Chait Manapragada <chaitanya.manapragada@monash.edu>

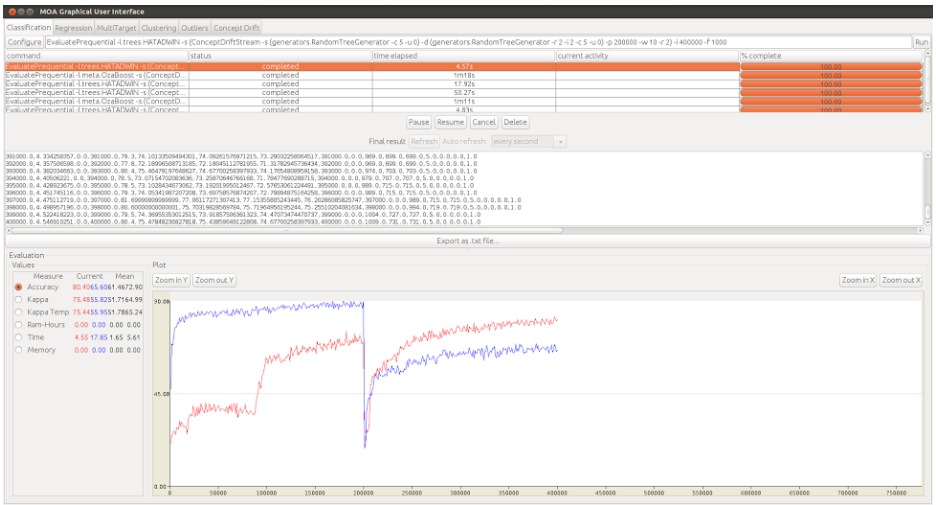
Structured data...

Chait Manapragada <chaitanya.manapragada@monash.edu> 17 June 2017 at 17:25
To: Geoff Webb <geoff.webb@monash.edu>

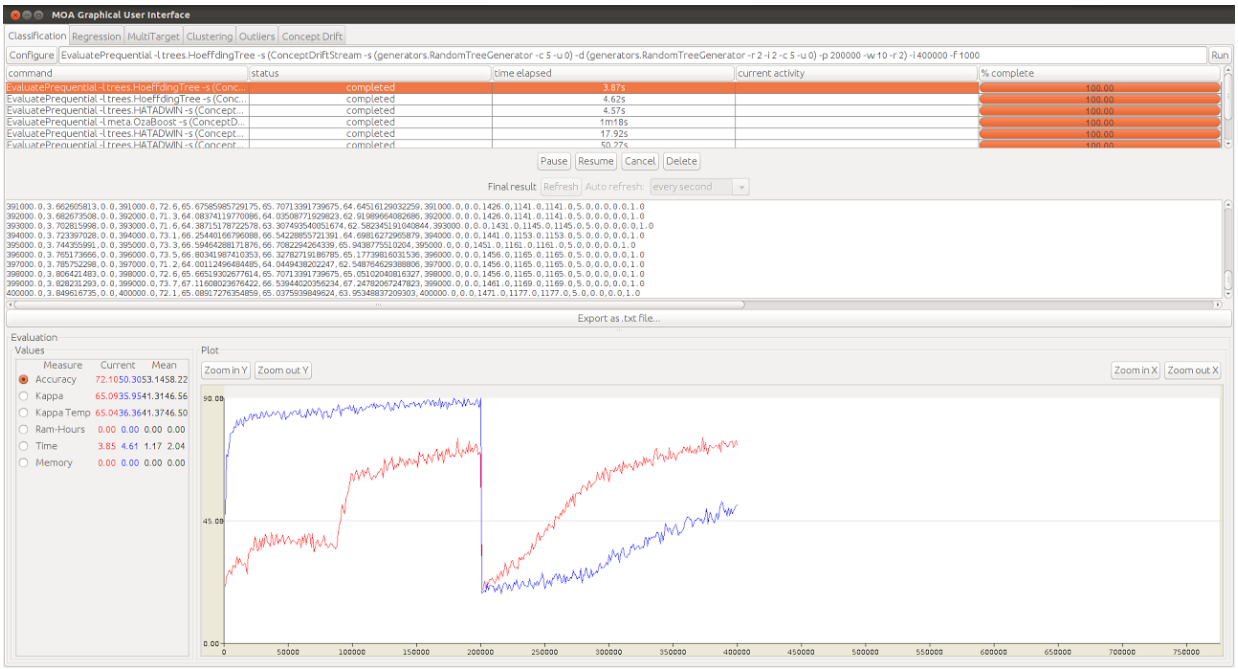
Before I decide it's an OzaBoost problem, I looked into the data stream- the MOA trees to figure if they treat ordinals and nominals equally. I don't think they do.

