ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

On aggregation in ranking median regression

Stephan Clémençon and Anna Korba *

Telecom ParisTech - LTCI, Université Paris Saclay
46 rue Barrault, 75634 Paris - France

Abstract. In the present era of personalized customer services and recommender systems, predicting the preferences of an individual over a set of items indexed by $[\![n]\!] = \{1, \ldots, n\}$, $n \geq 1$, based on its characteristics, modelled as a r.v. X say, is an ubiquitous issue. Though easy to state, this predictive problem refererred to as *ranking median regression* (RMR in short) is very difficult to solve in practice. The major challenge lies in the fact that, here, the (discrete) output space is the symmetric group $\mathfrak{S}_n$, composed of all permutations of $[\![n]\!]$, of explosive cardinality $n!$, and which is not a subset of a vector space. It is thus far from straightforward to build (non parametric) predictive rules taking their values in $\mathfrak{S}_n$, except by means of ranking aggregation techniques implemented at a local level, as proposed in [1] or [2]. However, such local learning techniques exhibit high instability and it is the main goal of this paper to investigate to which extent Kemeny ranking aggregation of randomized RMR rules may remedy this drawback. Beyond a theoretical analysis establishing its validity, the relevance of this novel ensemble learning technique is supported by experimental results.

1 Introduction

In an increasing number of modern applications, users are invited to declare their individual characteristics (*e.g.* socio-demographic features), taking the form of a r.v. X valued in a an input space $\mathcal{X} \subset \mathbb{R}^d$ say, and may express their preferences over a set of numbered services/products $[\![n]\!] = \{1, \ldots, n\}$. In this context, the goal pursued is to learn from historical data how to predict the preferences of any user based on her characteristics X. In the simplest formulation, the prediction takes the form of a permutation $s(X)$ on $[\![n]\!]$, mapping any item i to its rank $s(X)(i)$ on her preference list. Denoting by Σ the permutation that truly reflects the preferences of a user with characteristics X, the performance of any predictive rule, *i.e.* any measurable function $s : \mathcal{X} \to \mathfrak{S}_n$, can be measured by the expected Kendall τ distance between $s(X)$ and Σ

$$\mathcal{R}(s) = \mathbb{E}\left[d_\tau\left(s(X), \Sigma\right)\right], \tag{1}$$

where the expectation is taken over the (unknown) distribution of the pair (X, Σ) and $d_\tau(\sigma, \sigma') = \sum_{i<j} \mathbb{I}\{(\sigma(j) - \sigma(i)) \cdot (\sigma'(j) - \sigma'(i)) < 0\}$ for all $(\sigma, \sigma') \in \mathfrak{S}_n^2$, denoting by $\mathbb{I}\{\mathcal{E}\}$ the indicator function of any event $\mathcal{E}$. Stated this way, the objective is to build a mapping s that minimizes (1) and one may easily show

*This research is supported by a CRITEO Faculty Research Award and the chair Machine Learning for Big Data of Telecom ParisTech.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

with a straightforward conditioning argument that the optimal predictors are the rules that maps any point X in the input space to any (Kemeny) ranking median of P_X, Σ's conditional distribution given X. Recall that a Kemeny median of a probability distribution P on $\mathfrak{S}_n$ is any solution σ_P^* of the optimization problem

