ACTIVE LEARNING WITH LABEL PROPORTIONS

Rafael Poyiadzis, Raul Santos-Rodriguez, Niall Twomey

Intelligent Systems Laboratory, University of Bristol, UK

ABSTRACT

Active Learning (AL) refers to the setting where the *learner* has the ability to perform queries to an *oracle* to acquire the *true* label of an instance or, sometimes, a set of instances. Even though Active Learning has been studied extensively, the setting is usually restricted to assume that the oracle is *trustworthy* and will provide the actual label. We argue that, while common, this approach can be made more flexible to account for different forms of supervision. In this paper, we propose a new framework that allows the algorithm to request the label for a *bag of samples* at a time. Although this label will come in the form of proportions of class labels in the bags and therefore encode less information, we demonstrate that we can still learn effectively.

***Index Terms*—** Active learning, label proportions, label propagation

1. INTRODUCTION

It is often the case in Machine Learning that gathering labelled data is costly, either due to time constraints, limited resources or simply because of the complexity of the task. These have motivated research in learning from both labelled and unlabelled data within the semi-supervised learning paradigm. Active Learning can be seen as an extension of this framework, where the learner can perform queries to an oracle. These queries usually take the form of $\mathcal{Q}(\mathbf{x}_i)$, to which the oracle returns the label y_i corresponding to the instance $\mathbf{x}_i$. AL is based on the premise that if we can adequately select which samples to label, we can achieve better performance than otherwise. AL has been successful in many domains, ranging from image [1, 2] to text classification [3, 4]. However, the standard AL paradigm is not suitable for tasks that require extra degrees of flexibility. We depict two examples below.

Automatic Activity Recognition in Smart Home environments is essential for the monitoring of chronic health conditions. The task involves the classification of sensor data as belonging to a set of Activities of Daily Living. This is usually addressed in the supervised setting, assuming that an annotator (patient) has manually labeled enough examples. However, this approach is not realistic when deploying these systems in the wild, as patients are reluctant to provide detailed labels and an AL approach is more amenable [5]. This is still too strong a requirement in many cases. In this paper, we explore a novel annotation strategy where label aggregates are requested on demand over particularly relevant bags or groups of samples (e.g., 'What was the proportion of time spent in each of the target activities yesterday?').

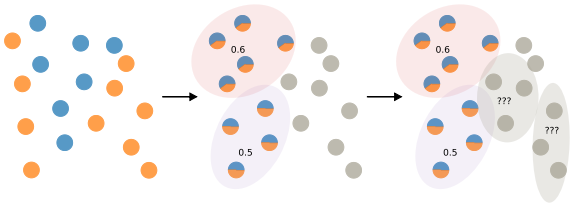


Fig. 1. Illustration of active learning with label proportions.

Similarly, land use classification from satellite images is also usually addressed using AL [6]. This involves the classification of pixels as belonging to either of the available categories (e.g., residential, commercial, vegetation, soil). Annotation is an arduous task as it requires labelling the individual pixels in each image. We argue that visually reporting on the aggregated label of a bag of pixels would greatly simplify the labelling procedure.

Both of these tasks would fall within the so-called *batch-mode* or top-k AL that proposes queries for sets of instances at a time [7]. In this case, queries take the form $\mathcal{Q}(\boldsymbol{x}_{[k]})$, to which the oracle responds with the set of labels $\boldsymbol{y}_{[k]}$. Differently, in our approach, we expect a single annotation consisting of the proportion of the classes within the bag to characterize the whole bag. This bag-level label provides a weak supervision with respect to the instance-level labels. Although this relaxes the assumption of infallibility of the oracle, we argue that exploring varying degrees of supervision can lead to an easier and simpler interaction in between the algorithm and the oracle. In this paper, we cast our problem as an instance of the *Learning from Label Proportions* (LLP) setting [8, 9, 10, 11, 12].

Figure 1 illustrates this idea. The sub-figure on the left depicts the (unavailable) ground truth data with the class membership depicted by the colour of the circles. The middle im-

Supported by EPSRC (SPHERE EP/R005273/1) and the MRC Momentum award (CUBOID MC/PC/16029).

age shows two bags of instances, that constitute the training set, along with the proportion of positive instances in each bag, also shown with a pie-chart on the instances to depict uncertainty on the labels. The final image presents two candidate query bags in grey ellipses. It should be noted again that the labels received will be bag-level, describing the class proportions within the bag, and not instance-level.

