Research Summary and Agenda

Pavel P. Kuksa
Machine Learning Department
NEC Laboratories America
pavel@pkuksa.org

November 20, 2013

My research interests focus on advancing state-of-the-art in both the theory and practice of machine learning, bioinformatics, natural language processing, and algorithms as applied to analysis, modeling, and understanding of complex and large data sets (e.g., text, biomedical, and time series data) and multi-disciplinary applications.

The core theme of my research is the development of computational methods, algorithms, and models that are applicable in a variety of distinct fields and domains by combining ideas from machine learning, algorithms, language processing, computational biology, pattern recognition and statistics.

In the following I will outline my main theoretical and practical contributions in the fields of sequence modeling and analysis, bioinformatics, and natural language processing using machine learning and algorithmic approaches.

General sequence analysis

Analysis of large scale sequence data has become an important task in machine learning and data mining inspired by numerous scientific and technological applications such as text and audio mining, biological sequence analysis, and time series modeling.

Sequence analysis tasks are made challenging by the variability in the sequence lengths and their content, absence of readily available representations/feature vectors (feature extraction problem), potential existence of important features on multiple scales, as well as the size of the sequence alphabets and datasets. Typical alphabet sizes can vary widely, ranging in size from 4 nucleotides in DNA sequences, up to thousands of words from a language lexicon for text documents, resulting in potentially very high-dimensional and computationally challenging representations. Sequences/strings within the same class, such as the proteins in one fold or documents about politics, can exhibit a wide variability in the primary sequence content. Moreover, important datasets continue to increase in size, easily reaching millions of sequences.

As a consequence, to accomplish many sequence analysis tasks such as classification, clustering, pattern extraction one needs accurate yet efficiently computable representations and algorithms. These representations have to be robust, multiscale, so that they can be computed quickly over a range of parameter settings and scale with the data set size and alphabets.

To address these issues, in my work I have introduced a number of novel computationally efficient and accurate methods, representations, and algorithms for a variety sequence analysis tasks, such as classification, clustering, and information extraction.

Similarity inference in k-mer sequence models. k-mer based sequence models are widely used in many bioinformatics, text (k-grams), and time series analysis applications. Sequence similarity

assessment, clustering, classification are all inherently rely on the ability to compute sequence similarity functions efficiently.

For approximate (e.g., with mismatches) sequence similarity computation, my work introduces novel algorithmic approaches that improve currently known time bounds for such tasks and show orders-of-magnitude running time improvements [KHP08e, KP12]. E.g., the complexity of the k-mer based computation with m mismatches improves from a much higher complexity $O(k^{m+1}|\Sigma|^m(|X|+|Y|))$ of existing algorithms to $O(c_{k,m}(|X|+|Y|))$, where $c_{k,m}$ is *independent* of the alphabet Σ.

These algorithms can be easily generalized to other sequence representations and families of string kernels, such as the spectrum, gapped, wildcard kernels, as well as to *semi-supervised* settings [KHP08e].

I demonstrate the benefits of our algorithms on many challenging classification problems from various application domains, such as detecting homology (evolutionary similarity) of remotely related proteins, recognizing protein fold, and performing classification of music samples, and music artist recognition.

Low computational complexity of these algorithms opens the possibility of analyzing very large datasets under both fully-supervised and semi-supervised settings as well using more complex sequence models to improve predictive performance. A systematic method for computation of large sequence kernel models and optimal selection of more complex models is described in [KP12]. In the biological domain, the above-mentioned advancements gave rise to large scale and accurate DNA-based taxonomy systems for species identification (DNA barcoding) [KP09a].

Sparse, spatial multi-dimensional sequence models. I have proposed a general approach for sequence classification based on a novel idea of multi-dimensional sequence embedding, *sparse spatial sample* (SSS) representation [KP10b, KHP08b, KHP08a, KHP08c]. This approach achieves state-of-the-art performance in both *discrete*- and *continuous*- valued sequence classification tasks including supervised and large-scale semi-supervised structural classification of proteins, text classification, music genre classification, and artist recognition.

Multivariate sequence analysis. While most existing sequence kernel methods are restricted to one-dimensional sequence data, my recent work [Kuk13b, KKP12] focuses on the problem of multi-dimensional sequence analysis using kernel methods. I have introduced novel computationally efficient multivariate kernel methods for analysis of multivariate sequences based on similarity-preserving binary embeddings and direct feature quantization. The proposed approaches demonstrate excellent prediction accuracy in many challenging biological sequence analysis and music classification problems [Kuk13b, KKP12].

