VILNIUS UNIVERSITY

Indrė Žliobaitė

# ADAPTIVE TRAINING SET FORMATION

Doctoral dissertation
Physical sciences, informatics (09P)

Vilnius, 2010

The dissertation work was carried out at Vilnius University from 2006 to 2010 in cooperation with Bangor University (UK) and Eindhoven University of Technology (the Netherlands) researchers.

Scientific supervisor:

prof. habil. dr. Šarūnas Raudys (Vilnius University, physical sciences, informatics - 09P)

VILNIAUS UNIVERSITETAS

Indrė Žliobaitė

# ADAPTYVUS MOKYMO IMTIES FORMAVIMAS

Daktaro disertacija
Fiziniai mokslai, informatika (09P)

Vilnius, 2010

Disertacija rengta 2006 - 2010 metais Vilniaus universitete bendradarbiaujant su Bangoro universiteto (Didžioji Britanija) ir Eindhoveno technologijų universiteto (Nyderlandai) mokslininkais.

Mokslinis vadovas:

prof. habil. dr. Šarūnas Raudys (Vilniaus universitetas, fiziniai mokslai, informatika - 09P)

# Acknowledgements

Acknowledgements go here.

Indrė Žliobaitė
Vilnius
25th December 2013

# Table of Contents

# Notation

$A$ amplitude caused by the feeding screw
$a$ acceleration of a mass change
$\alpha_1$ the weight of distance in space
$\alpha_2$ the weight of distance in time

# Chapter 1

# Introduction

We live in a dynamic world, where changes are a part of everyday life. When there is a shift in data, the classification or prediction models need to be adaptive to the changes. In data mining the phenomenon of change in data over time is known as *concept drift*.

In this thesis we consider supervised learning under concept drift. In particular, we are interested in the training set formation strategies, which lead to achieving adaptivity to concept drift.

In this chapter we depict the research area, give motivating application examples and present the research outline and methodology. We define and narrow down a specific research problem and formulate the research questions. We finish the chapter by outlining and depicting the structure of the thesis.

## 1.1 Application Examples

Changes in underlying data might occur due to changing personal interests, changes in population, adversary activities or they can be attributed to a complex nature of the environment. Consider three motivating examples illustrating *a concept drift* phenomenon.

**Example 1.1.1** (Fraud detection). The estimated loss of UK issued credit cards amounts to 610 million pounds in year 2008 [3]. These costs are born by the banking industry and indirectly by users via increased premium for card insurance. The financial institutions employ filters to data mine and monitor daily transactions.

The classification decisions need to be made online, and the behavior of both legitimate users and criminals is shifting. Moreover, the criminals might try using adversary tactics to overcome the fraud detection mechanisms. Therefore the filters used for monitoring need to be reactive to changing adversary behavior to keep the classification accuracy. Similar motivation applies to computer network security monitoring, e-mail spam filtering. □

**Example 1.1.2** (News recommendation). Kate reads news on the internet every day. She uses a news recommender system, which provides a ranked list of headlines of potential interest to Kate. The recommender model is constantly updated using the records of her browsing history.

Kate has different interests. She likes Formula-1 sport, thus she reads the overviews of the races every second or third Monday, but not in winter when there are no races (long term interest). Recently she got an assignment at work to write a review on meat prices in New Zealand (short term temporal interest). She is also thinking if it is the right time to purchase a flat, thus her interest in real estate market situation has recently been increasing (gradual increase in interest).

The learning models in the news recommender system need to be adaptive over time to take into account short and long term interests, sudden and gradual changes. □

**Example 1.1.3** (Navigation). The DARPA Grand Challenge is a prize competition for driverless cars [4].

> This event required teams to build an autonomous vehicle capable of driving in traffic, performing complex maneuvers such as merging, passing, parking and negotiating intersections. $< \ldots >$

> The autonomous vehicles have interacted with both manned and unmanned vehicle traffic in an urban environment. $< \ldots >$ Robots were also being judged on their ability to follow California driving rules.

The sensors in the vehicles are monitoring the road conditions and classifying them for the selection of a driving mode. The road is changing, the decisions need to be made in real time and it is not possible to account for every possible combination of road changes in advance. Therefore the winning entry in year 2005 'Stanley' [11] was equipped with an adaptive learner (adaptive Mixture of Gaussians).

