Project Proposal for CS791C

Ekrem Kocaguneli

Possible Title: Correlation of Software Metrics with Effort: An Individual and Collective Analysis

Project Details:

In his book Software Metrics, Norman Fenton groups the software related metrics into three categories: Product metrics, process metrics and resource metrics. Fenton claims that for a successful productivity model the dataset needs to cover all these 3 categories.

Although there is a significant academic effort in software measurement as well as software effort estimation, the commonly used datasets were collected according to the needs of specific companies (Cocomo datasets were based on the needs of NASA projects, whereas Desharnais dataset was based on the characteristics of Canadian software houses and Albrecht was based on the needs of IBM). Therefore, it is worth the effort to investigate the distribution of the metrics in various datasets into the categories of process, product and resource.

Another point that I will investigate in this project are the individual explanatory power of the attributes. In other words, I will take a look at the correlation between individual attributes of the datasets and the effort values in the datasets. This will show which attributes are most directly correlated with software effort.

Although individual correlation analysis of the attributes in different datasets may give an idea regarding the importance of attributes, software effort may have more complex relationship with more than one attribute. A useful approach to discover more complex relationships would be to use feature selection algorithms. Although different types of feature selection methods were proposed in the literature, Wrapper algorithm has the advantage that it does not make any prior assumptions about the data. It merely tries out all the possible 2^n-1 combinations of n attributes of a dataset and decides on the combination that yields the highest value. I am intending to compare the results of correlation analysis of single attributes to effort with the results of the Wrapper analysis and will comment on the differences and similarities.

The differences and similarities between correlation and wrapper analysis can give us an idea whether individually explanatory features come together to yield higher accuracy values, or whether individual attributes are non-relevant when combination of attributes are concerned for higher accuracy values. In both cases, the question arises: Why? I am hoping to find an answer to these questions in this research project.