Local tree methods for classification

Alice Cleynen, Louis Raynal and Jean-Michel Marin

2023-09-18

Comments from the editor

Dear Jean-Michel Marin,

I apologize for the delay in sending you the review reports.

Two external reviewers have thoroughly exeamined your manuscript entitled “Local tree methods for classification,” submitted for publication to computo

The reports of the two reviewers, which are complete and comprehensive, are generally very positive and suggest that your contribution should eventually be published in our journal. However, they point out a number of possible improvements (mostly edits to the text) and raise various questions that do not allow the work to be published as it stands.

You will find the detailed reviews below.

we would be delighted if you would propose a new version of your manuscript for resubmission. Please accompany this with a response text justifying how you have incorporated the reviewers’ comments into the new version of your manuscript. Please note that in the event of final acceptance, your exchanges with the reviewers will be published with the manuscript. Thank you for considering our journal to publish your work,

Nelle Varoquaux, associate editor of Computo

Reviewer 1

The paper focusses on classification methods based on trees that aim at predicting a specific instance in the classification setting. This is motivated by an approximate Bayesian computation application in mind, where prediction is to be done according to one observation only. Such classification methods are called local. The paper proposes a review of such local classification approaches. It also introduces two novel such methods which did not show conclusive results. I think that the authors make a nice job in reviewing the literature about local tree methods for classification, as well as in proposing an extensive implementation of them. I do not have major critical comments about the paper. I rather list some suggestions that could hopefully help with the overall paper presentation.

Thank you for your time, comments and your constructive suggestions. We have addressed your comments in the revised version of the manuscript, and give details below.

Major comments

1. I find the abstract to be slightly misleading as it does not really reflect the content of the paper: a good part of it is devoted to the ABC motivating example, although the article does not implement it, in the end. Two options would be to either downplay the description of this ABC application in the abstract (which I think should focus on the proposed methods), or to keep the abstract this way but adding an illustrative application to ABC.

Thank you for your comment. We have reformulated the abstract in order to put less focus on ABC and more on random forest.

2. The paper makes a honest statement about its own proposed methods which turned out being unsatisfactory, due to high computational cost and limited performance. I think this is totally acceptable and the fact that the journal considers such negative results potentially useful to the readership is a good thing. In this direction, I wonder if the authors could provide some more feedback/explanation about what goes wrong with their two approaches. For instance, it is written at the beginning of Section 8 that those approaches "were implemented and compared on a lower dimensional simulation study (same Gaussian examples with only 500 test data and 5 replications) but were dropped of the final comparison due to high computational cost despite poor results." I think that it would be unfortunate not to include at least some toy illustration of the proposed methods. Maybe the setup that was dropped (same Gaussian examples with only 500 test data and 5 replications) could be considered in the Appendix?

We have re-run the small example and included the results at the end of section 8.2. The results of our methods are not particularily worse than the other methods we compare, but they clearly do not improve the results at the price of a huge computational price (about 750k the runtime of classic RF). To be honest it is hard to understand why local methods fail to provide better results than global ones. Our best interpretation is in fact that methods based on classic trees (Bagged CARTs, Random Forests) are already quite local, as at each split only the data in the mother node is considered. Hence provided the initial cuts are smart enough, the successive cuts will already have a local flavor. Bagging a large number of trees will attenuate the results of poor trees (those with poor inital cuts), hence an overall good performance.

3. Following the previous comment: I wonder if some more specific title wouldn't be in order. Indeed, on top of making a review about local tree methods for classification, the paper also proposes approaches which did not happen conclusive. Would something like "Local tree methods for classification: a review and some dead ends" work?

You are absolutely right, we changed our title as suggested.

4. Sections 4.2 and 4.3: Regarding the kernel bandwidths, I wonder if the authors thought about using larger values than the maximum absolute value considered when alpha is set to 1? As it is presented, it looks like this value is a kind of maximum possible value to be used, but nothing precludes using larger values actually. Could it lead to better results? I think a discussion about this limiting value would be useful.

Indeed there is no reason alpha=1 should be the maximum value. However as alpha increases the kernel will be more flat, giving more weight to non-neighboring data and getting closer to a classic Random Forest (which can be interpreted as a kernel-based approach with alpha going to infinity), hence reducing the local effect aimed at. We have not tried larger values, especially considering the computational burden of the methods, but we would expect the performance to be similar, as on our small example Multi-K, Uni-K and Random forests have similar performance.

Minor comments

1. The Gini index and entropy are alluded to in Section 3 but not defined there, and then are used in later sections. It could be useful to recall their expression from Section 3.