$$\min_{\sigma \in \mathfrak{S}_n} L_P(\sigma) \tag{2}$$

where $L_P(\sigma) = \mathbb{E}_{\Sigma \sim P}\left[d_\tau(\sigma, \Sigma)\right]$. For this reason, the predictive problem formulated above is referred to as *ranking median regression* (RMR in abbreviated form). A theoretical analysis of the RMR problem, including a description of the set of optimal rules and a study of the generalization capacity of empirical risk minimizers (*i.e.* minimizers of a statistical version of (1)) over classes of mappings $s : \mathcal{X} \to \mathfrak{S}_n$ of controlled complexity, when learning is based on i.i.d. copies of the pair (X, Σ), has been carried out in [2]. Regarding the problem of minimizing (1), attention should be paid to the fact that, in contrast to usual (median/quantile) regression, the set $\mathcal{S}$ of predictive ranking rules is not a vector space, which makes the design of practical optimization strategies challenging and the implementation of certain methods, based on (forward stagewise) additive modelling for instance, unfeasible (unless the constraint that predictive rules take their values in $\mathfrak{S}_n$ is relaxed, see [3] or [4]). For this reason, a recent practical approach for ranking median regression relies on the concept of local learning and permits to derive practical procedures for building piecewise constant ranking rules from efficient (approximate) Kemeny aggregation, when implemented at a local level, nearest-neighbor techniques or decision trees typically, see *e.g.* [1], [2] or Chapter 10 in [5]. However, such methods exhibit high instability in practice, in the sense that the rules they produce can be much affected by small changes in the training data. It is the goal of this paper to investigate the ensemble learning approach in this original context and show that (approximate) Kemeny ranking aggregation again, when applied to bootstrap versions of the training sample, may remedy this severe drawback. A theoretical result stating the consistency is preserved under aggregation is proved and empirical evidence of the gain in stability is provided by numerical experiments.

The paper is structured as follows. In section 2, basic notions related to Kemeny aggregation and ranking median regression are briefly recalled. The bootstrap Kemeneny aggregating technique is presented and analyzed in section 3. Experimental results are displayed in section 4.

2 Background and Preliminaries

A few concepts involved in the description of the ensemble learning methods we promote here and in its analysis in the subsequent section are summarized below.

Probabilistic framework for Kemeny aggregation. Whereas problem (2) is NP-hard in general, exact solutions, referred to as *Kemeny medians*, can be explicited when the pairwise probabilities $p_{i,j} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\}$, $1 \leq i \neq j \leq n$,

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

fulfill the following property, referred to as *stochastic transitivity*: $\forall (i,j,k) \in [\![n]\!]^3 : \ p_{i,j} \geq 1/2$ and $p_{j,k} \geq 1/2 \ \Rightarrow \ p_{i,k} \geq 1/2$. If, in addition, none of the pairwise probabilities is equal to $1/2$ ($p_{i,j} \neq 1/2$ for all $i < j$), distribution P is said to be strictly stochastically transitive. When stochastic transitivity holds true, the set of Kemeny medians is the set $\{\sigma \in \mathfrak{S}_n : \ (p_{i,j} - 1/2)(\sigma(j) - \sigma(i)) > 0$ for all $i < j$ s.t. $p_{i,j} \neq 1/2\}$, and, if the strict version of stochastic transitivity is fulfilled, the Kemeny median is unique and given by the Copeland ranking: $\sigma^*_P(i) = 1 + \sum_{k \neq i} \mathbb{I}\{p_{i,k} < 1/2\}$ for $1 \leq i \leq n$ (see Theorem 5 in [6]). The set of strictly stochastically transitive distributions on $\mathfrak{S}_n$ is denoted by $\mathcal{T}$. Assume that we observe i.i.d. copies $\Sigma_1, \ldots, \Sigma_N$ of a generic r.v. $\Sigma \sim P$ and let $\widehat{P}_N = (1/N) \sum_{i=1}^N \delta_{\Sigma_i}$, where δ_x denotes the Dirac mass at any point x. Suppose that $P \in \mathcal{T}$ and satisfies the low-noise condition $\mathbf{NA}(h)$ for a given $h > 0$: $\min_{i<j} |p_{i,j} - 1/2| \geq h$. It is also shown in [6] that under the former hypotheses the empirical distribution $\widehat{P}_N \in \mathcal{T}$ as well, with overwhelming probability, and that the expectation of the excess of risk of empirical Kemeny medians (denoted by $\sigma^*_{\widehat{P}_N}$) decays at an exponential rate, see Proposition 14 therein. In this case, the nearly optimal solution $\sigma^*_{\widehat{P}_N}$ can be made explicit and straightforwardly computed based on the empirical pairwise probabilities $\widehat{p}_{i,j} = (1/N) \sum_{k=1}^N \mathbb{I}\{\Sigma_k(i) < \Sigma_k(j)\}$, $i < j$. Otherwise, solving the NP-hard problem $\min_{\sigma \in \mathfrak{S}_n} L_{\widehat{P}_N}(\sigma)$ is required to get an empirical Kemeny median, refer also to [7] and the references therein for a description of methods dedicated to approximate Kemeny median computation. However, as can be seen by examining Proposition 14's proof in [6], the exponential rate bound holds true for any candidate $\widetilde{\sigma}_N$ in $\mathfrak{S}_n$ that coincides with $\sigma^*_{\widehat{P}_N}$ when the empirical distribution lies in $\mathcal{T}$ and takes arbitrary values in $\mathfrak{S}_n$ otherwise.

