A Comparative Study of Active Learning Techniques

Balaji Udayagiri

gbu5048@psu.edu

Introduction

In this report, we discuss three papers spanning different periods, all focusing on active learning. We explore how active learning has evolved over the years, noting both similarities and differences. To avoid repetition, we first define active learning and relevance feedback here. **Active learning** is a subfield of machine learning that utilizes both labeled and unlabeled data for learning. The need for active learning arises because traditional supervised learning typically requires large amounts of labeled data, which is both expensive and time-consuming to collect. Active learning methods aim to leverage information from a limited amount of labeled data to gain insights from unlabeled data through intelligent sample selection.

In **relevance feedback**, the model presents the user with a list of unlabeled data to annotate based on similarity, relevance, etc. The data annotated by the user is then used to train the model, thereby reducing the amount of data that needs to be annotated while ensuring that the data chosen by the user is of high quality.

1 Paper 1: Incorporating Diversity and Density in Active Learning for Relevance Feedback [3]

1.1 Problem and Motivation

The authors discuss relevance feedback in the context of document retrieval. They highlight two main problems in relevance feedback. First, there is the quality of the documents used for feedback. Second, there is the issue of reforming the input to improve performance based on the feedback. The first issue is addressed in this paper, focusing on improving the quality of documents presented to the user for feedback.

Showing only the most relevant documents to the user for feedback, i.e., traditional uncertainty-based active learning, often results in duplication among the selected documents, leading to redundancy. There is no research that explicitly addresses this redundancy.

1.2 Solution

The authors propose a novel active learning algorithm called Active-RDD (Relevance, Diversity, and Density), which uses the following metrics:

- Relevance: The measure of closeness between the query (input) and the search space of documents.
- Diversity: The measure of difference among documents, aimed at avoiding redundancy.
- Density: A measure of whether the selected documents come from dense areas of the dataset or sparse areas. Documents are chosen from dense areas to ensure they are representative of the overall data distribution and to avoid outliers.

Active-RDD aims to maximize the learning benefit from user feedback by selecting a diverse and representative set of relevant documents. It frames this selection as an optimization problem, balancing relevance, density, and diversity. The algorithm iteratively updates the document set by calculating a score based on these parameters.

Formally, it is defined as:

$$d_i = \arg\max_{d_i \in I \setminus S} [\alpha \cdot \text{relevance}(d_i) + \beta \cdot \text{density}(d_i) + (1 - \alpha - \beta) \cdot \text{diversity}(d_i, S)]$$

where I is the set of unlabeled documents, S is the set of already selected documents, and α and β are weights that balance the contributions of relevance, density, and diversity.

1.3 Novelties/Contributions

- A novel active learning algorithm, Active-RDD, that integrates relevance, diversity, and density to select a diverse and representative set of feedback documents.
- A significant reduction in redundancy within the feedback documents, resulting in more effective learning from user feedback.

1.4 Downsides

- Although the authors optimize the algorithm to reduce computational complexity, both document density and diversity are high-cost measurements in large datasets. (Both diversity and density are high-computational KL-divergence-based metrics).
- The performance of Active-RDD depends on whether the parameters α and β are well-tuned. They control the balance between relevance, density, and diversity.

2 Paper 2: Semisupervised SVM Batch Mode Active Learning with Applications to Image Retrieval [1]

2.1 Problem and Motivation

The authors' work is in active learning within the context of Content-Based Image Retrieval (CBIR). They identify two main problems in SVM-based active learning. First, there is an inherent problem with SVMs where they do not perform well with small training datasets. They refer to this as the *"small training size problem"*. Second, at the time of the paper's publication, SVM active learning methods often resulted in very similar examples being passed to relevance feedback. They refer to this as the *"batch sampling problem"*.

2.2 Solution

To address these problems, the authors introduce the algorithm, Semi-Supervised SVM Batch Mode Active Learning (SVM^{SS}_{BMAL}). The semi-supervised component of the algorithm addresses the first problem, while the Batch Mode Active Learning addresses the second problem.

Small Training Size Problem: The authors use a unified kernel learning approach that incorporates both supervised and unsupervised kernels. For the unsupervised kernel, they adopt a kernel deformation method. Given an original kernel $K(x, x')$, the deformed kernel $\tilde{K}(x, x')$ is computed as:

$$\tilde{K}(x, x') = K(x, x') - \kappa_x^\top (I + MK)^{-1} M \kappa_{x'}$$

where: - $K(x, x')$ is the original kernel. - $\kappa_x = (K(x_1, x), \ldots, K(x_n, x))^\top$ is a vector of kernel evaluations. - M is a matrix capturing the geometric structure of both labeled and unlabeled data. - I is the identity matrix.

The matrix M is typically derived from the graph Laplacian L, which encodes the similarity between examples (both labeled and unlabeled).

