February 2020

Nearest-neighbor based Manifold Expansion Technique for Active Learning

Anonymous Anonymous

Nearest-neighbor based Manifold Expansion Technique for Active Learning

Abstract

The present disclosure describes a nearest-neighbor based manifold expansion technique integrated into an active learner for seeking human review. Initially, the active learner performs a sampling formulation in which an unlabeled dataset, including unlabeled examples, is provided as an input to the active learner. The unlabeled dataset is then divided into seed datasets (i.e. a positive seed dataset and a negative seed dataset) and a test dataset. The positive seed dataset includes positive seeds, the negative seed dataset includes negative seeds and the test dataset includes test examples. In a voting process, each of the positive seeds and the negative seeds votes to the test examples that are in a neighborhood of the positive seed or the negative seed. A ranked list of the test examples is prepared based on an overall score for each test example accumulated by votes. Top-k examples in the ranked list are sent to annotators for review. The annotators assign labels (i.e. positive or negative) to the top-k examples. The annotators can interpret why a particular example got a particular score and how much the positive seeds and the negative seeds contributed to that score. The examples labeled by the annotators are added to the seed datasets. The voting process is executed again based on the updated seed datasets. This way, the voting process is executed continuously, and the ranked list is updated in an incremental manner in real time.

Problem statement

While solving a classification problem, it is often found that datapoints of certain classes are quite rare in a large dataset that has unlabeled datapoints in abundance. For example, in classification problems such as, anomaly detection, or fake data detection, the datapoints corresponding to anomalies or fake data first need to be found in order to remove these datapoints. However, the datapoints corresponding to the anomalies, or the fake data are likely to be lesser in concentration than genuine datapoints in the dataset. Thus, annotating each and every datapoint in the large dataset for finding the datapoints corresponding to rare classes costs human effort, time and money. In order to find the datapoints corresponding to the rare classes, test datapoints from the dataset are selected and a ranked list of the test datapoints is prepared. Top-k test datapoints in the ranked list are sent for human review. However, the ranked list does not update incrementally after review of each batch of the top-k test datapoints.

The present disclosure proposes a novel solution to overcome the aforementioned problems.

System and working

The present disclosure describes a nearest-neighbor based manifold expansion technique for seeking human review (illustrated in Figure 1). The nearest-neighbor based manifold expansion technique is integrated into an online active learning platform (referred to as an active learner hereafter).

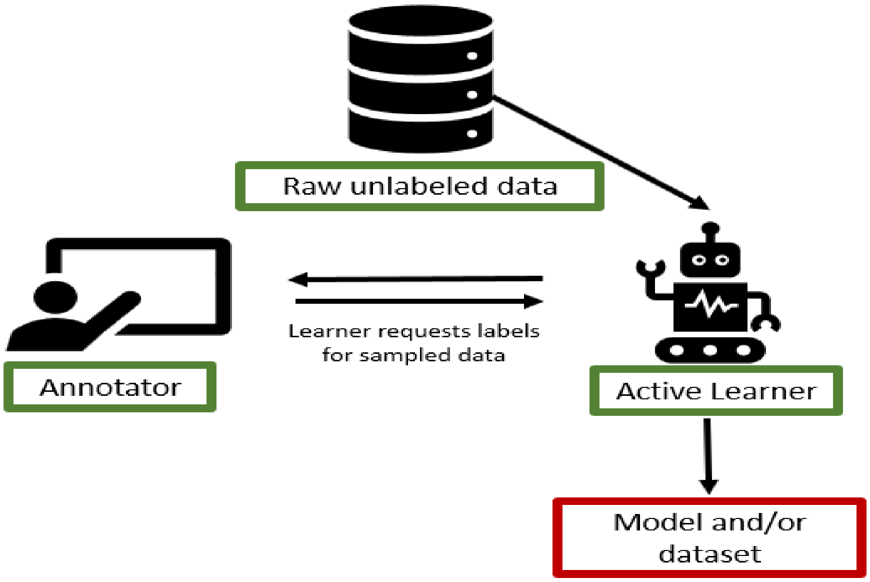

Figure 1: Nearest-neighbor based manifold expansion technique for seeking human review

