Database Systems CSE 414

Lecture 28: Database Techniques for Machine Learning

- Automating Machine Learning Model Building with Clinical Big Data

Announcements

- WQ7 is due tomorrow 11pm
- HW8 is due Friday 11pm

- Please complete course evaluations!
  - First 10 minutes of today's lecture

- Today's lecture is intended to help you understand how database techniques can be used in other computer science areas
  - The final exam will not test today's lecture material
  - Relax and enjoy ☺

Outline

- Predictive modeling on clinical big data
- Identification of challenges
- Our approach to address the challenges
  [HISS’15, HISS’16, HISS’17, JRP’15, JRP’16, JRP’17]

Clinical Big Data (Large Clinical Data Sets)

- Volume of healthcare data
  - Increase 50-fold in 8 years to 25,000 petabytes by 2020
- Diverse sources
  - Electronic medical records
  - Sensors
  - Mobile devices
- Opportunities to advance clinical care and biomedical research

Predictive Modeling

- Leverage these large, heterogeneous data sets to advance knowledge and foster discovery
- Facilitate appropriate and timely care by forecasting
  - Health risk: Put high-risk patients into care management
  - Clinical course: Guide appropriate admission of bronchiolitis patients in the emergency department
  - Outcome: Assist with timely asthma diagnoses in children with clinically significant bronchiolitis

Approaches to Predictive Modeling

- Statistical methods
  – E.g., logistic regression
- Machine learning algorithms that improve automatically through experience (model training)
  – E.g., support vector machine
  – Neural network
  – Decision tree
  – Random forest

Pros of Machine Learning

- Often achieves higher prediction accuracy than statistical methods
  - Sometimes doubles prediction accuracy
- With less strict assumptions on data distribution

Cons of Machine Learning

- Use in healthcare is challenging
- Requires many labor-intensive manual iterations and special computing expertise to select among complex algorithms and hyper-parameter values
- Most machine learning models give no explanation of prediction results
  - Explanation is essential for a learning healthcare system

My Contributions

- Identify and clarify two challenges faced by healthcare researchers when conducting machine learning on clinical big data
- Propose solutions to address these challenges

Outline

- Predictive modeling on clinical big data
- Identification of challenges
  - Challenge 1
- Our approach to address the challenges
  [HISS'15, HISS'16, HISS'17, JRP'15, JRP'16, JRP'17]

Parameters vs. Hyper-parameters

- Each machine learning algorithm has two types of model parameters:
  - Ordinary parameters: automatically optimized or learned in a model training phase
  - Hyper-parameters: typically set by the user of a machine learning software tool manually before training a model

Parameters vs. Hyper-parameters – Cont.

Machine learning algorithm	Example ordinary parameters	Example hyper-parameters
Random forest	the input variable used and threshold value chosen at each internal node of a decision tree	# of decision trees, # of input variables to consider at each internal node of a decision tree
Support vector machine	the support vectors, the Lagrange multiplier for each support vector	the kernel to use, the degree of a polynomial kernel
Neural network	the weight on each edge	# of hidden layers, # of nodes on each hidden layer

Traditional Method of Building Machine Learning Models

- Manually select a machine learning algorithm from a long list of applicable algorithms
  - 39 classification algorithms available in Weka: decision tree, random forest, support vector machine, neural network, …
  - Most of them are complex
- Manually set the chosen algorithm's hyper-parameter values

Traditional Method of Building Machine Learning Models – Cont.

- Train the machine learning model to automatically optimize the ordinary parameters of the chosen algorithm
- Check the model's prediction accuracy
  - High enough: Done
  - Low: Manually change the hyper-parameter values and/or the algorithm, re-train the model
- Often take hundreds or thousands of manual iterations

Challenge 1: Efficiently and Automatically Selecting Algorithms and Hyper-parameter Values

- The chosen algorithm and hyper-parameter values affect the resulting model's accuracy
  - Typical effect is >40% [Auto-Weka in KDD'13]
  - The effective algorithm and hyper-parameter values depend on the specific predictive modeling problem and data set

Challenge 1 – Cont.

- Traditional approach: Find a good algorithm and good hyper-parameter values through a long, iterative, manual process
  - Beyond the ability of users with limited computing expertise
  - Non-trivial task even for machine learning experts

Challenge 1 – Cont.

- Automatic selection methods for algorithms and hyper-parameter values have been developed
  - to help individuals with little computing expertise perform machine learning
  - but existing methods cannot efficiently handle clinical big data
  - Search can take several days on a data set with a moderate number of rows and attributes
    - E.g., several thousand rows and several dozen attributes

Challenge 1 – Cont.

