DEPARTMENT OF
COMPUTER
SCIENCE

Professor Marta Kwiatkowska
Fellow of Trinity College
Direct Line Tel: +44 (0)1865 283509
Email: marta.kwiatkowska@cs.ox.ac.uk

Personal Assistant: Anita Hancox
Direct Line Tel: +44 (0)1865 610754
Email: anita.hancox@cs.ox.ac.uk

30 July 2021

Prof. Pawel Idziak
Przewodniczacy Rady Dyscypliny Informatyka UJ
ul Lojasiewicza 6
30-348 Krakow
Poland

Recenzja dorobku naukowego w postępowaniu o nadanie stopnia doktora habilitowanego dr. Markowi Śmieji.

Introduction

This review is in response to the invitation nr 1198.5110.1-7.2021 by Prof. Pawel Idziak, Przewodniczacy Rady Dyscypliny Informatyka UJ. The aim is to evaluate the scientific achievements of Dr Marek Smieja, as well as Dr Smieja's teaching and related activities, according to the guidelines for awarding the habilitation degree, and specifically art. 221 ust 4 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie i nauce (Dz. U. z 2020 r. Poz. 85 z późn. Zm.). Dr Smieja was awarded a doctorate in 2015 by the Institute of Computer Science and Computational Mathematics, Jagiellonian University, where he is now Assistant Professor (adiunkt). I would like to declare that I have not met Dr Smieja and do not have a conflict of interest in writing this reference.

The documentation provided included (1) a summary and detailed description of the scientific achievements published in 10 thematically related papers that appeared in international peer reviewed journals and conferences (art. 219 para 1 point 2), (2) scientific activity as confirmed by internships, particularly abroad, and research grants (art. 219 para 1 point 3), and (3) teaching and other organizational activities (as per guidelines of Rady Doskonalosci Naukowej). I will comment on these in turn.

(1) Scientific achievements published in 10 thematically related papers on machine learning with incomplete data

The scientific achievements included in the case for habilitation were published in 10 papers (all CORE A/A*) written in English that appeared in international journals and conferences between 2018 and 2020. They all concerned formulating and implementing machine learning methods so that they can handle missing data in a principled way. Machine learning, and particularly deep learning, has made great strides into real-world applications in the past decade or so, recognized through the 2018 Turing Award for the inventors of deep learning. However, most of the established methods concern supervised learning, which rely on access to large sets of labelled data, often requiring a laborious, manual process. This habilitation addresses the more challenging semi-supervised learning problem, which aims to provide techniques to enable learning directly from incomplete datasets. This problem is important for real-world applications, for example healthcare records or molecular data, where providing completely labelled data sets may not be feasible. The challenge considered here is therefore dealing with missing information, such as missing attributes or incomplete class labels, without compromising accuracy and performance.

DEPARTMENT OF
COMPUTER
SCIENCE

Professor Marta Kwiatkowska
Fellow of Trinity College
Direct Line Tel: +44 (0)1865 283509
Email: marta.kwiatkowska@cs.ox.ac.uk

Personal Assistant: Anita Hancox
Direct Line Tel: +44 (0)1865 610754
Email: anita.hancox@cs.ox.ac.uk

At the technical level, the work can be essentially split into two main groups: (i) adaptation of discriminative clustering to deal with incomplete data with additional constraints (and specifically pairwise must/cannot-link constraints, grouping by similarity, or partial labels) and (ii) adaptation of kernel methods and neural networks so that missing values are replaced with appropriate estimates of expected values (induced by the kernel or from the neighbouring neurons).