Motif discovery. I have proposed novel algorithms for an important problem of finding representative patterns (motifs) in sequence databases [KP09b, KP10a]. We present new deterministic and exact algorithms for finding common patterns with the search complexity that scales well with the input length and size of the alphabet. Proposed algorithms improve motif search efficiency by focusing on the input instances that are more likely to be motif instances as opposed to using the entire input directly. Compared to existing exact algorithms for motif discovery our algorithms significantly improve search efficiency in the important cases of *large-alphabet* inputs (e.g., protein, or extended alphabet DNA sequences) and inputs of *large length*. This result extends applicability of the exact motif search algorithms to more complex problems requiring analysis of biological sequence data modeled as strings over large alphabets.

In summary, my work here spans a wide variety of problems in sequence analysis space where I proposed a number of approaches for efficient and accurate sequence matching, classification, and pattern extraction that (1) improve both in theory and practice efficiency/complexity of sequence matching/comparison, and (2) enhance performance on practical sequence classification tasks (text, music, and biological sequences).

Natural language processing

Many typical tasks in the natural language processing (e.g., part-of-speech tagging, named entity recognition, chunking, semantic role labeling) can be viewed as a task of assigning *labels* to words in the text sentences, i.e. reduce to a general sequence labeling or tagging problem. While words are a fundamental building block in the language, the *features* of these words serve as a fundamental building block in the natural language processing systems.

To improve supervised sequence tagging, I have proposed a novel, scalable *semi-supervised method*, *word codebook learning* (WCL) [KQ10, QKC$^+$09, QCK$^+$09]. WCL learns a class of word-level feature embeddings to capture word semantic meanings or word label patterns from a large unlabeled corpus. Words are then clustered according to their embedding vectors, and the code-words assigned to the words are treated as new word attributes and are added as features for entity tagging. Two types of word-codebook learning are proposed: (1) General WCL, where an unsupervised method uses contextual semantic similarity of words to learn accurate word feature representations; and (2) Task-oriented WCL, where for every word a semi-supervised method learns target-class label patterns from unlabeled data using supervised signals from a pre-trained model. WCL yields state-of-the-performance on many competitive benchmarks, including bio-named entity tagging, article classification in biomedical literature, etc ([KQ10, QKC$^+$09, QCK$^+$09]).

In [CWB$^+$11], we proposed a novel unified deep learning architecture and a learning algorithm that can be applied to a variety of natural language processing tasks. A *single learning system* (a deep learning neural network) is used to learn accurate word and sentence representations from mostly unlabeled training data. It generalizes over many NLP tasks avoiding the need for task-specific engineering, and shows excellent performance on multiple challenging benchmarks for semantic role labeling, chunking, part-of-speech tagging, named entity recognition, etc.

In [KQB$^+$10], I have introduced a simple, scalable semi-supervised approach that leverage large amounts of text from biomedical literature to improve entity relationship prediction (e.g., protein-protein interaction) and relevant article retrieval. The approach shows effective improvements over existing techniques on multiple benchmarks.

Summary of contributions

In summary, my main theoretical and practical research contributions in applied machine learning, modeling and analysis of sequences, large scale learning algorithms and inference and their applications in text analysis, bionformatics, natural language processing are as follows

- **Theoretical contributions**:
  - New computationally efficient methods for sequence similarity computation over high-dimensional feature spaces [KHP08e, KP12]. Proposed methods reduce the complexity for inexact sequence comparison by orders-of-magnitude. This allows to explore more

complex sequence models that are more accurate and improve over the prior art in sequence prediction. The utility of such more complex models was not possible to investigate before.

- New, faster algorithmic method for common pattern (motif) discovery in data modeled as symbolic strings ([KP09b], [KP10a]). The proposed method is computationally more efficient than existing exhaustive motif enumerators and its search complexity is independent of the cardinality of the sequence domain.
- New spatial sample sequence models and kernel algorithms [KHP08b, KHP08c, KP10b] that apply to both *discrete* and *continuous* sequence domains. Unlike the traditional k-mer models, the proposed model incorporates spatial arrangements of individual k-mers to improve accuracy traditional sequence models, while preserving linear computational complexity of the simpler k-mer models.
- A new systematic method for computing large sequence kernels ([KP12]). The method allows exactly evaluate kernels for long substrings with potentially many mismatches, thus enabling selection of optimal inexact kernels for a particular task.
- Novel semi-supervised learning method (WCL) and *discrete* distributed task-specific language models ([KQ10]). In contrast to existing methods, the proposed method efficiently learns *discrete* and task-specific word representations indicative of potential word label/class in sequence annotation tasks.

