Report from Dagstuhl Seminar 18291

Extreme Classification

Edited by

Samy Bengio[1], Krzysztof Dembczyński[2], Thorsten Joachims[3], Marius Kloft[4], and Manik Varma[5]

1 **Google Inc. – Mountain View, US,** bengio@google.com
2 **Poznan University of Technology, PL,** krzysztof.dembczynski@cs.put.poznan.pl
3 **Cornell University, US,** tj@cs.cornell.edu
4 **TU Kaiserslautern, DE,** kloft@cs.uni-kl.de
5 **Microsoft Research India – Bangalore, IN,** manik@microsoft.com

Abstract

Extreme classification is a rapidly growing research area within machine learning focusing on multi-class and multi-label problems involving an extremely large number of labels (even more than a million). Many applications of extreme classification have been found in diverse areas ranging from language modeling to document tagging in NLP, face recognition to learning universal feature representations in computer vision, gene function prediction in bioinformatics, etc. Extreme classification has also opened up a new paradigm for key industrial applications such as ranking and recommendation by reformulating them as multi-label learning tasks where each item to be ranked or recommended is treated as a separate label. Such reformulations have led to significant gains over traditional collaborative filtering and content-based recommendation techniques. Consequently, extreme classifiers have been deployed in many real-world applications in industry.

Extreme classification has raised many new research challenges beyond the pale of traditional machine learning including developing log-time and log-space algorithms, deriving theoretical bounds that scale logarithmically with the number of labels, learning from biased training data, developing performance metrics, etc. The seminar aimed at bringing together experts in machine learning, NLP, computer vision, web search and recommendation from academia and industry to make progress on these problems. We believe that this seminar has encouraged the interdisciplinary collaborations in the area of extreme classification, started discussion on identification of thrust areas and important research problems, motivated to improve the algorithms upon the state-of-the-art, as well to work on the theoretical foundations of extreme classification.

Seminar July 15–20, 2018 – http://www.dagstuhl.de/18291

2012 ACM Subject Classification Computing methodologies → Natural language processing, Computing methodologies → Supervised learning, Information systems → Collaborative filtering, Theory of computation → Machine learning theory

Keywords and phrases algorithms and complexity, artificial intelligence, computer vision, machine learning

Digital Object Identifier 10.4230/DagRep.8.7.62

Edited in cooperation with Marek Wydmuch

Extreme Classification, *Dagstuhl Reports*, Vol. 8, Issue 07, pp. 62–80
Editors: Samy Bengio, Krzysztof Dembczyński, Thorsten Joachims, Marius Kloft, and Manik Varma
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Samy Bengio (Google Inc. – Mountain View, US)
Krzysztof Dembczyński (Poznan University of Technology, PL)
Thorsten Joachims (Cornell University, US)
Marius Kloft (TU Kaiserslautern, DE)
Manik Varma (Microsoft Research India – Bangalore, IN)

The topic of this seminar is in the general context of machine learning [10] which concerns the study and development of algorithms that learn from empirical data how to make accurate predictions about yet unseen data without being explicitly programmed. Multi-class and multi-label learning are classical problems in machine learning. The outputs here stem from a finite set of categories (classes), and the aim is to classify each input into one (multi-class) or multiple (multi-label) out of several possible target classes. Classical applications of multi-class and multi-label learning include handwritten optical character recognition [8], part-of-speech tagging [11], and text categorization [7]. However, with the advent of the big data era, learning problems can involve even millions of classes. As examples let us consider the following problems:

- Person recognition in Facebook images (there are billions of Facebook users; given an image, we might want to predict the subset of users present in the image for such applications like security, surveillance, social network analysis, etc.).
- Predicting Wikipedia tags for new Wikipedia articles or webpages (Wikipedia has almost 2 million tags now).
- Recommending Amazon items where each of the 100 million items on Amazon is a separate label.
- Search on Google/Bing where each of the 100 million queries is a separate label.
- Language modelling – predicting the next word in a sentence from the millions of words available.

The problems of this type are often referred to as *extreme classification*. They have posed new computational and statistical challenges and opened a new line of research within machine learning.

The main goal of extreme classification is to design learning and prediction algorithms, characterized by strong statistical guarantees, that exhibit sublinear time and space complexity in the number of classes. Unfortunately, the theoretical results obtained so far are still not satisfactory and very limited. Moreover, the problems at this scale often suffer from unreliable learning information, e.g., there is no chance to identify all positive labels and assign them precisely to training examples. The majority of labels is used very rarely, which leads to the problem of the long-tail distribution. In practical applications, learning algorithms run in rapidly changing environments. Hence, during testing/prediction phase new labels might appear that have not been present in the training set [4, 2]. This is the so-called zero-shot learning problem. Furthermore, typical performance measures used to assess the prediction quality of learning algorithms, such as 0/1 or Hamming loss, do not fit well to the nature of extreme classification problems. Therefore, other measures are often used such as precision@k [9] or the F-measure [6]. However, none of the above is appropriate to measure predictive performance in the long-tail problems or in the zero-shot setting. Hence, the goal is to design measures, which promote a high coverage of sparse labels [5].

