Technology that uses machine learning to quickly generate predictive models from massive datasets

September 7 2015

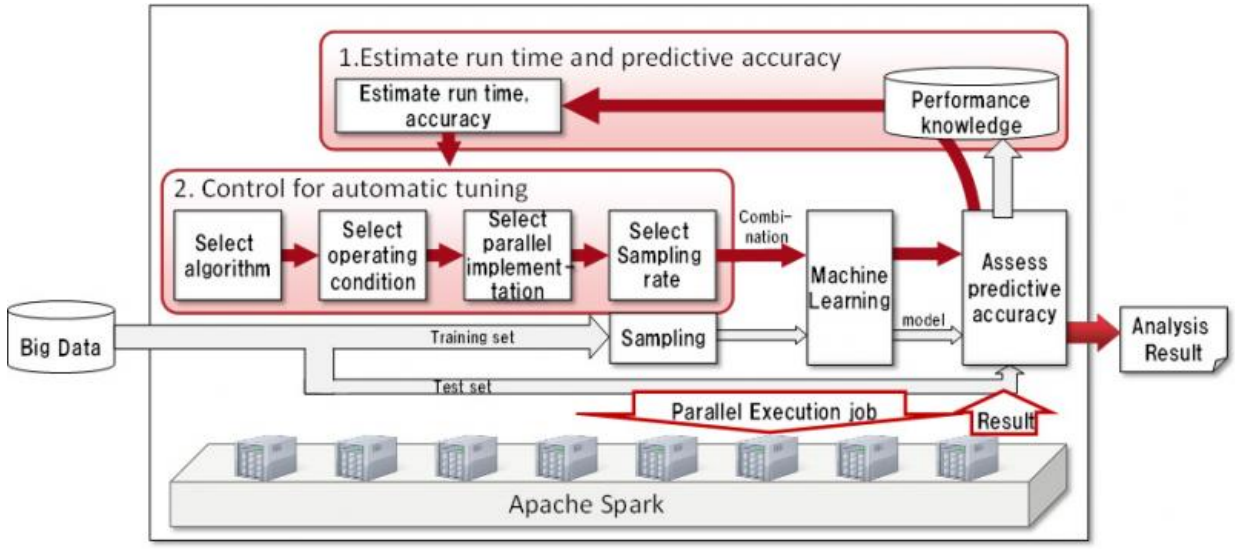

Figure 1. Schematic view of the technology

Fujitsu Laboratories today announced the development of a machine-learning technology that can generate highly accurate predictive models from datasets of more than 50 million records in a matter of hours.

Current techniques for generating highly accurate predictive models need to examine every combination of learning algorithm and configuration, taking more than one week to learn from a dataset

containing 50 million records. Fujitsu Laboratories has now developed a technology that estimates machine-learning results from a small set of sample data and the accuracy of past predictive models, extracts the learning algorithm and configuration combination that produce the most accurate result, and applies it to the larger dataset. This results in highly accurate predictive models from datasets of 50 million records in a few hours. Predictive models produced by this technology can work to quickly make improvements, such as minimizing membership cancellations on e-commerce websites and enhancing response times to equipment failures. Details of this technology are being presented at the meeting of the Information-Based Induction Sciences and Machine Learning (ISIMBL), opening Monday, September 14 at Ehime University in Japan.

The popularity of smartphones and other advances make it possible to gather massive quantities of sensor data, and machine learning and other advanced analytic techniques are being used extensively to extract valuable information from that data. Using the access logs of e-commerce websites, for example, it is possible to discover when people are most likely to cancel memberships on a given website, to identify those people quickly, and to take measures to discourage cancellation. Using detailed daily power-consumption data, it is possible to discover patterns of increased or decreased usage and to predict periods and times when power usage will increase. This can lead to a reduction in power costs by applying more precise controls over power generation, transmission, and storage. Developing predictive models by machine learning is considered an effective way to obtain accurate predictions. There are numerous methods for machine-learning algorithms, each for a different purpose, and they all differ in their predictive accuracy and run time. The algorithm that will produce the best accuracy will depend on the data being analyzed, and getting the most accurate predictions will also depend on fine-tuning its configuration. Therefore, generating an effective predictive model requires examining combinations of

algorithms and configurations.

Attempting to examine every possible combination of algorithm and conditions causes the number of combinations to balloon quickly. Furthermore, learning time of a combination can take days to examine, making it impractical to use machine learning extensively. Instead, algorithms and conditions are typically selected by analysts based on their experience, so the results ultimately depend heavily on the analyst's skill. In cases where the volume of data is great and analysis ends up taking more than one night, examinations are usually limited to a restricted number of combinations, or analysis can only be applied to a small portion of the data, and it is not possible to automatically derive accurate predictive models in a limited period of time.

Fujitsu Laboratories has developed a technology that estimates machine-learning results, able to generate and automatically tune an accurate predictive model from a small amount of sample data. It has prototyped this on Apache Spark, an open-source platform for parallel execution (Figure 1).

