Technische Universität Berlin

Report

Advanced Information Management III: Scalable Data Analytics and Data Mining

A Comparison of Online learning Naïve Bayes Classifier on RSS Feeds using SPARK

Authors:

Ahmet Anil Pala

Franziska Adler

January 28, 2015

Table of Contents

1 Theory

1.1 Motivation

In Text Classification Naïve Bayes Classifiers are popular due to their simplicity and find application in the field of spam email detection and descovery of specified web content (Ertel, 2008, p. 225). Classifier aim the prediction of a categorie like spam or ham on the basis of previous examples. This can be easily realized with a probalistic classifier like Naïve Bayes assuming that the data distribution is static over time, all data is always accessible and query time does not play a major role. For adapting real world problems and their dynamics this might not be sufficient. The developement of the internet and the need of processing huge amounts of available data and answering queries in time-critical situations needs to be taken in account for handeling model construction and model updating. In Online Machine Learning different methods for dealing with sequentially arriving data and updating the model with new incoming observations exist. This is called "Concept Drift" and referes to the problem of the variation of statistical properties (Astudillo and Gonzalez, 2013). Also aspects of scalability might to be considered since Concept Drift occurs mainly in situations of large data qantities arriving via stream (Tsymbal, 2004, p. 4) and the underlying data processing system is distibuted.

In this report we are using RSS feeds from BCC for classification on a distributed system. Focus is the evaluation of several Naïve Bayes classifiers which implement the online learning paradigm by using streaming techniques. The evalutation concentrates on the handling of Concept Drift. In a first theoretically part of this report challanges, concepts and methods as well as evaluation measures are described. The second part consits of the explanation of our implementation using SPARK to show how the different classifiers work. Furthermore the evaluation of their performances are analysed and presented.

1.2 Challanges/ Concepts

texttexttext

1.3 Methods

1.3.1 Sliding Window

texttexttext

1.4 Evaluation Measurse

texttexttext

2 Empirics

2.1 Data

texttexttext

2.2 Software

texttexttext

2.3 Data

texttexttext

2.4 Implementation

texttexttext

2.5 Evaluation

texttexttext

2.6 Results

texttexttext

2.7 Summary

texttexttext

References

Wolgang Ertel. *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung*. Vieweg + Teubner, Wiesbaden, 2008.

César A. Astudillo and Javier I. Gonzalez. Concept drift detection using online bayesian classifier. In *Proceedings of the XXXII International Conference of The Chilean Computer Science Society*, 2013. URL: http://jcc2013.inf.uct.cl/wp-content/proceedings/SCCC/Concept Drift Detection Using Online Bayesian Classifier.pdf,last accessed on 22.01.2015.

Alexey Tsymbal. The problem of concept drift: Definitions and related work. Technical report, Department of Computer Science, Trinity College Dublin, Ireland, 2004. URL: https://www.cs.tcd.ie/publications/tech-reports/reports.04/TCD-CS-2004-15.pdf, last accessed on 21.01.2015.