The seminar aimed at bringing together researchers interested in extreme classification to encourage discussion on the above mentioned problems, identify the most important ones and promising research directions, foster collaboration and improve upon the state-of-the-art algorithms. The meeting in this regard was very successful as participants from both academia and industry as well as researchers from both core machine learning and applied areas such as recommender systems, computer vision, computational advertising, information retrieval and natural language processing, were given the opportunity to see similar problems from different angles.

The seminar consisted of invited talks, working groups, presentation of their results, and many informal discussions. The talks concerned among others such topics as: common applications of extreme classification, potential applications in bioinformatics and biotechnology, neural networks for extreme classification, learning theory for problems with a large number of labels, approaches for dealing with tail labels, learning and prediction algorithms, extreme classification challenges in natural language processing, multi-task learning with large number of tasks, pitfalls of multi-class classification, recommendation systems and their connection to extreme classification, counterfactual learning and zero-shot learning. The short abstracts of these talks can be found below in this report. The four working groups focused on the following problems: loss functions and types of predictions in multi-label classification, deep networks for extreme classification, zero-shot learning and long tail labels, and generalization bounds and log-time-and-space algorithms. Short summaries of the results obtained by the working groups can also be found below.

During the seminar, we also discussed different definitions of extreme classification. The basic one determines extreme classification as a multi-class or multi-label problem with a very large number of labels. The labels are rather typical identifiers without any explicit meaning. However, there usually exists some additional information about *similarities* between the labels (or this information can be extracted or learned from data). From this point of view, we can treat extreme classification as a learning problem with a weak structure over the labels. This is in difference to structured output prediction [1], where we assume much stronger knowledge about the structure. The most general definition, however, says that extreme classification concerns all problems with an extreme number of choices.

The talks, working groups, and discussions have helped to gain a better understanding of existing algorithms, theoretical challenges, and practical problems not yet solved. We believe that the seminar has initiated many new collaborations and strengthen the existing ones that will soon deliver new results for the extreme classification problems.

References

1. Bakir GH, Hofmann T, Schölkopf B, Smola AJ, Taskar B, Vishwanathan SVN (eds) Predicting Structured Data. Neural Information Processing, The MIT Press, 2007
2. Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.
3. Yann Guermeur. l_p-norm Sauer-Shelah lemma for margin multi-category classifiers. *arXiv preprint arXiv:1609.07953*, 2016.
4. Bharath Hariharan, S. V. N. Vishwanathan, and Manik Varma. Efficient max-margin multi-label classification with applications to zero-shot learning. *Machine Learning*, 88(1-2):127–155, 2012.

5 Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 935–944, New York, NY, USA, 2016. ACM.

6 K. Jasinska, K. Dembczyński, R. Busa-Fekete, K. Pfannschmidt, T. Klerx, and E. Hüllermeier. Extreme F-measure maximization using sparse probability estimates. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning – Volume 48*, pages 1435–1444, 2016.

7 Thorsten Joachims. *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

8 Nei Kato, Masato Suzuki, Shin Ichiro Omachi, Hirotomo Aso, and Yoshiaki Nemoto. A handwritten character recognition system using directional element feature and asymmetric mahalanobis distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(3):258–262, 1999.

9 Yashoteja Prabhu and Manik Varma. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 263–272. ACM, 2014.

10 Vladimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

11 Atro Voutilainen. Part-of-speech tagging. *The Oxford handbook of computational linguistics*, pages 219–232, 2003.

2 Table of Contents

3 Overview of Talks

3.1 Discovering Tail-labels Through Robustness in Extreme Classification

Rohit Babbar (Aalto University, FI)

Joint work of Rohit Babbar, Bernhard Schölkopf
Main reference Rohit Babbar, Bernhard Schölkopf: "Adversarial Extreme Multi-label Classification", CoRR, Vol. abs/1803.01570, 2018.
URL http://arxiv.org/abs/1803.01570

The goal in extreme multi-label classification is to learn a classifier which can assign a small subset of relevant labels to an instance from an extremely large set of target labels. Datasets in extreme classification exhibit a long tail of labels which have a small number of positive training instances. The tail-labels exhibit a substantial change in their feature distribution within the training set and also from the training set to those encountered during prediction.

We, therefore, pose the learning task in extreme classification as learning in the presence of adversarial perturbations. By drawing connections to robust optimization, we show that this motivates the well-known l1-regularized SVM from an adversarial robustness perspective. For distributed training, the proposed method relies on one-vs-rest paradigm similar to DiSMEC [1], and resulting method leads to much better performance than state-of-the-art methods on publicly available datasets consisting of up to 670,000 labels.

References

1 Rohit Babbar, and Bernhard Schölkopf. DiSMEC – Distributed Sparse Machines for Extreme Multi-label Classification. In 10th ACM International Conference on Web Search and Data Mining, 2017.

