AIDEme: An active learning based system for interactive exploration of large datasets

Enhui Huang[†], Luciano Di Palma[†], Laurent Cetinsoy[†], Yanlei Diao[†][*], Anna Liu[*]
[†]École Polytechnique, France; [*]University of Massachusetts Amherst, USA
{enhui, lucianodipalma, yanlei.diao, cetinsoy}@lix.polytechnique.fr; anna@math.umass.edu

1 Introduction

There is an increasing gap between fast growth of data and limited human ability to comprehend data. Consequently, there has been a growing demand of data analytics tools that can bridge this gap and help the user retrieve high-value content from data more effectively.

Objectives. In this demo, we introduce AIDEme, a large-scale interactive data exploration system that is cast in a principled active learning (AL) framework: in this context, we consider the data content as a large set of records in a data source, and the user is interested in some of them but not all. In the data exploration process, the system allows the user to label a record as "interesting" or "not interesting" in each iteration, so that it can construct an increasingly-more-accurate model of the user interest. Active learning techniques are employed to select a new record from the unlabeled data source in each iteration for the user to label next in order to improve the model accuracy. Upon convergence, the model is run through the entire data source to retrieve all relevant records.

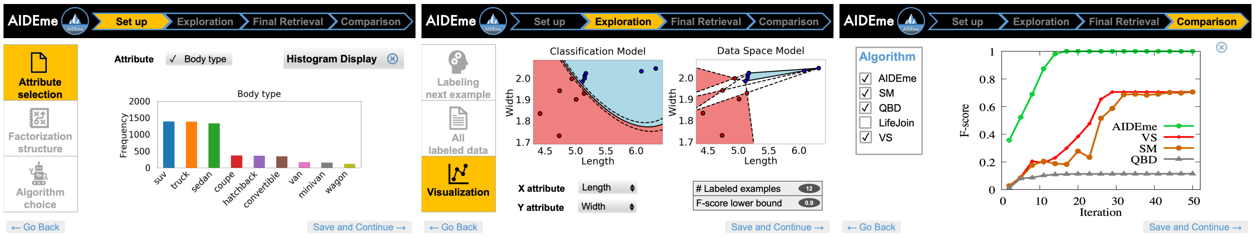
Novelty. A challenge in building such a system is that existing active learning techniques [3, 4] experience slow convergence in learning the user interest when such exploration is performed on large datasets: for example, hundreds of labeled examples are needed to learn a user interest model over 6 attributes, as we showed using a digital sky survey of 1.9 million records [1, 2]. AIDEme employs a set of novel techniques to overcome the slow convergence problem:

- **Factorization**: We observe that a user labels a data record, her decision making process often can be broken into a set of smaller questions, and the answers to these questions can be combined to derive the final answer. This insight, formally modeled as a factorization structure, allows us to design new active learning algorithms, e.g., factorized version space algorithms [2], that break the learning problem into subproblems in a set of subspaces and perform active learning in each subspace, thereby significantly expediting convergence.
- **Optimization based on class distribution**: Another interesting observation is that when projecting the data space for exploration onto a subset of dimensions, the user interest pattern projected onto such a subspace often entails a convex object. When such a subspatial convex property holds, we introduce a new "dual-space model" (DSM) [1] that builds not only a classification model from labeled examples, but also a polytope model of the data space that offers a more direct description of the areas known to be positive, areas known to be negative, and areas with unknown labels. We use both the classification model and the polytope model to predict unlabeled examples and choose the best example to label next.
- **Formal results on convergence**: We further provide theoretical results on the convergence of our proposed techniques [1, 2]. Some of them can be used to detect convergence and terminate the exploration process.
- **Scaling to large datasets**: In many applications the dataset may be too large to fit in memory. In this case, we introduce subsampling procedures and provide provable results that guarantee the performance of the model learned from the sample over the entire data source [1, 2].

Evaluation results using real-world datasets and user interest patterns show that AIDEme significantly outperforms state-of-the-art active learning techniques in accuracy while achieving desired efficiency for interactive performance [1, 2].

