Machine Learning Algorithms on HPCC/ECL Platform

The FAU's Center for Advanced Knowledge Enablement (CAKE) developed a wide range of machine learning algorithms on the High Performance Cluster Computing (HPCC) platform. This platform includes traditional algorithms such as Naïve Bayes and K-Nearest Neighbors, to more advanced techniques such as Deep Learning. This enables researchers and practitioners to apply machine learning algorithms on big data to extract patterns and perform predictive tasks. The HPCC architecture, written in conjunction with the ECL programming language, is LexisNexis's answer to applying machine learning methods on big data.

HPCC provides a platform for implementing parallel, distributed, and scalable machine learning algorithms. The general linear algebra and statistical operations implemented in HPCC along with the data structures provide an ideal platform for implementing the machine learning algorithms.

Machine learning algorithms using the vastly increasing volumes of personal, wearable health information to anticipate health risks and concerns given the user's historical and present lifestyle information. Image credit: Fitbit.

The amount of data produced from bioinformatics to social media to web documents, has exploded in recent years. Many traditional approaches fail when dealing with terabyte, petabyte, or larger datasets preventing consumers from fully benefiting from their data. Analyzing these large amounts of so-called big data opens up new research areas that were not possible 20 years ago. Major companies such as Google, Facebook, Twitter, and Amazon would be less effective without taking advantage of big data analytics. Big data analysis demands large-scale systems to both manage and process huge quantities of data. Big data with characteristics such as high volume, high complexity, and data heterogeneity require new ways of thinking and new paradigms for knowledge extraction.

This breakthrough is an improvement over more traditional analytics which rely heavily on human analysts. The sheer volumes of data that have high amounts of potential correlations and hidden patterns do not allow for comprehensive analysis using traditional data analytic tools. Machine learning and big data tools overcome these problems by leveraging properties intrinsic to the data to infer semantics and formats, deriving effective and general algorithms for data processing and analytics.

In comparison to traditional analytics, machine learning delivers the promise of extracting patterns in an automated fashion with far less reliance on human supervision. The methods are data driven. They thrive on and benefit from increased data because with more information, more can be learned. Given the limitations of human comprehension in the face of truly massive amounts of data, machine learning is able to discover hidden patterns on very large-scales.

End users/consumers receive multiple advantages from applying machine learning using big data to do many different kinds of predictive tasks. These are in various domains ranging from fraud detection and product recommendations to energy load forecasting to healthcare predictions. Example end-user products include: 1) Machine learning algorithms using the vastly increasing volumes of personal, wearable health information, via devices such as FitBit, to anticipate health risks and concerns given the user's historical and present lifestyle information; 2) HPCC big data processing and machine learning methods to analyze crime patterns to anticipate specific areas likely to have near term criminal activities and to adjust police resources accordingly; 3) Frameworks to collect and analyze historical and real-time power plant information to predict site-wide and individual component failures producing warnings and predictive work orders mitigating costs due to failures and equipment replacement, and; Analytics based on energy smart grid network data to optimize electrical loads and to anticipate specific failure points in order to avoid loss of service by fixing issues remotely or by dispatching crews prior to anticipated failures.

> **Economic impact**: Machine learning provides reliable and accurate predictions for a wide range of domains. Fraud detection and prevention has contributed to billions of dollars in recovered funds. In 2013, the IRS prevented or recovered $24.2 billion related to identity theft fraud. Predicting shopping patterns can allow stores to better anticipate customer needs and increase overall sales revenue; thus shifting the customer interaction paradigm from a mostly reactive process into a more proactive one wherein systems predict consumer needs and offer optimized services and products. Predicting diseases can reduce burdens on healthcare. Cyber security is critical for national security. Big data and machine learning approaches work synergistically to anticipate and prevent costly attacks. The HPCC architecture provides an established compute resource where big data and machine learning go hand-in-hand without new or additional development and storage costs. The use of ECL to implement the machine learning algorithms allows for better. Since the ECL language is automatically compiled into C++, a more traditional imperative/procedural lower level language, it's easier to measure efficiency gains by comparing the number of lines of code between those lines of code manually written in ECL versus lines of C++ code that are automatically generated by the same ECL codebase. This ratio is roughly 100/1, indicating two orders of magnitude improvements in coding efficiency by leveraging the higher level declarative dataflow paradigm that ECL offers.

For more information, contact Taghi Khoshgoftaar at Florida Atlantic University, khoshgof@fau.edu, Bio www.cse.fau.edu/~taghi/, 561.297.3994.