3.2 Insights on representational similarity in neural networks with canonical correlation

Samy Bengio (Google Inc. – Mountain View, US)

Joint work of Ari S. Morcos, Maithra Raghu, Samy Bengio
Main reference Ari S. Morcos, Maithra Raghu, Samy Bengio: "Insights on representational similarity in neural networks with canonical correlation", CoRR, Vol. abs/1806.05759, 2018.
URL http://arxiv.org/abs/1806.05759

Comparing different neural network representations and determining how representations evolve over time remain challenging open questions in our understanding of the function of neural networks. Comparing representations in neural networks is fundamentally difficult as the structure of representations varies greatly, even across groups of networks trained on identical tasks, and over the course of training. In this work, we present a new projection weighted CCA (Canonical Correlation Analysis) as a tool for understanding neural networks, building off of SVCCA [1], a recently proposed method. We first improve the core method, showing how to differentiate between signal and noise, and then apply this technique to compare across a group of convolutional networks (CNNs), demonstrating that networks which generalize converge to more similar representations than networks which memorize, that wider networks converge to more similar solutions than narrow networks, and that trained networks with identical topology but different learning rates converge to distinct clusters with diverse representations. We also investigate the representational dynamics of recurrent neural networks (RNNs), across both training and sequential timesteps, finding that RNNs converge in a bottom-up pattern over the course of training and that the hidden state is highly variable over the course of a sequence, even when accounting for linear transforms. Together, these results provide new insights into the function of CNNs and RNNs, and demonstrate the utility of using CCA to understand representations.

References

1 Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In Advances in Neural Information Processing Systems, 2017.

3.3 Extreme classification: applications and generalizations

Krzysztof Dembczyński (Poznan University of Technology, PL)

In this talk we shortly reviewed potential applications of extreme classification. We started with well-known applications that are usually referred to in research articles, such as tagging of Wikipedia articles [3] or predicting queries for which a given add will be clicked [2]. We then made a link to two types of dyadic problems [1] where predictions are made for pairs of objects. In the first type only ids of objects are known (e.g., collaborative filtering via matrix factorization), while in the second type also the feature descriptions of objects are given (e.g. link prediction, learning to rank, zero-shot learning). We showed that extreme classification is in-between these two types of dyadic prediction.

To be more precise, consider two types of objects coming from two different domains $\mathcal{X}$ and $\mathcal{Y}$. Assume that each object, either $x \in \mathcal{X}$ or $y \in \mathcal{Y}$, is identified by its id and/or described by a set of features. We are interested in determining the relation between a pair of objects (x, y). This relation could be a label (similar or not), an ordinal value (like stars), or real value (strength of the relation). Some of the objects are seen during training, while the others are not. In such setting, we can define four different learning scenarios: A, B, C and D. Scenario A corresponds to collaborative filtering in which all objects, xs and ys, are known. It is enough to use their ids to perform matrix factorization to obtain the final model. Alternatively, one can also try to use side information available for the objects. Scenarios B and C correspond to (extreme) multi-class or multi-label classification or multivariate regression where one type of objects is completely known during training and plays the role of labels or output variables. The features of the labels or the output variables can also be used to improve the final models. The most challenging scenario is D as the objects of interest have not been seen during training. It corresponds to zero-shot learning in classification, the cold-start problem in recommendation systems, and learning similarity functions in general. This generalized view shows that extreme classification is in fact strictly related to dyadic prediction. A similar observation is also behind the popular Star Space model [4].

This link to dyadic prediction shows new potential applications of extreme classification in a wide spectrum of problems with a large set of possible choices such as product recommendation, smart email replies, suggestion of related queries or assignment of experts to queries posted on Q&A platforms.

References

1 Aditya Krishna Menon and Charles Elkan. Predicting labels for dyadic data. Data Mining and Knowledge Discovery, (21) 2:327-343, 2010
2 Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. Proceedings of the 22nd international conference on World Wide Web, 13-24, 2013
3 Ioannis Partalas, Aris Kosmopoulos, Nicolas Baskiotis, Thierry Artières, George Paliouras, Éric Gaussier, Ion Androutsopoulos, Massih-Reza Amini, and Patrick Gallinari. LSHTC: A Benchmark for Large-Scale Text Classification, CoRR, abs/1503.08581, 2015
4 Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. StarSpace: Embed All The Things!, CoRR, abs/1709.03856, 2017

3.4 Extreme Classification Challenge – Seed Sorting

Matthias Enders (NPZ Innovation GmbH, DE)

