General:

The 2nd paragraph in 4.1 starting "In this research we propose.."

Sentence "TEAK does not use the overall performance of the whole dataset to decide on the number of analogies, TEAK does not use a fixed number of analogies for all test instances. Instead TEAK prunes all unnecessary analogies on the basis of performance-variance for single test instance…" should appear in the early sections of the paper describing the method.

Should be more explicit that the method is dynamically selecting the K value for each case using GAC trees algorithm..etc

What's missing is perhaps a discussion of computation complexity, how long does it take to compute a full dataset for example?

I personally think the method would work, but the paper needs a bit more work and restructuring the sections. And more samples are need to explain the idea.

Section 1

Paragraph 4, when you say .." when the dataset contains "discontinuities", define the meaning of "discontinuities" in here.

Also define "predictive performance"

Paragraph 4 can be split into two para. Split at "In this paper.. "

Section 2:

Paragraph 3, "software effort datasets are characteristically noisy datasets and CBR methods are more capable of handling noisy datasets than regression based models"

I don't think this statement is entirely true, unless you can provide an example to justify. Again it all depends on the dataset and method combination.

Paragraph 4 too long, break it down into smaller paragraphs

Last paragraph, last sentence, replace "come up with" with "identify" .

Section 3:

Para 1, consider rephrase the first few sentences.

Change to: “In this paper we propose a new approach…”

Delete sentence : “To the best of … “

“We would like to address the problem of (how many) analogies to use…”

Research Question 1, consider rephrase, ..”understand the dataset and the underlying characteristics of the dataset” sounds repetitive.

Research Questions2: “ease the procedure”?? I don’t think we are easing the procedure, we are trying to improve the procedure so the prediction accuracy can be improved.

…called TEAK on the basis of “performance-variance”. ? Need to explain this word. This paragraph needs a bit more work.

Section 3.1

I suggest using sub-sections ie. 3.1.1 COCOMO81 dataset, 3.1.2 NASA93 dataset etc..

ISBSG dataset, you need to mention what quality of the datasets were chosen, the selection criteria. For example in ISBSG they have quality rating = A , this is one of your criteria.

Section 3.2 TEAK Model

I suggest resurrecting the acronym TEAK again in here.

Needs an example in this section perhaps.

The sentence “ The popularity of making use of agglomerative clustering is its ability to use arbitrary dissimilarity or distances functions” is obvious, needs a better justification.

Paragraph starting “In this research we are building up two GAC trees…” needs a bit of work, shorter sentences.

Figure 2. GAC Pseudocode :

If I understand the algorithm correctly, then it will enter into an “infinite loop” when an odd number is reached.

For example, if Training_Instances = 6, function findClosest will produce two closest1 and closest2, Until size(currentLevelNodes) will be less than 2, but not equal to 0. It will repeat, when rerun findClosest, it will produce two sets (cloest1, closest2) again, in this case I am not sure what it will produce. The loop will go on forever.

Also nextLevelNode = null does nothing in the routine. Can it be removed?

Figure 3 Probabilistic Selection

Needs a better explanation, what is rand()^bias ? What does it do??

Then the following 3 paragraphs are unclear, needs a bit more work.

"Given (three subtrees) containing N = N1+N2+N3…" What are Ns ? Unclear in here.

Next paragraph:

"Variance" not "varince"

Figure 4: TEAK Pseudocode

Dataset = { Cocomo, SDR..etc } add brace symbol

Why 1 to 20 in the pseudocode in here? Why not 1 to N ?

"selectNode = start from root, …until , .. median (finalNode) " are very unclear and confusing.

Sentence "TEAK does not need any expert interference to discover dataset characteristics …."

is contradicting to a previous statement:

"TEAK enables its user to fine-tune GAC2 tree with the help of a variable called bias that represents the user's expert bias."

Needs example in the paragraph starting : "Regular k-based CBR methods start with single analogies…"

Paragraph "In that respect TEAK may resemble (Ada Boost) algorithm. " What's the purpose of including a discussion of Ada Boost in here?

Section 3.3 Experiment Design.

The last paragraph is unclear and rather confusing. Understand that you used 5 static k-values (1,2,4,8,16), and also one "dynamic K" value which is the best performing k value for each dataset? When a K value is selected and applies to individual case selection, it is not dynamic, and should be fixed/static. The best you could say it is the ideal-K for the dataset. Dynamic K is when Ks are different on each individual case selection (each Xi).

Section 3.4

May be a simple justification for the use of MdMRE is discussed in here.

I don't like the term "error based", it sounds very negative. Perhaps "residual based"?

Paragraph starting "In addition to MRE and Pred (25)…", sentence :"we first check if methodi and methodj are statistically different…"

Define : methodi, methodj, roundk, and tiei

Figure 6 Pseudocode for Win-Tie-Loss…

Shouldn't be WILCONXON (MREi, MREj) instead of (MREi, MREi) ?

Do you need to add (tiej=tiej+1)after tiei=tiei+1 ?

Mean = ?

Needs a bit explanation , it is really a counter for success / failure.

3.5 "training set" not "train set", please update similar error in other parts of the paper.

Section 3.5 Threats to validity

Perhaps move this section after results?

4.2 Results

"TEAK has a loss value of 0 for all datasets and this shows that TEAK has never been outperformed by any other method in all datasets for statistically significant cases"

This is too good to be true, ensure the code applied on the datasets are correct… I am not very confident about this statement..

Figure 9, the diagrams are too small to see, not sure showing all the plots add value to the paper.

I would like to see the result using MMRE based on the Desharnaise dataset.