Problem Formulation We assume we have access to a training set given by $\mathcal{L} = \{(\boldsymbol{x}_i, z_i)\}_{i=1}^{N} \in (\mathcal{X} \times \mathcal{Z})$, where $\mathcal{X} = \mathbb{R}^d$ denotes the feature space and $\mathcal{Z} = \{1, 2, \cdots, K\}$ denotes the discrete space of the bag assignments and where K is the number of bags. We also have access to an un-bagged[1] (and also unlabelled) training set denoted by $\mathcal{U} = \{\boldsymbol{x}_i\}_{i=1}^{M}$. We assume that the true label $y_i \in \mathcal{Y}$, where $\mathcal{Y} = \{-1, +1\}$, exists, but is not provided. For example, $z_i = k$, would mean that the pair $(x_i, z_i) \in \boldsymbol{B}_k$, where $\boldsymbol{B}_k$ denotes the set of points belonging to the k-th bag. We also assume that bags are non-overlapping, that is $\boldsymbol{B}_j \cap \boldsymbol{B}_k = \emptyset, \forall j, k \in [C]$ and also that $\cup_{j=1}^{C} \boldsymbol{B}_j = \mathcal{L}$, and that $\mathcal{U}$ is not preset to bags. Moreover, the class proportions of each bag are available. For the binary classification problem, we are provided with the class-proportion (per bag) matrix:

$$\boldsymbol{\Pi}_{LLP} = \begin{pmatrix} \pi_1 & 1-\pi_1 \\ \pi_2 & 1-\pi_2 \\ \pi_3 & 1-\pi_3 \end{pmatrix}$$

where π_1 is understood as the proportion of positive instances, in the first bag ($\boldsymbol{B}_1$). Furthermore, the learner can perform queries $\mathcal{Q}(\boldsymbol{x}_{[k]})$ to an oracle, and whose response takes the form of the bag proportions π_{c+1} for the selected set of points. In the case of querying just one point, $k = 1$, the response is the true label. It should be noted that we do not consider relabelling, and therefore for a query, $\mathcal{Q}(\boldsymbol{x}_{[k]})$, to be valid we need, $\boldsymbol{x}_{[k]} \in \mathcal{U}$.

The paper is structured as follows. First, we review the related work in Section 2. Then, we present general framework for active learning from label proportions and introduce a novel algorithm in Section 3. Section 4 describes the experimental work. We conclude in Section 5.

2. RELATED WORK

Active Learning comes primarily in two settings, the *offline* (also called *pool-based*) and the *online*. In the former, the practitioner already has access to a set $\mathcal{U} = \{\boldsymbol{x}_i\}_{i=1}^{M}$ and can select from within it which samples to query, while in the later the instances $\boldsymbol{x}_i$ come in the form of a stream and the decision of whether to query a particular sample, or not, needs to be made *in situ*. A second distinction comes from whether queries can, or even *have to*, depend on multiple data-points, $\mathcal{Q}(\boldsymbol{x}_{[k]}) = \mathcal{Q}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$, rather than just one. This is sometimes called top-k Active Learning. The setting arises either from the user's side in order to speed-up the process (as it might be costly to update the model after every query), or it might come in the form of a restriction from the oracle. In this paper we focus on pool-based top-k active learning, but refer the interested reader to [13] and the references therein for a comprehensive survey. The advantage of the pool-based setting over the on-line version is that the learner can rank the data points in the unlabelled training set and select the most informative one(s). The objective is usually to reduce the uncertainty of the model in general or over a specific region of the space. Approaches to Active Learning are usually closely related to the idea of Uncertainty Sampling [14], which instructs the learner to query the instances it finds more difficult to label. This could be derived from a purely intuitive point of view, by considering expected maximum model change or by considering reducing the set of all hypotheses compatible with the training set so far [7].