We agree and have added their definition after they are first mentionned in Section 3

2. The red & green colors used in Figure 1 are not color-blind friendly. I'm not sure about the other Figures. Maybe consider changing them to some more suitable palette?

Thank you for your suggestion. We have now modified the colors in all Figures using the safe color-blind palette in R.

3. Section 4.2 Unidimensional kernel approach: I would suggest to add "(per covariate)" to this section title in order to specify that unidimensional refers to treating covariates one at a time. Instead, there could be a misunderstanding that this section deals with unidimensional data.

Thank you for your suggestion. We have now modified the section's title

4. The quantile notation is quite straightforward, but still would need to be defined somewhere I guess.

We have now added the mathematical definition of the quantile in the manuscript.

5. The quantile order is often set to 1. Maybe just state that it amounts to choosing the maximum value of the absolute values considered?

Yes, this has been added

6. "We observed very few differences": specify. Maybe something like "We observed very few differences when using a a fixed or a varying bandwidth…"

We reformulated that sentence

7. "The first term is important and cannot be omitted contrary to the eager version, because it depends on the covariate index." This is not so clear at first sight since the notation does not depend on j. Maybe make both and notations depend on j?

You are right, we modified the notations to make the this statement more clear.

8. I would ask for some clarification about the bandwidths that are used throughout the paper. The one in Section 4.2 is associated to a univariate Gaussian kernel. I haven't seen it specified, but I guess it refers to the Gaussian standard deviation ? Then moving to Section 4.3, I wonder about some possible confusion on how the "scaling matrix" is used in the kernel. Shouldn't it be its inverse? Assuming the inverse is right, and in order to be coherent with Section 4.2, shouldn't the quantiles be squared in the definition of V? (since then it would be a covariance matrix, and not a matrix of standard deviations). I'm afraid that otherwise, the choices of quantile values made in both sections do not coincide.

Thank you for your careful reading. Indeed, the inverse and square of matrix V was missing in the text. We have corrected that, and confirm that the quantile values used in the univariate and multivariate kernels coincide.

9. Section 5.1 ends with "We tried various values of Nmin in our experiments." Could you add a conclusion sentence about the effect of the Nmin value?

We added a short comment at the end of the paragraph, and we discuss the effect of Nmin in the simulations sections.

10. Section 5.2: the NN acronym should be defined right after "nearest neighbors" is first used. I have to confess I first thought about "neural networks" when reading NN.

We appologize about the confusion, and have defined NN in the text now.

11. Numerical experiments: could the four tables be merged? The entries are always the same, it should be a question of adding new columns. Results would be more easily compared.

We have merged some of the tables, but have limited this to same example results as the html rendering of tables did not allow to clearly distinguish columns of different experiments. We hope this new version is a bit more readable.

12. Numerical experiments: "LDA axes": write in plain words? LDA may have multiple meanings.

We have now defined the LDA axes as the linear discriminant analysis axes.

13. Figure 4 caption: "blue" is used twice while "green" is missing. Also: the "black dashed lines" are not much visible. Try in white instead?

With your suggestion of using the safe color-blind palette the black dashed lines are now visible. We have also modified the error in the figure caption.

14. "In this example again, bagging CARTs outperforms a classic random forest, which itself outperforms all local approaches." It is not so clear-cut that "all local approaches" are outperformed.

You are right and we have mitigated this statement in the manuscript. However no local method reach the performance of bagging CARTs, and considering the computational cost of most of them, it is not clear they are worth using in practice.

Possible typos

1. Introduction: "To this effec" <- "To this effect"

2. Introduction: “we present/introduce”: choose only one verb?
3. It is often written “giving more weightS to”; I think the singular “weight” would be ok.
4. “Moreover, per tree a multidimensional kernel is used.” to be changed to “Moreover, a multidimensional kernel per tree is used.” or something like that.
5. Algorithm 1: “ends in the same leaf” <- “end in the same leaf”
6. After Algorithm 1: “The higher the” <- “The higher”
7. Section 6: Replace (Amaratunga, Cabrera, and Lee 2008) by Amaratunga, Cabrera, and Lee (2008) and (Maudes et al. 2012) by Maudes et al. (2012)
8. Breiman is the only author who’s first name (Leo) appears in the text, as well as in the list of References. First name to be removed?
9. Section 7: Add some space between citations “Robnik-Šikonja (2004);Tsymbal, Pechenizkiy, and Cunningham (2006)”
10. Section 8: “The classes have equal prior probabilities”: this may sound awkward to Bayesians... maybe just say “The classes have equal probabilities (or weights)”
11. Section 8: “The two first” <- “The first two” (other instances too)
12. “Note that during the preparation of the manuscript we detect [...] R package ranger and have to redo...” <- “Note that during the preparation of the manuscript we detectED [...] R package ranger and HAD to redo...”
13. “R package ranger”: use same font as with other packages?

