Project Deliverable 2

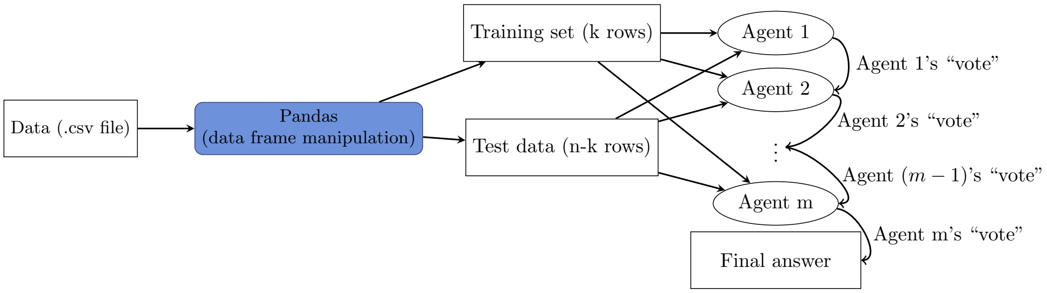
Andrew Berry

September 22, 2016

I pledge my honor that I have neither given nor received any unauthorized aid on this work.

1 Clarifications from Project Summary

The following diagram is intended to clarify the logistics of how the training and testing will occur in the project, how the agents will work together, and how the different tools will interoperate in the final system (n is the total number of rows in the original data set; k is the number of rows in the training set, which is a subset of the original; m is the number of agents):



To explain the above in more detail, the dataset (which is a csv file) will be loaded into pandas as a dataframe, where I will manipulate it to flatten out the data and break the set into the training and test sets. The training set will be given to each of m agents (created using the SPADE framework mentioned previously) to train a random forest classifier (taken from scikit-learn). Then, each row of the test set will be given to each of the agents. Each agent will have a randomly chosen size for their random forest; I'm currently considering the range of these to be from 5 to 20, but this range may change as experimentation progresses. Agent 1 will use its random forest to decide what its "vote" is for a classification.

Agent 1 will use SPADE's FIPA-ACL communication capabilities to convey its "vote" to agent 2, who will take that "vote" into account when making its own decision about voting for a classification of the row. The mechanics of how agent 2 makes a decision given agent 1's decision still need to be thought out thoroughly, but the system I envision has agent 2 finding an answer independently with its own decision tree, then comparing its answer to agent 1's answer. If the two agents have differing answers, then agent 2 will decide between the two with probabilities weighted based on the sizes of their respective decision trees. For example, if agent 1 used 10 decision trees in its random forest and agent 2 used 15, then there is a 0.6 probability that agent 2 will choose its own answer, since it used a larger forest to arrive at the answer. The voting passes on down the line in this fashion, with each agent taking into account the vote of its predecessor to inform its own decision, with the final agent (agent m) giving the final answer for the classification of the row.

2 Algorithms Considered

For this problem of multiagent classification, the algorithms of artificial neural networks, support vector machines, and random forest were considered. Neural networks, while flexible, have a tendency to overfit data.[1]

Support vector machines are generally regarded as more performant than neural networks.[1] However, one requirement of efficiently implementing SVMs is to have an appropriate choice of kernel, which is only possible if the programmer has an intimate knowledge of the subject matter of the data.[2]

A slight disadvantage of both neural networks and support vector machines is that they are black-box algorithms that lend little insight to exactly how they are making their decisions.[1]

Random forest runs efficiently on large databases, can handle a high number of variables, and is generally highly accurate. Its main disadvantage is a tendency to overfit for some noisy datasets.[3]

3 Potential Pitfalls

One aspect of this project that may cause issues is that there are 50 categories (one for each U.S. state) that the data can be classified as, which may make it difficult for the algorithm to distinguish different states from one another when they have similar metrics. For example, we may see all hospitals that perform well at treating stroke victims but poorly at treating COPD sufferers get mapped to Montana, when in fact Montana is only one of several states that have metrics that fit this pattern.

4 Contingency Plan

In the event that this project does not succeed as outlined, various aspects of the approach can be relaxed to increase probability of successful execution of the project. For example, the number of categories for the classification is large, so if that is found to be a difficult aspect to overcome, I can group the states into 6 distinct regions and have the algorithm predict regions rather than states.

5 Accomplishments Since Project Summary

Since the submission of the project summary, I have set up a working environment which has all necessary tools installed. I have reviewed the tutorials and documentation for the tools I plan to use and have run a test of the random forest algorithm on the “dummy” dataset to ensure that random forest classifies data as I expect it to.

I have begun learning about pandas functionality for pivoting data from being in multiple rows to multiple columns to better organize the dataset in such a way that each hospital has one row with multiple features to use for training and classifying. Given these accomplishments, I would consider myself on track with regard to the schedule I submitted for the initial deliverable. At this point, I don’t anticipate any major changes to the project schedule.

6 Multiagent Communication (FIPA)

The SPADE framework explicitly supports passing FIPA-ACL messages via the XMPP (Extensible Messaging and Presence Protocol) protocol. Since I intend to involve agent-to-agent communication in this project, I will be making use of these FIPA communication capabilities.

7 References

[1] Jacob Mick. (2012, June 8). Forum Post in Response to ”Neural networks vs support vector machines: are the second definitely superior?” [Online].
Available: http://stats.stackexchange.com/questions/30042/neural-networks-vs-support-vector-machines-are-the-second-definitely-superior

[2] Martin Sewell. (n.d.). ”Disadvantages of Support Vector Machines” [Online].
Available: http://www.svms.org/disadvantages.html . Accessed: Sept. 21, 2016

[3] Predrag Radenković. (n.d.). ”Random Forest” [Online].

Available: https://encrypted.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0ahUKEwiLjcO4_qHPAhUB6CYKHa3FD6sQFgghMAE&url=http%3A%2F%2Fhome.etf.rs%2F vm%2Fos%2Fdmsw%2FRandom%2520Forest.pptx&usg=AFQjCNEVo5hQOuo-6p2g3Tsa_snZfjlNnA&sig2=PYbl-X17PHv_VzqJrC-pjw&bvm=bv.133387755,d.eWE&cad=rja . Accessed: Sept. 21, 2016