We now review previous works that have considered the notion of an imperfect oracle. These usually model the oracle as providing the wrong label with a certain probability. In certain cases this probability is uniform across the input space, while in others it depends on the instance itself. In [15] the authors propose *Proactive Learning*, a framework that allows for possibly more than one oracle and that these do not always respond or are not always correct and where the cost of a label might change. Similarly, in [16] the oracle is a human annotator whose error rate is example-dependent. In [17], the authors explore constrained (also viewed as weak supervision) spectral clustering that aims to identify the noisy constraints through inconsistency with the remaining constraints. Finally, [18] presents a new framework where the oracle on top of the label of the queried instance, also returns its confidence on the prediction. A different line of work that also considers imperfect or weak labels is re-labelling [19]. In [20] the authors propose *Re-Active Learning*, as an extension of the AL framework, that also allows for querying data points that are already labelled. Their approach seeks the point that will have the most impact on the model. While a few works in the literature consider the noisy nature of the oracle, to the best of our knowledge, this is the first attempt to study the AL setting with an oracle whose responses take the form of label proportions. We will refer to such an oracle as an LLP-oracle.

3. ACTIVE LEARNING WITH BAG PROPORTIONS

In this section we present our algorithm for performing Active Learning with an LLP-oracle. Our approach is based on the method of Label Propagation (LP) which was developed for semi-supervised learning [21] and which was later adopted to cope with the Label Proportions level of supervision in [11]. We start with a very brief overview of LP, then proceed to

[1] We use this terminology for the equivalent of unlabelled data in Semi-supervised Learning. It refers to those points that do not belong to a bag yet, and therefore we do not have any knowledge about them, with regards to their true label.

present how it is adopted to learn with label proportions and finally how it could be employed within the Active Learning setting.

It should be noted that, similarly to how practically any learner capable of passive learning is also capable of active learning, any learner capable of learning with label proportion, could in principle be used for active learning with label proportions.

3.1. Label Propagation

Label Propagation solves the following constrained optimization problem:

$$Q(\boldsymbol{f}) = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}(f_i - f_j)^2 + \gamma\sum_{i=1}^{n}(f_i - y_i)^2 \quad (1)$$

where the first term encourages local smoothness while the second penalizes deviation from $\boldsymbol{y}$ and the balance between the two is controlled with γ. The terms s_{ij} are elements of a similarity matrix defined below. The solution of $\arg\min_{\boldsymbol{f}} Q(\boldsymbol{f})$ can be shown to give a solution of the form $\boldsymbol{f}^* = (\boldsymbol{I} - \alpha\boldsymbol{S})^{-1}\boldsymbol{y}$ [21], where $\boldsymbol{S} = \boldsymbol{D}^{-1}\boldsymbol{W}$ with $\boldsymbol{W}$ being a similarity matrix and $\boldsymbol{D}$ being a diagonal matrix, with the entry $\boldsymbol{D}_{ii}$ denoting the sum of the elements of the i-th row of $\boldsymbol{W}$. A popular choice, and the one we use in this paper, is $\boldsymbol{W}_{ij} = exp(-\sigma||\boldsymbol{x}_i - \boldsymbol{x}_j||^2)$, with $\sigma > 0$.

While it is usually presented as the solution of an iterative procedure, the look at the problem from a regularization perspective [21] will allow us to introduce the bag proportions as constraints in a principled manner.

3.2. Label Propagation with Label Proportions (LPLLP)

In the LLP setting *no label* is provided for any of the points. What LPLLP exploits is the idea that even if we do not actually know the true label of $\boldsymbol{x}_j$, we have access to its bag's label proportions, which we cast as prior probability of being assigned to a class.

In the binary classification case, our $\hat{\boldsymbol{y}}$ can now be defined as $\hat{y}_i = \frac{\pi_k - 0.50}{0.50}$, where $\boldsymbol{x}_i \in \mathbf{B}_k$ and π_k represents the proportion of positive labels (1) in bag k ($(1-\pi_k)$ is equivalently the proportion of negative labels (0) in the same bag). One could now compute $\boldsymbol{f}^* = (\boldsymbol{I} - \alpha\boldsymbol{S})^{-1}\hat{\boldsymbol{y}}$, but trivial decision making on $\hat{y}$ does not guarantee preservation of the class proportions.