- **Practical contributions**:
  - state-of-the-art remote sequence similarity inference systems for biological and music sequences ([KHP08a], [KHP08c], [KHP08b], [KHP09], [KHP08d], [KP07], [KP08], [Kuk12],[Kuk13b])
  - state-of-the-art semi-supervised named entity recognition systems ([KQ10],[QKC$^+$09])
  - state-of-the-art sequence labeling systems with both labeled and unlabeled data ([QKC$^+$09], [QCK$^+$09], [KQB$^+$10])
  - state-of-the-art end-to-end NLP systems for part-of-speech (POS) tagging, named entity recognition (NER), text chunking, etc. ([QKC$^+$09], [CWB$^+$11]) with *deep learning* neural network architectures.
  - efficient, large-scale semi-supervised conditional random field (CRF) models for sequence labeling and natural language processing ([KQ10])
  - state-of-the-art large-scale taxonomy system for DNA barcoding of life ([KP09a],[KP07])
  - efficient, scalable state-of-the-art implementations of kernel matching algorithms for structured objects ([KHP08e]) with orders-of-magnitude running time improvements over the previous art.
  - efficient and accurate semi-supervised entity relationship extraction system from text ([KQB$^+$10]).
  - state-of-the-art music classification system using multivariate kernel method ([KKP12, Kuk13a]).

Future Agenda

My future research goals are directed towards a long-term goal of understanding how to built fast, efficient, and robust end-to-end learning and pattern recognition systems that span multiple application fields and disciplines (natural language, biological domain, audio, visual perception, time series, etc) and are applicable to general types of data (such as sequences, sets, vectors, images, time series, etc.)

To this end, in my research I will continue to develop state-of-the-art algorithms, models, and systems that provide efficient means for analysis and modeling of complex and diverse data sets such as text, biological sequences, web, multimedia (music, video), time series, and biomedical data collections.

In the near term, I am particularly interested in exploring the following research directions:

- In the biomedical domain, scaling and improving computational modeling, algorithmic, and machine learning approaches to analysis of *massive* collections of next-generation genomic, metagenomic, transcriptomic sequence data, by combining ideas from algorithms, machine learning, computational biology, and statistics.

- For pattern analysis, continue exploring computationally efficient and general algorithms for similarity inference, in particular for multivariate time series and biological experimental data. Explore efficient kernel methods for information extraction and sequence tagging tasks.

- Parallel, distributed, and cloud computing machine learning models, algorithms, ensemble approaches and their applications in text and biological sequence analysis domains.

- Explore methods for multi-modal data analysis that combine diverse data types to build more accurate models, e.g., text, image, or audio in information retrieval or genomic sequence data, genome-wide assosiation studies (GWAS) data, and transcriptomic data in biological data analysis.

- In the text mining and the natural language domain, exploring and advancing areas of large scale matching and retrieval of text documents, similarity inference in content-recommendation systems, document similarity metrics, multi-level efficient deep learning architectures for document summarization and sentence level tasks. In the language modeling domain, explore methods for constructing more precise distributed real-valued and discrete language models and their applications in biological domain.

A higher goal here is to use computation to make sense of biological, text, etc., sequence information. This is one of central themes of the next decade in biological and text analysis research.

In the longer term, I am interested in multi-disciplinary algorithms, architectures, and learning methods for parallel, distributed processing, analysis, and efficient and accurate empirical inference from large, complex, and diverse data sets.

References

[CWB+11] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning*, 12:2493–2537, 2011.

[KHP08a] Pavel Kuksa, Pai-Hsi Huang, and Vladimir Pavlovic. Fast and accurate multi-class protein fold recognition with spatial sample kernels. In *Computational Systems Bioinformatics: Proceedings of the CSB2008 Conference*, pages 133–143, 2008. Acceptance rate: 30/135 (22%).

[KHP08b] Pavel Kuksa, Pai-Hsi Huang, and Vladimir Pavlovic. Fast protein homology and fold detection with sparse spatial sample kernels. In *19th International Conference on Pattern Recognition ICPR 2008*, 2008. Acceptance rate: 18% (oral). Best paper nominee.

