CIKM 2015
The 24th ACM International Conference on Information and Knowledge Management
19-23 October 2015, Melbourne, Australia

Reviews For Paper

Track	Knowledge Management
Paper ID	1085
Title	Clustering-based Active Learning on Sensor Type Classification in Buildings

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Is the submission relevant to the KM track?	Yes
What do you think of the ideas in the submission?	Standard
Is the writing clear?	Yes
Overall recommendation	-4: Should reject
How to improve the submission (for the camera-ready, if accepted or for resubmission to another venue, if rejected)? List your three suggestions.	W1. Novelty is limited: adaptive clustering and label propagation have been used before in active learning. W2. Experimental evaluation is incomplete and unfair. W3. The actual sensor data are not utilized in the classification.
	The authors propose an active learning algorithm to categorize sensors in buildings based on their type. They perform clustering of the unlabeled instances during active learning. The main assumption is that instances in the same cluster tend to share the same class label and thus the querying of similar instances is avoided. The acquired labels are propagated to their unlabeled neighbors within the cluster, which are also used in the training process. Experiments on real data show that the proposed technique can achieve high classification accuracy using less labeled examples in comparison with existing active learning algorithms. The paper is well-written and has nice flow. The authors explain the problem and the method in a good way. However, there are several problems: 1. The adaptive clustering strategy and label propagation are highlighted as the main contributions of the present work. However, it is not the first time that these techniques are used in active learning. [Dasgupta et. al., 2008] also performed adaptive hierarchical clustering on unlabeled data, where the number and the size of clusters are refined in every iteration. [Zhu et al., 2003] after querying some random points and building a neighborhood graph, spread the obtained labels to the nearest neighbors. Label propagation has been also used in [Dasgupta et. al., 2008], where

Detailed review	all examples inside a cluster are assigned the majority label, and then the fully labeled dataset is used to build a classifier. The above-mentioned papers are cited by the authors, but it's not clear how the present work is novel. It seems that the only new aspect is the utilization of an adaptive distance threshold for label propagation, which constitutes very minor novelty. 2. In the experimental evaluation while a linear SVM model is used as a classifier for all the active learning algorithms under consideration, a logistic regression model is used as a classifier for the Least Margin (LM) method. This raises serious concerns about the superiority of the proposed method since LM is the only baseline algorithm that goes head to head with the proposed method in terms of classification effectiveness. Either the authors should use the same classification model for all the methods under consideration or they should explain their choice. Furthermore, they compared with Pre-clustering (PC) method. However, they used Gaussian Mixture Model (GMM) for clustering and the original PC paper suggests using k-mediods for that. GMM for PC is problematic as the authors of this paper have shown in their classification results section. 3. There is no comparison with active learning algorithms that use adaptive clustering and label propagation (e.g., [Dasgupta et. al., 2008]). The authors compare their method with an active learning algorithm that performs static clustering of the instances, but they do not compare with active learning approaches that use adaptive clustering and label propagation. 4. For applications related to sensors, it would be more interesting and informative to use the actual data of the streams and not only the text features. For example, one would expect different readings from a temperature and a humidity sensor; the actual readings could distinguish the sensor type better.

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Is the submission relevant to the KM track?	Yes
What do you think of the ideas in the submission?	Some novelty
Is the writing clear?	Yes
Overall recommendation	+3: Should accept
How to improve the submission (for the camera-ready, if accepted or for resubmission to another venue, if rejected)? List your three	1) The use of k-mers for this problem is a nice innovation and should be moved to the earlier parts of the paper. 2) Sec. 3.4 claims that the classification algorithm proposed is of general use. To substantiate this claim, some set of experiments would be nice, and it would strengthen the paper overall.

suggestions.	
Detailed review	The problem treated in this paper is that of sensor classification by an active learing approach: an algorithm is proposed whereby a human operator provides training data as demanded from the algorithm (the sensor class) and a guides a clustering process based on the similarity of sensor "point names", or tags. The combination of active learning via clustering in this paper is combined with label propagation in a more-or-less standard way, although admittedly, I cannot think of a citation. A nice addition to the paper is the use of k-mers to construct the feature space from the (often rather cryptic and esoteric) sensor tags. This is a concept which I believe is not exploited enough outside of bioinformatics and I liked its use in this context. Overall, although the paper does not contain a major breakthrough, it is a simple and elegant combination of known techniques yielding a solution both simple and practical, and I believe it should be published.

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Is the submission relevant to the KM track?	Yes
What do you think of the ideas in the submission?	Original
Is the writing clear?	Yes
Overall recommendation	+6: Must accept
How to improve the submission (for the camera-ready, if accepted or for resubmission to another venue, if rejected)? List your three suggestions.	1. It would be great if the authors could introduce a simple running example that takes the reader through the intuition of the idea (rather than the details) at each step of the normalization process. 2. In the experiments section please plot the ROC curves instead of plotting only the accuracy. Also increase the size of the axis labels and the figure itself. One has to squint to see them clearly. 2. It would be nice if the authors drew a picture summarizing the whole process described in section 3.
	This paper address an important problem of metadata normalization in large-scale sensor installations. Day-by-day sensors are getting inexpensive, therefore it is possible to install them in large number. But the installation process of these sensors is largely manual. A sensor, in the installation process, is given a name/label by the technician who installs the sensor. There are well-established standards for the naming convention, but they are rarely followed in practice. This leads to an information integration problem. We have a large number of

Detailed review	arbitrarily named sensors and the task is to find groups of sensors that sense the same attribute. For eg., some would call a temperature sensor "ART" (ambient temperature sensor) or RMT (room temperature sensor). Both of these sensors do the same task of sensing temperature and therefore should be grouped together and should be named by a uniform tag. I liked the problem itself, more than the solution to the problem. I believe the solution could be simplified than its current form. The problem of meta-data normalization is one of core issues in large-scale sensor installations that has been largely overlooked. Until this issue is addressed further analytics cannot be executed because of this semantic discrepancy in sensor tags. This is a very well-written paper, with the problem and the solution clearly explained. Also the experimental evaluation is fairly complete. In the experiments section the various algorithms are not compared correctly. The accuracy is plotted against the number of labels. The appropriate method would be to plot the ROC curve, where each point on the curve shows the false positive rate and the true positive rate at a given number of labels.