Research Statement

Steven C.H. HOI
School of Information Systems
Singapore Management University
E-mail: chhoi@smu.edu.sg
http://mysmu.edu.sg/faculty/chhoi/

Last updated on 9 Jan, 2017

1. Background

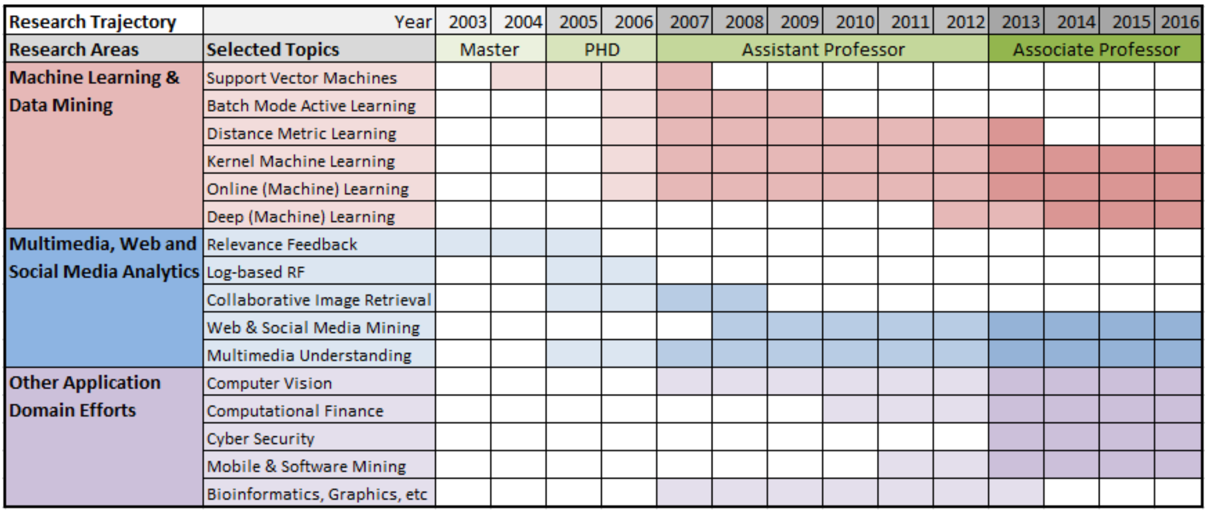
My primary research interests lie in the development of new machine learning and data mining methodologies for solving real-world problems. Machine learning is concerned with the design of learning algorithms and systems that can adapt to new scenarios by learning from past observations or experience. My early interest to machine learning can be dated back to my graduate study. At that time, I was working on content-based image retrieval (CBIR) projects where one challenging problem is to enhance CBIR systems by effectively exploiting user's relevance feedback while minimizing user's feedback effort. Very soon I realized such practical problem can be essentially formulated as a machine learning task, e.g., active learning, which learns a model to select informative unlabeled data in order to maximize the accuracy of the trained model while minimizing the amount of requested labeled data.

After the CBIR project, I found that machine learning techniques can be generally applied to a much wider range of real-world applications. This has motivated my research agenda of past few years to focus on two key aspects. On one hand, one major concern of my research agenda is to investigate **fundamental machine learning methodology** by developing novel learning methods and algorithms to address long-standing open challenges in machine learning. Some example challenges addressed in my study include online learning, deep learning, active learning, kernel learning, distance metric learning, etc. On the other hand, another concern of my research agenda is to explore emerging **applications of machine learning techniques** for solving real-world problems in different domains. Specifically, I have supervised an active research group with talented graduate students and research staff for conducting multidisciplinary research by applying machine learning to a wide spectrum of application domains, ranging from multimedia and computer vision, to web search and social media, computational finance, cybersecurity, bioinformatics, mobile and software mining, etc.

My research group has been well funded by different funding agencies from both academia and industry, partially thanks to my multidisciplinary experiences in tackling real problems from a unique machine learning perspective. Below I will describe in detail my research contributions and achievements as well as my ongoing research agenda and future directions.

2. Research Contributions and Achievements

My research contributions cover both **fundamental research** on *machine learning* methodology and **practical research** on real-world application areas, including *multimedia information retrieval, social media analytics, web search and mining, computational finance, computer vision and pattern recognition, bioinformatics and medical imaging, cybersecurity analytics, mobile and software mining, etc*. Below gives a summary of my major research contributions and achievements, including three categories: (i) foundation of machine learning and data mining methodology, (ii) multimedia search, and web & social media analytics, and (iii) Other application projects spanning a wide range of domains in information and intelligent systems. Figure 1 gives an overview of research trajectory of my academic career (2003-2016).



