Ann Macintosh,

Richard Ellis

and Tony Allen (Eds)

APPLICATIONS AND INNOVATIONS IN INTELLIGENT SYSTEMS XII

Proceedings of AI-2004, the Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence

BCS CONFERENCE SERIES

BCS

SGAI

AI 2004

Ann Macintosh,

Richard Ellis

and Tony Allen (Eds)

APPLICATIONS AND INNOVATIONS IN INTELLIGENT SYSTEMS XII

Proceedings of AI-2004, the Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence

Springer

BCS CONFERENCE SERIES

Applications and Innovations in Intelligent Systems XII

Ann Macintosh, Richard Ellis and
Tony Allen (Eds)

Applications and Innovations in Intelligent Systems XII

Proceedings of AI-2004, the Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence

Springer

Professor Ann Macintosh, BSc, CEng
Napier University, Edinburgh, EH10 5DT, UK

Richard Ellis, BSc, MSc
Stratum Management Ltd, UK

Dr Tony Allen
Nottingham Trent University

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 1-85233-908-X
Springer is part of Springer Science+Business Media
springeronline.com

Printed in Great Britain

Typesetting: Camera-ready by editors
Printed and bound at the Athenæum Press Ltd., Gateshead, Tyne & Wear
34/3830-543210 Printed on acid-free paper SPIN 11006725

APPLICATION PROGRAMME CHAIR'S INTRODUCTION

A. L. Macintosh, Napier University, UK

The papers in this volume are the refereed application papers presented at ES2004, the Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, held in Cambridge in December 2004. The conference was organised by SGAI, the British Computer Society Specialist Group on Artificial Intelligence.

This volume contains twenty refereed papers which present the innovative application of a range of AI techniques in a number of subject domains. This year, the papers are divided into sections on Synthesis and Prediction, Scheduling and Search, Diagnosis and Monitoring, Classification and Design, and Analysis and Evaluation

This year's prize for the best refereed application paper, which is being sponsored by the Department of Trade and Industry, was won by a paper entitled "A Case-Based Technique for Tracking Concept Drift in Spam Filtering". The authors are Sarah Jane Delany, from the Dublin Institute of Technology, Ireland, and Pádraig Cunningham, Alexey Tsymbal, and Lorcan Coyle from Trinity College Dublin, Ireland.

This is the twelfth volume in the *Applications and Innovations* series. The Technical Stream papers are published as a companion volume under the title *Research and Development in Intelligent Systems XXI.*

On behalf of the conference organising committee I should like to thank all those who contributed to the organisation of this year's application programme, in particular the programme committee members, the executive programme committee and our administrators Linsay Turbert and Collette Jackson.

Ann Macintosh
Application Programme Chair, AI-2004

ACKNOWLEDGEMENTS

AI-2004 CONFERENCE COMMITTEE

Dr. Tony Allen, Nottingham Trent University	(Conference Chair)
Dr. Robert Milne, Sermatech Intelligent Applications Ltd	(Deputy Conference Chair, Finance and Publicity)
Dr. Alun Preece, University of Aberdeen	(Deputy Conference Chair, Electronic Services)
Dr. Nirmalie Wiratunga, Robert Gordon University, Aberdeen	(Deputy Conference Chair, Poster Session)
Prof. Adrian Hopgood Nottingham Trent University	(Tutorial Organiser)
Prof. Ann Macintosh Napier University	(Application Programme Chair)
Richard Ellis Stratum Management Ltd	(Deputy Application Programme Chair)
Prof. Max Bramer University of Portsmouth	(Technical Programme Chair)
Dr Frans Coenen, University of Liverpool	(Deputy Technical Programme Chair)
Dr. Bob Howlett, University of Brighton	(Exhibition Organiser)
Rosemary Gilligan University of Hertfordshire	(Research Student Liaison)

APPLICATIONS EXECUTIVE PROGRAMME COMMITTEE

Prof. Ann Macintosh, Napier University (Chair)
Richard Ellis, Stratum Management Ltd (Vice-Chair)
Dr Robert Milne, Sermatech Intelligent Applications Ltd
Richard Wheeler, University of Edinburgh
Alan Montgomery, InferMed Ltd
Rosemary Gilligan, University of Hertfordshire

