FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Mitchell Borchers

Active learning in E-Commerce Merchant Classification using Website Information

Department of Theoretical Computer Science and Mathematical Logic

Supervisor of the master thesis:	Mgr. Marta Vomlelová, Ph.D.
Study programme:	Artificial Intelligence
Study branch:	IUIPA

Prague 2023

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

I would like to thank my supervisor, Mgr. Marta Vomlelová, Ph.D., for her guidance throughout the duration of my thesis, as well as Adam Dobiáš and his team for the opportunity to work on this project. Additionally, I extend a big thank you to my parents, Daniel and Mary, and sisters for their unending support and encouragement.

Title: Active learning in E-Commerce Merchant Classification using Website Information

Author: Mitchell Borchers

Department: Department of Theoretical Computer Science and Mathematical Logic

Supervisor: Mgr. Marta Vomlelová, Ph.D., Department of Theoretical Computer Science and Mathematical Logic

Abstract: Data and the collection and analysis of data plays an important role in everyday life even though it often goes unseen. In our case, our partner is using data to classify websites into different categories. We used active learning and other machine learning methods to help classify websites into these categories and to explore the data collection and classification process. We scraped text data from websites, translated the data to English, and then worked with machine learning tools to understand the data and classify it. We found that the xPAL active learning strategy and linear support vector classifers seemed to perform best with our data.

Keywords: active learning xPAL machine learning multi-class classification website classification data mining web crawler web scraper

Contents

Introduction

One of the main challenges of creating a successful machine learning model is obtaining labeled data. With easy access to a variety of modern tools, devices, and sensors, we are able to rapidly collect unlabeled data. But, in supervised learning, prediction models are trained using labeled data. The problem is that acquiring labels for the collected data can be expensive, time-consuming, or even impossibly difficult in some cases.

However, methods have been developed to help reduce the number of labeled data required to train the classifier. Active learning is a semi-supervised machine learning framework where the model is trained with a smaller set of labeled data but which also aims to exploit trends within the unlabeled data. It's a framework in which the learner has the freedom to select which data points are added to its training set (Roy and McCallum [2001]).

Active learning is different from other frameworks because it uses the unlabeled data and some evaluation criteria to determine which candidate could be the most beneficial to the model if it was given a label. The model requests the label from some oracle that provides the label then it takes this new labeled data and rebuilds the classifier. We describe it as semi supervised active learning because the model is initially trained on both the labeled and unlabeled data, and then active learning is used to select the most informative examples for labeling.

In our case we will provide a set of labeled data to the active learning framework (or sampling strategy). The sampling strategy will assume all the data is unlabeled and then choose a candidate from the unlabeled data pool. Then the label is revealed and the classifier is updated using the new data. The newly labeled data is then added to the labeled pool and the process repeats.

We have some data (website urls) for some company or business that are given to us from our partner. From this data our partner currently utilizes human labor to browse the website and then label the url with a category (23 labels) and a subcategory (234+ tags, that branch from the main category but still have some relation). This is a repetitive and expensive task that could be supplemented using active learning.

To reduce the amount of data required to train the classifier we consider a combination of tools and frameworks, namely: Scrapy, Postgres, a translation service, and an active learning sampling strategy paired with a classifier. We also explore the use of different classifiers to determine if there is some optimal classifier.

A website is required, then we use the Scrapy framework to navigate to the webpage, and collect then store the scraped data into the database. Next we access the data, translate the text, and add the translated data back into the database. During this process we also remove the html and numbers.

Once the the data is close to just pure text we use TF-IDF to transform it into a vectorized representation so we can use it with the classifiers. We experiment with different classifiers to determine if there is some optimal classifier for our data.

In the first section we introduce active learning and the different components of active learning. In the second section we look more into the details of xPAL and how it works. In the third section we discuss the data and the steps we took to collect and process the data. In the fourth and fifth sections we conduct a variety of experiments to explore the performance of the sampling strategies and alternative classifiers.

Our goal is to understand the entire process including the web scraping, translation, storage, and performance of the selection strategies and classifiers. This analysis will allow our partner to learn from our tests and experiments. It will also allow them to make an informed decision on which models and selection strategies may be best suited for their needs moving forward.