- In practice, search time can be up to thousands of times longer
- Machine learning is an iterative process
  - If a set of clinical parameters produces low prediction accuracy, the analyst is likely to consider other available but unused clinical parameters that may be predictive
  - Each iteration requires a new search for algorithms and hyper-parameter values

Challenge 1 – Cont.

- A data set can contain many rows
  - E.g., from multiple healthcare systems
- A data set can have many attributes
  - E.g., extracted from genomic and/or textual data
- A machine learning algorithm's execution time often grows
  - superlinearly with the number of rows
  - at least linearly with the number of attributes

Challenge 1 – Cont.

- To achieve personalized medicine, many predictive modeling problems must be solved for various diseases and outcomes
  - Search time will be a bottleneck here, regardless of whether it is an issue for a single problem
- To leverage clinical big data, automated approaches appealing to healthcare researchers are needed for selecting algorithms and hyper-parameter values
  - Completely automatic
  - Efficient

Outline

- Predictive modeling on clinical big data
- Identification of challenges
  - Challenge 2
- Our approach to address the challenges
  [HISS’15, HISS’16, HISS’17, JRP’15, JRP’16, JRP’17]

Challenge 2: Explaining Prediction Results

- Explanation is essential for clinicians to
  - Trust prediction results
  - Determine appropriate, tailored interventions
    - E.g., provide transportation for patients who live far from their physicians and have difficulty accessing care
  - Defend their decisions in court if sued for medical negligence
  - Formulate new theories or hypotheses for biomedical research

Challenge 2 – Cont.

- Most machine learning models give no explanation of prediction results
  - Most models are complex
- Prediction accuracy and giving explanation of prediction results are frequently two conflicting goals
- Need to achieve both goals simultaneously
  - Explain prediction results without sacrificing prediction accuracy

Outline

- Predictive modeling on clinical big data
- Identification of challenges
- Our approach to address the challenges
  [HISS'15, HISS'16, HISS'17, JRP'15, JRP'16, JRP'17]
  – Overview

Our Approach

- Develop a software system that can perform the following tasks in a pipeline efficiently and automatically
  - Select effective machine learning algorithms and hyper-parameter values to build predictive models
  - Explain prediction results to healthcare researchers
  - Suggest tailored interventions

Our Software System

- PredicT-ML (Prediction Tool using Machine Learning)
  - Developed using Spark, MLlib, and new techniques to address existing software's limitations
  - Can run on a cluster of commodity computers for fast parallel processing
- Goals: Healthcare researchers can use it to
  - Develop machine learning predictive models with clinical big data
  - Achieve similar prediction accuracy as computer scientists
  - Understand prediction results

Existing Big Data Software Systems

- Hadoop implements Google's MapReduce framework for distributed computing
  - Unsuitable for iterative and interactive jobs
    - Job execution usually requires repeated reading and writing of data from and to disk, incurring significant overhead
- Spark overcomes Hadoop's shortcomings
  - Executes most operations in memory and avoids disk inputs/outputs when possible
  - Improves performance
- MLlib is Spark's machine learning library

Outline

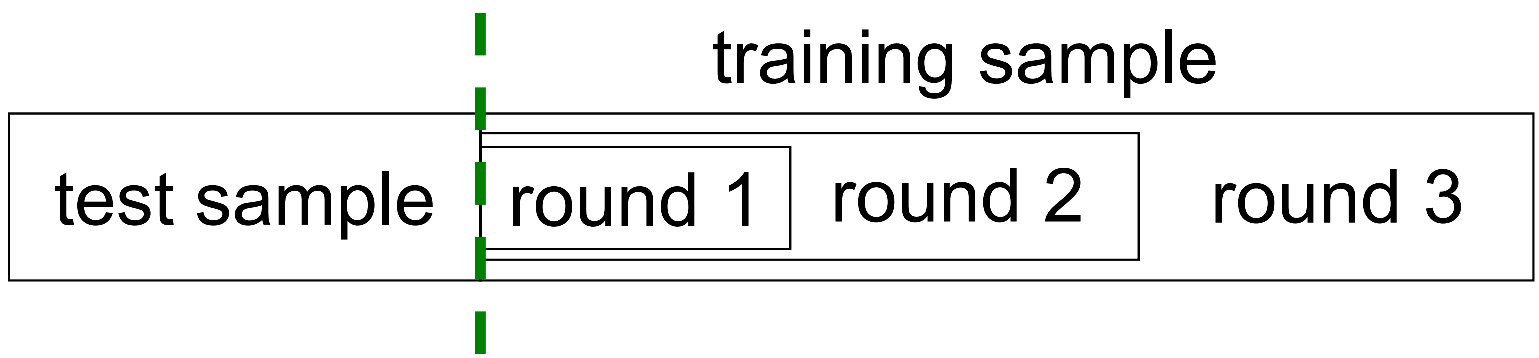
- Predictive modeling on clinical big data
- Identification of challenges
- Our approach to address the challenges
  [HISS'15, HISS'16, HISS'17, JRP'15, JRP'16, JRP'17]
  - Efficient and automatic selection of algorithms and hyper-parameter values