[KHP08c] Pavel Kuksa, Pai-Hsi Huang, and Vladimir Pavlovic. A fast, semi-supervised learning method for protein sequence classification. In *8th International Workshop on Data Mining in Bioinformatics (BIOKDD 2008)*, pages 29–37, 2008. Acceptance rate: 8/25 (32%).

[KHP08d] Pavel Kuksa, Pai-Hsi Huang, and Vladimir Pavlovic. On the role of local matching for efficient semi-supervised protein sequence classification. In *BIBM*, 2008. Acceptance rate: 38/156 (24%).

[KHP08e] Pavel Kuksa, Pai-Hsi Huang, and Vladimir Pavlovic. Scalable algorithms for string kernels with inexact matching. In *NIPS*, 2008. Spotlight Presentation. Acceptance rate: 123/1022 (12%).

[KHP09] Pavel Kuksa, Pai-Hsi Huang, and Vladimir Pavlovic. Efficient use of unlabeled data for protein sequence classification: a comparative study. *BMC Bioinformatics*, 10(Suppl 4):S2, 2009. Impact factor: 3.78.

[KKP12] Pavel P. Kuksa, Imdadullah Khan, and Vladimir Pavlovic. Generalized similarity kernels for efficient sequence classification. In *SDM*, 2012.

[KP07] Pavel Kuksa and Vladimir Pavlovic. Fast kernel methods for SVM sequence classifiers. In *WABI*, pages 228–239, 2007. Acceptance rate: 37/131 (28%).

[KP08] Pavel Kuksa and Vladimir Pavlovic. Approximate substructure matching for biological sequences. In *NIPS, Machine Learning in Computational Biology*, 2008.

[KP09a] Pavel Kuksa and Vladimir Pavlovic. Efficient alignment-free dna barcode analytics. *BMC Bioinformatics*, 10(Suppl 14):S9, 2009. Impact factor: 3.78.

[KP09b] Pavel Kuksa and Vladimir Pavlovic. Fast motif selection for biological sequences. In *IEEE International Conference on Bioinformatics and Biomedicine BIBM'09*, 2009. Acceptance rate: (44+37)/233 (35%).

[KP10a] Pavel Kuksa and Vladimir Pavlovic. Efficient motif finding algorithms for large-alphabet inputs. *BMC Bioinformatics*, 11(Suppl 8):S1, 2010.

[KP10b] Pavel P. Kuksa and Vladimir Pavlovic. Spatial representation for efficient sequence classification. In *ICPR*, 2010. Acceptance rate: 385/2140 oral (18%).

[KP12] Pavel P. Kuksa and Vladimir Pavlovic. Efficient evaluation of large sequence kernels. In *KDD*, 2012.

[KQ10] Pavel Kuksa and Yanjun Qi. Semi-supervised bio-named entity recognition with word-codebook learning. In *SDM*, 2010. Acceptance rate: 82/351 (23%).

[KQB+10] Pavel P. Kuksa, Yanjun Qi, Bing Bai, Ronan Collobert, Jason Weston, Vladimir Pavlovic, and Xia Ning. Semi-supervised abstraction-augmented string kernel for multi-level bio-relation extraction. In *ECML*, 2010. Acceptance rate: 106/658 (16%).

[Kuk12] Pavel P. Kuksa. 2d similarity kernels for biological sequence classification. In *BIOKDD*, 2012.

[Kuk13a] Pavel Kuksa. Computanionally efficient multivariate sequence kernels. In *In submission*, 2013.

[Kuk13b] Pavel P. Kuksa. Biological sequence analysis with multivariate string kernels. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, March 2013.

[QCK+09] Yanjun Qi, Ronan Collobert, Pavel Kuksa, Koray Kavukcuoglu, and Jason Weston. Combining labeled and unlabeled data with word-class distribution learning. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management CIKM 2009*, pages 1737–1740, 2009. Acceptance rate: (123+171)/847 (20% short paper).

[QKC+09] Yanjun Qi, Pavel P. Kuksa, Ronan Collobert, Kunihiko Sadamasa, Koray Kavukcuoglu, and Jason Weston. Semi-supervised sequence labeling with self-learned features. In *Proc. International Conference on Data Mining (ICDM'09)*. IEEE, 2009. Acceptance rate: 8.9% regular (70/786).