A survey of scikit-learn tree-related pull requests

Nelson Liu
University of Washington
nfliu@uw.edu

March 17, 2016

1 Introduction

While reviewing the various tree-related pull requests to evaluate which would make good candidates for a Google Summer of Code project, I figured it would be prudent to keep a written log of my thoughts regarding the various pull requests.
To find pull requests to work on, I ran a GitHub search with the keywords `'tree' is:pr is:open`. My search filters were extremely loose and they literally captured any reference of tree. This resulted in a lot of code regarding nearest neighbors / the source tree in general to pop up, but I figured that its better to cast a wider net to ensure that I don't miss any potential work to be done. The pull requests fetched are listed below in order from oldest to newest. I used this method of ordering because I figured that older pull requests are also more likely to be stale, which would make them good candidates for a GSoC project.

2 Tree-related pull requests and associated commentary

1. Title: **#911 WIP: Covariance Updates**
   Commentary: This pull request doesn't seem too relevant, although its development seems to have stalled. Last substantive comment was in July 2013.

2. Title: **#941 Adding a pruning method to the tree**
   Commentary: This pull request seems well suited to a Google Summer of Code project. Throughout the thread, the developers were in support of its implementation. It's also seem to have gone quite stale, the tree class has changed considerably since this PR was first written and the author seems to have given up development. **All in all, this seems like a great pull request to pick up for Google Summer of Code**

3. Title: **#1368 WIP: Implement ROC-SVM linear pairwise ranking loss with SGD**
   Commentary: This pull request deals with SGD, and is not quite relevant to my proposed theme of "tree" as a result. Seems to have gone stale, with little interest in revitalizing it from the outside.

4. Title: **#1454 MRG+1: Add resample to preprocessing**
   Commentary: This pull request doesn't deal with the tree module, although it seems mostly completed and is waiting for another review before merging (@agramfort already +1'd).

5. Title: **#1574 WIP: sample_weight support**
   Commentary: This pull request doesn't deal with the tree module, and seems to have gone stale with a fair amount of work left to be done. Many of its ideas seem to have been proposed / implemented in more recent pull requests, however (see the myriad of references at the bottom of the pull request).

6. Title: **#1742 MRG Training Score in Gridsearch**
   Commentary: This pull request doesn't deal with the tree module.

7. Title: **#1830 MRG: Evidence Accumulation Clustering**
   Commentary: This pull request doesn't deal with the tree module.

8. Title: **#1984 Implementation of OPTICS**
   Commentary: This pull request doesn't deal with the tree module.

9. Title: **#2043 OPTICS clustering**
   Commentary: This pull request doesn't deal with the tree module.

10. Title: **#2285 [WIP] Earth (MARS)**
    Commentary: This pull request doesn't deal with the tree module.

11. Title: **#2391 [MRG] Implemented Determinant ECOC**
    Commentary: This pull request doesn't deal with the tree module.

12. Title: **#2461 [MRG] Label power set multilabel classification strategy**
    Commentary: This pull request doesn't deal with the tree module.

13. Title: **#2530 [MRG] sklearn.tree.export_dict for converting Tree objects into a jsonable format**
    Commentary: This pull request is a proposal for a method for converting tree objects to JSON. The associated PR discussion eventually concluded that there have been no exports in the main project, and that it was unfit to be in the main repo. @GaelVaroquaux suggested that it was out of scope, but it might be useful to have PMML export of trees from scikit-learn. This has been since completed in a separate project with some relevant discussion at #1596.

14. Title: **#2567 Issue #2559: Added normalize option to LogisticRegression**
    Commentary: This pull request doesn't deal with the tree module.

15. Title: **#2580 [WIP] first cut at LambdaMART**
    Commentary: This pull request doesn't deal with the tree module. Seems to have gone stale, although there appears to be some lingering interest in its implementation. @mblondel did however express that he has had some bad experiences with LambdaMART, and that it "performs either worse or comparably to random forests or gradient boosting".

16. Title: **#2784 Liblinear Sample Weights**
    Commentary: This pull request doesn't deal with the tree module, although work seems to have gone stale. It has since been picked up by @mechcoder, and he has made a few PRs toward its completion.

