SDM2016
SIAM International Conference on Data Mining (2016)
May 5-7, 2016, Miami, FL, USA

Reviews For Paper

Paper ID	137
Title	Learning Hierarchically Decomposable Concepts with Active Over-labeling

---

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
The paper is relevant to SDM	Yes
Which of the following is the most accurate summary of the type of the work proposed in the paper	A new problem formulation
Which of the following is the best characterization of the paper contribution	Innovative work [idea] but requires more convincing
Is the abstract an accurate summary of the paper	Yes
Does the paper properly contextualize the "Research Question" [e.g., through the Introduction and or Related Work Section]	No
Which of the following is the most accurate description of the writing style of the paper	Unnecessarily verbose
Is the primary	

claimed contribution clearly reflected in the theoretical and/or experimental sections of the paper	Yes
Detailed Comments [Please be as specific as possible]	The problem formulation of Hierarchically Decomposable Concepts seems somewhat innovative, although its relationship to the established multi-label formulation is not discussed. The theoretical results are interesting but very limited and need some clarification, the presentation of the approach needs improvement. The introduction and problem definition are overall well written, although contextualization could be improved (see below). The theoretical results section needs some clarification. Proposition 2 seems to hold only if each box is a separate subconcept, which needs to be stated in the proposition, as it is an important limitation. I suggest to formally define the disjoint k-intervals problem before proposition 3 (it is explained informally in the text, though). In the proof of proposition 3, I am not convinced about the worst-case $\Omega(m_c)$. What is the argument against using binary search here like in the later proof of proposition 4? In section 3.2, when writing "it is similar to an active learning model" (and in the subsequent sentences), it should be stated more precisely that "it is similar to an pool-based active learning model", as in query synthesis the actively requested instances need not to be from a pre-specified set. I suggest to extend the discussion on proposition 5 and its the relation to results from [11]. The approach section is not described in sufficient detail. For example, what is the probabilistic estimate in the output of $C_{i,j}$ referring to, the probability of class j against all other (sub)classes, the conditional probability for the arc j in the parent node $i-1$ of the label tree, or something else? The definition and example are somewhat contradictory: in the example "building" of figure 1, the probability

	"that example x is positive for class j at level i" (if it refers to the probability P(Location, Building, Museum|x)) is different from the probability that instance x is a "building" (if the latter is to refer to P(Museum|x)). Furthermore, given the prominence of the term "active" in the paper's title and abstract, the active selection approach is very limited (the single paragraph on this component describes a confidence-based uncertainty sampling approach). Contextualization of the research question The authors should discuss the relationship between their Hierarchically Decomposable Concepts and established literature on multi-class and multi-label problems. In their example labeling tree, museum is a label in both the building and the attraction concept. What is the argument against a multi-label formulation? The authors should discuss the relation to ensemble approaches, in particular to those that manipulate output targets for diversity, and to active learning ensembles (Query-by-Committee approaches). The relationship to Query-by-Committee approaches for Error-Correcting Output Codes should be discussed and compared.
Overall Rating	Reject

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
The paper is relevant to SDM	Yes
Which of the following is the most accurate summary of the type of the work proposed in the paper	A new or improved algorithm, data structure or analysis for a known problem or a small variation of a known problem
Which of the following is the best characterization of the paper contribution	Incremental work [idea] but solid execution

Is the abstract an accurate summary of the paper	Yes
Does the paper properly contextualize the "Research Question" [e.g., through the Introduction and or Related Work Section]	Yes
Which of the following is the most accurate description of the writing style of the paper	Sharp and Precise
Is the primary claimed contribution clearly reflected in the theoretical and/or experimental sections of the paper	Yes
Detailed Comments [Please be as specific as	This paper is very well written. It clearly states its hypothesis and proceeds to solid theoretical and empirical backing of the hypothesis. Results are good. Overall the paper is of high quality and my main considerations are related primarily to its impact and secondarily on its limited novelty. The paper focuses on improving the performance on coarse-level labels. However, in most applications finer-grain labels are more challenging and building systems that can predict the finer-grain labels are more useful than systems that can predict coarse-grain labels only. Therefore improvements in coarse-grain labels are of less impact. Another consideration is that the main contribution is combining child label classifiers to predict the parent label. This has been discussed in the past to some extend, e.g. in the True-Path-Rule line of papers by Giorgio Valentini. So, novelty has some limitations. There is also some related work on subclass discovery