Batch Sampling Problem: They extend the min-max optimization of SVM, which uses a single example per iteration, to batch mode active learning. They use this approach to reduce the uncertainty in the model. To incorporate diversity in the batch, they introduce quadratic programming, which penalizes similarity among the batch. This balances uncertainty with diversity.

2.3 Novelties/Contributions

- Although neither semi-supervised active learning nor SVM batch mode are new, this paper is the first to combine both ideas to create a hybrid method (SVM^{SS}_{BMAL}) that enhances relevance feedback.
- Unlike other SVM active learning methods, this method provides multiple examples for feedback at once.

2.4 Downsides

- The authors mention that the quadratic programming (QP) solver used in the optimization process may not scale well for large datasets, which limits the algorithm's applicability in large-scale, real-world CBIR use cases.

- They base their work on the assumption that SVM performs well on most CBIR tasks, which might not be applicable in certain problem domains.
- The unlabeled data may not be similar to the labeled data, which can lead to the semi-supervised learning approach not performing well.

3 Paper 3: Uncertainty-Based Active Learning via Sparse Modeling for Image Classification [2]

3.1 Problem and Motivation

The authors address the inefficiencies associated with conventional uncertainty-based active learning in image classification. While uncertainty sampling is widely used, it often results in the selection of redundant samples, which diminishes the performance of the active learning process. This work also attempts to combine uncertainty with measures of density and diversity to avoid selecting duplicate samples, thereby enhancing classifier performance and reducing labeling efforts.

3.2 Solution

Building on similar ideas as the first paper, the authors propose a method that incorporates uncertainty, diversity, and density in the sample selection process (though this is in the context of images, whereas the first paper is in the context of text retrieval). The core idea is to simultaneously account for diversity and density while representing uncertainty through a sparse linear combination of Gaussian kernels. This approach ensures that the selected samples are both representative of the data distribution and diverse.

The sparse modeling optimization problem is formulated as follows:

$$\mathbf{f} = \underset{\mathbf{f}}{\operatorname{argmin}} \|\mathbf{Q}\mathbf{f} - \mathbf{s}\|_2^2, \quad \text{subject to } \|\mathbf{f}\|_0 = B_q,\ 0 \leq \mathbf{f} \leq 1, \tag{1}$$

where $\mathbf{f}$ denotes the adjusted uncertainty score vector, $\mathbf{Q}$ represents the similarity matrix among samples, and B_q signifies the batch size. This formulation aims to select a sparse subset of samples that captures the highest uncertainty and diversity within the dataset.

A kernel function is used to define the similarity between samples:

$$Q_{i,j} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right), \tag{2}$$

where σ determines the kernel width, and $\mathbf{x}_i$ and $\mathbf{x}_j$ are feature vectors.

3.3 Novelties/Contributions

- Introduces a sparse modeling approach that integrates uncertainty, diversity, and density for active learning.

- Utilizes Gaussian kernels to represent the uncertainty of unlabeled samples, ensuring both informativeness and diversity.
- Proposes a selective sampling technique, where samples with low uncertainty are excluded before optimization, reducing computational complexity.
- Develops efficient optimization techniques, including greedy search and quadratic programming, to solve the sparse representation problem effectively.

3.4 Downsides

- The reliance on Gaussian kernels and sparse modeling may limit generalizability to other problem domains.
- The method depends on a robust initial classifier; if the classifier is poorly trained, the active learning process may be less effective.

Conclusion

All three papers address the insufficiency of uncertainty as the sole criterion for active learning. Each paper presents techniques that incorporate additional criteria such as variety, density, and sparsity to overcome the limitations of conventional uncertainty sampling.

The first paper employs sparse modeling to effectively balance uncertainty, diversity, and density; the second paper integrates semi-supervised learning with batch-mode selection to improve sample efficiency in large-scale tasks; and the third paper uses a kernel density approach to ensure representative sample selection. Each of these methods has its own advantages. However, these techniques also have certain drawbacks, such as dependency on the quality of the unlabeled data.

Future research in active learning should aim to develop generalizable and scalable techniques that minimize computational complexity while maintaining efficiency across diverse domains.

References

[1] Steven C.H. Hoi et al. "Semi-supervised SVM batch mode active learning for image retrieval". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 2008, pp. 1–7. DOI: 10.1109/CVPR.2008.4587350.

[2] Gaoang Wang et al. "Uncertainty-Based Active Learning via Sparse Modeling for Image Classification". In: *IEEE Transactions on Image Processing* 28.1 (2019), pp. 316–329. DOI: 10.1109/TIP.2018.2867913.

[3] Zuobing Xu, Ram Akella, and Yi Zhang. "Incorporating Diversity and Density in Active Learning for Relevance Feedback". In: *Advances in Information Retrieval*. Ed. by Giambattista Amati, Claudio Carpineto, and Giovanni Romano. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 246–257. ISBN: 978-3-540-71496-5.