Having read the detailed summary and inspected the publications, I will comment on a few specific contributions in more detail. Within grouping (i), I would like to highlight a range of innovative techniques that were used to improve the performance of clustering, such as Siamese networks [A2] (neural network pairs that use the same weights to produce comparable output, here used to derive must-link/cannot-link constraints of unlabelled inputs) and the semi-supervised Gaussian mixture clustering model [A5] for partially classified data. In [A6], the focus is on generative models, and a semi-supervised generative modelling approach is developed for the case where labels are partial and continuous; by working with a mixture of Gaussians as a (joint) distribution in the latent space it is possible to achieve smooth interpolation between data points, which enables model exploration around selected data points and style transfer. The method is implemented in an autoencoder framework and optimized to work with closed-form expressions, improving computational efficiency.

Publications in grouping (ii) are focused on principled methods to replace the missing data points, and are based on a very natural and appealing idea that one should replace the loss function with the expectation derived from the probability density of the input data (conditioned on the observed values for the missing data). Since the precise input distributions are typically not available, instead of working with the loss functions Dr Smieja relies on well justified estimates derived, for example, from Gaussian mixture models. In [A7] this methodology is developed for kernel methods (SVM), where it is proved that the reformulated kernel yields the expected value under the assumption that the missing data follows a Gaussian distribution. In [A8], the methodology is adapted to neural networks, where, importantly, analytical formulations are given for selected activity functions (including ReLU) and proved to yield reliable expectations, thus also improving computational performance by reducing the sampling requirements.

Summarising, the 10 submitted publications represent an original and valuable scientific contribution to the field of machine learning, with a route towards industrial applications. It is a commendable strength of Dr Smieja's results that the proposed methods are theoretically justified, with proofs of convergence, and thus ensure improved robustness of the model predictions in addition to facilitating model training. Further, they are also implemented and experimentally validated, showing strong performance compared to state of the art. His summary (Section 4.3) serves as evidence that he has novel ideas for how to develop this line of work further, both theoretically and towards applications. The publications have appeared in journals and conferences of high international standing; a highlight is a publication in the top machine learning conference, NeurIPS 2018. They are mostly as first author (9/10) and the statement of contributions and bibliometric information are sufficiently high to warrant a positive evaluation of both his individual contribution, as well as

DEPARTMENT OF
COMPUTER
SCIENCE

Professor Marta Kwiatkowska
Fellow of Trinity College
Direct Line Tel: +44 (0)1865 283509
Email: marta.kwiatkowska@cs.ox.ac.uk

Personal Assistant: Anita Hancox
Direct Line Tel: +44 (0)1865 610754
Email: anita.hancox@cs.ox.ac.uk

international recognition of this research. The citations have been steadily increasing, particularly on Google Scholar, which is recognized in computer science as an indicative measure of scientific impact, given that the field has a high representation of conference publications. Dr Smieja has also completed several publications (total of 16) that have not been included as part of the thematically related work stream, covering topics in unsupervised learning, neural networks and cheminformatics. These were published in reputable international journals and conferences between 2016 and 2020. Together with the doctorate (art. 219 para 1 point 1), the 10 publications therefore clearly meet the required criteria for awarding habilitation (art. 219 para 1 point 2b).

(2) Scientific activity as confirmed by internships, particularly abroad, and research grants

Dr Smieja's case also includes successful internships and research visits to labs of high international standing abroad (IST Lisbon, TU Graz) and within Poland (Polish Academy of Science, TU Wroclaw). These led to 4 joint publications plus 1 under submission, three of them, [A2], [A4] and [A10], directly contributing to the habilitation. He has a very strong track record of research grants (4/12 as principle investigator) in data science and cheminformatics funded by Polish funding bodies (NCN, FNP and NCBiR), and has collaborated widely with institutions outside him home institution of the Jagiellonian University.

Summarising, Dr Smieja has engaged in high-quality research at the international level, as evidenced by successful execution of research grants and internships abroad and within Poland, and therefore his case for habilitation clearly meets the required criteria (art. 219 para 1 point 3).