Thank you for your careful reading of our manuscript. All the typos have been corrected

Some random comments and thoughts about differences between pdf and html rendering

1. Both Algorithm 1 & 2 appear twice in the pdf rendering of the manuscript (but only once in the html version).
2. Some code snippets appear in the pdf but not in html.
3. Some words in the text have hyperlinks associated (which is a cool feature since it adds some possible further reading). While the hyperlinks are visible in html (they appear in blue font), this is not the case for the pdf version. Maybe just add something like \hypersetup{colorlinks,citecolor=blue,linkcolor=red,urlcolor=blue}?

Thank you for your comments, we hope we have addressed all these issues in the revised manuscript.

Reviewer 2

Summary

The paper is concerned with model choice in the setting of intractable- likelihood models (i.e., when the likelihood of the model is unavailable but simulations from the hierarchical model can be drawn), also known as ABC model choice. To this aim, the authors consider the Random Forest (RF) methodology, following a previous work, which consists of generating simulations from the different models under consideration and building an RF to classify the model samples were drawn from. The authors attempt to improve this methodology by considering “local” RF methods, which guide the construction of the RF by considering the observation on which inference is required. This is intuitively good when, as in model choice, only a single real-world observation is present. The authors consider both existing local RF methods as well as introduce new ones. As the authors honestly point out, the performance of local methods on the considered examples is disappointing. Nonetheless, I believe it is good to have an illustration of these results in the literature and I think that running more experiments may discover some setups in which the proposed methodology outperforms standard RFs (see my comments below). Moreover, the paper is a good overview of local RF methods.

Thank you for your time, comments and your constructive suggestions. We have addressed your comments in the revised version of the manuscript, and give details below.

Main comments

Experiments

I think the range of experimental evaluation is slightly limited; I suggest the authors add one more experimental setup.

We added a novel experimental setup (section 8.3) where we tried to combine a fragmentation framework with spherical data distribution to challenge the splitting rules of standard random forests. Unfortunately, even in this setup local methods did not outperform bagging CARTs or random forests.

In the original paper on ABC model choice with RFs (Pudlo et al, 2016), a second RF is used to estimate the model probabilities, after the first one is used to assign the observation to a model. In the present paper, the authors only do the first step. Could the local RF methodology be applied to the second step too? Possibly, the local strategies may prove to outperform standard RF in that case.

This is a good but still open question. The second RF used to estimate the model probabilities in Pudlo et al. (2016) is a Regression random forest, for which implementation and comparison of local methods have not been performed. As we state in our discussion, many local approaches are already available for regression and the generalization of others should not be an obstacle, so that in principle the local RF methodology could very well be applied to the second step too.

Although some local methods are more costly than standard RFs (such as the ones requiring 2 RFs to fit), others are less (such as the one building RFs using only a subset of data). Do the authors think that considering an experimental setup where the same computational budget for, say, RF and NN-RF would change the results presented? That may lead to more trees being used for NN-RF and therefore to better results for that algorithm.

It is very difficult to evaluate the computational budget of each method. RF runtime is not linear in the number of point, and there is no rule linking the number of trees to the number of training data. Moreover, as now presented in Section 8.2 and 9.3, RF is the fastest method, the bottleneck of all other methods being the preprocessing to weigh the individuals or the variables. For instance the Nearest Neighbor approach takes 44 thousand times the runtime of RF, as identifying neighbors in large dimensions is very costly. Still we performed a comparison for NN-RF with 1000 neighbors using 100 or 200 trees, and the results were identical in both cases. With so few datapoints, 100 trees is enough to explore the variability of possible trees.

Text

I am a bit confused by the computational cost of the different local methods. Particularly, at the end of Section 4, the authors say that a tree built by local splitting rules may be much faster to construct and that seems reasonable. However, in the experimental section, all methods proposed in Section 4 are discarded saying they had "high computational cost". These two statements seem to disagree one with another. More in general, I advise the authors to explicitly say how the local methodology changes the computational cost of each of the proposed methods. Ideally, a table comparing all the proposed methods would be provided.