Obviously, unmanned vehicles are not limited to competitions. They are irreplaceable in situations where it is dangerous (e.g. ecological accidents), infeasible (e.g. in space) to employ a human driver. □

These application examples give a motivation for the learners to be equipped with the concept drift adaptation mechanisms. The need for adaptation mechanisms in data mining are discussed in a number of position papers [5–8, 10, 12] for about a decade. In the next section we take a closer look, what adaptation mechanisms mean.

# Chapter 2

# The Context of a Concept Drift Problem

In this chapter we present a context of concept drift problem. We focus on the issues relevant to adaptive training set formation. We present the framework and terminology, and formulate a taxonomy of concept drift learners design.

# Appendix A

# Algorithms

In this Appendix we present pseudo codes and the settings used for the peer algorithms, which we implemented and used in experimental evaluation through the thesis. For consistency, the algorithms were named using the first three letters of the surname of the first author.

# Appendix B

# Derivations and Computation Details

## B.1 Change Detection Using Hotelling T-test

Suppose that we have the labeled sequence of observations labeled in $c$ classes. To estimate the likelihood of a change at time $d$, where $1 \leq d \leq t$, we assume that the class means migrate independently of one another. Then the probability that there is a change at time $d$ is

$$P(\text{change}|d) = 1 - \prod_{k=1}^{c} P(\text{no change in } \mu_k|d),$$

where $\mu_k$ is the mean for class $k$, $k = 1, \ldots, c$. Given that the data lives in $\Re^n$, the value of $P(\text{no change in } \mu_k|d)$ can be estimated using the p-value of the Hotelling multivariate $T^2$-test. This test compares the means for class $k$ before and after the hypothetical change at $d$.

The mean before the change, $\mu_k^{(1)}$, is estimated from the streaming data from 1 to $d$ labeled in class $k$; the mean after the change, $\mu_k^{(2)}$, is estimated from the streaming data from $d+1$ to $t$ labeled in class $k$. Let $\hat{\Sigma}$ be the unbiased pooled covariance matrix estimated from the data and $N_1$ and $N_2$ be the respective sizes

of the two samples. The $T^2$ statistic is

$$T^2 = \frac{(N_1 + N_2 - n - 1)}{n(N_1 + N_2 - 2)} \frac{N_1 N_2}{(N_1 + N_2)} \hat{\delta}^2 \tag{B.1}$$

where $\hat{\delta}^2$ is the squared Euclidean distance $\hat{\delta}^2 = \left(\hat{\mu}_k^{(1)} - \hat{\mu}_k^{(2)}\right)^T \hat{\Sigma}^{-1} \left(\hat{\mu}_k^{(1)} - \hat{\mu}_k^{(2)}\right)$.

If the two means come from the same distribution, $T^2$ has $F$-distribution with degrees of freedom $(n, N_1 + N_2 - 1 - n)$.

If we use the notation $p_k(d)$ as the $p$-value returned by the Hotelling $T^2$-test comparing class $k$ samples before and after time moment $d$, the probability of change at $d$ can be estimated as

$$P(\text{change}|d) = 1 - \prod_{i=1}^{c} p_k(d). \tag{B.2}$$

To arrive at a probability distribution over $t_1, \ldots, t_i$, we can normalize by dividing the conditional probability (**??**) by

$$P(\text{change}, t_k) = \frac{P(\text{change}|t_k)}{\sum_{d=1}^{i} P(\text{change}|t_d)}. \tag{B.3}$$

In order to calculate the distribution, we need to check all possible split points between $t_1$ and $t_n$, as done in other studies performing retrospective change detection [1, 2, 9].

# Bibliography


[1] R. Adams and D. MacKay. Bayesian online changepoint detection. Technical report, University of Cambridge Technical Report, 2007.

[2] A. Bifet and R. Gavalda. Learning from time-changing data with adaptive windowing. In *Proc. of SIAM int. conf. on Data Mining (SDM'07)*, pages 443–448. SIAM, 2007.

[3] Card Watch. Card fraud overview. online, accessed July 15, 2009. URL `http://www.cardwatch.org.uk`.