Ranking median regression. As, for all $s \in \mathcal{S}$, $\mathcal{R}(s) = \mathbb{E}_{X \sim \mu}[L_{P_X}(s(X))]$, the optimal RMR rules s are those such that $s(X)$ is a median of P_X with probability one and the minimum risk is $\mathcal{R}^* = \mathbb{E}_{X \sim \mu}[L^*_{P_X}]$, where L^*_P denotes the minimum of (2) for any distribution P. In the case where P_X is strictly stochastically transitive with probability one, the optimal RMR rule is μ-almost surely unique, *i.e.* we μ-a.s. have $s^*(X) = \sigma^*_{P_X}$. Statistical learning being based on a finite sample $(X_1, \ \Sigma_1) \ \ldots, \ (X_1, \ \Sigma_N)$ of independent copies of the pair $(X, \ \Sigma)$, nearly optimal RMR rules cannot be built by trying to statistically recover a Kemeny median of P_x for each possible input point x, except when $\mathcal{X}$ is of finite cardinality, small compared to N. Nevertheless, under the additional assumption that the pairwise probabilities $p_{i,j}(x)$, related to the conditional distribution P_x, are Lipschitz, local learning techniques based on the solving of a few well-chosen Kemeny consensus problems have been proved to be consistent[1], nearest neighbor and decision tree methods namely. However, these methods exhibit high instability in practice. In the spirit of ensemble learning methods, it is proposed in the next section to apply Kemeny aggregation again, to randomized

[1] A sequence of RMR rules s_N is said to be consistent if $\mathcal{R}(s_N) \to \mathcal{R}^*$ in probability, as $N \to \infty$.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

versions of such RMR rules this time, in order to increase stability.

3 Aggregation of Ranking Median Regression Rules

We now investigate RMR rules that compute their predictions by aggregating those of *randomized RMR rules*. Let Z be a r.v. defined on the same probability space as (X, Σ), valued in a measurable space $\mathcal{Z}$ say, describing the randomization mechanism. A randomized RMR algorithm is then any function $S : \bigcup_{N\geq 1}(\mathcal{X} \times \mathfrak{S}_n)^N \times \mathcal{Z} \to \mathcal{S}$ that maps any pair $(z, \mathcal{D}_N)$ to a RMR rule $S(., z, \mathcal{D}_N)$. Given the training sample $\mathcal{D}_N$, its risk is $\mathcal{R}(S(., ., \mathcal{D}_N)) = \mathbb{E}_{(X,\Sigma,Z)}[d_\tau(\Sigma, S(X, Z, \mathcal{D}_N))]$. Given any RMR algorithm and any training set $\mathcal{D}_N$, one may compute an aggregated rule as follows.