17. Title: **#2567 Issue #2559: Added normalize option to LogisticRegression**
    Commentary: This pull request doesn't deal with the tree module.

18. Title: **#3306 [[MRG] Google Summer of Code 2014: Standard Extreme Learning Machines**
    Commentary: This pull request doesn't deal with the tree module, and seems mostly implemented.

19. Title: **#3346 [WIP] Categorical split for decision tree**
    Commentary: This pull request is extremely relevant to the tree module. The original author of this PR seems to have abandoned development, and it was subsequently reproposed in #4899 NOCATS: Categorical splits for tree-based learners, which is still in active development.

20. Title: **#3306 [MRG] Google Summer of Code 2014: Standard Extreme Learning Machines**
    Commentary: This pull request doesn't deal with the tree module, and seems mostly implemented.

21. Title: **#3436 ENH: Added Random Forest feature importance based on out of bag data**
    Commentary: This pull request doesn't have any commentary from core devs or even other developers, so I can't tell if it would be well received as an enhancement. However, it was referenced in issue #3455, which was subsequently picked up and seemingly abandoned in #3723. **I'm on the fence as to whether this would be a suitable Google Summer of Code project, as it isn't quite related to trees but seems somewhat useful nonetheless. It's difficult to tell without more comments on the matter from core devs, particularly @amueller and @larsmans as they designed the scorer API**

22. Title: **#3464 Added Example for Plotting Hierarchical Clustering Dendrogram**
    Commentary: This pull request doesn't quite deal with the tree module, but is nonetheless an interesting example that plots the hierarchical clustering dendrogram. This is useful because hierarchical cluster is one of the few clustering algorithms where adding structure results in speed improvements. **This might be out of the scope of my Google Summer of Code project on trees, but it's a very interesting project that has stalled due to missing (now present) functionality. Should be tagged as "Need Contributor", or I could possibly find some time on the side to work on it.**

23. Title: **#3907 [WIP] Adding tests for estimators implementing 'partial_fit' and a few other related fixes / enhancements**
    Commentary: This pull request doesn't quite deal with the tree module, and it seems like @rvraghav93 is still actively working on it.

24. Title: **#3922 [WIP] allow nearest neighbors algorithm to be an estimator**
    Commentary: @jnothman seems to have finished most of the work on this PR, but it is stalled in the review process. It seems most devs are +1 on adding this. I've pinged @jnothman about the current status of the pull request, and am waiting for his response.

25. Title: **#4191 Add Jensen-Shannon distance metric**
    Commentary: This pull request seems to have most of the core functionality in place, but it's missing docs and a convincing example as to why it should be included. Most of the PR discussion centers around whether the Jensen-Shannon distance metric would be useful for solving a real applied problem. The original contributor seems to have abandoned the project, although I'm not sure there's a clear consensus as to whether it would be well received.

26. Title: **#4215 [MRG] GBM & meta-ensembles - support for class_weight**
    Commentary: This pull request seems to have been abandoned, but most of the work is completed. Maybe just needs some refactoring before another review. Not quite relevant to the tree module, however.

27. Title: **#4217 [MRG] Neighbors refactor**
    Commentary: This pull request seems mostly completed, and the devs were discussing how it would affect performance (the answer seems to be minimally, but its not possible to tell completely yet) before it went stale. Again, not quite relevant to the tree module.

28. Title: **#4288 [WIP] Allow nan for userdefined metric in dbscan et al**
    Commentary: This pull request seems to have been abandoned. Regardless, it does not deal with the tree module.

29. Title: **#4354 [WIP] Make GridSearchCV and all estimators support sparse y**
    Commentary: @rvraghav93 is still working on this pull request, and is still working on it. There is a temporary standstill because he is waiting for sparse y (which he and @betatim) are working on to be supported in trees before continuing.

30. Title: **#4458 [MRG+1] Implement haversine metric in pairwise**
    Commentary: This pull request seems to have gone stale, but all that is needed for its completion is some benchmarking. Again, not quite relevant to the tree module.

31. Title: **#4501 [MRG] Add distance threshold on Hierarchical Clustering, see #3796**
    Commentary: This pull request seems to be mostly completed, but there are still a few outstanding comments to be addressed. Doesn't quite fit in the scope of a project on the tree module.