Research Trajectory	Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Research Areas	Selected Topics	Master		PHD		Assistant Professor						Associate Professor			
Machine Learning & Data Mining	Support Vector Machines		░	░	░	▒									
	Batch Mode Active Learning				░	▒	▒	▒							
	Distance Metric Learning				░	▒	▒	▒	▒	▒	▒	▓			
	Kernel Machine Learning				░	▒	▒	▒	▒	▒	▒	▓	▓	▓	▓
	Online (Machine) Learning				░	▒	▒	▒	▒	▒	▒	▓	▓	▓	▓
	Deep (Machine) Learning										▒	▒	▓	▓	▓
Multimedia, Web and Social Media Analytics	Relevance Feedback	░	░	░											
	Log-based RF			░	░										
	Collaborative Image Retrieval			░	░	▒	▒								
	Web & Social Media Mining						▒	▒	▒	▒	▒	▓	▓	▓	▓
	Multimedia Understanding			░	░	▒	▒	▒	▒	▒	▒	▓	▓	▓	▓
Other Application Domain Efforts	Computer Vision					░	░	░	░	░	░	▓	▓	▓	▓
	Computational Finance								░	░	░	▓	▓	▓	▓
	Cyber Security											▓	▓	▓	▓
	Mobile & Software Mining									░	░	▓	▓	▓	▓
	Bioinformatics, Graphics, etc					░	░	░	░	░	░	░			

Figure 1. Research Trajectory of My Academic Career (2003-2016)

2.1 Foundation of Machine Learning Methodology

- **Online Learning for Big Data Analytics**

I have been working actively in the area of online learning over the past years [10-11,14-16]. Online learning represents a family of efficient and scalable machine learning algorithms. We have proposed novel techniques to overcome several limitations of traditional online learning methods. First of all, unlike traditional methods that are often designed for optimizing mistake rate or classification accuracy, we proposed a novel *Online AUC Maximization* (OAM) method which is designed to optimize the Area under the ROC Curve (AUC). Second, to make second-order online learning methods robust under noisy observations, we proposed a novel *Soft Confidence-Weighed learning* method which can handle non-separable data and more effective for noisy training examples. Third, to improve the learning efficacy of traditional kernel-based online learning methods that often performs single updating when adding a new support vector, we proposed a novel *Double Updating Online Learning* (DUOL) method, which updates two support vectors simultaneously to boost the efficacy. We have also proposed a novel framework of *Online Multiple Kernel Learning* (OMKL), which learns models effectively from multiple heterogeneous sources. Last but not least, we have been actively developing LIBOL [16] (http://libol.stevenhoi.org/), an open-source library for online learning algorithms.

- **Batch Mode Active Learning.**

Active learning is an important machine learning methodology. It plays an important role in many real-world applications when unlabeled data is abundant but manually labeling can be very expensive. Most of previous research on active learning methodologies is limited to *selecting a single unlabeled example* at each learning iteration. This could be computationally inefficient since the model has to be re-trained for every labeled example. To address the critical limitation of conventional active learning methods, we have presented the novel framework of **"Batch Mode Active Learning"** (BMAL) [1], which aims to select *multiple informative examples* simultaneously at each iteration. Based on our framework, we proposed several variants of BMAL algorithms to address specific research challenges raised by different learning tasks, including the kernel logistic regression based BMAL method [1], and the semi-supervised SVM BMAL method [6]. Our seminal work of batch mode active learning has inspired many studies in the literature, as reflected by **350+ citations** for the series of our work according to Google scholar. Our article of semi-supervised SVM batch mode active learning [6] ranks **No. 1 of most cited articles** published in ACM Transactions on Information Systems (TOIS) over the last five years according to ISI journal citations.

- **Distance Metric Learning.**

The concept of distance metric or distance function is fundamental to many fields of science and engineering. Choosing an appropriate distance metric or distance function is vital to many real-world applications, including multimedia information retrieval and data mining tasks. In the past, I have worked extensively on the problem of *Distance Metric Learning (DML)* [3,8], with the focus on very tough scenarios where side information could be scarce, noisy, or even not given explicitly at all. We have proposed several novel DML algorithms, including the well-known Discriminative Component Analysis (DCA), semi-supervised distance metric learning by exploring both labeled and unlabeled data, probabilistic distance metric learning for handling uncertain side information, and nonlinear distance function learning, etc. Our research efforts in tackling the challenges of learning robust and effective distance metrics from various challenging situations have produced a lot of impact on many real-world applications, which is reflected by a total of **350+ citations** from a variety of application domains in literature.