Both curves below are accuracy for HAT-ADWIN. The blue curve has 5 extra numeric attributes in addition to 5 nominals. Notice how well it is learned.

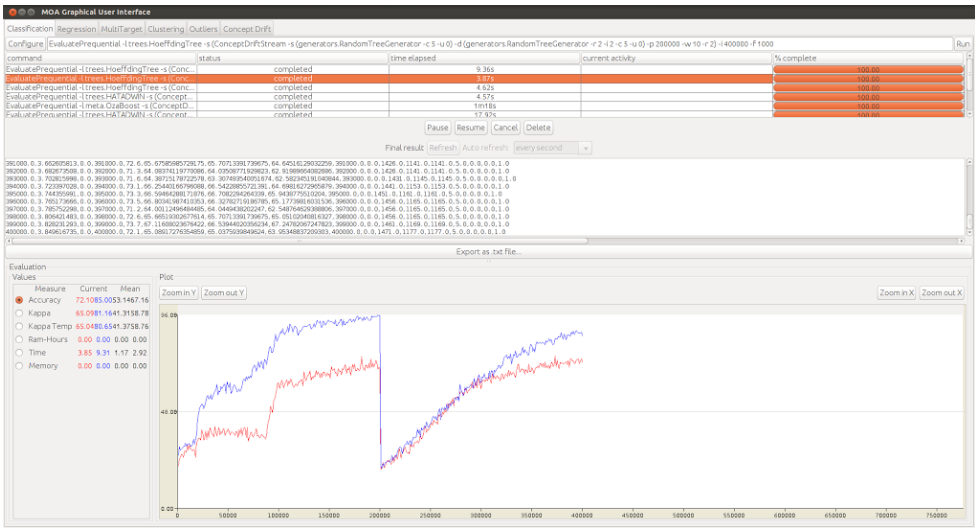
All I did to get HAT-ADWIN to get the red performance (low accuracy pre-drift) was... to take out those 5 ordinals. With just the 5 nominals that also previously existed, but without the ordinals... we get very slow initial learning.



Could it be an issue with HAT-ADWIN? Let's try VFDT. VFDT does something similar- taking out the ordinals and leaving in the nominals makes it a worse pre-drift learner (red). Blue has 5 ordinals, 5x5 nominal. Red has 0 ordinals, 5x5 nominal.



Let's add a single ordinal attribute (blue) and compare with no ordinals on VFDT. Blue has 1 ordinal, 5x5 nominal. Red has 0 ordinals, 5x5 nominal.



It's made it better! i.e. the blue curve with a single ordinal has done better than the red curve with no ordinals.

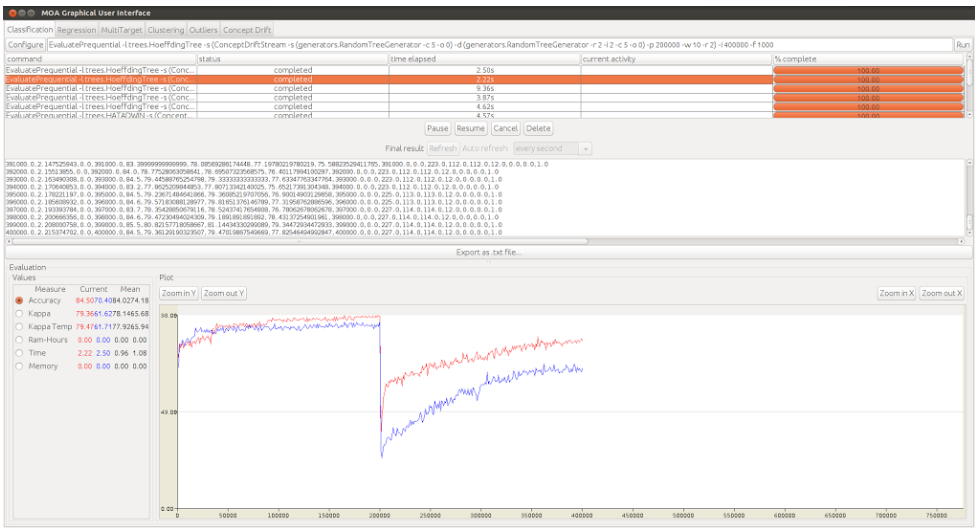
This could be due to:
(1) the specific random seeds chosen for building these trees cause a sharp increase in prior class purity when an ordinal is added (but why?)
(2) a bug in VFDT
(3) a bug that always biases the MOA trees towards having ordinals contain the greatest information relative to nominals, so much so that their addition causes an increase in prior class purity.
(4) A combination of (1) and (3)

