Expert Systems With Applications 114 (2018) 578–587

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Regression with re-labeling for noisy data

Youngdoo Son[a], Seokho Kang[b,*]

[a] Department of Industrial and Systems Engineering, Dongguk University-Seoul, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, South Korea
[b] Department of Systems Management Engineering, Sungkyunkwan University, 2066 Seobu-ro, Jangan-gu, Suwon 16419, South Korea

ARTICLE INFO

Article history:
Received 13 October 2017
Revised 5 August 2018
Accepted 15 August 2018
Available online 17 August 2018

Keywords:
Active learning
Re-labeling
Exploration-refinement sampling
Regression

ABSTRACT

Active learning, which focuses on building an accurate prediction model with a reduced cost by actively querying which instances should be labeled for training, has been successfully employed in several real-world applications involving expensive labeling costs. Although most existing active learning strategies have focused on labeling unlabeled instances, it has been shown that improving the quality of previously annotated labels is also important when the annotator produces noisy labels. In this study, we propose a novel active learning framework for regression, which is effective for the scenarios with noisy annotators, by providing a new sampling strategy named exploration-refinement (ER) sampling. The ER sampling performs two main steps: exploration and refinement. The exploration step involves finding unlabeled instances to be labeled, and the refinement step seeks to improve the accuracy of already labeled instances. The experimental results on several benchmark datasets demonstrate the effectiveness of the ER sampling with statistical significance.

1. Introduction

Active learning is a huge branch of machine learning, in which a prediction model is constructed using actively querying by which instances are labeled for training (Settles, 2010). As the labeling cost is usually expensive, it is common to have just a few labeled and many unlabeled instances in real-world scenarios. Unlike semi-supervised or transductive learning, which learns prediction models with given fixed labeled and unlabeled instances, active learning selects some informative unlabeled instances interactively and queries an information source, such as the user or expert annotators, to obtain the labels of the selected instances. Accordingly, the prediction model is efficiently constructed using a smaller number of labeled instances, thus incurring a lower labeling cost. Active learning has played a critical role when the labeling cost is prohibitively expensive, such as for texts and videos that require human annotators (Lang, 1995; Settles & Craven, 2008; Zha et al., 2012; Zhu, 2005), for scientific and engineering results that are obtained from costly and time-consuming experiments (Flaherty, Arkin, & Jordan, 2006; Guo, Chen, Sun, & Lin, 2004; King et al., 2009; 2004; Liu, 2004), and for survey results often used in the field of social sciences.

Because the performance of a prediction model constructed by active learning depends strongly on the quality of labels, one practical concern is that the annotator is naturally noisy and thus the labels could be inaccurate. Recent studies have shown that obtaining the labels of an instance repeatedly, *i.e.*, re-labeling, provides the possibility of improving the prediction model because the noise in the labels can be canceled out by averaging over the multiple labels (Ipeirotis, Provost, Sheng, & Wang, 2014; Kääriäinen, 2006; Sheng, Provost, & Ipeirotis, 2008). Lin, Mausam, and Weld (2016) addressed the problem of finding an instance that should be labeled or re-labeled next to improve the model, and suggested a new sampling method that simultaneously considers both labeling and re-labeling. However, this method is only applicable to the classification task, and to the best of our knowledge, a combination of newly labeling unlabeled instances and re-labeling labeled instances for the regression task has not been studied.

In this study, we present a novel active learning framework for regression with a noisy annotator by proposing a new sampling strategy, exploration-refinement (*ER*) sampling, which incorporates both the new labeling of unlabeled instances and re-labeling of already labeled instances. In the *ER* sampling, the labeling is performed using two kinds of steps: *exploration* and *refinement*. The *exploration* step selects an unlabeled instance to be labeled next as in conventional active learning. In the *refinement* step, a labeled instance is selected instead of an unlabeled one and the selected instance is labeled again to improve the accuracy of its label. Experimental results on several benchmark datasets and the subsequent statistical tests demonstrate the effectiveness of the proposed method compared with conventional active learning which

* Corresponding author.
E-mail addresses: youngdoo@dongguk.edu (Y. Son), s.kang@skku.edu (S. Kang).

https://doi.org/10.1016/j.eswa.2018.08.032
0957-4174/

only considers *exploration* in terms of the accuracy with the same number of labeling steps.

The remainder of this paper is organized as follows. In Section 2, we review the related work concerning active learning. In Section 3, we describe the proposed *ER* sampling strategy for active learning with re-labeling for regression. Section 4 reports the experimental results. Finally, we provide concluding remarks and describe future research directions for this work in Section 5.

2. Related work

In this section, we briefly review the milestones of active learning strategies from early work to the state-of-the-art and the recent research efforts on active learning with noisy labels.

2.1. Active learning strategies

Active learning has been an important branch in machine learning. There have been proposed numerous active learning strategies for both classification and regression tasks.