Definitions

In this section we define some terms and ideas that will be helpful in understanding the upcoming sections.

Definition 1 (Beta Prior)**.** *A beta prior is a conjugate prior for the binomial distribution. It is a continuous probability distribution defined on the interval [0, 1] and is parameterized by two positive shape parameters, α and β. The beta distribution is defined as:*

$$Beta(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where Γ is the gamma function and x is a random variable. The gamma function is defined as:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

The gamma function is used as a normalizing constant to ensure that the probability density function integrates to 1 over the simplex, which is the space of all probability vectors that sum to 1.

Definition 2 (Conjugate Prior)**.** *A conjugate prior is a prior distribution that is in the same family of distributions as the likelihood function. In other words, the posterior distribution will have a similar functional form to the prior distribution.*

Definition 3 (Decision-Theoretic)**.** *Decision-theoretic active learning is a framework that uses the expected performance gain of a candidate to determine which candidate to label. The expected performance gain is the expected performance of the classifier after labeling the candidate minus the expected performance of the classifier before labeling the candidate. The expected performance of the classifier is the expected value of the performance measure given the posterior distribution of the classifier.*

Definition 4 (Dirichlet Distribution)**.** *The Dirichlet distribution is a multivariate generalization of the beta distribution. It is a continuous probability distribution defined on the K-simplex, where:*

$$\Delta_K = \left\{ x \in \mathbb{R}^K : x_i \geq 0, \sum_{i=1}^{K} x_i = 1 \right\}$$

Definition 5 (Ground Truth)**.** *Ground truth is the true label of a data point.*

Definition 6 (Posterior Probabilities)**.** *Posterior probability is a type of conditional probability that results from updating the prior probability with information summarized by the likelihood via an application of Bayes' rule. The posterior probability is the probability of an event occurring given that another event has occurred.*

Definition 7 (Omniscient Oracles)**.** *Omniscient oracle is a hypothetical entity that has complete knowledge of the true labels of all data points in a given dataset. An omniscient oracle knows the ground truth labels of all data points.*

Definition 8 (TF-IDF). *TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus (large structured set or collection of speech or text data).*

1. Active Learning

1.1 Introduction

Russel and Norvig succinctly define an agent and different types of learning in their book "Artificial Intelligence: A Modern Approach" (Russell and Norvig [2009]), their definition is paraphrased here. They define an agent as something that acts and a rational agent as one that acts so as to achieve the best outcome. If there is uncertainty, then the agent tries to achieve the best expected outcome. Any component of an agent can be improved by learning from data. The improvements and techniques used to make them depend on four major factors:

1. Which component is to be improved.
2. What prior knowledge the agent already has.
3. What representation is used for the data and the component.
4. What feedback is available to learn from.

Here we will mostly be focused on the final point, "What feedback is available to learn from". However, we will also discuss the importance of the second and third points because we will use Bayesian learning. There are three main types of feedback that determine the three main types of learning, which are: unsupervised, reinforcement, and supervised.

In unsupervised learning an agents goal is to discover patterns in the data even though no feedback or labels are provided. In reinforcement learning, the agent learns from a series of rewards or punishments that are dealt out based on its decisions. In supervised learning, an agent learns from input-output pairs, which can be discrete or continuous, to find a function that maps the pairs as best as possible.

The goal of supervised learning, given a training set of N example input-output pairs:

$$(x_1, y_1), (x_2,y_2),...(x_N,y_N),$$

where each y_j was generated by some unknown function $y = f(x)$, is to find a function h that approximates the true function f.

In reality, the types of learning overlap. In semi-supervised learning, some data points are labeled, and some are not. The model is trained on the labeled data, and then the knowledge gained from that labeled data is used to improve the model's predictions on the unlabeled data.

Supervised learning models almost always get more accurate with more labeled data. Active learning is the process of deciding which data to select for annotation (Munro [2021]). In other words, the central component of an active learning algorithm is the selection strategy, or deciding which of the unlabeled data could be the most useful to the model if it was labeled. Active learning uses a selection strategy that augments the existing classifier, it is not itself a classifier but rather an evaluation methodology working with a classifier.

