Clustering a binary labeled dataset for partial classification and outlier detection (working title)

Jonas Faßbender

1 Problem

Supervised learning is the task of approximating the unknown function $y = f(x)$ with $h(x)$ where x generates y. Both x and y can be any object. If y is from a finite set this learning problem is called classification.

A classifier builds $h(x)$ based on a set T containing n example tuples:

$$T := \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

referred to as the training set.[RN12]

A classifier is trained to approximate $f(x)$ with $h(x)$ based on T, which makes the accuracy $a(h)$, $0 \leq a(h) \leq 1$ of $h(x)$ (how well $h(x)$ generalizes and predicts the corresponding y to unseen x from a test set similar to T) dependent on the quality of T.[TC17]

For some data sets the data quality is not high enough to be able to find $h(x)$ with a sufficient accuracy $a(h)$. In this case sufficient means that $a(h)$ is greater than or equal to a context dependent value s representing a threshold $s \leq a(h)$.

I am trying to find a method that first clusters a data set with an insufficient $h(x)$ and afterwards builds a subset $P_s \subseteq P$ of these clusters or partitions P of the data set so that for each partition in P_s exists a classifier with a sufficient $h(x)$. My goal is to maximize the space P_s covers.

2 Next steps

To start this project I want to begin with a literature research focusing on the following topics:

- common clustering methods, especially the k-means algorithm
- Nearest Neighbor algorithms, Voronoi Cells, LSH (Local Sensitive Hashing)
- Isolation Forrest and ensemble-based classification methods
- Randomized methods (Monte Carlo methods)

Literatur

[RN12] Stuart Russel and Peter Norvig. Künstliche Intelligenz: Ein moderner Ansatz. Pearson, Higher Education, München, 3 edition, 2012.

[TC17] Ophir Tanz and Cambron Carter. Why the future of deep learning depends on finding good data. https://techcrunch.com/2017/07/21/why-the-future-of-deep-learning-depends-on-finding-good-data/, 2017. 10.28.2018.