(3) Teaching and other organizational activities

Dr Smieja has been engaged in teaching since 2010. In the early phase, this included leading laboratory classes, supervision of Masters projects and tutoring of individual student. As of 2017, he started lecturing on information theory and machine learning and led a seminar on AI methods. Notably, he has also contributed to outreach by giving a number of science popularisation lectures between 2012 and 2015. Regarding conference organisation, he participated in programme committees and twice chaired the organising committee of a specialised TFML meeting and additionally acted as organising committee member for two more conferences.

Summarising, Dr Smieja has satisfactorily contributed to teaching and conference organisation, thus enhancing his case for habilitation.

Summary

The documentation provided by Dr Marek Smieja to support his habilitation reports a comprehensive and coherent body of work aimed at resolving difficulties with existing machine learning methods. The chosen aim– to make machine learning feasible and practical in a principled way for situations with missing data – is ambitious and the resulting methods outperform state-of-the-art techniques. The submitted publications develop the foundations and algorithms for semi-supervised learning, thus ensuring continuing relevance in future. The potential impact, via theoretical results and software tools, of the habilitation research is evident from the high number and quality of publications that have been reported, and through citations to this work.

DEPARTMENT OF
COMPUTER
SCIENCE

Professor Marta Kwiatkowska
Fellow of Trinity College
Direct Line Tel: +44 (0)1865 283509
Email: marta.kwiatkowska@cs.ox.ac.uk

Personal Assistant: Anita Hancox
Direct Line Tel: +44 (0)1865 610754
Email: anita.hancox@cs.ox.ac.uk

Recommendation

Overall, the quality of the research submitted for evaluation is very high, and is supported by innovative ideas, theoretical investigation, extensive experimental results, and important insights. Considering the declarations of contributions, I am fully satisfied that the work presented here indicates that Dr Smieja is able to lead an independent scientific research programme, and do so to a high standard. I note that he was recognized for his scientific achievements through several awards. Dr Smieja has also contributed to lecturing and conference organisation. I therefore conclude without reservations that the achievements merit habilitation in accordance with the criteria, and recommend to the panel that the title be awarded to Dr Marek Smieja.

Should you need any further information do not hesitate to get in touch with me.

Yours sincerely

Marta Kwiatkowska FRS
Professor of Computing Systems
Profesor Nauk Inżynieryjno-Technicznych

About the Writer

Marta Kwiatkowska graduated in computer science from the Institute of Computer Science (Instytut Informatyki), Jagiellonian University, in 1980 where was Assistant Professor for the period of 1980-88. She obtained her Doctorate from the University of Leicester in 1989, where she was a Lecturer. In 1994 she moved to the University of Birmingham, where she was promoted to Professor in 2003 before taking up a Statutory Chair and Fellowship of Trinity College at the University of Oxford in 2007. In 2020 Kwiatkowska was awarded the Title of Professor by the President of Poland. She is known for fundamental contributions to the theory and practice of model checking for probabilistic systems, focusing on automated techniques for verification and synthesis from quantitative specifications. She led the development of the PRISM model checker (www.prismmodelchecker.org), the leading software tool in the area and winner of the HVC Award 2016. Probabilistic model checking has been adopted in diverse fields, including distributed computing, wireless networks, security, robotics, healthcare, systems biology, DNA computing and nanotechnology, with genuine flaws found and corrected in real-world protocols. More recently, her focus has shifted towards developing automated (including probabilistic) verification and synthesis for machine learning and AI safety. Kwiatkowska is the first female winner of the Royal Society Milner Award, winner of the BCS Lovelace Medal and was awarded an honorary doctorate from KTH Royal Institute of Technology in Stockholm. She won two ERC Advanced Grants, VERIWARE and FUN2MODEL, and is coinvestigator of the EPSRC Programme Grant on Mobile Autonomy and the Turing Institute project FAIR (Framework for responsible adoption of artificial intelligence in the financial services industry). Kwiatkowska is a Fellow of the Royal Society, Fellow of ACM, EATCS, BCS and PTNO, and Member of Academia Europea.