Plant seeds are by far the most important source of human nutrition. All over the world large quantities are harvested and processed every year. Only covering five important crops (Barley, Maize, Rapeseed, Wheat, Rice) 7.4x1016 seeds where harvested in 2016. A major step in processing these seeds is cleaning. This is accomplished using the combination of some physical methods (sieving, blowing out lightweight objects, ...) and novel optical seed sorters, which imaging single seeds and sort them according to size, shape and color. This sorting and cleaning is of major importance, as some objects (e.g. see Ergotism / Claviceps purpurea) are hazardous and can contaminate large quantities after milling. On the other hand, the state-of-the-art setup of cleaning machinery has a rough false positive rate of about 0.01. Thus, at least 1% of all harvested seeds are sorted out mistakenly, summing up to more than 27 million tons per year for the five crops given above. Improving sorting systems thus may contribute to meet the necessary increase in food production required to cope with the growing world population. Extreme classification could be employed to tackle a range of challenges like the enormous number of species (classes) which could potentially be found in a seed lot, the bias in the numbers of objects (crop vs. weed) both in the training set, as well as later in the classification or the huge amount of variance within on class versus the small amount of variance between some classes. Furthermore, image acquisition technologies evolve further providing multiple, heterogeneous (multi-modal) data sources which could be combined with traditional imaging.

3.5 Extreme classification challenges in natural language processing

Edouard Grave (Facebook – Menlo Park, US)

Joint work of Edouuard Grave, Armand Joulin, Moustapha Cissé, David Grangier, Hervé Jégou, Nicolas Usunier
Main reference Edouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, Hervé Jégou: "Efficient softmax approximation for GPUs", in Proc. of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, Proceedings of Machine Learning Research, Vol. 70, pp. 1302–1310, PMLR, 2017.
URL http://proceedings.mlr.press/v70/grave17a.html

In this talk, we give a brief (and incomplete) overview of extreme classification problems which arise in the field of natural language processing. First, we discuss the problem of statistical language modeling, which has applications in machine translation, speech recognition and summarization. The goal of language modeling is to learn a probability distribution over sequences of words. This problem is usually framed as learning the conditional probability of word at position t, given the history of all words appearing up to time t-1. Nowadays, this conditional probability is usually estimated using neural networks, which implies computing a softmax over the full vocabulary. Many datasets contain hundreds of thousands of words in their vocabulary, and computing the softmax can be seen as an extreme classification problem. Different approaches have been proposed to speed up this computation bottleneck, such as negative sampling, hierarchical classifiers, or vocabulary selection. A second challenge for language models are out-of-vocabulary words, such as new named entities or unseen inflected forms of words for morphologically rich languages. We discuss potential solutions for this problem, such as: using character level information, copy mechanisms (e.g. in machine translation) or few shot learning with cache models. Finally, in the last part of the presentation, we briefly discuss extreme classification challenges in other applications of natural language processing, such as entity linking, learning word representation or text classification.

3.6 Combinatorial and Structural Results for gamma-Psi-dimensions

Yann Guermeur (LORIA & INRIA Nancy, FR)

Joint work of Yann Guermeur
Main reference Yann Guermeur: "Combinatorial and Structural Results for gamma-Psi-dimensions", CoRR, Vol. abs/1809.07310, 2018.
URL http://arxiv.org/abs/1809.07310

One of the main open problems of the theory of margin multi-category pattern classification is the characterization of the way the confidence interval of a guaranteed risk should vary as a function of the three basic parameters which are the sample size m, the number C of categories and the scale parameter gamma. This is especially the case when working under minimal learnability hypotheses. In that context, the derivation of a bound is based on the handling of capacity measures belonging to three main families: Rademacher/Gaussian complexities, metric entropies and scale-sensitive combinatorial dimensions. The scale-sensitive combinatorial dimensions dedicated to the classifiers of interest are the gamma-Psi-dimensions. This talk introduces the combinatorial and structural results needed to involve them in the derivation of guaranteed risks. Such a bound is then established, under minimal hypotheses regarding the classifier. Its dependence on m, C and gamma is characterized. The special case of multi-class support vector machines is used to illustrate the capacity of the gamma-Psi-dimensions to take into account the specificities of a classifier.

3.7 Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion

Armand Joulin (Facebook – Menlo Park, US)

Joint work of Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Edouard Grave, Hervé Jégou
Main reference Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Edouard Grave: "Improving Supervised Bilingual Mapping of Word Embeddings", CoRR, Vol. abs/1804.07745, 2018.
URL http://arxiv.org/abs/1804.07745

Continuous word representations learned separately on distinct languages can be aligned so that their words become comparable in a common space. Existing works typically solve a least-square regression problem to learn a rotation aligning a small bilingual lexicon, and use a retrieval criterion for inference. In this talk, we propose an unified formulation that directly optimizes a retrieval criterion in an end-to-end fashion. Our experiments on standard benchmarks show that our approach outperforms the state of the art on word translation, with the biggest improvements observed for distant language pairs such as English-Chinese.

