Finding useful information in a large data set to better predict consumption of electricity

Data describing the weather at different places in Scandinavia shows a lot of redundancy which may affect its usefulness in predicting future electricity consumption. This master thesis tests two methods for removing lots of useless or harmful information.

Predicting the consumption of electricity on a city-wide scale allow those who manage equipment, generate and store electricity, and buy and sell energy to better plan the maintenance of their equipment, and to ensure that there are enough electrons flowing through your wall socket when you plug in your new computer. The predictions are done using artificial intelligence methods that look for patterns in data that can be used to determine the magnitude of electric consumption in the future. One of the main problems in performing accurate predictions is finding the right data to use; Choosing the wrong variables may lead to poor predictions which in turn may lead to equipment failure or other costly decisions for the energy providers and utility companies.

The data sets typically used for these kinds of predictions describe different aspects of future weather. Since weather is a natural phenomenon that varies differently depending on how far between two points you look we may assume that there will be a lot of data showing basically the same thing; the weather in Lund is probably not very different from the weather in Malmö, while the weather in Umeå might differ much more from the other two cities. In this case, we call the data from Lund and Malmö **redundant** in the light of each other. The goal of this thesis has been to investigate methods that sort through the data set in order to find useful data, which we call **relevant**, and remove redundant information.

Two approaches are taken. First, we look at the properties of the data itself and measure relevancy and redundancy by seeing if there is a significant similarity between pairs of variables. For this purpose an algorithm called the **Fast Correlation-Based Filter** is implemented and evaluated. The filter searches through the data set without considering all possible combinations of variables in order to make it faster. Furthermore, we look at the possibility of being able to choose relevant data based on the geographical location. Motivated by the fact that weather data from places close to each other are very similar it is possible to sort through the data set just by using distance from the city where electrical consumption is being predicted. Both methods show promising results when tested on predicting the daily average electricity consumption on four areas, managing to remove over 99% of the data while still performing accurate predictions. Further tests should investigate the computations performed for the statistical measure used, as well as see how useful the methods are on data of higher resolution.

Oscar Utterbäck