- **Kernel Machine Learning.**

Kernel methods are an important family of machine learning techniques, where a well-known example is support vector machines (SVM). For kernel methods, choosing a good kernel is essential, and many kernel methods assume either fixed or certain parametric/semi-parametric forms for the kernels to be used. As one of the pioneering approaches, we present the seminal work of fully **Non-Parametric Kernel Learning (NPKL)** technique [4] that aims to learn the optimal kernel from data effectively. Recently, we have developed a family of efficient and scalable **SimpleNPKL** algorithms that make it possible to apply the proposed NPKL technique to very large datasets [8], a key step towards real-world applications. In addition to NPKL, I also studied novel algorithms for other kinds of kernel machine learning tasks, such as Multiple Kernel Learning (MKL) which aims to learn the optimal combination of multiple kernels. We have proposed new algorithms to tackle different challenges of MKL, including unsupervised MKL and multi-layer MKL. The series of our work had been cited for **250+ times**.

- **Parallel Deep Learning for Large-Scale Data Analytics**

Recently I have actively worked in the area of parallel deep learning, a family of state-of-the-art machine learning algorithms for a wide range of real intelligent applications. We have developed new techniques to overcome some limitations of traditional deep learning methods. For example, we have proposed a family of deep learning algorithms for large-scale content-based image retrieval tasks [13], in which we can train deep learning models (deep CNN specifically) from large amounts of weekly labeled training data (in the forms of pairwise constraints). This allows to apply the deep learning techniques to a broad range of multimedia applications, where it may not be easy to collect a large amount of high quality labeled training data. Besides, we are also exploring techniques for speeding up the training of large-scale CNN models using GPU-based parallel computation techniques.

1.2 Multimedia, Web and Social Media Analytics

Besides the fundamental research in machine learning methodology, I believe it is equally important to investigate the applications of machine learning techniques to address real-world challenges. One of my key application areas is to multimedia search and information systems. The following gives a summary of my major contributions in this area.

- **Collaborative Image Retrieval.**

A fundamental challenge in multimedia retrieval is the semantic gap between semantic meaning and low level features of multimedia data. To tackle this challenge, we proposed a novel paradigm of image retrieval, termed Collaborative Image Retrieval (CIR) [2,3,5], that explores machine learning techniques in bridging the semantic gap by mining the logs of user's search history. In particular, we developed two types of approaches for CIR. The first approach, termed *log-based relevance feedback* [3], explicitly utilizes the logs of user's relevance feedback in an online fashion,. The second approach explores users' logs in an offline fashion. It learns robust distance metrics from noisy user log data for multimedia retrieval, using DML methods, such as regularized metric learning, Laplacian regularized metric learning, and probabilistic metric learning, etc. Our seminal work in this area has made significant impact in multimedia retrieval, as reflected by a total of **350+ citations** for the series of our work.

- **Interactive Multimedia Retrieval.**

One way to close the semantic gap of content-based multimedia retrieval is to explore interactive retrieval paradigm via relevance feedback. However, traditional relevance feedback methods suffer from some critical drawbacks, such as poor learning efficacy, class-imbalance and insufficient labeled data, etc. I have attempted to overcome these limitations from a machine learning perspective. In particular, we proposed to apply batch mode active learning techniques to improve the learning efficacy by optimizing the selection of multiple examples for relevance feedback. We addressed the class imbalance and insufficient labeled data issues by applying semi-supervised active learning algorithms. These techniques have been applied to various multimedia retrieval tasks, including content-based image retrieval, multimodal news video retrieval, and medical image categorization and retrieval [1], with significant impact as reflected by a total of **300+ citations** in this series.

- **Social Media Search & Mining.**

Social media, an emerging new multimedia data, has raised many new research challenges [9]. Recently we have worked actively and tackled some of the key challenges in this area, including large-scale social image retrieval, automated photo tagging, and auto face annotation by mining web/social images. For these real-world applications, we have applied our proposed novel machine learning techniques to solve the emerging research challenges of social media search and data mining. To encourage researchers and practical engineers to make contributions to this new emerging area, I also *co-founded and co-chaired the series of ACM SIGMM Workshops on Social Media* (WSM) in conjunction with ACM Multimedia conferences. ACM Multimedia conference has accepted **"Social Media"** as a new track this year and invited me as one of **Area Chairs** to lead the new track in this top conference.

