Studies in Computational Intelligence, Volume 255

Editor-in-Chief
Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series can be found on our homepage: springer.com

Vol. 234. Aruna Chakraborty and Amit Konar
Emotional Intelligence, 2009
ISBN 978-3-540-68606-4

Vol. 235. Reiner Onken and Axel Schulte
System-Ergonomic Design of Cognitive Automation, 2009
ISBN 978-3-642-03134-2

Vol. 236. Natalio Krasnogor, Belén Melián-Batista, José A. Moreno-Pérez, J. Marcos Moreno-Vega, and David Pelta (Eds.)
Nature Inspired Cooperative Strategies for Optimization (NICSO 2008), 2009
ISBN 978-3-642-03210-3

Vol. 237. George A. Papadopoulos and Costin Badica (Eds.)
Intelligent Distributed Computing III, 2009
ISBN 978-3-642-03213-4

Vol. 238. Li Niu, Jie Lu, and Guangquan Zhang
Cognition-Driven Decision Support for Business Intelligence, 2009
ISBN 978-3-642-03207-3

Vol. 239. Zong Woo Geem (Ed.)
Harmony Search Algorithms for Structural Design Optimization, 2009
ISBN 978-3-642-03449-7

Vol. 240. Dimitri Plemenos and Georgios Miaoulis (Eds.)
Intelligent Computer Graphics 2009, 2009
ISBN 978-3-642-03451-0

Vol. 241. János Fodor and Janusz Kacprzyk (Eds.)
Aspects of Soft Computing, Intelligent Robotics and Control, 2009
ISBN 978-3-642-03632-3

Vol. 242. Carlos Artemio Coello Coello, Satchidananda Dehuri, and Susmita Ghosh (Eds.)
Swarm Intelligence for Multi-objective Problems in Data Mining, 2009
ISBN 978-3-642-03624-8

Vol. 243. Imre J. Rudas, János Fodor, and Janusz Kacprzyk (Eds.)
Towards Intelligent Engineering and Information Technology, 2009
ISBN 978-3-642-03736-8

Vol. 244. Ngoc Thanh Nguyen, Rados law Piotr Katarzyniak, and Adam Janiak (Eds.)
New Challenges in Computational Collective Intelligence, 2009
ISBN 978-3-642-03957-7

Vol. 245. Oleg Okun and Giorgio Valentini (Eds.)
Applications of Supervised and Unsupervised Ensemble Methods, 2009
ISBN 978-3-642-03998-0

Vol. 246. Thanasis Daradoumis, Santi Caballé, Joan Manuel Marquès, and Fatos Xhafa (Eds.)
Intelligent Collaborative e-Learning Systems and Applications, 2009
ISBN 978-3-642-04000-9

Vol. 247. Monica Bianchini, Marco Maggini, Franco Scarselli, and Lakhmi C. Jain (Eds.)
Innovations in Neural Information Paradigms and Applications, 2009
ISBN 978-3-642-04002-3

Vol. 248. Chee Peng Lim, Lakhmi C. Jain, and Satchidananda Dehuri (Eds.)
Innovations in Swarm Intelligence, 2009
ISBN 978-3-642-04224-9

Vol. 249. Wesam Ashour Barbakh, Ying Wu, and Colin Fyfe
Non-Standard Parameter Adaptation for Exploratory Data Analysis, 2009
ISBN 978-3-642-04004-7

Vol. 250. Raymond Chiong and Sandeep Dhakal (Eds.)
Natural Intelligence for Scheduling, Planning and Packing Problems, 2009
ISBN 978-3-642-04038-2

Vol. 251. Zbigniew W. Ras and William Ribarsky (Eds.)
Advances in Information and Intelligent Systems, 2009
ISBN 978-3-642-04140-2

Vol. 252. Ngoc Thanh Nguyen and Edward Szczerbicki (Eds.)
Intelligent Systems for Knowledge Management, 2009
ISBN 978-3-642-04169-3

