Expert Systems With Applications 140 (2020) 112907

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A divide-and-conquer approach to geometric sampling for active learning

Xiaofeng Cao

Advanced Analytics Institute, University of Technology Sydney, Blackfriars St, Chippendale, NSW 2008, Australia

ARTICLE INFO

Article history:
Received 12 June 2019
Revised 9 August 2019
Accepted 30 August 2019
Available online 31 August 2019

Keywords:
Active learning
Uncertainty evaluation
Geometric sampling
Cluster boundary

ABSTRACT

Active learning (AL) repeatedly trains the classifier with the minimum labeling budget to improve the current classification model. The training process is usually supervised by an uncertainty evaluation strategy. However, the uncertainty evaluation always suffers from performance degeneration when the initial labeled set has insufficient labels. To completely eliminate the dependence on the uncertainty evaluation sampling in AL, this paper proposes a divide-and-conquer idea that directly transfers the AL sampling as the geometric sampling over the clusters. By dividing the points of the clusters into cluster boundary and core points, we theoretically discuss their margin distance and hypothesis relationship. With the advantages of cluster boundary points in the above two properties, we propose a Geometric Active Learning (GAL) algorithm by knight's tour. Experimental studies of the two reported experimental tasks including cluster boundary detection and AL classification show that the proposed GAL method significantly outperforms the state-of-the-art baselines.

1. Introduction

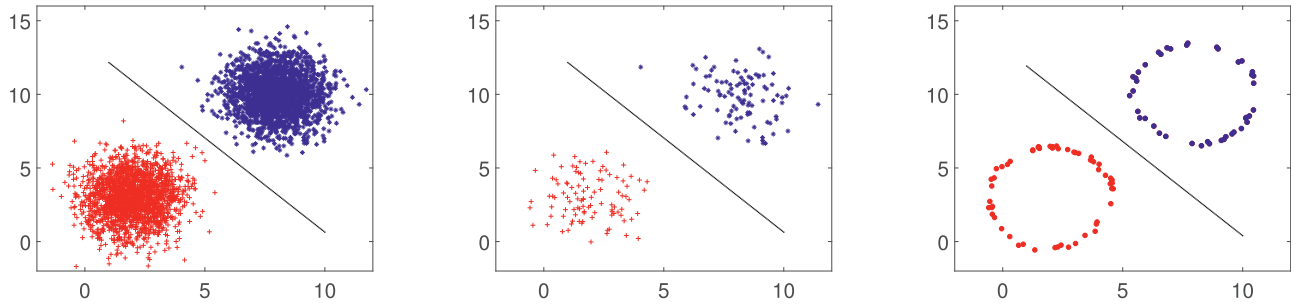
Active learning (Cohn, Atlas, & Ladner, 1994) is developed to further improve the prediction accuracy of the current classification model in supervised learning problems without sufficient labels. This study has been widely applied in various of learning scenarios when the unannotated data are abundant but annotating them is expensive and time-consuming, such as semi-supervised text classification (Hu, Mac Namee, & Delany, 2016), image annotation (Li, Shi, Liu, Hauptmann, & Xiong, 2012), transfer learning (Guo, Ding, Wang, & Jin, 2016), etc. Generally, the proposed AL algorithms focus on the construction of an uncertainty evaluation function which guides the subsequent sampling such as Lewis and Gale (1994) and Roy and McCallum (2001), etc. However, the label diversity and distribution features of the initial labeled set decide the performance of the uncertainty evaluation progress. When the initial labeled set only has a few data, performance degeneration of the subsequent sampling would be inevitable.

Geometric sampling shows its power in various of domains such as fast SVM training (Tsang, Kwok, & Cheung, 2005), Bayesian adversarial spheres algorithm (Bekasov & Murray, 2018), geometric deep learning (Fey, Eric Lenssen, Weichert, & Müller, 2018), etc. Especially in large scale classification issue, Core Vector Machine (CVM) (Tsang, Kwok, & Zurada, 2006) changed the SVM to a problem of minimum enclosing ball (MEB), which is popular in hard-margin support vector data description (SVDD) (Tax & Duin, 2004), and then iteratively calculated the ball center and radius in a $(1+\epsilon)$ approximation. In this process, the cluster boundary points located on the surface of each MEB are added into a special data collection called core sets. Trained by the detected core sets, the proposed CVM performed faster than the SVM and needed less support vectors. Especially in the Gaussian kernel, a fixed radius was used to simplify the MEB problem to the EB (Enclosing Ball), and accelerated the calculation process of the Ball Vector Machine (BVM) (Tsang, Kocsor, & Kwok, 2007). Without sophisticated heuristic searches in the kernel space, the training model, using points of high dimensional ball surface, can still be approximated to the optimal solution.

In this paper, we are motivated by the advantages of boundary points of CVM and propose a divide-and-conquer approach to geometric sampling for AL (see Fig. 1). Underlying MEB model, we divide the data of each class into two types: cluster boundary and core points. In geometric description, cluster boundary points are located at the surface of one cluster and core points are distributed inside the cluster. To study the properties of the two types of points, we compare them from two-fold: margin distance (w.r.t. Lemma 1) and hypothesis relationship (w.r.t. Lemma 2). The conclusion shows that cluster boundary points play more important role in the construction of the classification hyperplane compared to core points in a geometrical perspective.