In its original form, the problem would be an Integer Program, $\arg\min_{\boldsymbol{f}\in\{-1,1\}^N} Q(\boldsymbol{f})$, which is in general intractable. Building on this, one could enforce bag constraints through a system of linear equations $\boldsymbol{A}(\frac{\boldsymbol{f}+\mathbf{1}}{2}) = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{K\times N}$, with K being the number of bags and $\boldsymbol{b} \in \mathbb{R}^K$. $\boldsymbol{A}$ is defined as $\boldsymbol{A}_{ki} = 1$, if $\boldsymbol{x}_i \in \boldsymbol{B}_k$ and 0 otherwise, and $\boldsymbol{b}_k = \boldsymbol{N}_{k,1}$, where $\boldsymbol{N}_{k,c}$ corresponds to the number of instances of class c in $\boldsymbol{B}_k$.

We now have a constrained Integer Program and proceed by relaxing it to a constrained Linear Program. The constraint would then be $\boldsymbol{f} \in [-1,1]^N$ instead, giving our final problem formulation:

$$\boldsymbol{f}^* = \arg\min_{\boldsymbol{f}\in[-1,1]^N} \boldsymbol{Q}(f)$$
$$s.t.\ \boldsymbol{A}\left(\frac{\boldsymbol{f}+\mathbf{1}}{2}\right) = \boldsymbol{b} \quad (2)$$

The proposed algorithm solves Eq. 2 in two steps, by first solving the unconstrained problem $\boldsymbol{f} = (\boldsymbol{I} - \alpha\boldsymbol{S})^{-1}\hat{\boldsymbol{y}}$ (with $\hat{y}_i = (\frac{\pi_k - 0.50}{0.50})$ where $\boldsymbol{x}_i \in \mathbf{B}_k$), and then normalizing the predictions, $\boldsymbol{f}^*$, for the instances of each bag such that their sum is equal to the provided bag proportions, such that $\boldsymbol{A}\left(\frac{\boldsymbol{f}^*+\mathbf{1}}{2}\right) = \boldsymbol{b}$. These two steps are repeated until convergence, with $\boldsymbol{f}^{(t+1)} = (\boldsymbol{I} - \alpha\boldsymbol{S})^{-1}\boldsymbol{f}^{(t)}$, and $\boldsymbol{f}^{(0)} = \hat{\boldsymbol{y}}$. The procedure for solving the optimization problem in Equation 2 is depicted in Algorithm 1.

Algorithm 1: LPLLP (binary case)

Input : Bag assignment matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$ with $\boldsymbol{b}_k = \boldsymbol{N}_{k,1}$

Output: Estimated labels $\hat{\boldsymbol{f}}$

1. Compute similarity matrix $\boldsymbol{W}$ and then $\boldsymbol{S} = \boldsymbol{D}^{-1}\boldsymbol{W}$
2. Compute $\boldsymbol{f}^{(t+1)} = (\boldsymbol{I} - \alpha\boldsymbol{S})^{-1}\boldsymbol{f}^{(t)}$, where $\boldsymbol{f}^{(0)}$ is defined as: $f_i^{(0)} = \frac{\pi_k - 0.50}{0.50}$ for $\boldsymbol{x}_i \in \boldsymbol{B}_k, \forall i$.
3. Solve for $f^{(t+1)^*}$ using Alternating Projections.
4. Repeat steps (2) and (3) until convergence.
5. Estimate labels based on $\hat{f}_i = sgn(f_i^*)$

3.3. Active Learning with LPLLP

In this section we describe our methodology for performing Active Learning in the LLP setting. It should be noted that the set of unlabelled data $\mathcal{U}$ is not already grouped, i.e. it is not a problem of having to choose between pre-determined bags of points, but rather a problem of constructing the bag of points to query for.

Motivated by getting as much information as possible out of the noisy label returned by the oracle, our first approach aims at querying for a *pure bag*. We use the term 'pure' for bags whose instances belong to the same class. If we were in a position to do that, then we would be back in the batch-mode Active Learning setting, where our one label describing the class proportions of our bag, would actually be adequate to describe the labels for all the instances in the bag. Unfortunately, this is not possible, since we do not have access to the true labels and therefore we need to resort to heuristics. These are based on the assumption of local smoothness and exploiting the global structure of the data; that is, points that are close to each other, or lie on the same structure, are likely to be labelled similarly. Acquiring a pure bag would be beneficial but we would also want to maximize the information gain, with respect to selecting instances that would also reduce the uncertainty of the model.