We are sorry this was indeed confusing and we hope our new formulation is now clearer. Localizing trees has two aspects. The first is that indeed, at each split only the leaf containing the local instance is considered, hence the number of criterion to compute is greatly reduced. Comparing the computational cost of local and global trees with identical criterion should lean in favor of the local tree. However, the second aspect is the criterion itself. Classif RF use a Gini-based criterion which requires only to estimate the proportion of each class in each leaf, computations that are highly efficient in any standard language. When the criterion is modified to information gain or kernel-based Gini criterion, it requires the computation of one weight per training data in the leaf, which can be very burdensome. This is particularly true since given our first results, we have not optimized our codes to allow faster computations.

As suggested we have now included runtime comparison at two locations in the paper. In the small second Gaussian example (table 8.3), we have included the runtime (in seconds) of each method in the last column. In the spherical fragmented example (table 8.4), we have compared each method tested on only one test-data with respect to the classic RF runtime. This allows to be more fair to methods which perform one pre-processing per test data since in practice local methods are of interest when only one or very few data are to be classified.

Comments on specific parts

Abstract

"performances are quite good even if no tuning is performed": I would change this text to something like "they perform well even with little or no tuning"

We have modified this sentence as suggested.

the last sentence of the abstract seems to refer only to the new methods proposed by the authors, but I think, after reading through the paper, it refers to all local methods.

You are right we have modified that. We have also modified the abstract to lighten the weight of ABC and focus directly on random forests.

Introduction

In the first paragraph, the authors talk about "eager learning". I never heard this nomenclature before, can they provide a reference to where that is introduced?

It is hard to trace back to the introduction of the eager learning nomenclature. It was likely defined by opposition to instance-based or lazy learning, for which we refer to Aha (1997) for the first introduction.

Fourth paragraph: remove "The" at the start.

This has been corrected.

Section 4

Maybe introduce this section with a sentence like "we now turn to discuss local tree methods" or similar.

We have modified this sentence as suggested.

second paragraph: "putting aside interesting data" -> "discarding data relevant for the considered example at early stages of the tree." or similar

We have modified this sentence as suggested.

"It is interesting to note that building a local tree by modifying its internal construction results in building a path." I don't get this statement. It seems to say that building a local tree corresponds to identifying a path inside a global tree, but I don't think that is the case, even according to the following text.

What we mean is that a traditional tree will split each node until a certain stopping criteria is reached (typically minimal entropy, or number of data points in the leave). This is not the case of a local tree, since at each recursion the algorithm will split a node into two leaves, discard entirely the leave that does not contain the data of interest, and continue the splitting process on the node of interest only. Hence only one path, or one trajectory, is visited within the tree. We have modified the paragrpah in hope of making our statement more clear.

Section 4.1

I think K refers to the number of models, but I don't recall it being introduced before.

Thank you for your careful reading, we have now defined this notation.

Third paragraph: I think the discussion on impurity measures not using the class is important. I think however the idea of using weights taking into account the class for determining a split can also be used for global trees, or am I missing something? If that is the case, maybe, the authors could point this out here, just for the sake of clarity.

In theory it would of course be possible to use weights in global trees as well. However, the crux of the matter with local trees is that the information gain only takes into account the mother node and one of the daughter nodes. In a global tree framework, the information gain of both daughter nodes is used. Hence in our 20-80 switch proportion example, such a situation could only be reached if the second daughter was almost pure, hence of very low impurity. Over every possible cuts, this one would be very competitive, so there is no need for extra weight introduction.

First bullet point (about discretisation) I am confused here, especially by the discussion of being located on the borders. Can the authors please revise the text in this paragraph?

What we meant is that splits on noise variable may lead to pure leaves even though the variable is not informative, and this may in turn lead to false classification. An example of such situation is depicted in the figure below, where x1 is a

noise variable whereas x2 is very informative. Yet a split along x1 will result in a pure leaf with sky-blue label, and thus an early stop of the algorithm, even though the data of interest should most likely have been predicted as purple label. Discretizing the variable would most likely result in bigger, and hence non-pure leaves, so that the algorithm would continue over a next iteration of split along another variable.