E-mail address: xiaofeng.cao@student.uts.edu.au

https://doi.org/10.1016/j.eswa.2019.112907
0957-4174/



(a) Original data space (b) Cluster core points (c) Cluster boundary points

Fig. 1. Motivation of our active learning work. In each sub-figure, the black line denotes the generated SVM classification model based on the data points in the figure. (a) Training the original data space. (b) Training the cluster core points. (c) Training the cluster boundary points. We observe that the generated classification lines of (c) are similar to the models of (a) and (b).

Our conquer step is to obtain the cluster boundary points. By setting a knight in the geometric space, the path disagreement of the tour helps us to differ from cluster boundary and core points. We assume the tour path is decided by the update process of traversing 1 to k nearest neighbors (kNN) of the current tour position (data point). Their geometric disagreement in path length become the key of our detection method, i.e., the average tour path of boundary points are longer than that of the core points. With the above divide-and conquer analysis, we finally propose a Geometric Active Learning (GAL) algorithm by training the geometric cluster boundary points. The contributions of this paper are described as follows.

- We propose a divide-and-conquer idea to geometric AL sampling. It transfers the uncertain sampling space of AL into a set of the cluster boundary points.
- We provide the geometric insights for cooperating cluster boundary points in AL under the assumption of geometric classification.
- An AL algorithm termed GAL is developed in this paper. It samples independently without iteration and help from the labeled data.
- We break the theoretical curse of uncertainty evaluation sampling by GAL algorithm since it is neither a model-based nor label-based strategy with the fixed time and space complexities of $\mathcal{O}(NlogN)$ and $\mathcal{O}(N)$, respectively.
- A lot of experiments are conducted to verify that GAL can be applied in multi-class settings to overcome the binary classification limitation of many existing AL approaches.

The remainder of this paper is structured as follows. The related work is reported in Section 2. The preliminaries are described in Section 3 and the geometric insights on cluster boundary points in AL are presented in Section 4. The divide-and-conquer approach of knight's tour is presented in Section 5. The experiments and results are reported in Sections 6. The discussion is presented in Section 7. Finally, we conclude this paper in Section 8.

2. Related work

In this section, we present the related work on active learning and cluster boundary research.

2.1. Active learning

The learning goal of AL is to obtain a descried error rate by annotating as fewer queries as possible. To improve the performance of the current classification model, the AL learner (human expert) is allowed to pick up a subset from an unlabeled data pool. Those data, which may largely affect the subsequent update of the learning model, are the primary goals of the learner. As a policy, accessing the unlabeled data pool to sample and querying their true labels with a given budge are approved. However, all the learners would face an awkward and difficult situation: how to fast select the descried data from the massive unlabeled data in the pool.

To resolve the above challenges, uncertainty evaluation (Lewis & Gale, 1994) was proposed to guide AL by selecting the most informative or representative instances in a given sampling scheme or distribution assumption, such as margin (Tong & Koller, 2001), uncertainty probability (Roy & McCallum, 2001), maximum entropy (Melville & Mooney, 2004), confused votes by committee (Seung, Opper, & Sompolinsky, 1992), etc. For example, Tong and Koller (2001) proposes to select the data which is nearest to the current classification hyperplane, Roy and McCallum (2001) selects the data which can maximize the error rate change, Melville and Mooney (2004) selects the data with the maximum entropy of prediction probability, etc. Basically, these uncertainty-based AL algorithms aim to reduce the number of queries or converge the classifier quickly. Accompanied by multiple iterations, querying stops when the defined sampling number is met or a satisfactory model is found. It is thus these algorithms still need to traverse the whole data set repeatedly in this framework, although this technique performs well. However, they always suffer from one main limitation, that is, heuristically searching the whole data space to obtain the optimal sampling subset is impossible because of the unpredictable scale of the candidate set.

In practice, incorporating the unsupervised learning in the sampling process shows powerful advantages such as Nguyen and Smeulders (2004), Kang, Ryu, and Kwon (2004), and Urner, Wulff, and Ben-David (2013). It makes the learner solve the previous limitation be possible. One classical method (Dasgupta & Hsu, 2008) is performing the hierarchical clustering before sampling to improve th lower bound of the subsequent training performance. By setting up a probability condition, the learner is allowed to confidently annotate a number of subtrees with the label of the root note. When the clustering structure is perfect, it wold be positive for the sampling. However, an improper clustering results will mislead the annotation process. Then, performance degeneration of the subsequent sampling is inevitable.

2.2. Cluster boundary

Cluster boundary points are a set of special objects distributed in the margin regions of each cluster. Their labels are given by the cluster structure and guide the clustering partition. However, those label assignations are uncertain. Nowadays, the practical advantage of the cluster boundary has been widely used in the latent virus