Initially, in the nearest neighbor based manifold expansion technique, the active learner performs a sampling formulation, as illustrated in Figure 2. In the sampling formulation, an unlabeled dataset including unlabeled examples is provided as an input to the active learner. The unlabeled dataset is then divided into a positive seed dataset, a negative seed dataset and a test dataset. The positive seed dataset includes positive seeds (shown in green color in the Figure 2) and the negative seed dataset includes negative seeds (shown in red color in the Figure 2), and the test dataset includes test examples (shown in black color in the Figure 2). The positive seed dataset and the negative seed dataset are used to train the active learner like a machine learning classifier. Each of the positive seeds vote +1 to the test examples that are in a neighborhood of the positive seed within a threshold area. Similarly, each of the negative seeds vote -1 to the test examples that are in a neighborhood of the negative seed within the threshold area. In terms of scalability, each of the positive seeds and the negative seeds votes to about 500-1000 test examples in the unlabeled dataset of about 100 million. Thus, a voting process is a function of a distance between the positive seed or the negative seed and each of the test examples. All positive votes and all negative votes allocated to each of the test examples accumulate into an overall score for the test example.

Based on their overall scores, the test examples are categorized into four categories: easy positives, hard positives, hard negatives and easy negatives. The easy positives are samples, which are closest to most of the positive seeds, and thus having maximum overall scores. The easy negatives are samples, which are closest to most of the negative seeds, and thus having minimum overall scores. The hard positives are samples which lie near to a decision boundary of the positive seeds. All the test examples within the decision boundary have a positive overall score. The hard negatives are samples that have the overall score of zero. After categorizing the test examples, these are sorted based on their overall scores in a ranked list (shown in the Figure 2) in which examples with higher overall scores are ranked higher.

Thereafter, top-k examples in the ranked list are sent to annotators for review. The annotators assign labels (i.e. positive or negative) to the top-k examples. The annotators can interpret why a particular example got a particular score and how much the positive seeds and the negative seeds contributed to that score. This is the advantage offered by the nearest-neighbor based manifold sampling formulation since it is a non-parametric model, and hence it does not depend on a finite set of parameters to perform categorization. In addition to that, the decision boundary and the voting process are well defined and transparent. This ensures interpretability of the nearest-neighbor based manifold sampling formulation. So, the annotators may change the category of any example in the top-k examples if the annotators think that the example was wrongly categorized previously. The examples labeled as positive by the annotators are added to the positive seed dataset and the examples labeled as negative are added to the negative seed dataset. The voting process is executed again on the basis of the updated positive seed dataset and the updated negative seed dataset. This way, the voting process is executed continuously, and the ranked list is updated in an incremental manner in real time, which does not require reconstruction of the ranked list each time.

Therefore, the active learner supports both the interpretability and the incremental scoring for the nearest-neighbor based manifold expansion technique.