Vol. 253. Akitoshi Hanazawa, Tsutom Miki, and Keiichi Horio (Eds.)
Brain-Inspired Information Technology, 2009
ISBN 978-3-642-04024-5

Vol. 254. Kyandoghere Kyamakya, Wolfgang A. Halang, Herwig Unger, Jean Chamberlain Chedjou, Nikolai F. Rulkov, and Zhong Li (Eds.)
Recent Advances in Nonlinear Dynamics and Synchronization, 2009
ISBN 978-3-642-04226-3

Vol. 255. Catarina Silva and Bernardete Ribeiro
Inductive Inference for Large Scale Text Classification, 2009
ISBN 978-3-642-04532-5

Catarina Silva and Bernardete Ribeiro

Inductive Inference for Large Scale Text Classification

Kernel Approaches and Techniques

Springer

Catarina Silva
School of Technology and Management
Polytechnic Institute of Leiria
Alto do Vieiro, 2401-951 Leiria
Portugal
E-mail: catarina@estg.ipleiria.pt

Centre for Informatics and Systems
University of Coimbra
Polo II, 3030-290 Coimbra
Portugal
E-mail: catarina@dei.uc.pt

Bernardete Ribeiro
Centre for Informatics and Systems
Department of Informatics Engineering
University of Coimbra
Polo II, 3030-290 Coimbra
Portugal
E-mail: bribeiro@dei.uc.pt

ISBN 978-3-642-04532-5 e-ISBN 978-3-642-04533-2

DOI 10.1007/978-3-642-04533-2

Studies in Computational Intelligence ISSN 1860-949X

Library of Congress Control Number: Applied for

Typeset & *Cover Design:* Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed in acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

To Nuno and Miguel
To my family
Catarina Silva

To Miguel and Alexander
To my family
Bernardete Ribeiro

Preface

Motivation and Scope

Text classification is becoming a crucial task to analysts in different areas. In the last few decades the production of textual documents in digital form has increased exponentially. Their applications range from web pages to scientific documents, including emails, news and books. Searching for a digital text in Google is now more than a reality, it is a commonplace. In the near future, with the advent of intelligent text classification methods, people will have even more access to a large variety of enhanced digital text services, viz. filtering, searching and filing.

Despite the widespread use of digital texts, handling them is inherently difficult - the large amount of data necessary to represent them and the subjectivity of classification complicate matters. Earlier research has addressed the extraction of information from relatively small collections of well-structured documents such as news wires and scientific publications.

Intelligent text classification methods, which rely heavily on machine learning algorithms, have the potential to supersede existing information retrieval techniques and provide superior facilities that will save time and money for users and companies, while providing a vital tool for dealing with the proliferation of digital texts they are faced with.

The book is based on the PhD thesis of the first author and gives a concise view on how to use kernel approaches for inductive inference in large scale text classification; it presents a series of new techniques to enhance, scale and distribute text classification tasks. The approaches combine high performance and robustness with an extensive assessment of the suggested techniques.

The book is not intended to be a comprehensive survey of the state-of-the-art of the whole field of text classification. Its purpose is less ambitious and more practical: to explain and illustrate some of the important methods used in this field, in particular kernel approaches and techniques.

Challenges and Contributions

The relevant kernel approaches and techniques used to respond to some of the most prominent challenges are unfolded in this book, covering several facets of the whole problem.

Automatic text classification is an effervescent field of research. Many scientific and industrial fields generate enormous amounts of text data, such as news wires, microarray gene data and web pages. This trend seems to be spreading and there is no end in sight. Users are overwhelmed with the amount of information and thus need efficient, reliable and mostly intelligible text classification methods that they can relate to and understand. Another great challenge is the available knowledge that can be integrated by users and engineers in processing, learning and evaluation procedures.

