Culminating Experience Project Opportunities with Prof. D. Petkovic

November 2012
Opportunities overview

• Web Engineering and Machine Learning for molecular and drug development – FEATURE - joint NIH funded project with Stanford Helix Group
• Use of Machine Learning for SE teamwork assessment and prediction – NSF funded

• Real projects, some with top world institutions and scientists (e.g. Prof. Russ Altman, Stanford)
• Exciting, real projects
• Visibility (papers, WWW sites running from Stanford)
• Research papers, real app/WWW development
• Great for resume (real projects, papers, NSF/NIH funded)
• Funding: $15/h, part time during the year, full time summer

• Competitive, must apply and interview
What is FEATURE?

• FEATURE is software developed by the Stanford Helix Group in the 1990s
• It provides three major breakthroughs
  – Analyzes biomolecular structure to produce computable vectors
  – Learns the relationship between biomolecular structure and function
  – Uses what it has learned to predict function given a new structure
Stanford Helix Group

• Prof. Russ Altman
• http://helix-web.stanford.edu/
• SFSU has NIH Sub-grant with Prof. Altman 2 more years for FEATURE project
  – Explore various Machine Learning methods and compare them with existing ones
  – Develop WWW sites to support FEATURE applications
FEATURE Uses Machine Learning to Find the Relationship Between Structure and Function

We can use the computable vectors to see which ML approach is best for the problem, using the strengths of each approach to cover weaknesses in others.

Naïve Bayes

Support Vector Machines (SVM)

Random Forests

Other (e.g. Hidden Markov Models, Neural Networks)
WebFEATURE Is A User-Friendly Scientific Interface to FEATURE

http://feature.stanford.edu/webfeature/
Structure-based methods to study drug target relationships

**Observations**

Set A
- Drug 1
- Drug 2
- Drug 3

Set B
- Drug 2
- Drug 4
- Drug 5

**Target space**

Target A binding

Target B binding

**Prediction and Discovery**

- If targets A and B are similar (structural and physiochemical properties), targets A and B are predicted to bind to a same drug.

- If drug sets A and B are similar (chemical groups), targets A and B are predicted to bind to a same drug.

- Useful in:
  - Drug repositioning
  - Multi-targeting (A and B)
  - Side effect identification
• Microenvironment refers to the local, spherical region in the protein structure. Specifically, we calculate a set of 480 physicochemical properties collected over the predefined functional centre.
• Match microenvironments with geometric flexibility, allowing for conformational change or dynamics in both the ligand and the pocket
• Identify microenvironments for recognizing ligand chemical groups.
Database of Ligand Substructures (Fragments)

Small molecule (ATP) from the Protein Data Bank

Example fragments derived from the small molecule

Design Requirements
• Large scale computing required to determine the fragments for a given small molecule
• Storage approach should enable rapid querying for a given small molecule ID or fragment ID
• Architecture must permit incremental updates
Projects and skills

• Try new machine learning methods on FEATURE data
  – Needs to know ML, took or taking Prof. Okada’s class
  – Good with analytics and big data
  – C++, Python, Java, Amazon cloud services
  – GPA > 3.5 GPA

• Develop WWW sites to support FEATURE applications
  – WWW engineering
  – Good UI and SW Engineering skills – took or taking SE class
  – Python, java
  – Visualization, UI
  – Industry experience a plus

• Develop SW for big data analytics
  – WWW engineering
  – Good UI and SW Engineering skills – took or taking SE class
  – Python, java
  – DB
  – Performance optimization
  – Cloud computing (Amazon)
  – Industry experience a plus
A Machine Learning Approach for Assessment and Prediction of Teamwork Effectiveness in Software Engineering Education

NSF Funded
Problem we are focusing on

• Motivation: How to effectively teach and train practical SW engineering to address main impediments in successful SW product delivery → focus on SE teamwork and “soft skills”

• Problem we are addressing:
  – How to objectively assess the progress in learning
  – Can we predict team failures and learning problems so that early interventions can be applied
  – Develop easy to use and effective methods that can be widely applied by practitioners
Our approach – use Machine learning on class extracted data

• Use existing SE class to create *training database*
  – Teamwork behavior data measured using only qualitative data (e.g. times spent on certain activities)
  – Instructors’ objective observations (counts of issues and problems etc.,)
  – Data gathered automatically from instrumented tools for SW development and communication
  – Independently assessed SE process and SE product learning achievement ➔ outcomes

• Use *Machine Learning* (e.g. Random Forrest) on training database to develop predictive models that can predict teamwork learning achievement from observed data
SE classes

Student teams develop their SW products using SE Development and Comm. tools and answer surveys

Instructors grade each Team(i) Outcomes in Process and Product

Surveys to team members

SW Development Tools

SW Comm. Tools

Student Activity Measures

Training Database for ML data analysis

For each TEAM(i):
- Team Outcomes for PROCESS and PRODUCT (A,B,C)
- Team Activity Measures

Observe

Grade

Combine
Machine learning Technology used – Random Forest

- Tree based classifiers very popular today
- Achieve excellent recognition accuracy
- Allow extraction of simple rules from trained RF classifiers
- Open Source SW available ("R")
- Training DB ➔ RF ➔ Rules ➔ a) assessment and teaching recommendations; b) prediction software

Project

• Apply Machine Learning and data analytics to the data
  – Collect and Organize data
  – Analyze data
  – Extract rules and recommendations

• Skills:
  – Python, C++, Java
  – Data analytics, machine learning, data mining. Took or taking AI or DM class
  – Took or taking SE class
  – GPA > 3.5
To Apply:

- Send e-mail to Prof. Petkovic petkovic@sfsu.edu
  - What you would like to do
  - Why
  - GPA
  - Confirm that you took or are taking required classes
  - Availability
  - CV