Many different sampling strategies exist. First we will discuss query functions then we will briefly define three basic sampling strategies: uncertainty, diversity, and random sampling to get an idea of sampling. We will then discuss some other more advanced sampling strategies that are used in our experiments. When sampling the unlabeled data an ordered list is returned and the top candidate is the candidate that is expected to be most valuable for the model, but we are not strictly limited to taking just one candidate.

1.2 Query Function Construction

There are various techniques used to construct the querying functions. We will focus on pool-based active learning, but a number of interesting and relevant ideas appear within other active-learning frameworks that are worth mentioning.

1.2.1 Pool-Based

In pool-based active learning, a fixed set of unlabeled examples is provided at the start of the learning process, and the active learner iteratively selects a subset of these examples for annotation (Huang and Lin [2016]). The selection of the subset is based on a query strategy that aims to maximize the information gain from each annotation. Pool-based active learning is useful in situations where all the data is available in advance, such as in document classification or image classification.

1.2.2 Stream-Based

In stream-based active learning, data arrives in a continuous stream, and the active learner must make real-time decisions about which examples to label (Baram et al. [2004]). This is common in settings such as sensor networks or social media feeds. The selection of examples for annotation is based on a query strategy that takes into account the current state of the model, as well as the uncertainty and informativeness of each incoming example. The stream-based model can be viewed as an online version of the pool-based model.

1.2.3 Membership Queries

In membership-query-based active learning, the active learner can make queries to an oracle or construct a point in input space and requests its label from an oracle, such as a human expert, to obtain labels for specific examples (Baram et al. [2004]). The goal is to select the examples for which obtaining a label is most informative, in order to minimize the number of queries required to achieve a high accuracy. Membership-query-based active learning is useful when labeling each example is expensive or time-consuming, such as in medical diagnosis or legal document review.

1.3 Sampling Strategies

Sampling strategies, also referred to as selection strategies, are the core of the active learning process. The goal of sampling is to select the most useful data points from the unlabeled pool to label. The most useful data points are those that are expected to improve the classifier the most.

1.3.1 Random Sampling

Random sampling is self explanatory as it randomly selects an unlabeled data point from the pool and requests to have it labeled then it uses this newly selected data point to update the model. Random sampling is good to use as a baseline to compare other sampling strategies with.

1.3.2 Diversity Sampling

Diversity sampling is the set of strategies for identifying unlabeled items that are underrepresented or unknown to the machine learning model in its current state (Munro [2021]). The items may have features that are unique or obscure in the training data, or they might represent data that are currently underrepresented in the model.

Either way this can result in poor or uneven performance when the model is applied or the data is changing over time. The goal of diversity sampling is to target new, unusual, or underrepresented items for annotation to give the algorithm a more complete picture of the problem space.

1.3.3 Uncertainty Sampling

Uncertainty sampling is based on the idea that the most informative examples to query are the ones that the current model is most uncertain about. For example, in binary classification, an uncertain example might be one that is close to the decision boundary, or one that has a low predicted probability for the majority class (Munro [2021]). The idea is that by querying these uncertain examples, the model can better learn the boundary between the classes and improve its accuracy. Uncertainty sampling is simple given a classifier that estimates $P(C|w)$ (Lewis and Gale [1994]). On each iteration, the current version of classifier can be applied to each data point, and the data with estimated $P(C|w)$ values closest to 0.5 are selected, since 0.5 corresponds to the classifier being most uncertain of the class label.

These items are most likely to be wrongly classified, so they are the most likely to result in a label that differs from the predicted label, moving the decision boundary after they have been added to the training data and the model has been retrained.

1.3.4 EER

Monte Carlo estimation of error reduction (EER) estimates future error rate by log-loss, using the entropy of the posterior class distribution on a sample of the unlabeled examples, or by 0-1 loss, using the posterior probabilities of the most probable class for the sampled unlabeled examples (Roy and McCallum [2001]).

Basically, the goal is to estimate the expected reduction in error for each unlabeled example by randomly sampling from the model's predictions and comparing the performance of the model with and without the example included in the training data.

1.3.5 PAL

Probabilistic Active Learning (PAL) follows a smoothness assumption and models for a candidate instance both the true posterior in its neighborhood and its label as random variables (Krempl et al. [2014]). By computing for each candidate its expected gain in classification performance over both variables, PAL selects the candidate for labeling that is optimal in expectation. PAL shows comparable or better classification performance than error reduction and uncertainty sampling, has the same asymptotic linear time complexity as uncertainty sampling, and its faster than error reduction based on the tests from the paper.