Main Ideas

- Major obstacle: A long time is needed to examine a combination of an algorithm and hyper-parameter values on the entire data set
  - E.g., it takes two days on a modern computer to train a champion ensemble model once on 10K patients with 133 independent variables
  - The entire space of algorithms and hyper-parameter values is extremely large
- Solution: Perform progressive sampling, filtering, and fine-tuning to quickly narrow the search space

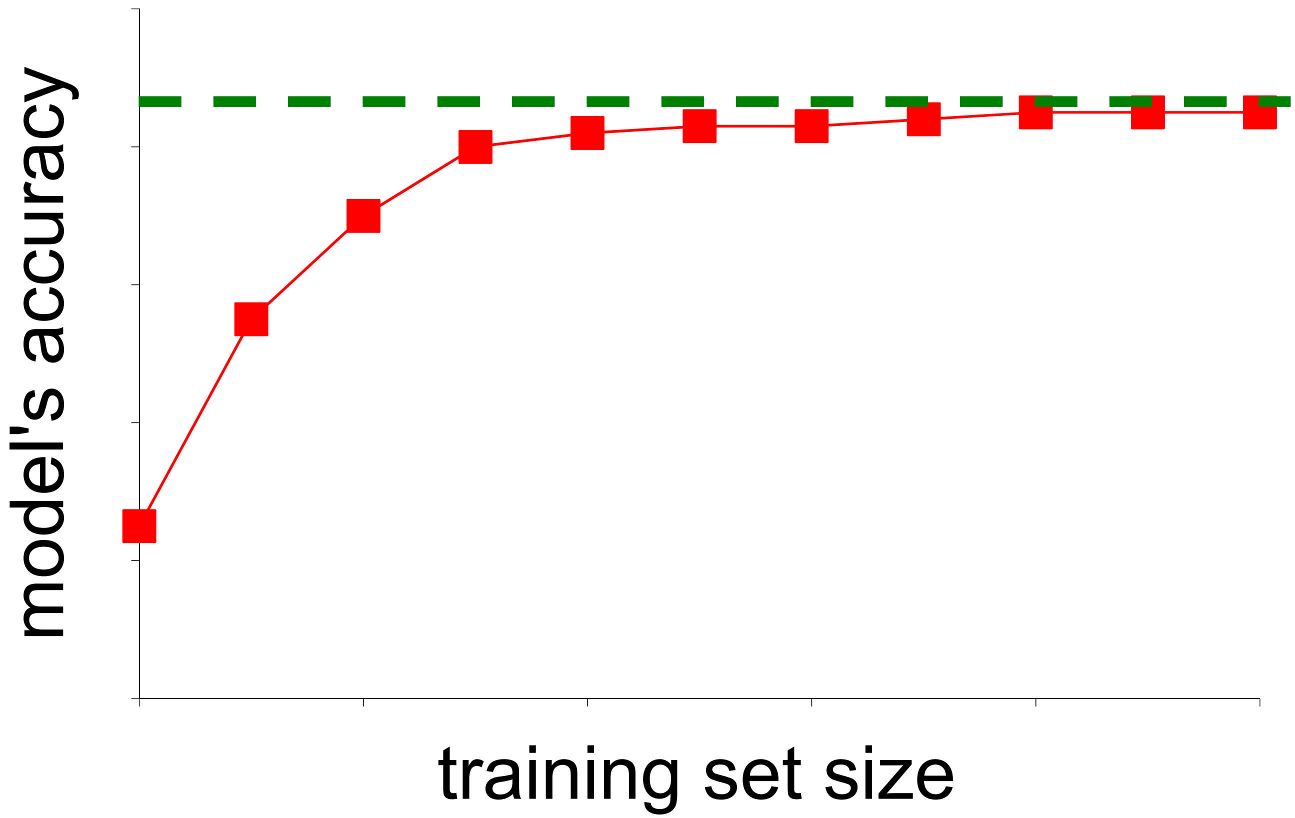
Main Ideas – Cont.