1.3 Other Application Domains for Information & Intelligent Systems

The advent of big data age has presented a number of challenges and opportunities for the applications of machine learning techniques to large-scale knowledge discovery and intelligent systems. Recently, we have investigated both theoretical and practical issues in mining big data. We have been investigating a variety of practical machine learning techniques for mining massive amount of data in real-world applications. Examples of our related work include peer-to-peer machine learning in distributed environments, collaborative online learning, multi-view semi-supervised learning, multi-kernel boosting classification, and semi-supervised clustering, etc. Besides studying these novel algorithms, we have also applied **machine learning techniques to multidisciplinary research** across several real-world application domains, including *computer vision, bioinformatics, medical imaging, and computational finance*. The following gives a summary of our contributions to multidisciplinary research in some areas.

- **Computer Vision & Pattern Recognition.**

We have investigated novel techniques for large-scale computer vision and recognition systems. Examples include visual recognition for image recognition using deep learning, and its application to food image recognition project. We have also studied face alignment, tracking and annotation, in which we built new technologies of face alignment, tracking, and annotation by machine learning for various applications in computer vision and augmented reality. We have also investigated the face annotation problem extensively by exploring machine learning techniques to overcome insufficient and noisy labeled data. In addition, we have investigated several effective machine learning methods for solving the challenge of 3D object modeling and tracking. In particular, we proposed efficient 3D deformable techniques for modeling implicit surfaces to tackle the non-rigid shape recovery, and geometry image and geometry video approaches for compressing and streaming objects of large-scale 3D meshes efficiently.

- **Bioinformatics & Medical Imaging.**

In bioinformatics, we addressed several open challenges by developing computational methods for the prediction of binding hot spots, an important task towards understanding protein-protein interactions. For example, we proposed novel approaches for binding hot spot predictions by predictive learning methods to identify binding hot spots at the epitope sites of the HA1 proteins and at the paratope sites of the 2D1 antibody, and also computational methods for identifying B-cell epitopes to understanding basic recognition mechanism of immune response, which in turn guides disease diagnosis, vaccine design and drug development. Last but not least, we also proposed machine learning to tackle real-world challenges in medical imaging domain, including medical image categorization [1] and medical image retrieval tasks.

- **Computational Finance.**

Last but not least, machine learning and data mining techniques have also emerged as one of promising directions for solving many open challenges in computational finance. My major research contribution in this area is focused on the open problem of *On-line Portfolio Selection* [12-13], a critical component of many real-world intelligent financial systems. As a machine learning and data mining researcher, I take a very different perspective to address this challenge. In particular, we have developed new strategies, based on machine learning techniques, for online portfolio selection. We proposed a family of new online trading algorithms for on-line portfolio selection by exploiting the mean reversion principle using state-of-the-art online learning techniques. Our promising results from comprehensive empirical studies on a variety of large-scale real testbeds showed that the proposed novel strategies outperform the state-of-the-art strategies for intelligent portfolio management in literature.

3. Ongoing Research Agenda and Future Directions

My current research agenda has been focused on massive-scale machine learning techniques for resolving emerging big data analytics. Currently I am interested in exploring a smart fusion of online learning and deep learning techniques for resolving many emerging big data and AI applications. My future research agenda include but not limited to:

- New machine learning algorithms for enhancing learning efficacy and capacity;
- Novel machine learning paradigms for resolving many emerging challenges of living data analytics problems in different real-world applications;
- New AI capabilities for real-world applications with machine learning technologies, particularly for addressing the open challenge in the context of smart nation.