APPLICATIONS PROGRAMME COMMITTEE

Nick Adams (Napier University)

Dan Allsopp (University of Portsmouth)

Victor Alves (Universidade do Minho)

Li Bai (University of Nottingham)

Euan Davidson (University of Strathclyde)

Argiris Dentsoras (University of Patras)

Richard Ellis (Stratum)

Florentino Fdez-Riverola (ESEI)

Rosemary Gilligan (University of Hertfordshire)

John Gordon (Applied Knowledge Research Institute)

Daniel Kerr (BUPA)

John Kingston (University of Edinburgh)

Paul Leng (University of Liverpool)

Shuliang Li (University of Westminster)

Ann Macintosh (Napier University)

Rob Milne (Sermatech Intelligent Applications Ltd)

Alan Montgomery (InferMed Ltd)

Pavlos Moraitis (University of Cyprus)

Kushan Nammun

Gilbert Owusu (BT)

Paul Thomas (Dstl)

Botond Virginas (BT)

Cornelius Weber (University of Sunderland)

CONTENTS

BEST APPLICATION PAPER

A Case-Based Technique for Tracking Concept Drift in Spam Filtering

Sarah Jane Delany[1], Pádraig Cunningham[2], Alexey Tsymbal[2], Lorcan Coyle[2]
[1]Dublin Institute of Technology, Kevin St., Dublin 8, Ireland.
[2]Trinity College Dublin, College Green, Dublin 2, Ireland.

Abstract

Clearly, machine learning techniques can play an important role in filtering spam email because ample training data is available to build a robust classifier. However, spam filtering is a particularly challenging task as the data distribution and concept being learned changes over time. This is a particularly awkward form of concept drift as the change is driven by spammers wishing to circumvent the spam filters. In this paper we show that lazy learning techniques are appropriate for such dynamically changing contexts. We present a case-based system for spam filtering called ECUE that can learn dynamically. We evaluate its performance as the case-base is updated with new cases. We also explore the benefit of periodically redoing the feature selection process to bring new features into play. Our evaluation shows that these two levels of model update are effective in tracking concept drift.

1 Introduction

With the cost of spam to companies worldwide estimated to be ca. $20 billion a year and growing at a rate of almost 100% a year [1], spam is a problem that needs to be handled. It is a challenging problem for a number of reasons. One of the most testing aspects is the dynamic nature of spam. Something of an arms race has emerged between spammers and the spam filters used to combat spam. As filters are adapted to contend with today's types of spam emails, the spammers alter, obfuscate and confuse filters by disguising their emails to look more like legitimate email. This dynamic nature of spam email raises a requirement for update in any filter that is to be successful over time in identifying spam.

Lazy learning is good for dynamically changing situations. With lazy learning the decision of how to generalise beyond the training data is deferred until each new instance is considered. In comparison to this, eager learning systems determine their generalisation mechanism by building a model based on the training data in advance of considering any new instances. In this paper we explore the application of Case-Based Reasoning (CBR), a lazy machine learning technique, to the problem of spam filtering and present our CBR system – Email Classification Using Examples (ECUE). We concentrate in this paper on evaluating how ECUE can assist with the concept drift that is inherent in spam.

CBR offers a number of advantages in the spam filtering domain. Spam is a disjoint concept in that spam selling cheap prescription drugs has little in common with spam offering good mortgage rates. Case-based classification works well for disjoint concepts whereas Naïve Bayes, a machine learning technique that is popular for text classification, tries to learn a unified concept description.