KEMENY AGGREGATED RMR RULE

Inputs. Training dataset $\mathcal{D}_N = \{(X_1, \Sigma_1), \ \ldots, \ (X_N, \Sigma_N)\}$. RMR randomized algorithm S, randomization mechanism Z, query point $x \in \mathcal{X}$. Number $B \geq 1$ of randomized RMR rules involved in the consensus.

1. (RANDOMIZATION.) Conditioned upon $\mathcal{D}_N$, draw independent copies $Z_1, \ \ldots, \ Z_B$ of the r.v. S and compute the individual predictions

$$S(x, Z_1, \mathcal{D}_n), \ \ldots, \ S(x, Z_B, \mathcal{D}_n).$$

2. (KEMENY CONSENSUS.) Compute the empirical distribution on $\mathfrak{S}_n$

$$P_B(x) = \frac{1}{B}\sum_{b=1}^{B} \delta_{S(x,Z_b,\mathcal{D}_n)}$$

and output a Kemeny consensus (or an approximate median):

$$\bar{s}_B(x) \in \operatorname*{argmin}_{\sigma\in\mathfrak{S}_n} L_{P_B(x)}.$$

The result stated below shows that, provided that P_X fulfills the strict stochastic transitivity property and that the $p_{i,j}(X)$'s satisfy the noise condition $\mathbf{NA}(h)$ for some $h > 0$ with probability one (we recall incidentally that it is shown in [2] that fast learning rates are attained by empirical risk minimizers in this case, see Proposition 7 therein), consistency is preserved by Kemeny aggregation, as well as the learning rate.

Theorem 1. *Let $h > 0$. Assume that the sequence of RMR rules $(S(., Z, \mathcal{D}_N))_{N\geq 1}$ is consistent for a certain distribution of (X, Σ). Suppose also that P_X is strictly stochastically transitive and satisfies condition $\mathbf{NA}(h)$ with probability one. Then, for any $B \geq 1$, any Kemeny aggregated RMR rule $\bar{s}_B$ is consistent as well and its learning rate is at least that of $S(., Z, \mathcal{D}_N)$.*

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Proof. Recall the following formula for the risk excess: $\forall s \in \mathcal{S}$,

$$\mathcal{R}(s)-\mathcal{R}^* = \sum_{i<j} \mathbb{E}_X[|p_{i,j}(X)-1/2|\mathbb{I}\{(s(X)(j)-s(X)(i))(\sigma^*_{P_X}(j)-\sigma^*_{P_X}(i)) < 0\}]$$
$$\leq \mathbb{E}_X[d_\tau(s(X), \sigma^*_{P_X})] \leq (\mathcal{R}(s) - \mathcal{R}^*)/h,$$

see section 3 in [2]. In addition, the definition of the Kemeny median combined with triangular inequality implies that we a.s. have:

$$Bd_\tau(\bar{s}_B(X), \sigma^*_{P_X}) \leq \sum_{b=1}^{B} d_\tau(\bar{s}_B(X), S(X, Z_b, \mathcal{D}_N)) + \sum_{b=1}^{B} d_\tau(S(X, Z_b, \mathcal{D}_N), \sigma^*_{P_X})$$
$$\leq 2\sum_{b=1}^{B} d_\tau(S(X, Z_b, \mathcal{D}_N), \sigma^*_{P_X}).$$

Combined with the formula/bound above, we obtain that

$$\mathcal{R}(\bar{s}_B) - \mathcal{R}^* \leq \mathbb{E}[d_\tau(\bar{s}_B, \sigma^*_{P_X})] \leq \frac{2}{B}\sum_{b=1}^{B} \mathbb{E}_X[d_\tau(S(X, Z_b, \mathcal{D}_N), \sigma^*_{P_X})]$$
$$\leq (2/h)\frac{1}{B}\sum_{b=1}^{B}(\mathcal{R}(S(., Z_b, \mathcal{D}_N)) - \mathcal{R}^*).$$