The biggest issue with using the MOA trees would arise if (3) were true, so let's test that prognosis first. Please note that we're examining the pre-drift curve for now- I've kept the post-drift portion because it could be useful information.

Let's start by looking at how adding nominals affects accuracy. Our expectation is that adding dimensions should make the distribution generally harder to learn.

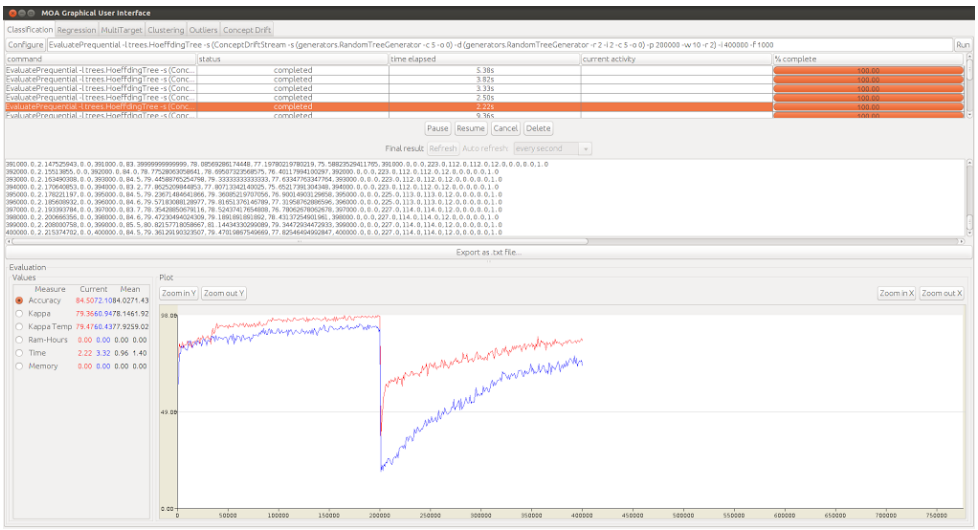
We start with 5 ordinals (red).

Then a single 5-valued (1x5) nominal (blue) is added:

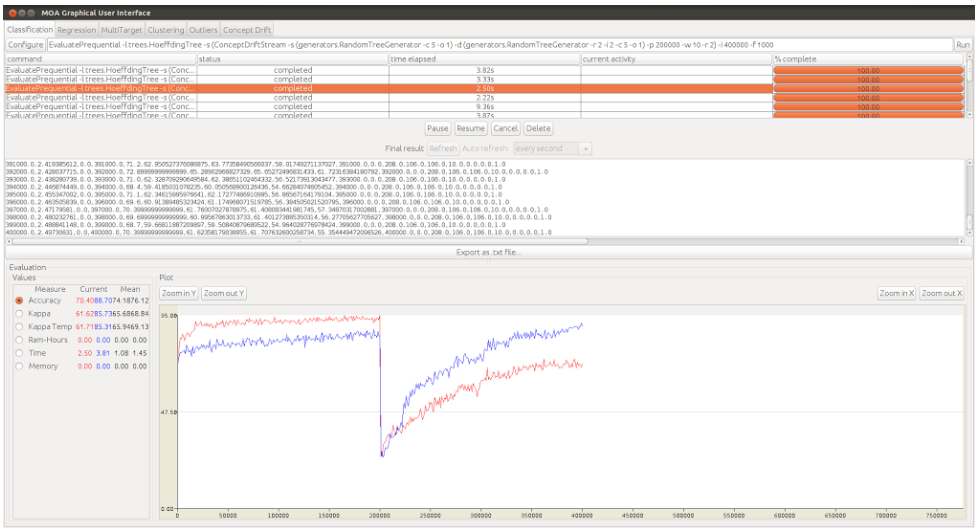


Please note, again, that it is the pre-drift learning behaviour we are currently examining.

