IRI Voracity
An Insatiable Appetite for Data

Machine Learning in Analytics and Anonymization

THE INCREASING NUMBER OF applications for machine learning testify to its ability to improve the speed and accuracy of informational assessment from ever larger sources of data. Users of the IRI Voracity data management platform can leverage two machine learning modules: one for predictive analytics, and another for protecting sensitive data. Many more are possible.

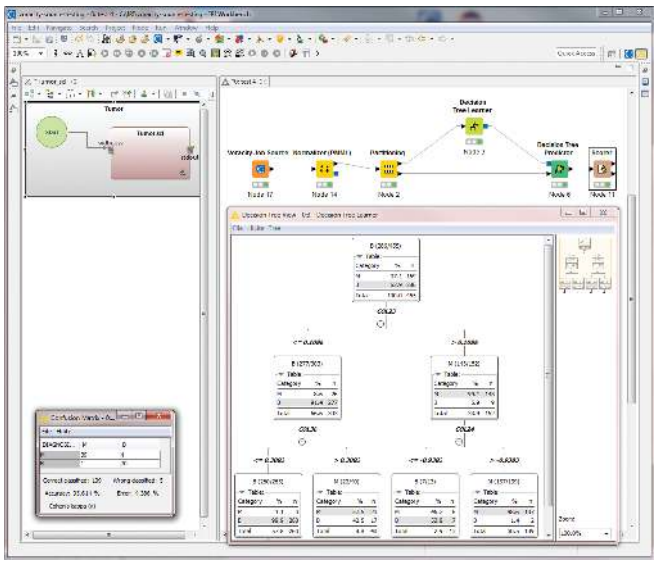
PREDICTING MALIGNANCIES

A common use of machine learning involves training a computer to evaluate data sets and create prediction models from trends in that data. Machine learning builds off traditional statistics and rapidly creates larger and more advanced models.

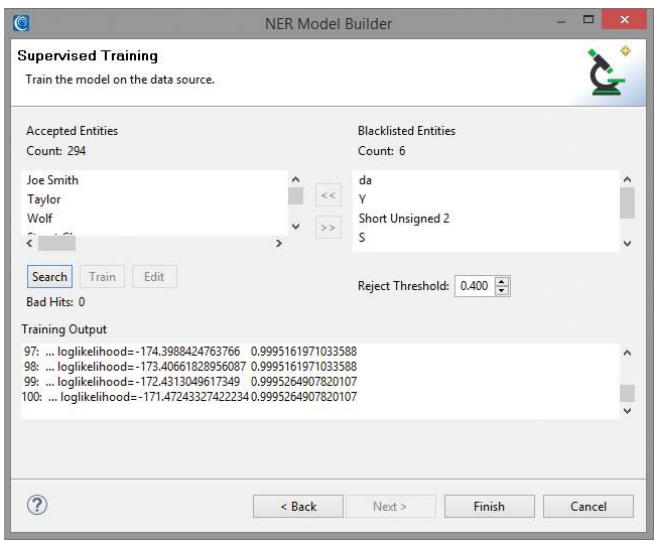
Many machine learning-related modules are included in KNIME, a popular open source data science platform that runs with Voracity in Eclipse. In this KNIME workflow, a Voracity data wrangling node feeds tumor measurement data into a KNIME decision tree node to improve breast cancer prediction accuracy:

Here, Voracity prepared raw data containing 20 different measurements of breast tumors, including their overall size, shape, and features of the cells' nuclei. Within seconds, the wrangled results flow into a decision tree to help determine if a tumor is likely to be malignant or benign.

The "Decision Tree Learner" node goes through different variables and creates multiple binary trees. Each tree determines if a given factor is likely to be a cause for a malignant tumor before it tries the next variable. Once the tree is built, the predictive model using those variables is tested for accuracy. In this case, it was about 95%, so the model should continue to be a reliable predictor future for data sets, too.



KNIME decision tree analysis of tumor data wrangled in IRI Voracity node



Using Machine Learning in IRI Workbench to train IRI DarkShield NER models

FINDING AND MASKING PII

Personally identifiable Information (PII) in documents like Word and PDF can be difficult to discover, delete, or de-identify en masse. This is particularly true when items like names or addresses — which do not match patterns or lookup values — can only be found in their Natural Language (Processing) context.

To address this challenge, IRI DarkShield technology in Voracity supports the use, and training, of Named Entity Recognition (NER) models to find those items. NER models are built from custom training data to improve search results.

The graphical front-end for DarkShield (and the rest of Voracity), called IRI Workbench, includes a wizard for either creating a NER model from existing annotated training data, or for training a model with actual documents that DarkShield parses. The latter employs a semi-supervised, iterative machine learning and annotation process.

Model training in this way improves the accuracy of PII search results when performed on a representative subset of documents. 15,000 sentences are considered a good minimum for teaching the machine to find named entities.

For more information email voracity@iri.com.

www.iri.com/voracity