1. Estimating machine-learning run time and predictive accuracy

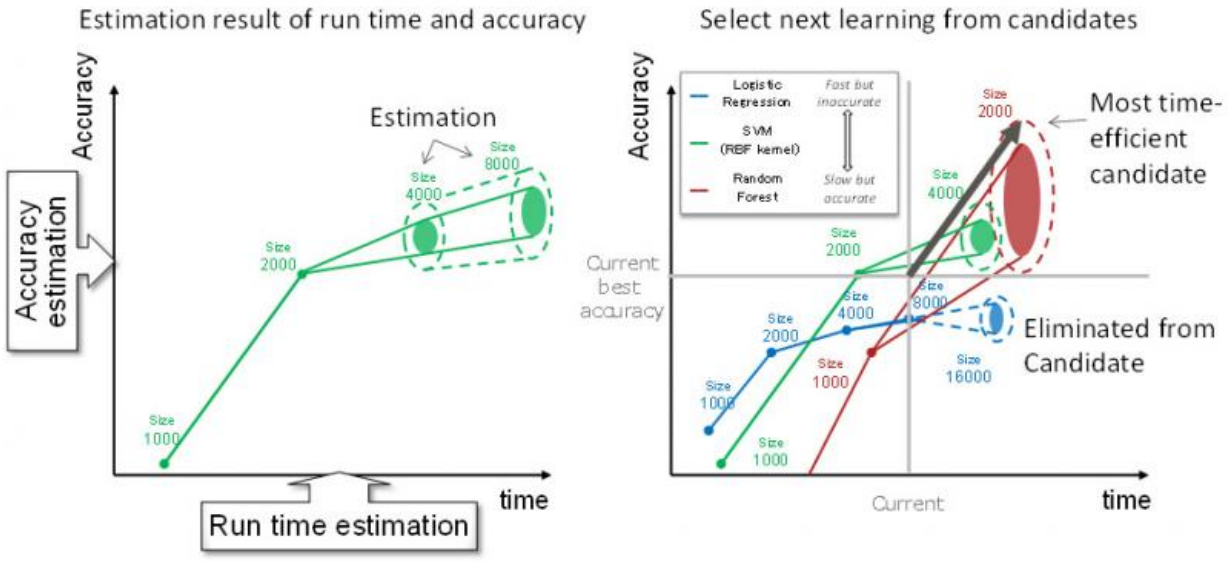

Figure 2. Control technology to automatically tune machine-learning algorithms

For each standard machine-learning algorithm, Fujitsu Laboratories measured actual machine-learning run times while varying the number of records in a dataset and the number of attributes used to represent the data, and built a run-time estimation model based on those measurements. Additionally, to improve the accuracy of those estimates, actual on-the-fly run-time measurements are used for correction. The company built up a database of combinations of previously used algorithms and configurations, along with the accuracy of the predictive model they produced, and uses this to estimate the predictive accuracy of new combinations. This makes it possible to make an assessment based on the smallest amount of data possible without sacrificing predictive accuracy. Estimating the run time and the accuracy of a predictive model produces accurate predictive models quickly. Techniques for estimating the predictive accuracy of a single machine-learning algorithm do exist, but there has been no such technology that can be applied to multiple algorithms and multiple dataset sizes. Because

this technique incorporates actual run-time measurements into estimates based on the conditions for each machine-learning run (including the algorithm, number of records, number of attributes, infrastructural information, and so forth), it gets more accurate the more it is used.

2. Control technology to automatically tune machine-learning algorithms (Figure 2)

This technology selects time-efficient candidates from among all the candidate combinations, and iterates over them efficiently and in parallel. In existing techniques, there is no way to determine which combination of machine-learning conditions is best according to any ranking system; instead, they depend on the know-how of an analyst manually picking conditions in order for the analysis to proceed. This technology combines estimates of run time and predictive accuracy to select candidate combinations of algorithms and configurations that are expected to provide high improvements of predictive accuracy in return for short run time. Each selected combination is then run in a distributed manner. Taking run time into consideration when selecting candidates makes it possible to execute each algorithm in an optimal order and quickly obtain the most accurate machine learning model. Because this technique automatically focuses on the most effective combinations, it does not depend on the know-how of an analyst.

The company ran internal tests using a dataset of 50 million records and eight servers with 12 processor cores each. Existing techniques would take roughly one week to develop a predictive model with 96% accuracy; Fujitsu Laboratories confirmed that this technique reached that level in slightly more than two hours. It is also demonstrated that this technology would make the practical application of machine-learning possible when used for access-log analysis with 30 million lines of web access logs. This technology could, for example, also be used to provide

services such as predicting electrical-power demand for every household in an area the size of Tokyo metropolitan area, or detecting early-warning signs of intent to cancel among users of online services with hundreds of thousands of members.

Fujitsu Laboratories is conducting field trials of this technology in Fujitsu Analytic's solutions using big data, with a goal of a practical implementation during fiscal 2015.

Provided by Fujitsu