[4] DARPA. Urban challenge. online, accessed July 15, 2009. URL `http://www.darpa.mil/grandchallenge/index.asp`.

[5] G. Dong, J. Han, L. Lakshmanan, J. Pei, H. Wang, and P. Yu. Online mining of changes from data streams: Research problems and preliminary results. In *Proc. of the SIGMOD2003 workshop on Management and Processing of Data Streams*, 2003.

[6] J. Han and J. Gao. Research challenges for data mining in science and engineering. In *Next Generation of Data Mining*. Chapman & Hall, 2009.

[7] D. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006.

[8] M. Kelly, D. Hand, and N. Adams. The impact of changing populations on classifier performance. In *KDD '99: Proc. of the 5th ACM SIGKDD int. conf. on Knowledge discovery and data mining*, pages 367–371. ACM, 1999.

[9] D. Kifer, Sh. Ben-David, and J. Gehrke. Detecting change in data streams. In *Proc. of the 30th int. conf. on Very Large Data Bases (VLDB 2004)*, pages 180–191, 2004.

[10] H. Kriegel, K. Borgwardt, P. Kroger, A. Pryakhin, M. Schubert, and A. Zimek. Future trends in data mining. *Data Mining and Knowledge Discovery*, 15(1):87–97, 2007.

[11] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel,

P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L.-E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. van Niekerk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehler, A. Nefian, and P. Mahoney. Winning the darpa grand challenge. *Journal of Field Robotics*, 23(9):661–692, 2006.

[12] G. Webb, M. Pazzani, and D. Billsus. Machine learning for user modeling. *User Modeling and User-Adapted Interaction*, 11(1-2):19–29, 2001.

# Publications by the Author

## Periodic

1. Kuncheva, L.I. and Žliobaitė, I. (2009). On the Window Size for Classification in Changing Environments. Intelligent Data Analysis 13(6), p. 861-872. ISSN:1088-467X [ISI]

2. ...

## Peer reviewed conference publications

9. Bakker, J., Pechenizkiy, M, Žliobaitė, I., Ivannikov, A. and Karkkainen, T. (2009). Handling Outliers and Concept Drift in Online Mass Flow Prediction in CFB Boilers. Proc. of the 3rd int. workshop on Knowledge Discovery from Sensor Data (SensorKDD-09), p. 13-22, ACM. ISBN:978-1-60558-668-7 [The Best Paper Award]

10. ...

# Curriculum Vitae

Indrė Žliobaitė graduated from ...

# Vocabulary - Žodynėlis

base learner - bazinis klasifikatorius
baseline - bazinis metodas
change point - pokyčio taškas
concept drift - koncepcijos pokytis
context aware - kontekstinis
data mining - duomenų gavyba
data source - duomenų šaltinis
gradual drift - palaipsnis pokytis
instance - vektorius
instance based learning - mokymas pagal vektorius
label - klasė
moving average - slenkantis vidurkis
peer methods - lyginamieji metodai
recurring concepts - pasikartojantis pokytis (pasikartojančios koncepcijos)
sequential learning - mokymas paeiliui
source - šaltinis
sudden drift - staigus pokytis
supervised learning - mokymas su mokytoju
training window - mokymo langas
unsupervised learning - mokymasis

# Summary in Lithuanian (Santrauka)

Tiriamojo darbo objektas yra adaptyvūs mokymo metodai, kurie remiasi kryptingu mokymo imties formavimu. Patobulintos žinomos mokymo strategijos esant staigiems, palaipsniams ir pasikartojantiems pokyčiams. Sukurti ir eksperimentiškai aprobuoti keturi adaptyvaus mokymo imties formavimo algoritmai, kurie leidžia pagerinti klasifikavimo bei prognozavimo tikslumą besikeičiančiose aplinkose, esant atitinkamai kiekvienam iš trijų pokyčių tipų. Naudojant generuotus bei realius duomenis, eksperimentiškai parodytas klasifikavimo bei prognozavimo tikslumo pagerėjimas, lyginant su visų istorinių duomenų naudojimu mokymui, bei žinomais šioje srityje naudojamais adaptyviais mokymo algoritmais. Sukurta metodika pritaikyta pramoninio katilo atvejui, jungiančiam kelis aplinkos pokyčių tipus.

...