32. Title: **#4522 [WIP] Metrics Testing**
    Commentary: This pull request doesn't deal with the tree module, but instead works with neighbors / clustering.

33. Title: **#4525 [MRG+1] Listed valid metrics for neighbors algorithms**
    Commentary: This pull request doesn't work with the tree module. However, it seems completed and is just waiting for a few more reviews before merge.

34. Title: **#4535 [WIP] Fortran c order test**
    This PR seems to have gone stale, and is not quite relevant for the tree module. Seems like a significant amount of work would be required to get this working.

35. Title: **#4801 Discretization using Fayyad's MDLP stop criterion**
    This PR seems mostly complete, but it seems like most of the devs do not know of suitable applications for the code. The original developer is still willing to work on the code, in the case that it is deemed relevant enough for inclusion.

36. Title: **#4844 [MRG] Add KNN strategy for imputation**
    This PR seems to have gone stale, however it seems as if its ready for review and merging. Regardless, it deals more with NN than the tree module.

37. Title: **#4848 [MRG+1] Multioutput bagging**
    This PR seems complete, and is just waiting for a few more reviews before merging.

38. Title: **#4899 NOCATS: Categorical splits for tree-based learners**
    Currently in active development (mostly by @rvraghav93), and very relevant to the tree module.

39. Title: **#4926 Hdbscan**
    Comments on the PR indicate that it will not be merged into master, and is best served as a related project.

40. Title: **#4950 Add monotonicity parameter to Gradient Boosting and Decision Trees**
    This PR seems mostly complete, and doesn't cause many changes in time complexity to the tree code. However, it is still debatable whether this niche case is worth the added complexity in the tree code. Seems like some more discussion is needed, but seems mostly complete otherwise.

41. Title: **#4951 DOC add official install guide in README.rst**
    Not relevant to the tree module.

42. Title: **#5099 [WIP] Modified LSHForest to check candidates only in relevant sections**
The code seems mostly completed, and is just needing benchmarks that demonstrate its benefit over the current implementation of LSHForest. @jnothman ran some initial tests, however, and they "don't seem to be as positive as I'd hoped." Regardless, it is more related to neighbors code than the tree module.

43. Title: **#5181 Balanced Random Forest**
Seems to have gone stale, but is needing documentation and tests before merge. The code seems mostly complete though? This isn't quite relevant to the tree module, but it **could be an interesting PR to pick up on the side**.

44. Title: **#5279 WIP: LOF algorithm (Anomaly Detection)**
This PR seems to still be in active development.

45. Title: **#5333 [MRG] pairwise_distances outputs Nan and negative values**
This PR seems to have gone slightly stale, but it seems mostly complete and is waiting for review.

46. Title: **#5333 [MRG] pairwise_distances outputs Nan and negative values**
This PR seems to be stale, and doesn't seem too relevant to the tree code.

47. Title: **#5414 [MRG+1] Elkans K means**
This PR is still in active development, and just needs some reviews before it can be merged.

48. Title: **#5460 [MRG] convert to boolean arrays for boolean distances (ex: jaccard)**
This PR seems to have gone slightly stale, but generally looks complete. Maybe just needs a few more reviews before merging? Not too relevant to the tree module.

49. Title: **#5491 [WIP] Gaussian Process-based hyper-parameter optimization**
This PR is still in active development.

50. Title: **#5532 Add 'return_std' option to ensembles**
This PR seems to have slightly stale, but @glouppe might still be working on it. Not too relevant to the tree module, moreso for ensemble.

51. Title: **#5564 [WIP] Default dict for GridSearch**
This PR seems to have gone stale, and isn't too relevant to the tree code. **Not too relevant to a Google Summer of Code proposal on trees, but could be something interesting to pick up on the side, though.**

52. Title: **#5593 [MRG] ENH: Support threshold='auto' in Birch**
@mechcoder did say he plans on finishing this PR, but there has been no activity since January. Regardless, it is more relevant to the cluster module.