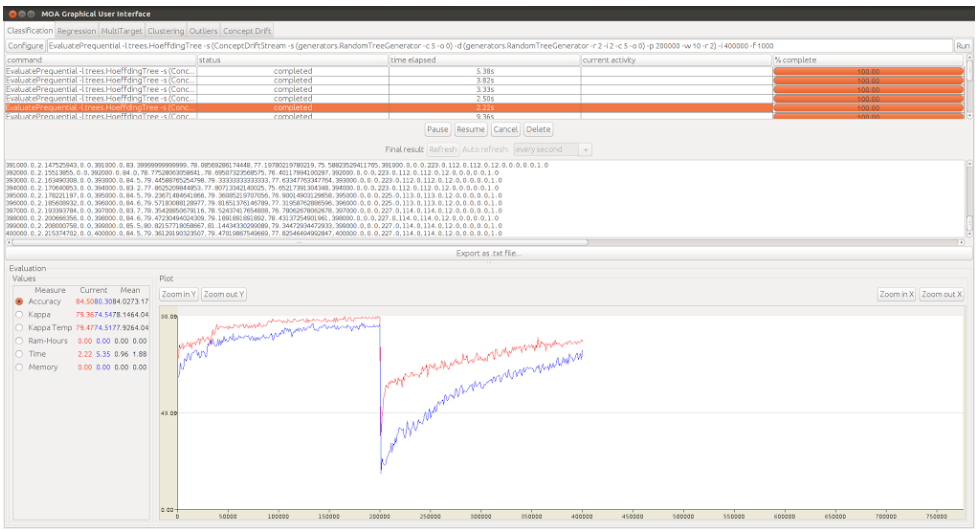
Two nominals added (2x5) to existing 5 ordinals.



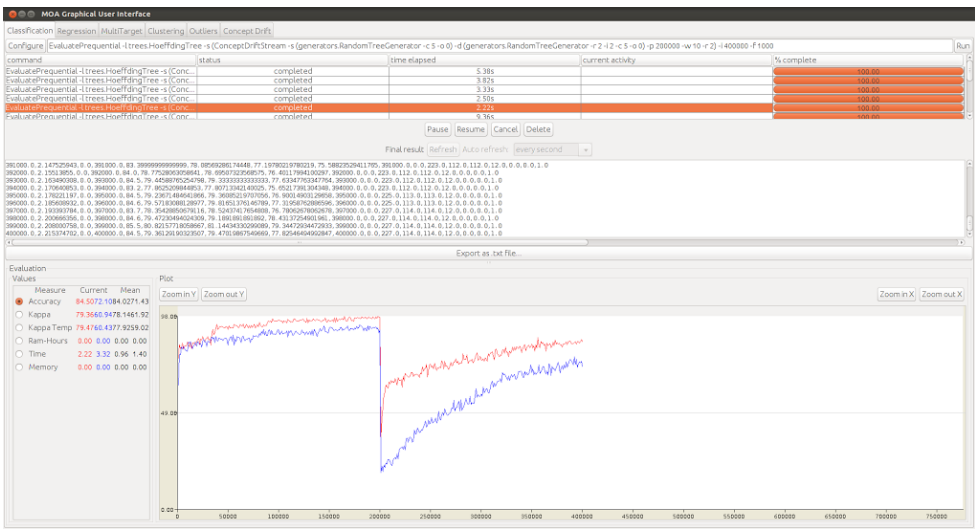
Three 5-valued nominals added (3x5) to existing 5 ordinals.