From an academic point of view, putting together a probabilistic formulation which integrates the underlying knowledge of the problem at hand can also present an immense challenge. Acknowledging there is no free lunch i.e. that any two algorithms are equivalent when their performance is averaged across all possible problems, research can instead be focused on methods to tackle their inherent high dimensionality.

From an engineering perspective, dealing with the curse of dimensionality of text representations and learning is an interesting problem.

The content of this book sets out the importance of the challenges mentioned above and explains and illustrates the main approaches and techniques for dealing with them. The most relevant original contributions of this book are related to the challenges described above:

- **Empirical Evaluation of Text Pre-processing Methods.** We have undertaken an empirical study to compare the influence and relative importance of standard pre-processing dimensionality reduction methods in text classification performance. Low frequency word removal, stopword removal and stemming were tested and stopword removal was found to be the most useful technique to apply since it yields the best performing classifiers in all tested conditions. Stemming also plays an important role, especially in the precision of the classifiers. While stopword removal significantly alters the content of input data, stemming only alters its shape, i.e. the reduction of information is not significant. We can therefore say that stemming is more relevant in terms of efficiency of the learning machine (there is less redundancy in data). Low frequency word removal has little influence on classification performance, but it can decrease training complexity by reducing the number of features. A general conclusion is that the evaluated dimensionality reduction techniques can strengthen classifier performance.

- **Knowledge Integration in SVMs.** We have investigated the introduction of unlabeled data in the support vector machine (SVM) learning stage and the potential of using several learning machines organized in a

committee. We have presented two margin-based approaches to introduce unlabeled document information into the learning stage: background knowledge and active learning. We have also proposed an SVM ensemble with a two-step learning strategy using the separating margin as a differentiating factor in positive classifications.

Both the proposed enhancements to SVMs in text classification integrate new knowledge in the learning procedures and show improvements over the baseline SVMs. The separating margin plays a crucial role in both techniques and can be used in further enhancements.

- **Reducing the Dimensionality of RVMs.** Due to the poor scaling capabilities of the relevance vector machine (RVM) algorithm when faced with high-dimensional data sets, like text classification training sets, it is crucial to limit the training set dimension to a minimum. We have examined ways to reduce the dimensionality, viz. active learning and similitude measure between terms. We introduced an active learning RVM method based on the kernel trick. Using a kernel distance metric we have defined a higher-dimensional space where active examples are selected. Complexity escalation was controlled since the number of added documents was fixed and the kernel trick provides a simple strategy to determine those active documents. To reduce the number documents in the training set, we presented a two-step RVM : the first stage selects which training documents go to the next level, using a similitude measure between documents, based on the co-occurrence of words; the second step gathers all remaining documents and infers an RVM classifier. While maintaining the RVMs' sparseness, we still show competitive accuracy, as long as training examples are carefully established.

- **Divide-and-Conquer RVM Approaches.** To keep pursuing the scalability of RVMs, we have focused on three divide-and-conquer RVM methods: incremental, boosting and ensemble strategies. These methods rely on a selection of small working chunks from the training set and then explore different combining strategies that permit the use of all training examples in RVM expansion to large datasets. We demonstrated that it is possible to make use of an RVM's advantages, such as predictive distributions for testing instances and sparse solutions, while maintaining and even improving the classification performance. The proposed methods adapt RVMs to large scale text sets, maintaining their probabilistic Bayesian nature and providing sparse solutions.

- **Hybrid SVM-RVM.** SVMs and RVMs constitute two state-of-the-art learning machines that are currently the focus of cutting-edge research. SVMs present accuracy and complexity preponderance, but are surpassed by RVMs when probabilistic outputs or kernel selection come into the discussion. We have proposed a two-level hierarchical hybrid SVM-RVM

model to combine the best of both learning machines. The first level of the proposed model uses an RVM to determine the less confident classified examples and the second level uses an SVM to learn and classify the tougher examples. The hierarchical approach outperforms both baseline learning machines.