The proof is then immediate. □

4 Experimental Results

For illustration purpose, experimental results based on simulated data are displayed. Datasets of full rankings on n items are generated according to p=2 explanatory variables. We carried out several experiments by varying the number of items ($n = 3, 5, 8$) and the "level of noise" of the distribution of permutations. For a given setting, one considers a fixed partition on the feature space, so that on each cell, the rankings/preferences are drawn from a certain Mallows distribution centered around a permutation with a fixed dispersion parameter ϕ. We recall that the greater ϕ, the spikiest the distribution (so closest to piecewise constant and less noisy in this sense). In each trial, the dataset of $N = 1000$ samples is divided into a training set (70%) and a test set (30%). We compare the results of (a randomized variant of) the CRIT algorithm (refer to [6] for a description of this decision tree algorithm) vs the aggregated version: in our case, the randomization is a boostrap procedure. Concerning the CRIT algorithm, since the true partition is known and can be recovered by means of a tree-structured recursive partitioning of depth 3, the maximum depth is set to 3 and the minimum size in a leaf is set to the number of samples in the training set divided by 10. For each configuration (number of items n and distribution

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Level of Noise	Number of items		
	n=3	n=5	n=8
$\phi = 2$	0.534 +/ - 0.167	1.454 +/ - 0.427	3.349 +/ - 0.952
	0.385 +/ - 0.085*	1.001 +/ - 0.232*	2.678 +/ - 0.615*
	0.379 +/ - 0.057**	0.961 +/ - 0.218**	2.281 +/ - 0.589**
$\phi = 1$	0.875 +/ - 0.108	2.346 +/ - 0.269	5.638+/ - 1.688
	0.807+/ - 0.061*	2.064 +/ - 0.130 *	4.499 +/ - 0.574*
	0.756 +/ - 0.063**	2.011 +/ - 0.110**	4.061 +/ - 0.259**

Table 1: Empirical risk averaged on 50 trials on simulated data.

of the dataset parameterized by Φ), the empirical risk, denoted as $\widehat{\mathcal{R}}_N(s)$, is averaged over 50 replications of the experiment. Results of the aggregated version of the (randomized) CRIT algorithm (one star * indicates the aggregate over 10 models, two stars over 30 models **) and of the CRIT algorithm (without stars) in the various configurations are provided in Table 1. In practice, for $n = 8$, the outputs of the randomized algorithms are aggregated with the Copeland procedure so that the running time remains reasonable. The results show notably that the noisier the data (smaller ϕ) and the larger the number of items n to be ranked, the more difficult the problem and the higher the risk. In a nutshell, and as confirmed by additional experiments, the results show that aggregating the randomized rules globally improves the average performance and reduces the standard deviation of the risk.

References

[1] P. L. H. Yu, W. M. Wan, and P. H. Lee. *Preference Learning*, chapter Decision tree modelling for ranking data, pages 83–106. Springer, New York, 2010.

[2] S. Clémençon, A. Korba, and E. Sibony. Ranking median regression: Learning to order through local consensus. In *Submitted*, 2017.

[3] S. Clémençon and J. Jakubowicz. Kantorovich distances between rankings with applications to rank aggregation. In *Machine Learning and Knowledge Discovery in Databases*, pages 248–263. Springer, 2010.

[4] F. Fogel, R. Jenatton, F. Bach, and A. d'Aspremont. Convex relaxations for permutation problems. In *Advances in Neural Information Processing Systems*, pages 1016–1024, 2013.

[5] M. Alvo and P. L. H. Yu. *Statistical Methods for Ranking Data*. Springer, 2014.

[6] A. Korba, S. Clémençon, and E. Sibony. A learning theory of ranking aggregation. In *Proceeding of AISTATS 2017*, 2017.

[7] Y. Jiao, A. Korba, and E. Sibony. Controlling the distance to a kemeny consensus without computing it. In *Proceeding of ICML 2016*, 2016.

[8] C. L. Mallows. Non-null ranking models. *Biometrika*, 44(1-2):114–130, 1957.