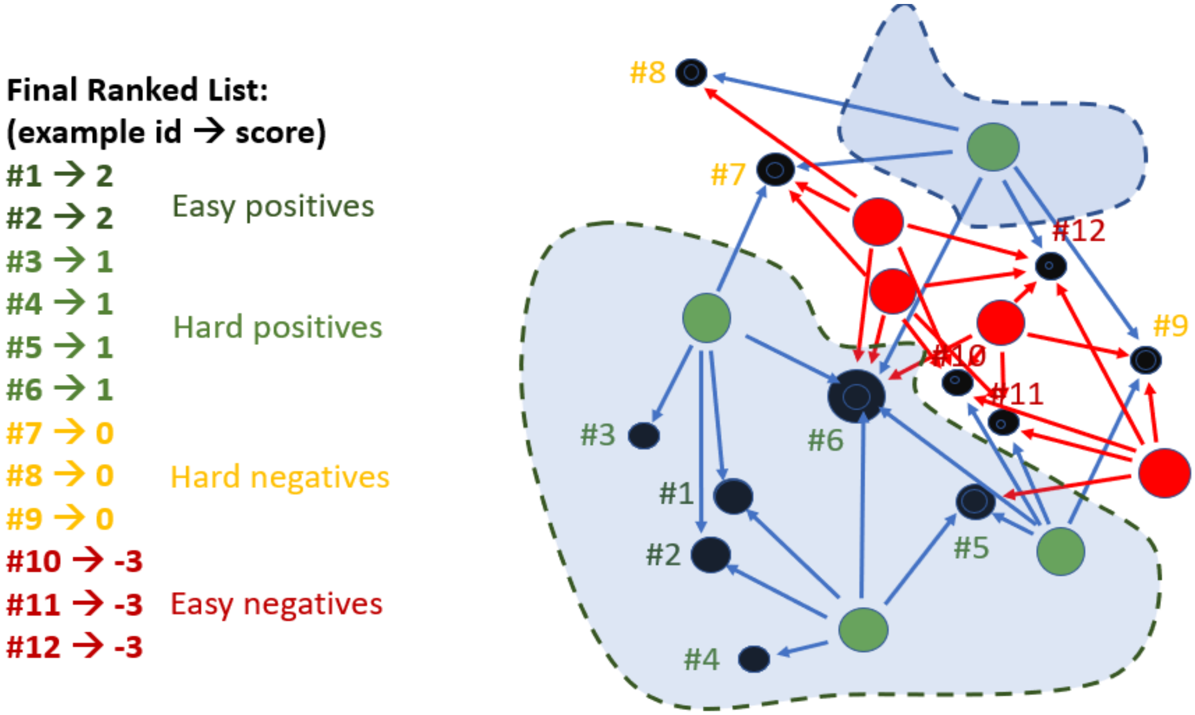

Figure 2: Nearest-neighbor based manifold sampling formulation for active learning

Additional embodiments

In an additional embodiment, aspects of the present disclosure may be utilized to find policy violating ads on a social media website. Live ads are provided as the unlabeled examples to the active learner. Top-k ads in the ranked list are sent to the annotators for labeling. Thus, the top-k ads are labeled/annotated and the policy violating ads amongst the top-k ads are removed from the live ads and are added to the positive seed dataset. The ads that are not violating policy are added to the negative seed dataset. This way, the positive seed dataset and the negative seed dataset are continually updated, and newer live ads are ranked based on the updated positive seed dataset and the negative seed dataset.

In another embodiment, aspects of the present disclosure may be utilized to find civic engagers on the social media website during elections. The unlabeled examples include user profiles having group posts, which constitute more than 1% of activities of users on the social media website. The positive seed dataset includes the user profiles (for example, 500 user profiles), which had civic engagements. Top-k user profiles in the ranked list are sent for review to the annotators. Annotation results for the top-k user profiles are utilized by the active learner to grow the positive seed dataset and the negative seed dataset, and to monitor the user profiles that have civic engagements.

In yet another embodiment, aspects of the present disclosure may be utilized to find hate speech from images that are to be uploaded on the social media website. The unlabeled examples include the images that are being uploaded by the users on the social media website. The positive seed dataset includes image examples (for example, 40K image examples) having hate speech. Top-k images in the ranked list are sent for review to the annotators. The images amongst the top-k images having hate speech, as labeled by the annotators, are removed from the social media website and are utilized to grow the positive seed dataset and the negative seed dataset.

Conclusion

In many machine learning and statistical tasks, collecting data is time-consuming and costly. Thus, finding ways to selectively collect relevant data is beneficial. In most of these cases, active learning may be utilized. The active learning allows us to select future training data, or reject irrelevant data based on the data that we have previously seen. The present disclosure proposes an incremental nearest-neighbor based manifold expansion technique that allows an active learner to iteratively select relevant unlabeled samples for human review.