Text Classification: Concepts and Methods
Abstract

Weiwei Cheng
University of Marburg / Deutsche Bank
www.chengweiwei.com

Text classification is a supervised learning task, in which a free-text document is assigned with one or more pre-defined category labels based on the information suggested by the set of labeled training documents. Text classification technique is ubiquitous in our everyday life and commonly applied in business. For example, news stories can be typically grouped into sub-domains such as “sports”, “politics”, “entertainment” and so on; in email spam filtering, spam emails are separated from non-spam emails, which is done with text classification as well; another example is the autoresponding system used in large and medium-sized companies. It automatically answers the received messages based on their contents.

Given its great need in practice, text classification has been intensively studied in artificial intelligence in general and machine learning in particular. After two decades of research and practice, it is now fair enough to say that text classification is already a well established topic in the field. Quite a number of easy to use, carefully maintained software packages are public available. Even for non-technical users, it is not very challenging to follow their documentations and apply text classification straightforwardly in practice.

However, to fully exploit the knowledge in the text data and further increase the classification accuracy, one should have at least a rough idea of the text classification algorithms they used, and understand the pros and cons of their choices. Moreover, one should know how to choose a suitable classifier given different learning scenarios.

In this talk, we will give a short introduction on text classification and have a quick overview of several state-of-art text classification algorithms, including naïve Bayes, k-nearest neighbor, and support vector machine. We will focus on the big ideas behind these algorithms and discuss their strong and weak points, respectively.

About the author

程蔚蔚　德国马尔堡大学(University of Marburg)数学与计算机系Knowledge Engineering & Bioinformatics实验室成员，科研助理、博士研究生，德意志银行(Deutsche Bank)数据挖掘项目咨询。主要研究方向为机器学习、数据挖掘，并在相关领域的国际重要期刊及会议上发表论文多篇。担任多个国际学术期刊、会议委员会成员，审稿人。德国马格德堡大学(University of Magdeburg)计算机硕士学位，郑州大学计算机与工商管理双学士学位。曾获得2009年第十九届欧洲机器学习大会最佳学生论文奖、2009年第二十六届国际机器学习大会奖学金、2008年马格德堡大学最佳毕业生奖、2008年国际机器学习Summer School奖学金、2006年德国萨哈森安哈特州教育与文化部优秀国际学生奖学金、2002年中国河南大专辩论赛最佳辩手、1999年建国五十周年演讲比赛安徽省安庆市三等奖。
电子邮件：roywwcheng@gmail.com　个人主页：www.chengweiwei.com