguity term is also equal to $\frac{1}{d}\sum_{f=1}^{d}\sigma_f^2$ *and where* $V_r = \frac{d^2-1}{12}$. *Therefore, the error of the ensemble ranker is guaranteed to be less or equal than the one of the individual rankers on average.*

Theorem 4 provides a decomposition of the squared error of the mean ranking $\bar{\mathbf{r}}$ into two *positive* terms: the average MSE of the K rankers (which is the average of the MSE over the d features of the K rankings) and the ambiguity term, which is a linear function of the stability estimate $\hat{\Phi}^{all}$. Since these two terms are positive, we can see that for a given MSE, having a higher ambiguity term (which corresponds to having a less stable set of rankers) will result in a lower average SE for the mean ranking $\bar{\mathbf{r}}$. This decomposition shows two things:

1. The mean rank $\bar{\mathbf{r}}$ is guaranteed to be *closer* to the true ranking than would be an individual ranker on average.

2. The use of Spearman's ρ to estimate stability is a sensible choice since it can be interpreted as the ambiguity term of this decomposition.

Naturally, this decomposition does not show that the aggregated ranker will be more stable then the individual ranker, which is the topic of the next section.

4.3 Does Mean Rank Aggregation Always Increase Stability?

We aim at giving an explanation of why we should expect a higher stability when performing mean rank aggregation. The mean ranking is not a permutation of the integers 1 to d any more: since $\bar{r}_f = \frac{1}{K}\sum_{f=1}^{d} r_{i,f}$, the mean rank of the f^{th} feature can be any real number in the interval $[1, d]$. Similarly to Equation (5) where the true stability Φ of a FR is a linear function of the average variance of X_f, we can define the stability of the mean ranking $\bar{\mathbf{r}}$ as a linear function of the average variance of mean rankings $\bar{r}_f$ over the d features as follows:

$$\Psi(\bar{\mathbf{r}}) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} \mathrm{Var}(\bar{r}_f)}{V_r}. \tag{8}$$

Theorem 5 derives the stability Ψ of the mean ranking $\bar{\mathbf{r}}$ as a linear function of the true stability Φ of the individual FR. This theorem shows that Ψ increases with the number of FRs in the ensemble and that eventually, as we keep adding FRs to the ensemble, the ensemble will be fully stable (as Ψ converges to 1 when K goes to infinity). Figure 4.3 illustrates the value of Φ against the number of FRs in the ensemble K for different values of Φ. We can see that this value converges to 1 as K increases.

Theorem 5. *Assuming the K rankings in the ensemble are independent and identically distributed (i.i.d), the stability of the mean ranking is reduced by $\frac{1}{K}$ compared to the stability of the individual FR:*

$$\Psi = \frac{K-1}{K} + \frac{\Phi}{K}. \tag{9}$$

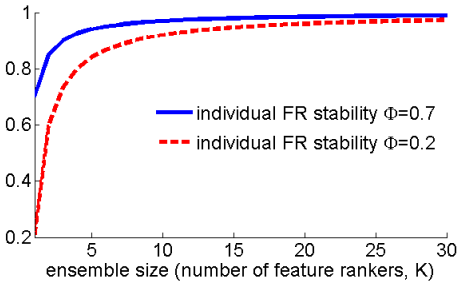

Fig. 1. Stability of the mean rank aggregation against the number of ensemble members K when the individual ranker has a stability $\Phi = 0.7$ and $\Phi = 0.4$.

One could question the choice of Ψ as a stability measure for the aggregated *mean* ranking. As we can see in Equation (8), Ψ is a linear function of the variance of the mean rank of each feature. In the literature, a threshold τ is often applied to the mean ranking to obtain a feature set. For example, we could decide to select all the features for which the mean rank $\bar{r}_f < \tau$ and discard the other features. Therefore, if the mean rank $\bar{r}_f$ of each feature has a low variance (which corresponds to a high value of Ψ), it is more likely that the same features would be selected when small perturbations are applied to the training set, hence producing a stable feature set.

5 Experiments

In this section, we aim at illustrating the results of Section 4. To be able to illustrate the result of Theorem 4, we need to know the *true* ranking $\mathbf{r}^*$. For this reason, we generate an artificial dataset consisting of $d = 20$ binary features with different degrees of dependency with the target class Y [18]. To create the data, firstly we generate the values of Y, by taking n samples from a Bernoulli distribution with $p(y = 1) = 0.50$. Then, for each feature X, we randomly choose the parameters $p(x|y)$ that guarantee the desired degree of dependency expressed in terms of $I(X;Y)$ and we use these parameters to sample the values of X. The mutual information $I(X;Y)$ population values for each features are:

$$[9 \;\; 9.5 \;\; 8.5 \;\; 8 \;\; 7.5 \;\; 7 \;\; 6.5 \;\; 6 \;\; 5.5 \;\; 5 \;\; 4.5 \;\; 4 \;\; 3.5 \;\; 3 \;\; 2.5 \;\; 2 \;\; 1.5 \;\; 1 \;\; 0.5 \;\; 0.1] \times 10^{-2},$$

where a high mutual information translates into a high rank of the feature. We repeat the experiment taking different sample sizes n as given in Table 1. Then, we take $M = 100$ bootstrap samples of each one of these datasets and estimate the mutual information on each bootstrap, thus getting M rankings.