3.8 Multi-task Learning with A Very Large Number of Tasks

Christoph H. Lampert (IST Austria – Klosterneuburg, AT)

Joint work of Anastasia Pentina, Christoph H. Lampert
Main reference Anastasia Pentina, Christoph H. Lampert: "Multi-task Learning with Labeled and Unlabeled Tasks", in Proc. of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, Proceedings of Machine Learning Research, Vol. 70, pp. 2807–2816, PMLR, 2017.
URL http://proceedings.mlr.press/v70/pentina17a.html

We study a multi-task learning setting in which a learning system is given a very large number of supervised learning tasks and needs to solve all of them. A typical example is a personalization task, where individual predictors should be constructed for many users. In contrast to previous work, which required that annotated training data must be available for all tasks, we consider a new setting, in which for some tasks, potentially most of them, only unlabeled training data is provided. Consequently, to solve all tasks, information must be transferred between tasks with labels and tasks without labels. Focusing on an instance-based transfer method we analyze two variants of this setting: when the set of labeled tasks is fixed, and when it can be actively selected by the learner. We state and prove a generalization bound that covers both scenarios and derive from it an algorithm for making the choice of labeled tasks (in the active case) and for transferring information between the tasks in a principled way.

3.9 Contextual Memory Trees

John Langford (Microsoft Research – Redmond, US)

Joint work of Wen Sun, Alina Beygelzimer, Hal Daumé III, John Langford, Paul Mineiro
Main reference Wen Sun, Alina Beygelzimer, Hal Daumé III, John Langford, Paul Mineiro: "Contextual Memory Trees", CoRR, Vol. abs/1807.06473, 2018.
URL http://arxiv.org/abs/1807.06473

We design and study a Contextual Memory Tree (CMT), a learning memory controller that inserts new memories into an experience store of unbounded size. It is designed to efficiently query for memories from that store, supporting logarithmic time insertion and retrieval operations. Hence CMT can be integrated into existing statistical learning algorithms as an augmented memory unit without substantially increasing training and inference computation. We demonstrate the efficacy of CMT by augmenting existing multi-class and multi-label classification algorithms with CMT and observe statistical improvement. We also test CMT learning on several image-captioning tasks to demonstrate that it performs computationally better than a simple nearest neighbors memory system while benefitting from reward learning.

3.10 Statistical models of genotype-phenotype associations

Christoph Lippert (Max-Delbrück-Centrum – Berlin, DE)

Joint work of Christoph Lippert, Riccardo Sabatin, M. Cyrus Maher, Eun Yong Kang, Seunghak Lee, Okan Arikan, Alena Harley, Axel Bernal, Peter Garst, Victor Lavrenko, Ken Yocum, Theodore Wong, Mingfu Zhu, Wen-Yun Yang, Chris Chang, Tim Lu, Charlie W. H. Lee, Barry Hick
Main reference Christoph Lippert, Riccardo Sabatini, M. Cyrus Maher, Eun Yong Kang, Seunghak Lee, Okan Arikan, Alena Harley, Axel Bernal, Peter Garst, Victor Lavrenko, Ken Yocum, Theodore Wong, Mingfu Zhu, Wen-Yun Yang, Chris Chang, Tim Lu, Charlie W. H. Lee, Barry Hicks, Smriti Ramakrishnan, Haibao Tang, Chao Xie, Jason Piper, Suzanne Brewerton, Yaron Turpaz, Amalio Telenti, Rhonda K. Roby, Franz J. Och, and J. Craig Venter: "Identification of individuals by trait prediction using whole-genome sequencing data", In Proceedings of the National Academy of Sciences 114.38, 2017: 10166-10171.
URL https://doi.org/10.1073/pnas.1711125114

Technological advances in clinical measurement devices based on sequencing, imaging, and wearables promise to accurately diagnose diseases in their earliest stages when they can be readily treated. Machine learning is central to this vision of personalized medicine, where each individual is monitored based on their medical history, as well as their own genetic and environmental disease risk. While today, medicine is still centered around treating symptoms rather than personalized treatment of disease mechanisms, current prospective cohort studies such as the UK Biobank and the German NaKo that pair deep phenotyping, genetics and detailed longitudinal recordings of occurrence and progression of disease in large numbers of individuals will serve as reference populations to assess and predict disease risk and progression of an individual in a data-driven way. With these large cohorts comprising multi-modal structured data types coming online, the need for Machine Learning methods for extracting, quantifying, and integrating high-dimensional disease phenotypes from multiple data sources is mounting. Accurate statistical models that take into account confounding, data biases and multiple testing are essential to determine robust associations and derive precise risk models for diseases in the presence of environment, lifestyle, medication, and molecular measurements that ultimately will serve as an empirical footing for personalized predictive medicine. While high-throughput methods have been simplifying the process of screening enormously large cohorts for genomic variation and imaging phenotypes, the ability to obtain accurate quantitative phenotypic information is becoming the next bottleneck to closing the genotype-phenotype gap. In my talk I will present a proof-of-concept study, where we applied whole-genome sequencing, detailed phenotyping, and statistical modeling to predict a wide range of phenotypes, including height, weight, BMI, age, and 3D facial images [1].

References

1 Lippert, C., et al. *Identification of individuals by trait prediction using whole-genome sequencing data*. Proceedings of the National Academy of Sciences 114.38 (2017): 10166-10171.