53. Title: **#5596 [MRG] Fix NearestNeighbors algorithm='auto' to work with all supported metrics by default**
@mechcoder gave a brief review, but the original author has yet to address the comments. **This PR looks almost complete, but it isn't quite relevant to the tree module. Another thing to possibly work on on the side**.

54. Title: **#5653 [WIP] Generalized partial dependence plots**
This PR is slightly stale as the original author is a bit busy. At the moment, its more suited for ensemble / boosted methods than just the decision tree. This project also seems fairly big, so it is probably out of scope.

55. Title: **#5689 [MRG] Gradient Boosting Classifier CV**
    This PR seems to have gone slightly stale, but appears to have been well received by many contributors. Not relevant to the the tree module, however.

56. Title: **#5757 [MRG+1] fixed IsolationForest(max_features=0.8).predict(X) fails input validation**
    This PR fixes the bug at hand, but causes a memory explosion on large datasets, causing it to be 5x slower. Probably needs some more discussion / review before it can be merged. Again not quite relevant to the tree module.

57. Title: **#5815 #5789 Expose base estimator iforest**
    PR seems to have gone stale, but seems like there is some disagreement about how to move forward with the PR in terms of API / confusion with the meaning of various parameters. Not quite relevant to the tree module.

58. Title: **#5825 [WIP] Adding Fixed Width Discretization**
    PR seems to have gone slightly stale, and it does not quite apply to the tree module.

59. Title: **#5862 Add clarification on random forest regressor default params**
    Docfix that isn't relevant to the tree module. Needs reviewing, however.

60. Title: **#5963 Added subsampling to RandomForest**
    PR seems to have gone stale. Hasn't been quite reviewed yet, so it's hard to gauge the level of developer acceptance of the feature. Regardless, this PR doesn't apply to the tree module.

61. Title: **#5974 [MRG] ENH Add support for missing values to Tree based Classifiers**
    @rvraghav93 is actively working on this PR.

62. Title: **#6029 Adding Time Series Regressors**
    Time Series is generally out of scope in scikit-learn, and this PR is unlikely to be merged into master as a result.

63. Title: **#6039 Add mae regression tree**
    The original maintainer is not working on the PR anymore, but @vene and @mechcoder are interested in taking it on. **This is pretty relevant to the tree module, and perhaps would be something useful to take on if @vene or @mechcoder are not already working on it.**

64. Title: **#6114 Fix docstring signature mismatch in cython code**
    Docfix PR, so not quite relevant to the tree module. This is probably a quick fix though, and can be pushed through pretty quickly if someone decides to take it up.

65. Title: **#6169 BestFirstTreeBuilder should ignore tree.max_depth**
    This is a quick bugfix in the tree module, there should probably be some clearer documentation about `max_depth` and `max_leaf_node` and how `max_depth` would not be ignored in that case. **Probably a pretty quick fix if the original auathor decides to abandon the PR**.

66. Title: **#6274 [MRG] MAINT: Make sure all tests are included with installation**
    This is a build PR, and is not relevant to the tree module.

67. Title: **#6376 [MRG+2] Fix for issue #6352**
    This PR is relevant to the tree module, but it is pretty much done. Maybe one more review + merge?

68. Title: **#6380 WIP: Add decision tree plotting**
    This PR is relevant to the tree module, but would require fairly wide modification in the API to enable functions like `DecisionTreeClassifier/Regressor().plot()` and would thus have to go through a full RFC and enhancement proposal. Not really any strong opinions about whether it would be worth it to go through with submitting a proposal yet.

69. Title: **#6400 [MRG] ENH use n_jobs in brute-force radius_neighbors**
    This PR is more relevant to the neighbors module than the tree module. There haven't been any reviews on it yet.

3 Summary

Issues that seem fit for a project would be:

1. **#941 Adding a pruning method to the tree**
2. **#6039 Add mae regression tree**
3. **#6169 BestFirstTreeBuilder should ignore tree.max_depth**

Out of these, **#941** and **#6039** are the large issues, and I believe that I could tackle **#6169** as well if it's not complete by the time Google Summer of Code comes around. Additionally, I'm tentatively planning on taking on a few misc PR's during the summer (e.g. **#5099**) and participating in code reviews / discussions per usual.