Novi Quadrianto

Research Statement

Approximately 1.5 billion people use the *Internet*. People write news articles, blogs, and reviews; people upload videos, audios, and photos. People become web content creators. This directly translates to the availability of half a Zettabyte of data. Synergistically with rapid progress in machine learning models and algorithms, as well as rapid rises in computing power and storage, the challenge of the 21$^{\text{st}}$ century consists of finding way to transform this complex massive yet noisy and sparse Internet data coming from a variety of sources into insights in support of knowledge creation. My research aims to address data to knowledge transformation in the context of *machine learning*.

Machine learning techniques have become prevalent for drawing inference and making prediction from massive scale data. Given input-output data pairs, the goal of learning is to infer a latent function that maps inputs to outputs. This function will then be used to predict an output for a given unseen input. Consider as an illustrative example, a task of categorising web videos (user-generated videos from video sharing websites). Here the inputs are the web videos and the outputs are the categories such as entertainment, music, news and politics, science and technology, among others.

The complex nature of Internet data manifests itself along both the input (feature) and output (label) dimensions. On the input dimensions, we deal not only with potentially millions of features but also the features might come from multiple modalities or data sources. Web videos admit the conventional representation of audio-visual features, the associated text (the filenames, titles, or descriptions) and even the intricate social network representation (the relationship among videos through the users, links, or recommendations). On the label dimensions, the information is sparse. For instance, there might be millions of web videos but only a few by a particular user or labelled with a particular tag. Adding to the sparsity challenge, Internet data tend to have multiple sets of different labels. In the case of our illustrative example, the categorisation of web videos from several different video sharing services depends heavily on the editors of each web service. Different editors have very different perception of video categories, thus the label categories are often 'inconsistent'.

Research Contribution–Learning for the Internet: Kernel Embeddings and Optimisation

My research aims to address research challenges for Internet applications in the context of machine learning. My present work focuses on addressing Internet complexity on *output label dimensions*. We introduce non-standard machine learning problems, and we present scalable solutions for several existing machine learning problems. My current solutions all centre around two main mathematical ingredients. First, I am employing Hilbert Space embeddings of distributions via averages. This allows distance computation between distributions in terms of distances between averages, which, in turn, yield elegant ways to deal with distributions without the need of estimating them as an intermediate step. Second, I am heavily drawing on recent advances in field of optimisation, in particular, in the area of online stochastic optimisation to address the sheer size and the non-convex nature of mostly Internet problems.

Formulation and Solution for New Machine Learning Problems

Weak label supervision
Traditional classification setting infers a statistical model based on observed input-output data pairs. We introduce a new problem where instead of each input is supervised with an output, we are given groups of unlabelled inputs [1,2]. Each group is endowed with information on class label proportions. The number of group is at least as many as number of class labels. This seemingly contrived setting

has plethora of applications such as in areas like politics, spam filtering, e-commerce, and improper content detection. We also introduce a learning framework where a set of inputs and a set of outputs are given however they are not paired [3-5]. The goal of learning is now to infer a correspondence or a permutation that maps each input to its output. This has applications in data visualisation, image search browsing, photo album summarisation, cross-domain matching, to name a few.

Label inconsistency
In machine learning it is folk knowledge that if several prediction tasks are related, then learning them simultaneously can improve performance. For instance, a web videos categoriser trained with data from several different video sharing sites is likely to be more accurate than one that is trained with data from a single video site. We introduce a new setting of jointly learning several related tasks where each task has potentially distinct label sets and label correspondences are not readily available [6]. This is in contrast with existing settings which either assume that the label sets shared by different tasks are the same or that there exists a label mapping oracle.

Scalable Solution for Existing Machine Learning Problems

Transductive learning
Internet data, while very large, are very sparse on its label, i.e. only a minute amount of them are human annotated. In the transductive setting, this unlabelled data is harnessed to improve the performance of classifier simply trained on annotated data. We present a transductive algorithm exploiting a simple fact that the distributions over the outputs on annotated and unannotated data should match [7]. As our solution is amenable to an online optimisation method, it can process received data one at a time and then discard them in an excess data stream.

Storage and indexing management
Finding a needle in a haystack best describes a process of locating relevant data points in monstrous Internet space. Thus, each data point needs to be attached a label index before it is stored for a later efficient retrieval. We propose an algorithm for webpage tiering for search engine indices that can process billion of webpages in seconds [8]. Our presented algorithm solves an integer linear program in an online fashion. This indexing and storage problem is related to a larger class of parametric maximum flow problem and therefore our algorithm has potential applications also in those problems.

Future Work

I am interested in developing machine learning models and algorithms that will address Internet complexity issues on *input feature dimensions*, such as multi-modality of Internet data [9]. In term of research techniques, I am keen to learn a Bayesian view of machine learning as some of the Internet challenges are more naturally tackled under this framework. As mentioned, sparsity is an inherent nature of Internet data. Ideally for such a setting, Bayesian statistics provide a robust approach to drawing inferences and making predictions from very sparse information. I am interested in both modelling aspect of nonparametric Bayesian methods in solving Internet challenges and in scaling up learning and inference of nonparametric Bayesian methods to handle Internet-scale data.

References

[1] Novi Quadrianto, Alex J. Smola, Tiberio S. Caetano, Quoc V. Le. *Estimating Labels from Label Proportions*, JMLR 2009.

[2] Novi Quadrianto, Alex J. Smola, Tiberio S. Caetano, Quoc V. Le. *Estimating Labels from Label Proportions*, ICML 2008.

[3] Novi Quadrianto, Alex J. Smola, Le Song, Tinne Tuytelaars. *Kernelized Sorting*, PAMI 2010.

[4] Novi Quadrianto, Kristian Kersting, Tinne Tuytelaars, Wray L. Buntine. *Beyond 2D-Grids: A*

Dependence Maximization View on Image Browsing, MIR 2010.

[5] Novi Quadrianto, Le Song, Alex J. Smola. *Kernelized Sorting*, NIPS 2008.

[6] Novi Quadrianto, Alex J. Smola, Tiberio S. Caetano, S.V.N. Vishwanathan, James Petterson. *Multitask Learning without Label Correspondences*, NIPS 2010.

[7] Novi Quadrianto, James Petterson, Alex J. Smola. *Distribution Matching for Transduction*, NIPS 2009.

[8] Gilbert Leung, Novi Quadrianto, Alex J. Smola, Kostas Tsioutsiouliklis. *Optimal Web-scale Tiering as a Flow Problem*, NIPS 2010.

[9] Novi Quadrianto and Christoph H. Lampert. *Learning multi-view neighborhood preserving projections*, ICML 2011.