possible]	and exploitation, which is a bit different, but with the same (and more difficult - hierarchy is not given) goal: S. Escalera, D. M. Tax, O. Pujol, P. Radeva, and R. P. Duin, "Subclass problem-dependent design for error-correcting output codes," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 6, pp. 1041–1054, Jun. 2008. Then, there is the contribution of using active learning (AL) on the finer-grain labels, but since using the finer-grain labels leads to better results, then doing AL on the finer labels is to some extended expected to help. The discussion of the varying costs at the different labels is really interesting and elegant and deals successfully with a potential negative point of the approach.
Overall Rating	Neutral

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
The paper is relevant to SDM	Yes
Which of the following is the most accurate summary of the type of the work proposed in the paper	A new or improved algorithm, data structure or analysis for a known problem or a small variation of a known problem
Which of the following is the best characterization of the paper contribution	Incremental work [idea] and weak execution
Is the abstract an accurate summary of the paper	Yes
Does the paper properly contextualize the "Research Question" [e.g., through the	Yes

Introduction and or Related Work Section]	
Which of the following is the most accurate description of the writing style of the paper	Unnecessarily verbose
Is the primary claimed contribution clearly reflected in the theoretical and/or experimental sections of the paper	No
	Summary --- The paper looks at the problem of solving classification problems by decomposing classes into finer subclasses and using these subclasses to form sub-learning problems. The solutions to these subproblems are them combined to solve the original problem. The candidate applications are multi-classification tasks where the classes lie in a hierarchy. The algorithm proposed passively samples labels at various levels of the hierarchy, based on the uncertainty in the algorithm's current prediction for them. Classifiers are trained for each label at each level of the tree and the final output is obtained by a max-pooling style operation. The paper has little to no novelty to offer in terms of theoretical results. Proofs are written rather shoddily. Experimental work is there but is not compared to any other algorithm. Algorithms for multi-label learning form clear competitors for this approach since they can also exploit label correlations. I did not find the experimental work up to the mark which is important since the paper does not have any significant theoretical contributions. There are actually several papers that use similar tricks to address intra class diversity. This has been used in object detection as well but with limited success.

Detailed Comments [Please be as specific as possible]	A statistical approach to texture classification from single images. International Journal of Computer Vision, vol. 62, pp. 61–81, 2005. Group-sensitive multiple kernel learning for object recognition. IEEE Trans Image Process. 2012 May;21(5):2838-52. Multimodal Object Recognition Using Random Clustering Trees. Volume 9117, Lecture Notes in Computer Science pp 496-504. Comments ---------- 1) The work of Blumer et al *defines* PAC learnability in a certain way. It is not an if and only if result the way the proof of Proposition 2 makes it look like. 2) Proposition 4 is a simple extension of the toy problem used to illustrate the benefits of active learning. 3) I am not sure if the proof of Proposition 1 qualifies as a proof since it simply appeals to a result of Blumer et al which negates the possibility of finding the minimum cover using axis aligned rectangles. This negates consistent learning but not PAC learnability which only requires an approximate cover to be learnt and that too with some confidence. 4) Proposition 3 is not proved properly - the term "worst case" is used without even defining what class of learning algorithms are being used. Indeed, if the distribution over the entire instance space is uniform then constantly many queries would yield the location of a single negative example between two adjacent positive intervals after which the binary search algorithm would yield the boundaries in logarithmically many more queries. 5) The function Purchase is not defined. What is the cost of purchasing labels at level i? How does this affect total cost? Given a budget B, how does this affect the performance of the learnt classifier? 6) Given the "uncertainty" of various points, are instances chosen deterministically or randomly?
Overall Rating	Reject