In addition, there is a natural hierarchy of learning available to a CBR system where the simplest level of learning is to simply update the case-base with new instances of spam or legitimate email. The advantage of CBR in this first level of learning is that it requires no rebuild of the model as is necessary with other machine learning solutions to spam filtering. The second level of learning is to retrain the system by re-selecting features that may be more predictive of spam. This level of retraining can be performed infrequently and based on newer training data. The highest level of learning, performed even more infrequently than feature selection, is to allow new feature extraction techniques to be added to the system. For instance, when domain specific features are used in the system, new feature extraction techniques will allow new features to be included. In ECUE we use a similarity retrieval algorithm based on Case Retrieval Nets (CRN) [2] which is a memory structure that allows efficient and flexible retrieval of cases. The benefit of using the CRN for implementing the second and third levels of learning is that it can easily handle cases with new features; the fact that these features may be missing on old cases is not a problem.

This paper begins in Section 2 with an overview of other work using machine learning techniques for filtering spam. Section 3 discusses the problem of concept drift and existing techniques for handling concept drift. Our case-based approach to spam filtering, ECUE, is presented in Section 4, while Section 5 presents our evaluation and results of using CBR to combat the concept drift in spam. Section 6 closes the paper with our conclusions and directions for future work.

2 Spam Filtering and Machine Learning

Existing research on using machine learning for spam filtering has focussed particularly on using Naïve Bayes [3-7]. In addition there has been work using Support Vector Machines (SVMs) [3,8,9] and Latent Semantic Indexing [10]. There has also been research using memory based classifiers [4,11]. However, this work does not address the issue of concept drift which is inherent in spam and the evaluations have been on static datasets.

One technique used for tracking concept drift is ensemble learning (see Section 3.2) which uses a set of classifiers whose individual results are combined to give an overall classification. There has been some work on ensemble learning in spam filtering using boosting [1,12] and also a combination of Naïve Bayes and memory based classifiers [13]. These evaluations use static data sets and do not attempt to track concept drift.

3 The Problem of Concept Drift

This section defines concept drift and discusses approaches to handling concept drift.

3.1 Definitions and Types of Concept Drift

A difficult problem with learning in many real-world domains is that the concept of interest may depend on some *hidden context*, not given explicitly in the form of predictive features. Typical examples include weather prediction rules that may vary radically with the season or the patterns of customers' buying preferences that may change, depending on the current day of the week, availability of alternatives, inflation rate, etc. Often the cause of change is hidden, not known in advance, making the learning task more complicated. Changes in the hidden context can induce changes in the target concept, which is generally known as *concept drift* [13]. An effective learner should be able to track such changes and to quickly adapt to them.

Two kinds of concept drift that may occur in the real world are normally distinguished in the literature: (1) sudden (abrupt, instantaneous), and (2) gradual concept drift. For example, someone graduating from college might suddenly have completely different monetary concerns, whereas a slowly wearing piece of factory equipment might cause a gradual change in the quality of output parts [15]. Stanley [15] divides gradual drift into moderate and slow drifts, depending on the rate of change.

Hidden changes in context may not only be a cause of a change in the target concept, but may also cause a change in the underlying data distribution. Even if the target concept remains the same but the data distribution changes, a model rebuild may be necessary as the model's error may no longer be acceptable. This is called *virtual concept drift* [16]. Virtual concept drift and real concept drift often occur together or virtual concept drift alone may occur, e.g. in the case of spam categorization. From a practical point of view it is not important what kind of concept drift occurs as in all cases, the current model needs to be changed.

3.2 Approaches to Handling Concept Drift

There are three approaches to handling concept drift: (1) instance selection; (2) instance weighting; and (3) ensemble learning (or learning with multiple concept descriptions). In instance selection, the goal is to select instances relevant to the current concept. The most common concept drift handling technique is based on instance selection and involves generalizing from a *window* that moves over recently seen instances and uses the learnt concepts for prediction only in the immediate future. Examples of window-based algorithms include the FLORA family of algorithms [13], FRANN [17], and Time-Windowed Forgetting (TWF) [18]. Some algorithms use a window of fixed size, while others use heuristics to adjust the window size to the current extent of concept drift, e.g. "Adaptive Size" [19] and FLORA2 [13]. Many case-base editing strategies in case-based reasoning that delete noisy, irrelevant and redundant cases are also a form of instance selection [20]. Batch