3.11 Gravity: Efficient Training on Very Large Corpora via Gramian Estimation

Nicolas Mayoraz (Google Research – Mountain View, US)

Joint work of Walid Krichene, Nicolas Mayoraz, Steffen Rendle, Li Zhang, Xinyang Yi, Lichan Hong, Ed Chi, John R. Anderson
Main reference Walid Krichene, Nicolas Mayoraz, Steffen Rendle, Li Zhang, Xinyang Yi, Lichan Hong, Ed H. Chi, John R. Anderson: "Efficient Training on Very Large Corpora via Gramian Estimation", CoRR, Vol. abs/1807.07187, 2018.
URL http://arxiv.org/abs/1807.07187

We study the problem of learning similarity functions over very large corpora using neural network embedding models. These models are typically trained using SGD with sampling of random observed and unobserved pairs, with a number of samples that grows quadratically with the corpus size, making it expensive to scale to very large corpora. We propose new efficient methods to train these models without having to sample unobserved pairs. Inspired by matrix factorization, our approach relies on adding a global quadratic penalty to all pairs of examples and expressing this term as the matrix-inner-product of two generalized Gramians. We show that the gradient of this term can be efficiently computed by maintaining estimates of the Gramians, and develop variance reduction schemes to improve the quality of the estimates. We conduct large-scale experiments that show a significant improvement in training time and generalization quality compared to traditional sampling methods.

3.12 Extremely Fast Extreme Classification

Alexandru Niculescu-Mizil (NEC Laboratories America, Inc. – Princeton, US)

Joint work of Alexandru Niculescu-Mizil, Ehsan Abbasnejad
Main reference Alexandru Niculescu-Mizil, Ehsan Abbasnejad: "Label Filters for Large Scale Multilabel Classification", in Proc. of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA, Proceedings of Machine Learning Research, Vol. 54, pp. 1448–1457, PMLR, 2017.
URL http://proceedings.mlr.press/v54/niculescu-mizil17a.html

With the advent of big data, the number of extreme classification problems, as well as the number of labels per problem, is bound to dramatically increase. One consequence of the explosion in the number of labels is a significant increase in the test-time (production time) computational burden. Most approaches to multiclass and multilabel classification, such as the very popular one-vs-all scheme or the Crammer-Singer multiclass SVM, have to systematically evaluate the match between each label and the test instance in order to make a prediction, leading to a test-time complexity linear in the number of labels.

As the number of labels grows the systematic evaluation of all labels becomes prohibitive for applications where the constraints on computational resources and response time are very stringent in production. Examples of such applications are interactive tag recommendation or real-time bidding where a real-time response is required in production; high volume streaming problems such as ad placement where a large volume of data has to be processed in production; or applications where classifiers must be deployed on restricted hardware such as laptops, smartphones or satellites. In all these types of applications, reducing the computational burden in production while maintaining top performance is critical.

In my talk at Dagstuhl I will give a quick overview of existing techniques for reducing the test-time computational burden of multilabel classifiers and I will discuss remaining challenges in this direction.

3.13 Structural Assumptions for Extreme Classification

Pradeep Ravikumar (Carnegie Mellon University – Pittsburgh, US)

Joint work of Ian En-Hsu Yen, Xiangru Huang, Kai Zhong, Inderjit S. Dhillon, Wei Dai, Eric P. Xing, Satyen Kale, Felix X. Yu, Daniel Holtmann-Rice, Sanjiv Kumar

Extreme classification problems, either multiclass or multilabel, have such a large number of classes, that even training or prediction costs that are linear in the number of classes become intractable. State-of-the-art methods aim to reduce this complexity by imposing structural constraints among the labels, or the classifier itself, either implicitly or explicitly. One class of methods exploit correlations among the labels, such as low-rank matrix structure, or a balanced tree structure over the set of labels. A related class of methods aim to compress the space of labels, that in turn imposes implicit constraints on the set of labels. Lastly, some methods impose either primal or dual sparsity on the classifier estimation problem.

We briefly discuss these varied assumptions that have been proposed in the literature, and pose the open question of which assumptions might be most natural in practical extreme classification settings. A related question is understanding the dependence of statistical complexity, specifically the generalization properties of extreme classification methods, on such structural assumptions.

References

1 PD-Sparse: A Primal and Dual Sparse Approach to Extreme Multiclass and Multilabel Classification. I. En-Hsu Yen, X. Huang, P. Ravikumar, K. Zhong, I. Dhillon. In International Conference on Machine Learning (ICML) 33, 2016.
2 PPDSparse: A Parallel Primal-Dual Sparse Method for Extreme Classification. I. En-Hsu Yen, X. Huang, W. Dai, P. Ravikumar, I. S. Dhillon and E. P. Xing. In ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 23, 2017.
3 Loss Decomposition for Fast Learning in Large Output Spaces. I. En-Hsu Yen, S. Kale, F. Yu, D. Holtmann-Rice, S. Kumar, P. Ravikumar. In International Conference on Machine Learning (ICML) 35, 2018.