Most previous studies have focused on the classification task. The most basic and widely employed strategy is uncertainty sampling, which queries the unlabeled instance that is the most uncertain. Regarding the binary classification task, Lewis and Catlett (1994) proposed a method for binary decision trees involving querying the instance whose posterior probability is closest to 0.5. Similarly, Fujii, Tokunaga, Inui, and Tanaka (1998) and Lindenbaum, Markovitch, and Rusakov (2004) implemented this strategy for nearest neighbor classification, where the posterior probability is decided by the voting of neighbors. Tong and Koller (2001) applied this strategy to support vector machines by querying the instance nearest to the decision boundary. For the multi-class classification task, Settles and Craven (2008) proposed querying the instance with the least confident prediction among the unlabeled instances. Scheffer, Decomain, and Wrobel (2001) considered the difference between the highest posterior probability and the second highest one. Hwa (2004) exploited the entropy of each unlabeled instance as the degree of uncertainty, in which the unlabeled instance with the highest entropy is queried. Other famous active learning strategies to be proposed include query-by-committee (Abe & Mamitsuka, 1998; McCallum & Nigam, 1998; Seung, Opper, & Sompolinsky, 1992), expected error reduction (Guo & Greiner, 2007; Moskovitch et al., 2007; Roy & McCallum, 2001; Zhu, Lafferty, & Ghahramani, 2003), and total expected variance minimization (Settles, 2010; Settles & Craven, 2008). The query-by-committee strategy employs multiple prediction models that are trained using the currently labeled instances and make predictions on every unlabeled instance. Then, the unlabeled instance with the most disagreeing predictions is queried. The expected error reduction strategy finds the unlabeled instance that will minimize the future generalization error. The total expected variance minimization strategy involves reducing the generalization error indirectly, and has been extensively applied with a closed-form solution (Cohn, 1996; Cohn, Ghahramani, & Jordan, 1996).

Active learning strategies have also been actively applied to the regression task. For the uncertainty sampling strategy, the prediction variance is usually considered as the degree of uncertainty, because it has monotone relationship with entropy-based uncertainty under the Gaussian assumption (Settles, 2010). The unlabeled instance with the largest prediction variance is queried. Seo, Wallat, Graepel, and Obermayer (2000) implemented this strategy for Gaussian process regression. Demir and Bruzzone (2014) considered the degrees of uncertainty and informativeness simultaneously for support vector regression by applying uncertainty sampling to the representative instances already chosen by clustering. In addition, Son and Lee (2016) suggested the use of both uncertainty and informativeness for relevance vector machine regression. The query-by-committee strategy can also be employed for the regression task by using regression values instead of class labels and querying the unlabeled instance with the most variable predictions (Burbidge, Rowland, & King, 2007; Douak et al., 2012; Pasolli, Melgani, Alajlan, & Bazi, 2012). The total expected variance minimization strategy has also been widely employed with several models including Gaussian process regression (Seo et al., 2000), artificial neural networks (MacKay, 1992), mixture of Gaussians, and locally weighted regression (Cohn et al., 1996). The expected model change strategy, which queries an unlabeled instance that maximizes the expected model change when it is labeled, is another strategy that has been recently studied for regression. Cai, Zhang, and Zhang (2017) proposed a method of querying unlabeled instances to maximize the expected change of the regression parameters. Ceperic, Gielen, and Baric (2012) and Ceperic, Gielen, and Baric (2014) suggested methods based on the expected error reduction scheme by querying an unlabeled instance with the largest prediction error for multi-kernel support vector regression and ε-support vector regression, respectively. Appice, Loglisci, and Malerba (2018) proposed a disagreement-based method for network regression problems, where an unlabeled node with the largest scarcity of the correlation matrix among the linked nodes is determined as the most disagreed instance and thus is selected to be labeled next.

2.2. Active learning with noisy labels

One of the latest trends in active learning research is the analysis on data with noisy labels due to the rise of crowdsourcing data. It has been studied in various ways including improving the label quality by cleansing noises or inferring the ground truths of labels (Karger, Oh, & Shah, 2013; Liu, Peng, & Ihler, 2012; Zhang, Chen, Tong, & Liu, 2015; Zhang, Sheng, Li, & Wu, 2018) and investigating the characteristics of data with noisy labels (Dawid & Skene, 1979; Kleindessner & Awasthi, 2018; Ma, Olshevsky, Szepesvari, & Saligrama, 2018).

There have been a number of studies on the active learning strategies considering noisy labels. Balcan, Beygelzimer, and Langford (2006) modified uncertainty sampling to be more robust to noisy annotation by identifying the uncertainty region with several instances. Zhao, Sukthankar, and Sukthankar (2011) proposed an uncertainty sampling method for support vector machines which is robust to noise by considering the data clusters simultaneously. An instance is more likely to be labeled if it is far from those labeled instances that are agreed with the cluster labels. Golovin, Krause, and Ray (2010) suggested a method of Bayesian active learning with noise labels based on the equivalence class determination and edge cutting. Donmez and Carbonell (2008) and Yan, Rosales, Fung, and Dy (2011) considered the situation that several imperfect annotators exist. They focused on the selection of the annotator in active learning. However, they did not take account of re-labeling for already labeled instances. Zhang and Chaudhuri (2015) studied the annotator selection method in active learning when two imperfect annotators with different costs are given. They constructed a classifier to determine a disagreed region of two annotators and used the high-cost annotator only for the disagreed region to minimize the total labeling cost.

There have been studied re-labeling approaches, which annotate already labeled instances repeatedly to reduce the effect of noises from annotators. Sheng et al. (2008) and Ipeirotis et al. (2014) mentioned the trade-off between re-labeling the existing instances and gathering new labeled instances, considering how to select the instance that should be re-labeled next.