

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.

© Copyright 2019 Donna Xu

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 enable me to develop an understanding of machine learning and help me to acquire 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

Certificate	ii
Abstract	iii
Acknowledgments	v
List of Figures	x
1 Introduction	1
1.1 Background	1
1.2 The 4Vs Challenges of Multiple Outputs	3
1.3 Research Topics	4
1.3.1 Semantic Retrieval	5
1.3.2 Zero-shot Learning	6
1.3.3 Efficiency of Nearest Neighbours	6
1.4 Thesis Contributions	6
1.4.1 Sub-problem of Multi-output Learning: Concept-based Semantic Retrieval via K-nearest Neighbours	7
1.4.2 Sub-problem of Multi-output Learning: Concept-based Zero-shot Learning via K-nearest Neighbours	8
1.4.3 Improving Nearest Neighbours Search Efficiency in Dynamic Database Environment	8
1.5 Thesis Organization	8
1.6 Publications	9

2 Literature Survey	12
2.1 Life Cycle of Output Labels	12
2.1.1 How is Data Labeled	13
2.1.2 Forms of Label Representations	14
2.1.3 Label Evaluation and Challenges	15
2.2 Problem Definition of Multi-output Learning	16
2.2.1 Zero-shot Learning	17
2.2.2 Semantic Retrieval	18
2.3 Model Evaluation Metrics	19
2.4 Semantic Retrieval	23
2.5 Zero-shot Learning	24
2.6 Hashing and Online Hashing	25
 3 Concept-based Semantic Retrieval with K-nearest Neigh- bours	 29
3.1 Motivation	29
3.2 Concept-based Retrieval System	32
3.2.1 M1. General Concept Extraction	33
3.2.2 M2. Domain-Specific Concept Mining via Distance Learning .	34
3.2.3 M3. Retrieval	41
3.2.4 Module-Task Alignment	42
3.2.5 Advantages of the Proposed System	43
3.3 Experiments	44
3.3.1 Data sets	45
3.3.2 User Search Experience Enhancement	47

3.3.3	Relevant Document Recommendation	48
3.3.4	Document Discovery with Boolean Queries	54
3.4	Conclusions	61
4	Concept-based Zero-shot Learning with K-nearest Neighbours	62
4.1	Motivation	62
4.2	Distance Learning Model with Semantic Concepts Embedding for Zero-shot Learning	64
4.2.1	Notations and Preliminaries	64
4.2.2	Concept-based Zero-shot Learning	64
4.3	Experiments	65
4.3.1	Baselines and Settings	65
4.3.2	Evaluation Metrics	66
4.3.3	Generalized Zero-shot Learning for Document Categorization	67
4.4	Conclusion	68
5	Online Product Quantization for Nearest Neighbour Search	70
5.1	Motivation	70
5.2	Online Product Quantization with Budget Constraints	74
5.2.1	Preliminaries	74
5.2.2	Online Product Quantization	76
5.2.3	Mini-batch Extension	78
5.2.4	Partial Codebook Update	78
5.2.5	Complexity Analysis	81
5.3	Loss Bound	81

5.4	Online Product Quantization over a Sliding Window	85
5.4.1	Online Product Quantization with Data Insertion and Deletion	86
5.4.2	Connections among Online PQ Algorithms	87
5.5	Experiments	89
5.5.1	Datasets and evaluation criterion	89
5.5.2	Convergence	90
5.5.3	Online vs mini-batch	92
5.5.4	Update time complexity vs search accuracy	92
5.5.5	Baseline methods	94
5.5.6	Object tracking and retrieval in a dynamic database	94
5.5.7	Continuous querying on dynamic real-time data	102
5.6	Conclusion	106
6	Conclusions and Future Works	107
6.1	Conclusions	107
6.2	Future Works	108
6.2.1	Challenges by Variety	108
6.2.2	Challenges by Veracity	111
	Bibliography	112

List of Figures

1.1	The organization of the thesis.	10
2.1	The life cycle of the output label.	12
3.1	System diagram of the proposed concept-based retrieval system with three modules.	32
3.2	In the Mac Hardware domain, a learned metric allows documents in general concept representation to be transformed to a domain-specific concept space, in which the importance of each concept has been re-weighted.	34
5.1	Hashing vs PQ in online update. The hash codes of the data points in the reference database will get updated if the hash functions get updated by the new data. <i>The index of the codewords in the PQ codebook, on the other hand, will remain the same even though the codebook gets updated by the new data.</i> Thus online PQ is able to save severely much time by avoiding codewords maintenance of the reference database. (Best viewed in colors)	72
5.2	A general procedure for Online Product Quantization update. At each iteration, the new codebook gets updated by the streaming data. Searching can be performed using the latest codebook.	73

5.3	A schematic figure of online product quantization with budget constraints. There are two subspaces where each subspace has two sub-codewords. After the codebook adapting to the new data, two of the four sub-codewords get hugely changed (highlighted in a red dashed rectangle) and the rest two sub-codewords barely changed. . . .	79
5.4	Approaches of Handling Data Streams. Streaming: data streams one at a time. Mini-batch: a mini-batch of data with size 3 is processed by the model at each iteration t . Sliding window: a moving window with size 3 applied on continuously changing data.	86
5.5	Convergence of online PQ using ImageNet dataset. Effective iterations are shown on the x-axis.	91
5.6	The left figure shows the update time for each iteration of update. The time of the online version for each iteration sums up the update time of the streaming data corresponding to the ones in the mini-batch. The right figure shows the recall@1, 20 and 100 for each iteration.	91
5.7	Trade-offs between update time cost and the search accuracy. The first column shows the impact of the subspace update constraint. The second column shows the impact of the sub-codeword update constraint. The red line is the reference line for the scatter plot. . . .	93
5.9	Results for news and image retrieval in a dynamic database comparison against online hashing methods. Recall@20 performance (1st row) and Update time cost (2nd row). 1st column: News20. 2nd column: Caltech-101. 3rd column: Sun397. 4th column: Half dome. Time cost is in log scale.	98
5.10	99
5.11	101

5.12	Online PQ over a sliding window approach between deletion and no deletion of the expired data to the model for Sun397. Recall@20 (1st row) and Update time cost (2nd row).	103
6.1	The conclusions of the thesis.	109