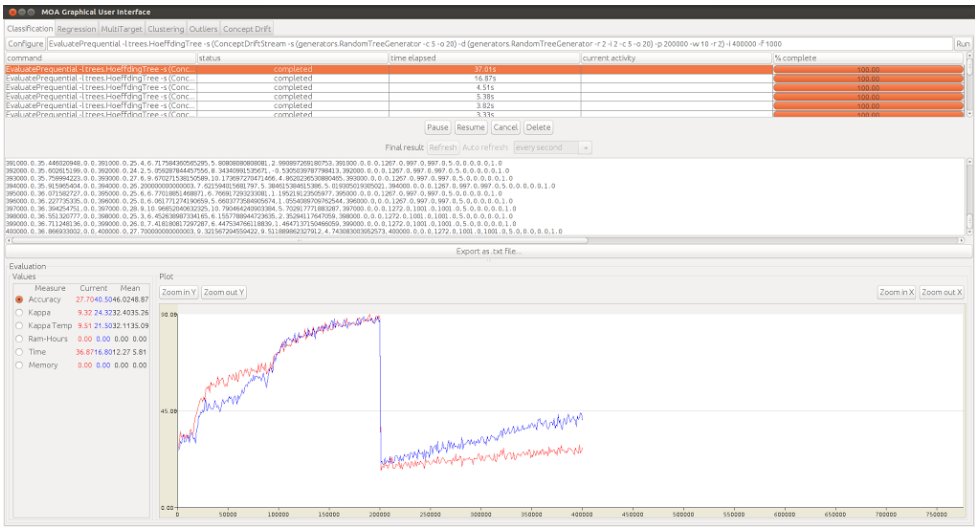
Let's go to 4 nominals added. (4x5). Looks like this is doing better than with 3x5.



And now... a 5x5 ordinal is added.

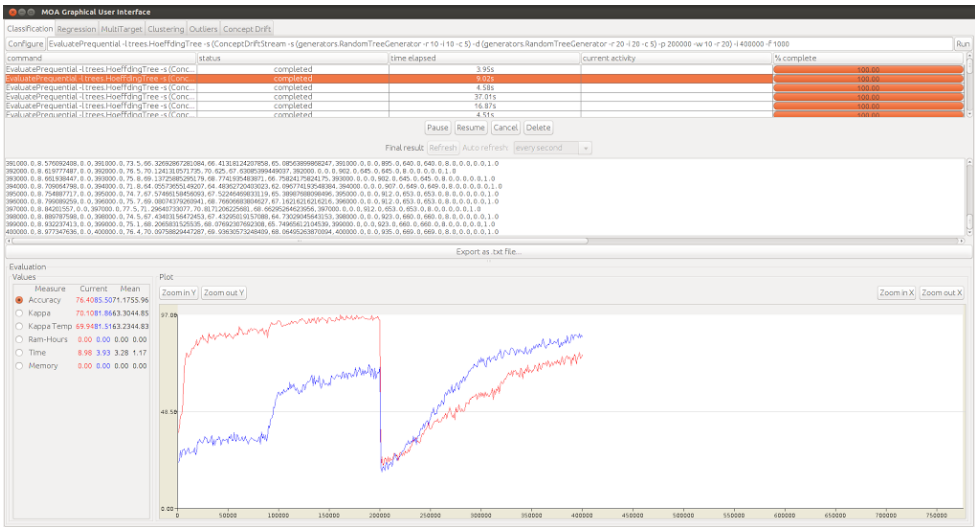


We've just added a massive number of dimensions to the data and learning has stopped getting worse? Let's try 10x5(blue) and 20x5(red).



The good news is that nominal attributes have enough information that adding them affects accuracy. That doesn't mean that ordinals are not given disproportionate information.

In order to prove it's not an effect from the seeds chosen, I'll have to do a large number of experiments with different seeds and show that removing the ordinals always pushes down accuracy, suggesting ordinals are given disproportionate information. However, the indiciations are that it's not the seeds at fault, there is a bias towards ordinals in the stream being associated with constructing trees with purer priors.



So directions are:

(a) Investigate this potential ordinal/nominal discrepancy (taking out ordinal attributes shouldn't increase performance so much)
(b) Continue building my own tree

My goal is to compare HAT-ADWIN and OzaBoost, not get stuck with data stream issues. Both above tasks will take focus away from the task at hand, so I need to devise an efficient strategy that avoids both.

I just need a reliable data source for now, so inducing some sort of simple structure for the data will probably do. Such as a relatively simple Bayes' net. Even better if I can find a reliable existing data generator.

At this point, I'm a little worried about MOA implementations in general.

Regards,

Chait

[Quoted text hidden]