Dear reviewers,

Thank you kindly for your comments on our manuscript. We value your input highly and have done our best to revise our manuscript in light of your comments. In the following, we address your comments in the order in which they were given in the letter from the editor to the authors.

With kind regards,

Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj and Sašo Džeroski

Reviewers' comments:

Reviewer #1: This paper presents a comparison of 12 multi-label learning methods. The methods are evaluated over 11 datasets from different application domains using a large number of evaluation measures (16 measures divided into example-based, label-based and ranking-based) and the statistical significance of the obtained results is also analyzed. It is, as the authors put it, an extensive amount of work. Furthermore, the paper is well written and organized and clearly motivated. There are however, several issues that should be addressed, in order to improve the quality of the manuscript.

Major points

1. Being such an exhaustive comparison, the number of tables and graphs produced is extremely large and difficult to analyze. I realize many results are already only in appendix. However, it is still essential that the authors make a very strong effort to make the results more compact and more enlightening in the final manuscript. One hypothesis, which is already mentioned by the authors in the conclusion, is to present graphs of precision versus recall instead of tables of precision and tables of recall or simply to present the values of area under the curve. Another idea is to remove accuracy. Is it really necessary when precision and recall are already calculated? Concerning the results using only the methods that finished and all the methods but with a penalty for the ones that did not finish. Is this duplication really necessary? Are the conclusions from the two any different? On the other hand, the authors should expand sections 5.1 to 5.3 where these results are discussed, providing a more detailed and profound analysis that may be more instructive for the readers.

>> Done. We have completely reorganized the results and their presentation in the Results and discussion section and Appendix B. In the updated version of the manuscript, only the results of the statistical evaluation for the datasets on which all algorithms finish are included in the main text: The tables, summary of the Friedman test and critical diagrams for all the datasets are given in Appendix B. Considering the precision-recall curves, we would like to point out that these curves are used in the task of hierarchical multi-label classification (e.g., see C. N. Silla Jr. & A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*. Vol. 22, No. 1-2, pp. 31-72, 2011), and can provide additional viewpoint on the task of multi-label learning. However, since this measure has not been previously used in this community we omit it from the results and just note that we plan to further investigate it. Finally, we report both statistical evaluation results (on all datasets, and on the ones that all algorithms produced a predictive model under a week of running), because the results on all datasets can give further insight onto the performance of the algorithms. For example, RF-PCT and BR perform better if we evaluate on all datasets, according to the example-based measures, and RF-PCT becomes statistically significantly better than HOMER on all ranking-based evaluation measures.

2. In section 4.1 the authors describe the datasets used in the experiments. Table 1 indicates the application domain, number of features, number of labels and label cardinality. Although a reference is provided for each dataset, for the sake of completeness and, more importantly, for motivation purposes, there should be a short description of each dataset explaining what the labels are (not only their number and cardinality) and what is the sense of muti-labelling in each case.

>> Done. We have provided short descriptions for each dataset.

Minor points

1. Give the expression of the Gini indices in section 3.4 and indicate what T denotes in the expression of Var(E)

>> Done. We have provided the expression and a reference.

2. Provide a reference for the critical distance at the end of the penultimate paragraph of section 4.4

>> Done. We have provided a reference.

3. In table 1, the numbers of training and testing examples used for each dataset are presented. Why this division and not another?

>> Done. The division for the (majority) datasets is from the literature that first presented them. For the remainder, we divided them 2/3 vs. 1/3. This is now clearly stated in the manuscript.

4. Before table 2 indicate what DNF stands for

>> Done. In each table caption, we state what DNF stands for (DNF – Did Not Finish).

5. What is the penalty given for a method that did not finish?

>> Done. The penalty for not providing a predictive model for a week is that the algorithm for the given dataset is ranked last (final paragraph of Section 4.4 - Statistical evaluation).

Reviewer #2: 1 - This paper presents an experimental comparison of multi-label (ML) learning methods, in the context of ML classification and ML ranking. It focus the evaluation of algorithms of the different families of approaches using several benchmark problems and evaluation measures.

2 - There is a problem with the structure of the article. Section 3 (Methods for multi-label learning) have many repetition of ideas presented on section 2.2 (An overview of methods for multi-label learning).

>> Done. We have edited the text and removed the repetitive parts.

3 - The authors propose a new categorization for ML algorithms, extending the categorization proposed by Tsoumakas. This new class is entitled "Ensemble Methods", and includes the RF-PCT algorithm, that, according to authors conclusions the best performing. The inclusion of Rakel in this class of algorithms should be better explained, as it is a problem transformation approach.

>> Done. We have revised the text concerning RAkEL and made it clearer why we believe is an ensemble method. RAkEL draws *m* random subsets of size *k* from all labels *L* and trains *m* label power-set classifiers using these labels. A simple voting process determines the final classification set.

4 - The paper aim is to be considered a reference survey that identifies the best performing algorithms based on experimental evaluation? I so, I think that the algorithms that showed better performance should be explained with further detail, allowing a easier reading to a new reader in the topic of ML learning.

>> The manuscript is written without prejudice of the outcome of the analysis. Moreover, we believe that the text should not have a tone favoring some methods over others, since all of them have equal chances at the begining. If the tone of the manuscript is changed, then the paper will compare some selected methods against others. Our intention was to provide an unbiased and extensive evaluation of all the methods over a variety of domains. Furthermore, for each method, we provide a reference and a download location (when it is available), and enough information for the reproducibility of the experiments.

5 - Figure 1 (pag 40) - does not match with text! The Homer algorithm is categorized as Pairwise algorithm, instead of Label power set method; Calibrated label Ranking (CLR) and Quick Weighted Voting Method (QWML) are categorized as label power set methods instead of pair-wise methods

>> Done. We have updated the figure.

6 - The statistic evaluation of the differences between the algorithms is assessed using the non-parametric corrected Friedman test, and a post-hoc Nemenyi test. As these information is relevant for the conclusions of the article, these tests should be summarized and the conclusions that can be extracted from them better explained. One of the critical diagrams should be explained in detail in the text.

>> Done. We have summarized the outcomes of the Friedman test on each evaluation measure. We also revised the text to make the explanation of the test more clear.

7- Appendix A existence is questionable. In [29] some of the measures (as mentioned) are extensively presented. The evaluation measures that are not so well defined in previous texts should be integrated in the text.

>> Although the majority of the evaluation measures are given in Tsoumakas et al. (2010), we believe that the evaluation measures should remain a part of the paper (in Appendix) for completeness. Namely, if a reader wants to check some information concerning a given evaluation measure, he should not be forced to download an additional paper.

8 - Tables should have captions. The reading of the article would become much easier if the tables (integrated at the end of the document) would have information about what they are describing, and what is highlighted.

What information is represented in tables B.6 to B.19? They are referred in appendix A and B, and not discussed in the main text?!?

>> Done. This version of the manuscript did not contain the caption because the submission system puts each table and figure at the end of the manuscript. All tables and figures come with an appropriate caption.

9 - Minor corrections: pag 10, 3rd paragraph: These are given in Figure 4 to REF?

>> Done. We have updated the reference.