- **Deployment in Distributed Environments.** In cooperation with researchers from the Laboratory of Adaptive Systems and Parallel Processing of the University of Ljubljana, Slovenia, we have deployed text classification in distributed environments. This work was carried out both in a cluster in the University of Ljubljana, Slovenia, and in the Center for Informatics and Systems of the University of Coimbra, Portugal. The proposed deployment employs a combination of the text classification task (and data) decomposition, configuration evaluation through the modeling of the design phases, and a high performance distributed computing model. We have shown that it is not only possible, but also advantageous to deploy text classification in a cluster environment, while using available middleware distributed platforms and existing sequential code.

- **Text Classification Framework.** In the last chapter of the book, we propose and discuss a framework for inductive inference-based text classification using kernels methods. The main components of this development, beyond the generally good performances of kernel methods on real-world problems and ease of use provided by current implementations, involve active learning, ensembles, incremental learning, boosting and knowledge integration. Finally, some especially promising lines of work will open windows to further research in the field.

Plan and Organization

This book has six chapters and two appendices. The chapters are organized into two parts: the first part relating to fundamental topics encloses the first two chapters and the second part concerning approaches and techniques includes the other four chapters.

Chapter 1 contains background material on text classification. In particular we review document corpora, representations, reduction methods, classifiers, and evaluation techniques.

Chapter 2 introduces the concept of kernel methods, and summarizes into a single framework the foundations of two paradigmatic techniques: support vector machines and relevance vector machines. Both approaches are introduced in a text classification perspective, along with results and comparisons of their application to benchmark corpora.

Chapter 3 discusses the learning techniques developed to integrate knowledge in the classification task in order to improve the performance of support vector machines (SVMs) in text classification applications.

Chapter 4 explores relevance vector machines (RVMs) and their application to text classification. We propose new approaches to tackle RVMs' scaling problems. In particular we examine techniques that reduce the dimensionality of the problem and we introduce incremental, boosting and ensemble divide-and-conquer strategies. Finally, a hybrid RVM-SVM combination is presented that substantially improves baseline results.

Chapter 5 describes the deployment of text classification in cluster environments, using a distributed system to optimize the procedures involved. In this Chapter we look at various ways to deploy complete text classification systems in distributed environments, employing a combination of the text classification task (and data) decomposition, configuration evaluation through the modeling and the design phases, and a high performance distributed computing model.

In the final Chapter (6) we propose a framework for inductive inference text classification and present a unified view of the field across four stages of design: sources, preprocessing, learning and evaluation. We outline each of the framework phases in a coherent result which can be a guide to real-world applications. In addition we present research trends for future work with special focus in the area of web applications and information technology.

Audience

The book is designed for practitioners and researchers and is suitable for postgraduate students in computer science, engineering, information technology and other related disciplines. Some knowledge in the area of machine learning and computational intelligence will be beneficial.

Acknowledgments

We would like to acknowledge and thank all those who have contributed to bringing this book to publication for their help, support, and input.

We would also like to acknowledge and thank Professor Andrej Dobnikar and Dr. Uroš Lotrič and everyone else in the Laboratory for Adaptive Systems and Parallel Processing of the Faculty of Computer and Information Science, University of Ljubljana, Slovenia, especially Dr. Branko Šter, for the fruitful cooperation and stimulating discussions on distributed text classification systems. They have made an invaluable contribution to the book.

We also wish to thank the support of the School of Technology and Management of the Polytechnic Institute of Leiria and of the Centre of Informatics and Systems of the Informatics Engineering Department, Faculty of Science and Technologies, University of Coimbra, for the means provided during the research.

Our thanks also to Jean Burrows who reviewed the syntactic aspects of the book.

Our special thanks and appreciation to our editor, Professor Janusz Kacprzyk, of Studies in Computational Intelligence, Springer, for his essential encouragement.

Lastly, to our families and friends for all their love and support.