Table 1. Demonstration that unstable FRs (in the sense of pairwise Spearman's ρ) provide a better ranking when aggregated together by their mean. The difference between the error of the mean rank and the mean error of the individual rankings is larger for lower stability values.

n	error of the mean rank $\bar{\mathbf{r}}$	mean error of the K rankers	ambiguity	Stability $\hat{\Phi}^{all}$
30	29.5	46.0	16.5	0.505
50	36.9	49.4	12.6	0.622
500	2.91	7.77	4.85	0.854
1000	2.05	4.23	2.19	0.934
10000	0.149	0.52	0.366	0.989

We can see in Table 1 that as we increase the sample size n, the stability of the FR increases and therefore, the ambiguity term decreases. This is expected since the mutual information estimates become better as we increase the sample size and thus, the ranking become more accurate. We can observe that for lower

stability values, the improvement in terms of error (which corresponds to the difference between the average SE of the mean ranking and the average MSE of the M rankings and therefore to the ambiguity term) is larger; as we expected from Theorem 4. This follows the idea that an unstable set of FRs will yield better results once aggregated.

We now aim at illustrating the result of Section 4.3. Since we are considering small sample sizes, we use the jackknife resampling technique (which corresponds to a leave-one-out resampling) to get several ensembles of FRs and we estimate the value of Ψ. Table 2 shows the evolution of the stability of the aggregated ranking as we increase K. As expected, the estimated stability of the aggregated mean rank $\hat{\Psi}$ increases with K and converges to 1.

Table 2. The estimated stability $\hat{\Psi}$ of the mean rank increases with the number of FRs K aggregated. This illustrates the result of Theorem 5.

n	K	stability of the individual ranker $\hat{\Phi}^{all}$	stability of the mean ranker $\hat{\Psi}$
30	2	0.505	0.738
	5		0.887
	10		0.931
	50		0.975
500	2	0.854	0.926
	5		0.971
	10		0.985
	50		0.997

6 Conclusions

In this work, we showed that the stability of feature rankings using Spearman's ρ is in fact a random variable. Therefore, when we "calculate" stability, we are only estimating the *true* stability of a FR for a specific dataset. Our work also derives a set of properties deemed useful for interpretation and comparison of stability estimates. To the best of our knowledge, this is the first work proposing a statistical perspective on the stability values obtained when using Spearman's rho. We further provide an theoretical guarantees on the *error* and the true stability of the mean rank aggregation. Future work could include the derivation of asymptotic distribution for the stability estimate, which would allow to derive such tools as hypothesis testing for stability values and/or confidence intervals.

References

1. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., Saeys, Y.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics (2010)

2. Berry, K.J., Mielke, Jr, P.W., Johnston, J.E.: Permutation statistical methods: an integrated approach. Springer (2016)
3. Boulesteix, A.L., Slawski, M.: Stability and aggregation of ranked gene lists. Briefings in Bioinformatics (2009)
4. Brown, G., Wyatt, J.L.: The use of the ambiguity decomposition in neural network ensemble learning methods. In: Fawcett, T., Mishra, N. (eds.) ICML (2003)
5. Brown, G., Wyatt, J.L., Tiňo, P.: Managing diversity in regression ensembles. Journal of Machine Learning Research 6 (2005)
6. Dessì, N., Pes, B.: Stability in biomarker discovery: Does ensemble feature selection really help? In: IEA/AIE 2015, Proceedings (2015)
7. Dittman, D.J., Khoshgoftaar, T.M., Wald, R., Napolitano, A.: Classification performance of rank aggregation techniques for ensemble gene selection. In: FLAIRS Conference. AAAI Press (2013)
8. Dwork, C., Kumar, R., Naor, M., Sivakumar, D.: Rank aggregation methods for the web. In: International Conference on World Wide Web, Proceedings (2001)
9. He, Z., Yu, W.: Stable feature selection for biomarker discovery. Computational Biology and Chemistry (2010)
10. Jurman, G., Merler, S., Barla, A., Paoli, S., Galea, A., Furlanello, C.: Algebraic stability indicators for ranked lists in molecular profiling. Bioinformatics (2008)
11. Jurman, G., Riccadonna, S., Visintainer, R., Furlanello, C.: Algebraic comparison of partial lists in bioinformatics. PloS one 7 (2012)
12. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. (2007)
13. Kuncheva, L.I.: A stability index for feature selection. In: Artificial Intelligence and Applications (2007)
14. Nogueira, S., Brown, G.: Measuring the stability of feature selection. In: ECML/PKDD (2016)
15. Saeys, Y., Abeel, T., de Peer, Y.V.: Robust feature selection using ensemble feature selection techniques. In: ECML/PKDD (2) (2008)
16. Schmid, F., Schmidt, R.: Multivariate extensions of spearman's rho and related statistics. Statistics & Probability Letters 77 (2007)
17. Sculley, D.: Rank aggregation for similar items. In: Proceedings of the Seventh SIAM International Conference on Data Mining (2007)
18. Sechidis, K.: Hypothesis Testing and Feature Selection in Semi-Supervised Data. Ph.D. thesis, School of Computer Science, University Of Manchester, UK (2015)
19. Voorhees, E.M.: Evaluation by highly relevant documents. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '01, ACM (2001)
20. Wald, R., Khoshgoftaar, T.M., Dittman, D.J., Awada, W., Napolitano, A.: An extensive comparison of feature ranking aggregation techniques in bioinformatics. In: IRI. IEEE (2012)