Selected References

1. **Steven C.H. Hoi**, Rong Jin, Jianke Zhu and Michael R. Lyu, "Batch Mode Active Learning and Its Applications to Medical Image Classification", In **ICML**, Penn, US, 2006. (acceptance rate = 18%)
2. **Steven C.H. Hoi**, Wei Liu, Michael R. Lyu, and Wei-Ying Ma, "Learning Distance Metrics with Contextual Constraints for Image Retrieval," In *Proc IEEE* **CVPR**, 2006. (acceptance rate: ~20%)
3. **Steven C.H. Hoi**, Michael R. Lyu, and Rong Jin. "A Unified Log-based Relevance Feedback Scheme for Image Retrieval," *IEEE Tran. Knowledge and Data Engineering* (**TKDE**), vol. 18, no.4, pp. 509-524, 2006.
4. **Steven C.H. Hoi**, Rong Jin and Michael R. Lyu,"Learning Non-Parametric Kernel Matrices from Pairwise Constraints,", In *Intl. Conference on Machine Learning* (**ICML**), Corvallis, OR US, 20-24 June, 2007.
5. **Steven C.H. Hoi**, Wei Liu, and Shih-Fu Chang,"Semi-Supervised Distance Metric Learning for Collaborative Image Retrieval," In *Proc. IEEE* **CVPR**, Alaska, US, June 24-26, 2008.
6. **Steven C.H. Hoi**, R. Jin, J. Zhu, M.R. Lyu, "Semi-Supervised SVM Batch Mode Active Learning with Applications to Image Retrieval," *ACM Trans. on Information Systems* (**TOIS**), 27(3), 2009.
7. Peilin Zhao, **Steven C.H. Hoi**, Jialei Wang, Bin Li. Online Transfer Learning. *Artificial Intelligence **(AI)***, 2014, 216: 76-102.
8. Jinfeng Zhuang*, Ivor Tsang, **Steven Hoi**, "A Family of Simple Non-Parametric Kernel Learning Algorithms from Pairwise Constraints", *Journal of Machine Learning Research* (**JMLR**), 2011.
9. **Steven C.H. Hoi**, Jiebo Luo, Sussane Boll, Dong Xu, Rong Jin, Irwin King, "Social Media Modeling and Computing", book series of *Advances in Pattern Recognition*, Springer Press, 2011.
10. Jielei Wang, Peilin Zhao, **Steven C.H. Hoi**, "Exact Soft Confidence-Weighted Learning," **ICML**, Edinburgh, Scotland, June 26 - July 1, 2012.
11. **Steven C.H. Hoi**, Rong Jin, Peilin Zhao*, Tianbao Yang, "Online Multiple Kernel Classification", Machine Learning (**ML**), vol. 90, no. 2, 289-316, 2013.
12. Bin Li, **Steven C. H. Hoi**, Online Portfolio Selection: Principles and Algorithms, CRC Press, Nov 5, 2015
13. Bin LI, **Steven C.H. Hoi**, "Online Portfolio Selection: A Survey", *ACM Computing Surveys* (**CSUR**), 2014
14. Ji Wan, Dayong Wang, **Steven C.H. Hoi**, P. Wu, J. Zhu, Y. Zhang, J. Li , "Deep Learning for Content-Based Image Retrieval: A Comprehensive Study" ACM Multimedia Conference (**MM2014**), November 3-7, 2014
15. Doyen Sahoo, **Steven C. H. Hoi**, Bin Li, "Online Multiple Kernel Regression", ACM SIGKDD Conference (**KDD**), New York, USA, August 24 - 27, 2014.
16. **Steven C.H. Hoi**, Jialei Wang, Peilin Zhao, LIBOL: A Library for Online Learning Algorithms", *Journal of Machine Learning Research* (**JMLR**), 2014.
17. Bin Li, **Steven C.H. Hoi**, Doyen Sahoo, Zhi-Yong Liu, Moving Average Reversion Strategy for On-Line Portfolio Selection, *Artificial Intelligence (AI)*, 222 104-123 2015.
18. Jialei Wang, Ji Wan, Yongdong Zhang, **Steven C. H. Hoi**, SOLAR: Scalable Online Learning Algorithms for Ranking, *The 53rd annual meeting of the Association for Computational Linguistics (ACL2015)*, Beijing, China July 26-31, 2015
19. LU, Jing; **HOI, Steven C. H.**; WANG, Jialei; ZHAO, Peilin; LIU, Zhi-Yong "Large Scale Online Kernel Learning", *Journal of Machine Learning Research* **(JMLR)**, 17(47) 1-43 2016.
20. Chenghao Liu, **Steven C.H. Hoi**, Peilin Zhao, Jianling Sun, Online ARIMA Algorithms for Time Series Prediction , *AAAI Conference on Artificial Intelligence (AAAI-16)*, Phoenix, Arizona, February 12–17, 2016.
21. Jing Lu, Peilin Zhao, **Steven C. H. Hoi**, Online Passive-Aggressive Active learning, Machine Learning 103(2) 141-183, 2016
22. Chenghao Liu, Tao Jin, **Steven C.H. Hoi**, Peilin Zhao, Jianling Sun, Collaborative Topic Regression for Online Recommender Systems: An Online and Bayesian Approach, *Machine Learning,* 2017. To appear.