UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Advances in Multi-output Learning via Nearest Neighbours

by

Donna Xu

**A Thesis Submitted
for the Degree of
Doctor of Philosophy**

Sydney, Australia

January, 2019

Certificate of Authorship/Originality

I certify that the work in this thesis has not been previously submitted for a degree nor has it been submitted as a part of the requirements for other degree except as fully acknowledged within the text.

I also certify that this thesis has been written by me. Any help that I have received in my research and in the preparation of the thesis itself has been fully acknowledged. In addition, I certify that all information sources and literature used are quoted in the thesis. This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

ABSTRACT

Multi-output learning aims to simultaneously predict multiple outputs given an input. It is an important learning problem due to the pressing need for sophisticated decision making in real-world applications. Inspired by big data, the 4Vs characteristics of multi-output imposes a set of challenges to multi-output learning, in terms of the *volume*, *variety*, *velocity* and *veracity* of the outputs. *Volume* refers to the explosive growth of output labels that have been generated and it leads to two challenges, large output dimensions and unseen outputs. *Variety* refers to heterogeneous nature of output labels and it results in complex structures of the output. *Velocity* refers to speed of output label acquisition including the phenomenon of concept drift and update to the model. The challenge imposed by velocity could be the change of output distributions, where the target outputs are changing over time in unforeseen ways. The nearest neighbours is one of the most classic frameworks in handling multi-output problems. In this thesis, I focus to overcome the challenges encountered by the first three of the 4Vs characteristics of multi-output, using nearest neighbours-based methods.

The first work of this thesis deals with the challenges imposed by *volume* and *variety* of multi-output. It focuses on the nearest neighbours-based semantic retrieval and zero-shot learning, which are sub-problems of multi-output learning. I propose a novel concept-based information retrieval system that combines general semantic feature representation and a metric learning model. It achieves better semantic retrieval performance for domain-specific information retrieval problems. Together with the better learned semantic representation, the distance metric can be generalized to unseen output labels and can be applied to zero-shot learning applications. The second work of the thesis handles the challenge of changing of output distribution caused by velocity of multi-output. Nearest neighbours cannot be successfully adapted to deal with this challenge due to the inefficiency issue. This work focuses on improving the nearest neighbours efficiency for multi-output learning problems.

An online product quantization (online PQ) model is developed to accommodate to the streaming data with time and memory requirements. A loss bound is derived to guarantee the performance of the model.

Acknowledgements

I would like to express my deep appreciation to my supervisor, Prof. Ivor W. Tsang, for his help and guidance provided to me throughout my PhD study. His help has been immense in every aspects. He has taught me the basics and advances of machine learning to eable me to develop an understanding of machine learning and help me to acquaire a solid foundation of knowledge in this area; he has taught me how to conduct research and industry projects and write academic papers step by step; he has been very patient to explain every question I asked and always provided me valuable feedbacks on my works. He focuses on what students have learned and improved and how students grow as a research scientist. His ways of supervision and encouragement make me become more confident during my study. His enthusiasm, devotion and attitude to research have greatly inspired me. I thank him for his positive influence not only to my PhD study but also to my future career life. I could not have imagined having a better supervisor for my PhD study.

I would also like to express my gratitude to my fellow students in my group, for discussing research works with me and generously sharing their knowledge and advice. I feel very fortunate to be a member in this group in my PhD study and am very grateful for all the help the group has offered to me.

Finally, a big thank to my parents and my husband for all their love, care and constant support and encouragement. Without their precious support it would not be possible for me to complete this thesis.

Donna Xu

Sydney, Australia, 2019.

Contents

List of Figures