Research Synopsis, Vikram Pudi

Key citations and products have been directly linked from the synopsis below.

Over the last decade, data science has emerged as a key global enabler, upgrading every domain of human endeavour. Data science is normally used to refer to the application of data science techniques to guide and automate decision making. On a deeper level, data science includes not just the skill of applying ready-made techniques available in popular toolkits and libraries, but also the innovation of data-driven techniques for decision making.

In this context, the focus of my research has been on both innovation of core data science techniques and on their applications, leading to building large, usable systems, involving interdisciplinary efforts.

The algorithms we have built have largely been designed to be *simple*, *modular*, *customizable*, *generic (domain independent) and data driven*, and have been demonstrated to achieve these desirable traits without sacrificing on *speed* and *accuracy.* Our algorithms cover the entire spectrum of data-driven tasks and techniques, as listed later in this section.

The applications and systems we have built have been designed to have minimal clutter-free interfaces, improve productivity, and be easily maintainable.

Core Algorithms Developed

Classification: Developed a classifier named ACME which is theoretically robust (as it relies on maximum entropy principle) and practical (as it uses frequent itemsets as features). It is fast and accurate even with 1000s of features, so the data analyst does not need to be very picky during feature selection.

Regression: Developed 5 fast and accurate regression algorithms – PAGER, GEAR, BINGR, BINER and CLUEKR. It is not necessary to pre-select a mathematical *form* for the regression curve. So, unlike most regression algorithms, our algorithms are generic and don't need to be tweaked for individual applications.

Feature learning: Developed Autolearn – a regression-based data-driven feature generation and selection technique. In recent years, the importance of feature engineering has been confirmed by the exceptional performance of deep learning techniques, that automate this task for some applications. For others, feature engineering requires substantial manual effort in designing and selecting features and is often tedious and non-scalable. Being data-driven, Autolearn eliminates this manual effort.

Classifying imbalanced data: Although *k*NN is a ubiquitous classification tool with good scalability, it suffers from the drawback of using only local prior probabilities to predict instance labels, and hence it does not take into account the class distribution around the neighborhood of the query instance. This leads to undesirable performance on imbalanced data, which we solve in our approach.

Sequence Classification: Our algorithm, RBNBC, uses a novel formulation of Naive Bayes that uses *maximal frequent subsequences* as features and incorporates *repetitions of subsequences* within each sequence while computing feature probabilities. Application on biological sequences showed this generic approach to perform as well as extensively tuned domain-based classifiers.

Frequent Pattern Mining: Developed a frequent itemset miner called ARMOR and showed its speed to be within a factor of two of an optimal lower-bound. It was also a challenge to identify and implement a practically useful lower-bound.

Uniqueness Mining: We identified the problem to determine the manner in which *each record in a dataset differs* from the general trend, and provided an efficient algorithm for it. In the real world, records typically represent objects or people and it is often worthwhile to know what special properties are present in each object or person, so that we can make the best use of them.

Similarity Computation: Developed a data-driven similarity computation technique called DISC that doesn't require human experts. It works for both numeric and categorical attributes. DISC outperformed all competing approaches when used in classification and clustering algorithms that need similarity computation.

Clustering: Used frequent itemsets as features to cluster text-documents into a hierarchy of clusters. We enhanced the obtained clusters by using Wikipedia as external knowledge and also to name the clusters.

Active-learning: Developed active-learning classifiers based on naïve bayes, kNN and association rules that interact with the user. These classifiers recognize difficult-to-classify records and ask the user to confirm their true class. We showed that the accuracy of the resulting classifiers improved significantly, even when using less initial training data. Moreover, the number of queries posed to the user reduced exponentially as the classifier ran.

Applications

We have applied data science and analytics in diverse, broad and potentially high-impact applications. Our applications primarily lie in the broad areas of:

Mining research data: Our work includes several techniques to help explore trends in the growing collection of research papers, topics, conferences and authors. With the explosion of researchers and research articles in the recent past, it is becoming increasingly difficult to keep track of one's own field and the wider area of which it is a part.

We have designed efficient page-rank style algorithms to rank research papers, topics, papers, authors and conferences based on the citation network, identify past, current and emerging research topics, evolution of topics, hierarchies of topics, and determine dispersion-based similarity of papers. We can also automatically recommend emerging topics to authors and recommend possible research collaborations. We have also mined intellectual influence associations and developed methods to automatically extract structured information such as problems, competing algorithms, aims, methods and results from research papers. Using these we can identify grand challenges and saturated problems. Finally, we also have neural network embedding based approaches for creating vector representations of authors (author2vec) and scientific papers (paper2vec).