3.14 Fixing biases in extreme classification

Adith Swaminathan (Microsoft Research – Redmond, US)

Joint work of Thorsten Joachims, Adith Swaminathan, Maarten de Rijke
Main reference Thorsten Joachims and Adith Swaminathan and Maarten de Rijk: "Deep Learning with Logged Bandit Feedback, in International Conference on Learning Representations, ICLR, 2018
URL https://openreview.net/forum?id=SJaP_-xAb

Datasets for extreme multi-class classification and extreme multi-label learning often have severe biases – for instance, manually annotated data-points may have several relevant labels missing – that preclude standard supervised machine learning methods. Propensity-scored loss functions address this bias [1], but training high capacity models (e.g. deep neural networks) with these losses often suffers from propensity over-fitting [2]. Self-normalized estimators remain resistant to such over-fitting but it was not clear how they can be optimized over massive datasets in a scalable way. In recent work [3], we develop a trick to optimize self-normalized estimators using stochastic gradient descent and show how deep neural networks can be trained to fit propensity-scored loss functions reliably.

References

1 Jain, H., Prabhu, Y., and Varma, M. (2016). Extreme multi-label loss functions for recommendation, tagging, ranking and other missing label applications. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'16, pages 935–944, New York, NY, USA. ACM.
2 Swaminathan, A. and Joachims, T. (2015). The self-normalized estimator for counterfactual learning. In Proceedings of the 28th International Conference on Neural Information Processing Systems Volume 2, NIPS'15, pages 3231–3239, Cambridge, MA, USA. MIT Press.
3 Joachims, T., Swaminathan, A., and de Rijke, M. (2018). Deep learning with logged bandit feedback. In Proceedings of the International Conference on Learning Representations.

3.15 Is zero-shot learning possible without side information?

Willem Waegeman (Ghent University, BE)

Joint work of Peter Rubbens, Bac Nguyen, Willem Waegeman

In the talk I discussed a novel challenge in extreme classification, about zero-shot learning without side information. This setting was motivated by applications in species identification, for which often a lot of species are not observed during the training phase. When no side information is available, one can question whether zero-shot learning is still possible. In the talk I proposed a first approach to tackle this challenging problem. The approach consisted of (1) learning a metric that can be transferred to zero-shot classes (2) applying an unsupervised peak detection algorithm to spot novel classes.

4 Working groups

4.1 Generalization bounds and log-time-and-space algorithms

Krzysztof Dembczyński (Poznan University of Technology, PL) and Yann Guermeur (LORIA & INRIA Nancy, FR)

During this working group we discussed the existing results concerning the generalization bounds for multi-class classification with a very large number of categories. We mainly referred to the recent results [2, 1] which suggest square-root or even logarithmic dependence between the error and the number of classes. These results, however, concern only the 0/1 loss. In case of multi-label classification a multitude of loss functions is used and there are still no concrete theoretical results concerning these measures. As the result of the working group we emphasized the need of research in this direction.

We also discussed the possibility of involving the space and time complexity into the confidence interval of a guaranteed risk. We even derived *the first extreme classification bound* that bounds the 0/1 error by the logarithm of the number of classes for any algorithm with logarithmic space and time complexity in the number of classes. Unfortunately, this bound is completely uninformative as it is always greater or equal to 1. Nevertheless, its goal is to show this interesting new research direction in learning theory.

References

1 Y. Guermeur. Combinatorial and Structural Results for gamma-Psi-dimensions. *ArXiv e-prints*, September 2018.
2 Y. Lei, U. Dogan, D.-X. Zhou, and M. Kloft. Data-dependent Generalization Bounds for Multi-class Classification. *ArXiv e-prints*, June 2017.

4.2 Loss functions and types of predictions in multi-label classification

Eyke Hüllermeier (Universität Paderborn, DE)

The discussion in this working group centered around the extension of loss functions from standard classification to multi-label classification to extreme multi-label classification (XMLC). While the step from standard multi-class classification to multi-label classification is characterized by a significant increase of the number of reasonable loss functions that can be used, the step from multi-label classification to XMLC is more concerned with the question of which loss functions are still meaningful in settings with an extremely large label space.

To structure the discussion in a systematic way, a distinction was made between the type of ground truth that can be assumed and the type of prediction produced by a learning algorithm; these two do not necessarily coincide. Examples for assumptions on the ground truth include subsets and graded subsets (in the latter, the relevance of a label is a matter of degree). In this regard, there was also an interesting discussion about factual versus counterfactual ground truth. Indeed, there are many applications in which the existence of a (unique) ground truth is not obvious, or in which the "truth" is not independent of the prediction itself (as an example, the case of recommender systems was discussed). As for the predictions, there is an even larger spectrum of possibilities, including subsets, graded subsets, rankings, stratified rankings, scored rankings, etc. The case of rankings appears to be of specific importance in XMLC. All these predictions can be generalized further. For example, rankings could be partial instead of total orders. Moreover, predictions can be equipped with information about the uncertainty of the learner.