2 Demonstration

This demo aims to showcase the interactive application of our active learning algorithms for scalable data exploration, with real-time visualization of learned models. To enable the demo, we have designed the following exploration tasks on real-world datasets, and built a prototype system to offer live interactive experience of the user:



(a) Initial Setup

(b) Interactive Exploration

(c) Comparison

Figure 1: AIDEme front-end interface

- *Cars dealer scenario*: The user plays the role of a cars dealer trying to find all the cars matching a specific customer profile. In this scenario, the user will interact with a cars dataset (extracted from teoalida.com), containing 5622 examples of vehicles. Each car is described by 27 attributes, including price, model, length, width, color, horsepower, etc.
- *Job search scenario*: The user plays the role of a data scientist searching for the most suitable jobs that match her expertise and career goals. In this scenario, the user will interact with a job postings dataset (provided by a kaggle challenge), containing 5715 Data Science-related job postings across the United States. Each job contains 42 attributes, such as job title, company, company's revenue, salary, location, and job description.

Once the user chooses one of the above scenarios, the data exploration process begins. The **live experience** of the user includes understanding *the attributes and their data distributions*, *the evolution of the learned model with each labeled example*, *how the factorization of her decision making process makes a difference*, and *how the model evolves differently when various AL algorithms are used*. More details on the **interactivity** are given below and illustrated in Figure 1.

1. Initial Setup: At the beginning of the exploration process, the user inputs information including: (a) *Attribute selection and filtering*: The user selects a subset of attributes that are potentially relevant to her task, which may be a superset of the attributes finally used in the learned model. Attribute filters, such as ranges (min, max), can also be specified. (b) *Factorization structure*: Optionally, our system allows the user to explore a factorization structure for effective learning if she can break her decision-making process into a set of simpler, independent "yes" or "no" questions. Our system does not expect the user to specify these questions precisely, which is often hard given the complexity of the task, but only asks which attributes compose these questions. Once such high-level intuition is given, our algorithms can leverage it to significantly improve convergence speed. In addition, histograms and heatmaps are available to depict data distribution in various forms and help the user with the above two tasks. (c) *Algorithm choice*: The user may also specify a particular AL algorithm she has in mind. Otherwise, a default algorithm will be used.

2. Iterative Exploration: Then the demo will proceed to iterative exploration, where the system engages the user in a series of interactions, with a new example presented for the user to label in each iteration, until the user wishes to stop or our system has detected convergence. An important aspect of our demo is to allow the user to trace the evolution of the model, via real-time visualization, which provides insights into questions such as (a) what is the current decision boundary? (b) what is the estimated accuracy of the current model? (c) how does the recent labeled example change the decision boundary? (d) how does the evolution of the model vary when different algorithms are chosen?

3. Final retrieval: Once the iterative exploration terminates, our system retrieves all the records in the data source that are predicted as relevant, and displays them for the user to review.

4. Comparison: By using the final retrieval result as ground truth, our system will simulate the performance of existing AL algorithms including Version Space [4], Simple Margin [3], and Query by Disagreement [3]. This way, users can achieve an in-depth understanding of the tradeoffs between different AL methods in convergence and interactive speed.

References

[1] Huang, E., Peng, L., Di Palma, L., Abdelkafi, A., Liu, A. & Diao, Y. Optimization for active learning-based interactive database exploration. *Proceedings of the VLDB Endowment* (PVLDB), **12**(1), 71-84, September 2018.

[2] Di Palma, L., Diao, Y. & Liu, A. A Factorized Version Space Algorithm for "Human-In-the-Loop" Data Exploration. *IEEE International Conference on Data Mining* (ICDM), November 2019, to appear.

[3] Settles, B. Active Learning. Morgan Claypool Publishers, 2016

[4] Gonen, A., Sabato, S.& Shalev-Shwartz, S. Efficient Active Learning of Halfspaces: an Aggressive Approach. Journal of Machine Learning Research, vol. 14, no. 1, pp. 2583–2615, 2013