1.3.6 xPAL

Extended probabilistic gain for active learning (xPAL) is a decision-theoretic selection strategy that directly optimizes the gain and misclassification error, and uses a Bayesian approach by introducing a conjugate prior distribution to determine the class posterior to deal with uncertainties (Kottke et al. [2021]). Although the data distribution can be estimated, there is still uncertainty about the true class posterior probabilities.

These class posterior probabilities can be modeled as a random variable based on the current observations in the dataset. For this model, a Bayesian approach is used by incorporating a conjugate prior to the observations. This produces more robust usefulness estimates for the candidates.

1.3.7 ALCE

Active Learning with Cost Embedding (ALCE) is a non-probabilistic uncertainty sampling algorithm for cost-sensitive multiclass active learning (Huang and Lin [2016]). First a cost-sensitive multiclass classification algorithm called cost embedding (CE) was designed, which embeds the cost information in the distance measure in a special hidden space by non-metric multidimensional scaling. Then a mirroring trick was used to let CE embed the possibly asymmetric cost information in the symmetric distance measure.

It works by augmenting the example space with an additional dimension that represents the cost of labeling each example. This cost embedding can be learned from previous labeling efforts or estimated based on domain knowledge. The cost embedding can then be used to guide the active learning process by selecting examples that are not only informative but also cost-effective to label.

1.3.8 QBC

Query By Committee (QBC) uses an ensemble of classifiers that are trained on bootstrapped replicates of the labeled set (Seung et al. [1992]). The idea is to train a committee of classifiers on the available labeled data and then use the committee to select the most informative unlabeled data for labeling (Freund et al. [1997]). The committee consists of several classifiers, each trained on a slightly different subset of the available labeled data.

The QBC algorithm measures the disagreement of the committee's predictions on each unlabeled data point. The intuition is that if the committee members disagree then it is likely to be a difficult data point for the current classifier and thus informative for labeling.

The algorithm selects a fixed number of the most informative examples and requests their labels. The labeled examples are then added to the labeled pool, and the committee is retrained on the expanded labeled pool. This process is repeated until the algorithm achieves a desired level of accuracy or the available labeling budget is exhausted.

1.4 Classifiers

The classifier integrated into the active learning sampling strategy repository we used is the Parzen Window Classifier (PWC). It is a non-parametric method used for classification and density estimation in machine learning. It works by estimating the probability density function of a given class using a kernel density estimator, and then using Bayes' theorem to classify new instances based on their estimated probability densities.

We will also explore using other classifiers from Scikit-Learn and TensorFlow and compare their performance on the data without using active learning to see if there is any improvement beyond the PWC classifier.

1.5 Summary

It should now be more clear how the sampling strategy is the major component of active learning. The query function construction is also important but it is just a means of routing the data to be sampled. In the next chapter we will look into the specifics of xPAL.

Bibliography

Yoram Baram, Ran El Yaniv, and Kobi Luz. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5(Mar):255–291, 2004.

Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133, 1997.

Kuan-Hao Huang and Hsuan-Tien Lin. A novel uncertainty sampling algorithm for cost-sensitive multiclass active learning. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 925–930. IEEE, 2016.

Daniel Kottke, Marek Herde, Christoph Sandrock, Denis Huseljic, Georg Krempl, and Bernhard Sick. Toward optimal probabilistic active learning using a bayesian approach. *Machine Learning*, 110(6):1199–1231, 2021.

Georg Krempl, Daniel Kottke, and Myra Spiliopoulou. Probabilistic active learning: A short proposition. In Torsten Schaub, Gerhard Friedrich, and Barry O'Sullivan, editors, *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI2014)*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 1049–1050, Prague, Czech Republic, August 2014. IOS Press. Short Paper.

David D Lewis and William A Gale. A sequential algorithm for training text classifiers. *arXiv preprint cmp-lg/9407020*, 1994.

Robert Munro. *Human-in-the-loop machine learning*. Manning Publications, New York, NY, July 2021.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2:441–448, 2001.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2009. ISBN 9780136042594.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

List of Figures

List of Tables

A. Attachments