- Use progressive sampling to generate a sequence of random samples of the data set, one nested within another



Learning Curve

- For a specific combination of an algorithm and hyper-parameter values, a model's accuracy increases more and more slowly as the training set expands

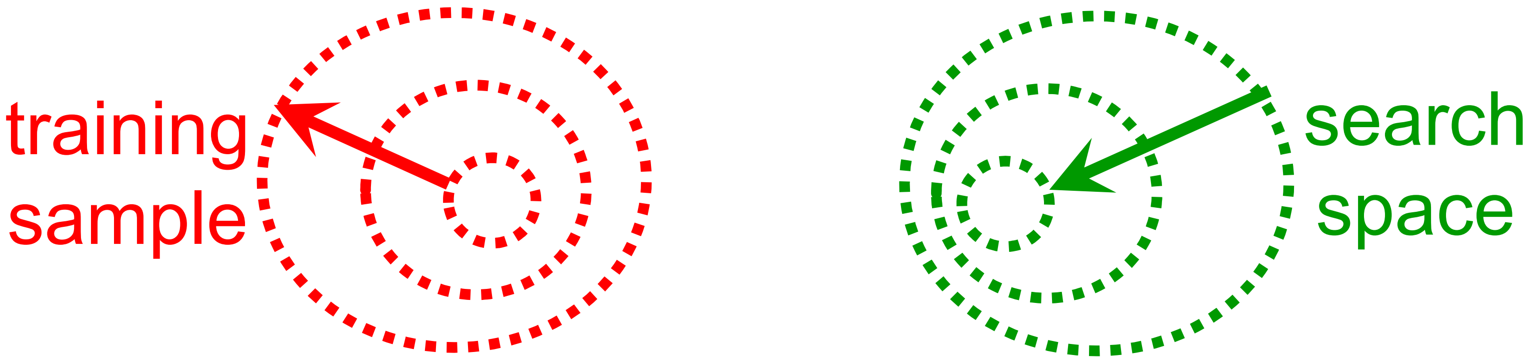


Main Ideas – Cont.

- Conduct inexpensive tests on small samples of the data set to eliminate unpromising algorithms and identify unpromising combinations of hyper-parameter values as early and as much as possible
- Devote more computational resources to fine-tuning promising algorithms and combinations of hyper-parameter values on larger samples of the data set

Main Ideas – Cont.

- The search process is repeated for one or more rounds
- As the sample of the data set expands, the search space shrinks



- In the last round, (a large part of) the entire data set is used to find an effective combination of an algorithm and hyper-parameter values

Some Results

- Compared to the state of the art Auto-WEKA automatic selection method on
  - 27 prominent machine learning benchmark data sets
  - A single computer
- On 27 data sets, on average our method
  - Reduces search time by 28 fold
  - Reduces the classification/prediction error rate by 11%

Outline

- Predictive modeling on clinical big data
- Identification of challenges
- Our approach to address the challenges
  [HISS’15, HISS’16, HISS’17, JRP’15, JRP’16, JRP’17]
  – Automatically explain prediction results and suggest tailored interventions

Main Ideas

- A model achieving high accuracy is usually complex and gives no explanation of prediction results
- Challenge: Need to achieve high prediction accuracy as well as explain prediction results
- Key idea: Separate prediction and explanation by using two models concurrently
  - The first model makes predictions and targets maximizing accuracy
  - The second model is rule-based
    - Used to explain the first model's results rather than make predictions

Main Ideas – Cont.

- The rules used in the second model are mined directly from historical data
- Use one or more rules to explain the prediction result for a patient
- Suggest tailored interventions based on the reasons listed in the rules

Some Results

- Test case: Predicting type 2 diabetes diagnosis within the next year
- Electronic medical record data of 10K patients
- Can explain prediction results for 87% of patients who were correctly predicted by a champion machine learning model to have type 2 diabetes diagnosis within the next year

Example Rule

- The patient had prescriptions of angiotensin-converting-enzyme (ACE) inhibitor in the past three years AND the patient's maximum body mass index recorded in the past three years is ≥35 → the patient will have type 2 diabetes diagnosis within the next year
  - ACE inhibitor is used mainly for treating hypertension and congestive heart failure
  - Obesity, hypertension, and congestive heart failure are known to correlate with type 2 diabetes
- Example intervention: Enroll the patient in a weight loss program