The way we solve this double-objective optimization problem is by first selecting an instance based on the notion of Uncertainty Sampling, and then building up a neighbourhood around it until the desired bag size is satisfied. The neighbourhood is built by making use of the matrix $\boldsymbol{L} = (\boldsymbol{I} - \alpha \boldsymbol{S})^{-1}$ from step (2) in Algorithm 1. (This matrix is already computed from training the model.) We will also use the notation $\boldsymbol{L}_{\mathcal{U}}$ to imply the sub-matrix indexed by the instances belonging to $\mathcal{U}$.

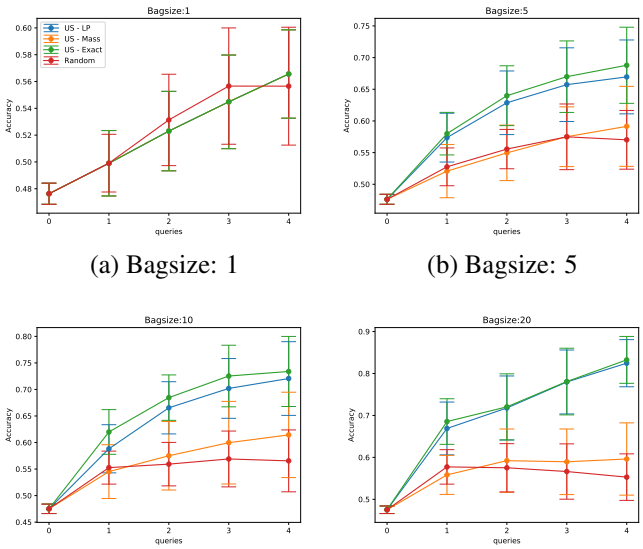
The procedure for solving this problem, for a bag of size k is summarized as follows. Using the predictions generated in Algorithm 1 find the one closest to the boundary. Using the matrix $\boldsymbol{L}_{\mathcal{U}}$ (also a product of Algorithm 1), find the $k-1$ points closest to the first. We will refer to this approach as `US-Mass`.

Our second approach ignores the nature of the oracle and just queries for the k most uncertain points. We will refer to this approach as `US-LP`. To avoid any confusion, it should be stated that while this approach is merely Uncertainty Sampling and the nature of the oracle is not taken directly into account, the learner should still be able to learn from label proportions, as this will be the only supervision.

4. EXPERIMENTS

The accelerometer data from the HAR Dataset [22] was used for our empirical evaluation. This comprises of a smart phone on the waist with six annotated activities: walking, walking up stairs, walking down stairs, sitting, standing and lying down. The acceleration was sampled at 50 Hz on tri-axial accelerometers, and statistical features extracted from a 5 second sliding window following [23, Sec 3.3.1]. In our experiments we transform the problem of activity recognition to a binary classification task by augmenting the first three targets to 'walking' and last three targets to 'not moving'.

In the experiments we start with an initial training set consisting of two bags, each of size 16, with bag proportions $[(0.75, 0.25), (0.25, 0.75)]$, while the rest remain unlabelled and do not belong to a bag. We have considered four different sizes of bags: 1, 5, 10, 20, and for each we perform four queries. It should be noted that in the case of the bag being of size 1, the oracle returns the true label. For the experiments we fix parameter α to 0.50 (see Algorithm 1 step 2) and choose σ (for $\boldsymbol{W}$, with $\boldsymbol{W}_{ij} = exp(-\sigma||\boldsymbol{x}_i - \boldsymbol{x}_j||^2)$) by running the algorithm on a grid of values and then choosing based on the heuristic of highest score in terms of $\boldsymbol{f}^T \boldsymbol{S} \boldsymbol{f}$. We consider the two approaches mentioned in the previous section; `US-Mass` and `US-LP`, and two baselines. The first one is based on random sampling where the bags are formed at random and the label returned is in the form of bag proportions, and the second one is where the the oracle returns the *true* labels, that is in a query of size k, all k true labels will be returned. In Figure 2 we plot the classification accuracy over the test set, averaged over the 21 subjects we consider and how it changes over time.



(a) Bagsize: 1 (b) Bagsize: 5

(c) Bagsize: 10 (d) Bagsize: 20

Fig. 2. Accuracy over time for different sizes of queried bags.