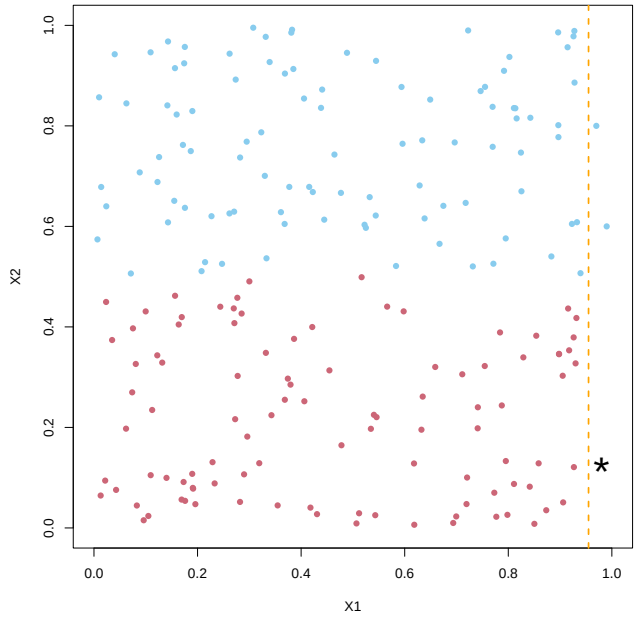

Figure 1: An illustrative classification problem with 2 classes (purple and sky blye), containing an informative covariate (x2) and a non-informative covariate (x1) and an unlabeled data to classify (black star). Splitting along x1 will result in a pure leaf with sky-blue label.

Third bullet point: I don't get this, is this meant to say that the covariates of x^* end up being the thresholds of the different cuts? That does not seem to make much sense, as how would follow along the tree if its entries exactly correspond to the thresholds?

Thank you for your careful reading, as precisely the covariates of the data of interets should NOT end up being the thresholds of the different cuts. Instead, as the covariates are now categorical, the only allowed splits are along the category values that are different from that of the data of interest, and each of those possible splits induce two leaves: one leaf with datapoints whose covariate is equal to the threshold, and one lead with the other datapoints. Then the local information gain is only computed using the mother node and this second leaf, as by definition the data of interest cannot belong to the first leaf.

Section 4.2

last paragraph: after introducing Armano and Tampone (2018), I believe the authors should use "that work" instead of "this work" to refer to Armano and Tampone (2018). Otherwise, I am confused as to whether they are talking about the current work or Armano and Tampone (2018).

This has been modified.

Section 4.3

after the expression for : "the node (4.2)" I think that should be "the daughter node"?

Yes thank you, we meant that the weights in each daughter node were modified as in (4.2).

"The major benefit of such weights is that they do not depend on the covariate index, thus the usual tree prediction, i.e. the majority class at the leaf where falls, can be replaced by a more coherent strategy with the tree construction,

using as a prediction the class with the maximal weighted class proportion at the leaf." Why can't this be done for the univariate kernel approach as well?

In the univariate approach, the weighted frequency of a given class label depends on the covariate. When the final prediction is made at the final leaf there is no reason to chose a covariate over another. It would however be possible to compute a multivariate weighted frequency at the final leaf at the price of additional computations, but this is something we have not tried.

Overall, sections 4.2 and 4.3 describe alternative ways to take into account info about x^* to build the tree, but could those be combined with 4.1? I am unsure

In theory there could be ideas that could be borrowed from one strategy to the other, but these two approaches are radically different and we are not sure it would make sense. For instance LazyDT works with categorical covariates, whereas a Gaussian kernel seems more appropriate for continuous variables. We could modify the information gain (4.3) of the kernel approaches to consider only the daughter node where the data of interest falls, but we would have to propose new normalisation to avoid negative gains and induce discriminatory power, but that would likely lead to more computations implying kernels. As they are kernel and lazyDT approaches are the most computationally expensive approaches we have tested, so it is likely that combining them would result in even more computational burden

Section 5.2

The authors use "NN" but I don't recall this abbreviation being defined before (although I understand it refers to Nearest Neighbors).

We apologize for the confusion and have now defined NN in the text.

Section 7.1

after Eq 7.1, some ??? are present.

We apologize about that and have corrected the reference in the manuscript.

Section 8

In the last paragraph, can you please cite the docs for the R package used for random forests?

We apologize about that, their seems to have been a citation missing for the ranger R package.

Section 8.1

In the caption of Table 8.2, I think the authors should say something like "100 additional noise variables" as, if I understood correctly, 20 noise variables are already present in the setup used for Table 8.1, and here we are adding 100 more. The same holds for Table 8.4

Thank you, we have clarified.

Section 10

The authors summarize the findings and properties of the local methods in the second and third paragraphs; I suggest they link the sections where they were introduced when discussing each method, as I find myself having to do some back-and-forth to recall which method is which.

We apologize for the inconvenience and thank you for the suggestion. Links to each relevant sections have now been added.

References

Aha, D. W., ed. 1997. *Lazy Learning*. Norwell, MA, USA: Kluwer Academic Publishers.

Pudlo, P., J.-M. Marin, A. Estoup, J.-M. Cornuet, M. Gautier, and C. P. Robert. 2016. "Reliable ABC Model Choice via Random Forests." *Bioinformatics* 32 (6): 859–66.