Natural language processing (NLP): We proposed *associative context classification*, a generic associative classification approach which uses common contexts to group potentially related items. In our research, we have demonstrated the application of this proposed approach in developing solutions to a few representative NLP tasks such as part-of-speech (POS) tagging and word-sense disambiguation.

Our methods are semi-supervised and require only a small amount of annotated data. Being generic, our methods perform well even without using extensive language-specific domain knowledge. They are especially suitable for resource-poor languages which lack domain resources.

Smart Power Systems: Buildings account for about 40% of total global power consumption. Energy feedback information provided by smart meters can enable consumers to reduce consumption by 5-15%.

We have used data analytics on power consumption signatures, for automatic identification of electrical appliances (plug-loads) and non-intrusive load-monitoring (NILM). Our techniques are novel in being able to work on low-quality power with varying and fluctuating voltages – conditions which are prevalent in developing countries like India. Our published results show significantly better accuracy than state-of-the-art.

Cricket: We have used data analytics in the sport of cricket to predict the outcome of One Day International (ODI) cricket matches and for the quantitative assessment of player performance. Our work suggests that the relative team strength between competing teams forms a distinctive feature for predicting the winner. Modeling the team strength boils down to modeling individual player's batting and bowling performances, forming the basis of our approach. We use career statistics as well as the recent performances of a player to model him.

Music recommendation: People often have implicit preferences while listening to music, though these preferences might not always be the same while they listen to music at different times. For example, a user might be interested in listening to songs of only a particular artist at some time, and the same user might be interested in the top-rated songs of a genre at another time. We model the implicit short-term user preferences for music recommendation using a novel concept of subsessions, which can be identified as short windows in which user-preferences are constant. Experiments on the user listening histories taken from Last.fm indicate that this approach beats existing methodologies in predicting the next recording a user might listen to. Deep-learning enhancements to this methodology produces even better results.

Systems Built

Metabolomics platform: Metabolomics is an area in bio-informatics related to the study of small molecules participating in networks of chemical reactions (or metabolic pathways) in living systems. We built an end-to-end platform for metabolomics researchers to upload, analyze, and share metabolite datasets and studies. Our platform is available at http://metabolomics.iiit.ac.in/. I was the chief architect of this system. The effort spanned over 5 years, with multiple student teams in succession, starting with vague requirements, as is typical of large interdisciplinary projects. It was built on web2py, a python-based web framework for the backend. This project is funded by DBT, Govt of India to the tune of Rs.4.6 crores. Our experience and learning in successfully handling interdisciplinary projects of this scale has been captured in this paper on "Computational Core for Plant Metabolomics: A Case for Interdisciplinary Research".

Log analysis to debug mobile protocols: A project from Qualcomm – the task was to identify software bugs in mobile protocols by analyzing software logs marked as pass or fail. This is an important problem because, currently, engineers manually identify software bugs by analyzing failed logs. This is time-consuming, inefficient and expensive. An automated or semi-automated solution can minimize the involvement of engineers. The key challenge was to scale to large data: Each log is in the range of gigabytes consisting of about 3 lakh messages. Each year, we may expect about 5000 test-cases generating about 3000 fail logs and 30000 pass logs. This is in the range of 30 terabytes per year. A suffix-tree based solution was designed and deployed.

Simulating infection spread: While there is a definite space for recent advances in AI and deep learning in tackling the COVID-19 pandemic, such as in drug discovery, medical imaging and diagnosis, answering basic questions such as "How much does social movement impact viral spread?" and "How long will this epidemic last?" does not require a big hammer. It is important to be able to answer such questions with *repeatable certainty*, with *interpretable* tools that are *simple*, *accessible* and *clear* to the general public. In the age of information explosion, where the reliability of information is always questionable, and where it requires authority to convince, it would be pure and simple if the information presented can be directly tested by the viewer.

These criteria require a simulation tool that is accessible over the web, on both desktops and mobile phones, and is simple to use. This was the motivation to develop this "Viral Simulator". By being available in a timely manner, this simulator attracted wide media attention and was also featured in the ACM SIGMOD blog.

Radiology Platform: We designed and developed RadSlice, a complete and active platform for online radiology reporting, with about 100+ registered radiologists and 30+ registered hospitals. The platform can be directly connected to radiology equipment (MRI/CT/X-ray modalities) through a PACS server. So, study images can be directly loaded from the equipment to the platform. The loaded cases can be assigned to radiologists for reporting.

To respect privacy and reduce overall bandwidth usage and carbon footprint of the platform, the images are directly transferred from the hospital to the radiologist and not stored on any central server. Hospitals may also choose to make their images available for research purposes, after getting patient consent. The images are then and stored centrally and made available after stripping out identifying (DICOM) information.