July 2009 Catarina Silva
Coimbra, Portugal Bernardete Ribeiro

Contents

Acronyms

ARD	Automatic Relevance Determination
AUC	Area Under the Curve
BEP	Break-Even Point
BOW	Bag-Of-Words
DAG	Direct Acyclic Graph
DF	Document Frequency
DNF	Disjunctive Normal Form
EM	Expectation-Maximization
ERM	Empirical Risk Minimization
FN, c	False Negative
FNR	False Negative Rate
FP, b	False Positive
FPR	False Positive Rate
FTP	File Transfer Protocol
HTC	High Throughput Computing
HTTP	Hypertext Transfer Protocol
IDF	Inverse Document Frequency
K-NN	K-Nearest Neighbour
KDA	Kernel Discriminant Analysis
KPCA	Kernel - PCA
LDA	Linear Discriminant Analysis
LR	Logistic Regression
LSI	Latent Semantic Indexing
MeSH	Medical Subject Headings
NASA	National Aeronautics and Space Administration
NB	Naïve Bayes
NE	Named-entities
NLP	Natural Language Processing
NN	Neural Network
OSH	Optimal Separating Hyperplane
P	Precision

PC	Personal Computer
PCA	Principal Components Analysis
POS	Part-of-speech
QBC	Query-By-Committee
R	Recall
RBF	Radial Basis Function
RCV1	Reuters Corpus Volume 1
ROC	Receiver Operating Characteristic
RV	Relevance Vector
RVM	Relevance Vector Machine
SETI	Search for ExtraTerrestrial Intelligence
SGML	Standard Generalized Markup Language
SRM	Structural Risk Minimization
SV	Support Vector
SVM	Support Vector Machine
TF	Term Frequency
TFIDF	Term Frequency - Inverse Document Frequency
TN, d	True Negative
TP, a	True Positive
TSVM	Transductive SVM
VC-dimension	Vapnik-Chervonenkis dimension

Notation

$\|.\|$	Number of elements in a set
$\|\|.\|\|$	$L_2 - norm$
$[.]$	Integer part of .
$.^T$	Transpose
α	Lagrange multipler in SVMs optimization
α	Precision hyperparameter in RVMs optimization
$\boldsymbol{\alpha}$	Set (vector) of α
c_j	Category
$\mathscr{C}$	Set of categories
C	Regularization parameter in SVMs
b	bias
$\mathbf{d}, \mathbf{d}_i$	Document
$\mathscr{D}$	Set of documents
$F()$	Distribution function
$F\beta, F1$	van Rijsbergen's measure
h	Hypothesis
H	Hypothesis space
$i.i.d.$	Independent and identically distributed
$k(.), \phi(.), \Phi(.)$	Kernel functions
λ, β, γ	Parameters for Rocchio's method
Neg	Negative examples, not belonging to a category
$O(.)$	Order of
p	Number of computing nodes
$P(.)$	Probability
$p(.)$	Probability function
$p(.\|.)$	Conditional probability
$\boldsymbol{\Psi}$	Design matrix
Pos	Positive examples, belonging to a category
$\boldsymbol{\varphi}$	Set of RVs
$\boldsymbol{\varphi}_i$	RV
q	Query

$\mathbf{Q}$	Hessian matrix
ρ	Separating margin of SVMs
s_i	Complexity parameters for phases in distributed environment
S_{ij}	Similitude measure between documents $\mathbf{d}_i$ and $\mathbf{d}_j$
$\sigma(.)$	Sigmoid function
t	Target
τ_{task}	Time to complete a task
$\mathbf{u}_j$	Unlabeled document
$\mathbf{U}$	Set of unlabeled documents
$\mathscr{W}$	Set of features (words or terms), dictionary
w_k	Word or term
w_{ik}	Value representing word w_k in a document $\mathbf{d}_i$
ω_k	Weight of term w_k for a given model
ω_{ik}	Weight of term w_k in document $\mathbf{d}_i$ for a given model
ω_{qk}	Weight of term w_k in query q
$\boldsymbol{\omega}$	Set of weights that define a model
y	Output of a model