CSE 4088 Introduction to Machine Learning

Term Project

Project Plan

1. Title and Team members:

"Comparison of Consistency-based and Correlation-Based Feature selection algorithms"

Project members:

Ozan Gülhan – 150114013

Zafer Emre Ocak – 150113075

2. Abstract

In machine learning algorithms, one of the biggest problems is scarcity of time and requirement of huge datasets. Feature selection provides optimization for these matters. Reducing time and amount of data required can change the fitness of the model significantly. By aiming this, feature selection algorithms are very important.

In order to see those effects, we consider implementing the algorithms which are "consistency-based" and " correlation-based" feature selection. In addition to that, we can also compare the required amount of time and data to reach the same fitness levels for the model.

To pick up a model, we have had a lot of different options but among them, we have decided on Bank Marketing Dataset. One of the most desired features of our project was finding an appropriate data. We have searched on the internet and it has two essential properties that first, high number of features and second, a lot of instances available.

In the model, the classification goal is to predict if the client will subscribe (yes/no) a term deposit.

Overall, we would like to reduce number of attributes and to decrease least amount of time to train the model for the same fitness levels.

3. Project Schedule

- **Finding Dataset**

  Considering finding the convenient dataset, we have a decent progress. This step is not going to be a big issue for us, we believe in.

- **Train data without Feature selection**

  Training the model would be a struggling but if we overcome this step then next subtasks will get easier for us in the future. It will probably take five to six days.

- **Run feature selection algorithms**

  This is the hardest and the most dangerous step ahead of our plan. If we cannot deal with the algorithms, the project can be stressful. We consider allocating most of the work and time for this subtask.

- **Reducing features based on the results**

  Basically, by a guidance of the selection algorithms, the corresponding features will be thrown away and re-train the model with the new dataset.

- **Re-train the model with new dataset attributes**

  This task is quite alike the second one because, at this time, we will already have great experience about training our model. It will take at most for one to two days.

- **Interpret and compare results**

  This is the final step and it is going to illustrate the improvement. We are planning to show how the required time and number of instances reduced after.

4. References

1. Manoranjan Dash, Huan Liu, Consistency-based search in feature selection, Artificial Intelligence 151 (2003) 155–176
2. Mark A. Hall, Correlation-based Feature Selection for Machine Learning, 1999