In principle, a loss function could be defined for each combination of ground truth and type of prediction. As for the reasonableness of such combinations, there was an agreement that this strongly depends on the purpose of a prediction and the type of application. A longer discuss centered around the idea of abstention, also known under the notion of "eject option" in standard classification. Abstention seems to be useful and highly relevant in XMLC, even if it did not attract much attention so far. It even appears to be more interesting in XMLC than in standard classification, because in XMLC it can be partial (i.e., the learner may abstain on some but not all labels). Obviously, allowing for (partial) abstention again calls for a proper adaptation of loss functions. Finally, there was a discussion of the case where the label space is equipped with a structure, i.e., where labels are not simply identical or different from each other; for example, it might be possible to define a natural measure of similarity between labels, or an order relation like in ordinal classification. Needless to say, loss functions should take such a structure in account. Again, in spite of the importance and practical relevance, there is only little work on this issue so far.

4.3 Deep eXtreme Classification

Marius Kloft (TU Kaiserslautern, DE)

In the working group Deep eXtreme Classification we have considered the central question: How can we learn good representations for XC using deep learning methodologies?

We found that typically the label matrix (# instances x # label classes) is very sparse, so when applying deep learning we have a misfit of a high number of parameters to learn, yet a sparse target. This raises the question of developing models that explore the given structure of the target efficiently and effectively.

Furthermore, already for shallow models, XC induces a substantial computational burden. Some authors address this by using extensive CPU parallelization (cf. Dismec). On the hand, vanilla (non-XC) deep learning requires substantial computational (GPU) resources. Combining XC with deep learning raises the problem of developing efficient architectures and computational infrastructures to train deep XC models.

The workgroup participants agreed that potentially deep XC may offer further boosts in accuracy, but further breakthroughs into that direction might be necessary to get it working.

4.4 Zero-Shot Learning and Long-Tail Labels

Alexandru Niculescu-Mizil (NEC Laboratories America, Inc. – Princeton, US)

In this working group we discussed the about few and zero-shot learning in the context of extreme classification. Due to the extremely large number of labels, the majority of them are bound to have very few training examples, and many might not even appear in the training set. Thus dealing with tail labels and with new labels is one of the most important challenges facing extreme classification.

In the working group we talked about whether it would be more appropriate to tread head and tail labels differently, for example, by using different techniques that might be more suitable for different conditions. We discussed the possibility of using "label features" to enhance the accuracy and/or speed for tail labels, and about the necessity of using them for zero-shot learning. Label features convey information about the labels themselves, rather then individual examples. Such information may come in the form of label descriptions, label taxonomies, properties associated with labels, etc. and it is prevalent in real applications. Finally, we have remarked the unreasonable effectiveness of one vs. all classification.

Participants

- Maximilian Alber
TU Berlin, DE
- Rohit Babbar
Aalto University, FI
- Samy Bengio
Google Inc. –
Mountain View, US
- Alexander Binder
Singapore University of Technology and Design, SG
- Evgenii Chzhen
University Paris-Est –
Créteil, FR
- Kunal Dahiya
Indian Institute of Technology –
New Dehli, IN
- Krzysztof Dembczyński
Poznan University of Technology, PL
- Urun Dogan
Microsoft Research UK –
Cambridge, GB
- Matthias Enders
NPZ Innovation GmbH, DE
- Asja Fischer
Ruhr-Universität Bochum, DE
- Johannes Fürnkranz
TU Darmstadt, DE
- Thomas Gärtner
University of Nottingham, GB
- Edouard Grave
Facebook – Menlo Park, US
- Yann Guermeur
LORIA & INRIA Nancy, FR
- Eyke Hüllermeier
Universität Paderborn, DE
- Christian Igel
University of Copenhagen, DK
- Himanshu Jain
Indian Institute of Technology –
New Dehli, IN
- Kalina Jasinska
Poznan University of Technology, PL
- Armand Joulin
Facebook – Menlo Park, US
- Nikos Karampatziakis
Microsoft Research –
Redmond, US
- Matthias Kirchler
HU Berlin, DE
- Marius Kloft
TU Kaiserslautern, DE
- Christoph H. Lampert
IST Austria –
Klosterneuburg, AT
- John Langford
Microsoft Research –
Redmond, US
- Antoine Ledent
TU Kaiserslautern, DE
- Christoph Lippert
Max-Delbrück-Centrum –
Berlin, DE
- Nicolas Mayoraz
Google Research –
Mountain View, US
- Jinseok Nam
TU Darmstadt, DE
- Alexandru Niculescu-Mizil
NEC Laboratories America, Inc.
– Princeton, US
- Yashoteja Prabhu
Indian Institute of Technology –
New Dehli, IN
- Pradeep Ravikumar
Carnegie Mellon University –
Pittsburgh, US
- Adith Swaminathan
Microsoft Research –
Redmond, US
- Manik Varma
Microsoft Research India –
Bangalore, IN
- Willem Waegeman
Ghent University, BE
- Marek Wydmuch
Poznan University of Technology, PL