From Figure 2 we see that `US-Mass`, that aims to combine Uncertainty Sampling with querying for a pure bag does not perform significantly better than random sampling. A possible explanation for this is that if the instances within a bag are already well-packed then we have an issue with redundancy. In other words, had fewer labels been queried for the same amount of information would be provided, as the algorithm (LPLLP) already exploits the local smoothness and structure of the data. On the other hand, we see that while `US-LP` is only based on label proportions, it performs better than `US-Mass` and compares favourably against `US-Exact`, that is provided with the true labels.

These findings serve as justification for revising the initial intuition of going for a pure bag, but instead these suggest going for uncertain points only, since the learner is capable of learning effectively with this level of supervision. In future work, we wish to explore the performance of algorithms from the batch-mode literature that aim at reducing redundancy within the queried set. It would be interesting to see whether in these settings, where the instances are necessarily not close to each other, performance with an LLP-oracle is still comparable with an exact-oracle.

5. CONCLUSSION

In this paper we have extended the framework of Active Learning to the setting of querying for a group of points, at the same time, and in return the *oracle* would return the proportion of the classes in the bag. We believe that this is a promising direction of research, given the rising interest in the field of weak supervision.

6. REFERENCES

[1] Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu, "Batch mode active learning and its application to medical image classification," in *ICML, 2006*.

[2] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2017.

[3] Simon Tong and Daphne Koller, "Support vector machine active learning with applications to text classification," *JMLR, 2001*.

[4] Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen, "Effective multi-label active learning for text classification," in *SIGKDD*, 2009.

[5] Tom Diethe, Niall Twomey, and Peter Flach, "Active transfer learning for activity recognition," in *ESANN*, 2016.

[6] Claudio Persello and Lorenzo Bruzzone, "Active learning for domain adaptation in the supervised classification of remote sensing images," *IEEE IEEE Trans Geosci Remote Sens*, 2012.

[7] Jan Kremer, Kim Steenstrup Pedersen, and Christian Igel, "Active learning with support vector machines," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 4, pp. 313–326, 2014.

[8] Novi Quadrianto et al., "Estimating labels from label proportions," *JMLR*, vol. 10, pp. 2349–2374, 2009.

[9] Felix Yu, Dong Liu, Sanjiv Kumar, Jebara Tony, and Shih-Fu Chang, "$\propto$-svm for learning with label proportions," in *ICML*, 2013.

[10] Felix X Yu et al., "On learning from label proportions," *arXiv preprint arXiv:1402.5902*, 2014.

[11] Niall Twomey Rafael Poyiadzi, Raul Santos-Rodriguez, "Label propagation for learning with label proportions," in *IEEE MLSP*, 2018.

[12] Rafael Poyiadzi, Raul Santos-Rodriguez, and Tijl De Bie, "Ordinal label proportions," in *ECML-PKDD*, 2018.

[13] Burr Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[14] Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K. Tsou, "Active learning with sampling by uncertainty and density for word sense disambiguation and text classification," in *ICCL*, 2008.

[15] Pinar Donmez and Jaime G. Carbonell, "Proactive learning: cost-sensitive active learning with multiple imperfect oracles," in *CIKM*, 2008.

[16] Jun Du and Charles X Ling, "Active learning with human-like noisy oracle," in *ICDM*, 2010.

[17] Xiatian Zhu, Chen Change Loy, and Shaogang Gong, "Constrained clustering: Effective constraint propagation with imperfect oracles," in *ICDM*, 2013.

[18] Eileen A Ni and Charles X Ling, "Active learning with c-certainty," in *PAKDD*, 2012.

[19] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *SIGKDD*, 2008.

[20] Christopher H Lin, "Re-active learning: Active learning with relabeling.," in *AAAI, 2016*.

[21] Fei Wang and Changshui Zhang, "Label propagation through linear neighborhoods," *TKDE, 2008*.

[22] Davide Anguita et al., "A public domain dataset for human activity recognition using smartphones," in *ESANN*, 2013.

[23] Niall Twomey, Tom Diethe, Xenofon Fafoutis, Atis Elsts, Ryan McConville, Peter Flach, and Ian Craddock, "A comprehensive study of activity recognition using accelerometers," *Informatics*, vol. 5, no. 2, 2018.