Pat Junghenn

Thomas Keeley

**Data Mining Project Proposal**

**Group 3**

**What problem did you select and why did you select it?**

The topic selected for this project is occupancy detection. Occupancy detection is utilized to gauge the occupancy level of a building. Buildings create one-third of worldwide emissions. Many of these buildings use far more power than necessary. The energy utilized by a building is largely dependant on the number of individuals in that building. Occupancy detectors give valuable information on the optimal energy allocation in a building. Ventilation and electricity could be powered off for Vacant rooms and offices. Occupancy detectors utilize readings taken from the surrounding environment to determine whether or not space is occupied.

Measurements are taken from light, temperature, humidity and CO2 levels. Using these four features, occupancy detectors determine if there is a presence of a human. Accurate occupancy detectors would greatly lower the emissions created from buildings around the world. The development of statistical learning software to determine occupancy status would be a powerful tool for reducing building emissions. Furthermore, if this learning software were to be integrated into a graphical user interface, it could easily be utilized by building engineers or staff to make these energy allocation decisions.

**What database/dataset will you use? Does it need to be cleaned?**

The occupancy detection dataset was acquired from the UCI Machine Learning Repository. The data is already broken down into separate train and test datasets but all of the data will be combined to achieve the desired split between training and testing. There is very little cleaning that needs to be done as it is already in the desired format and contains zero null values. The only variable that may be modified is the date. Our target variable (occupancy) is already in binary form and will be the main focus of our modeling. The predictor variable are all numerical variables and are in the desired format. The data is clean and ready for modeling.

**What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?**

Multiple data mining algorithms will be used in this project to achieve the desired results. To begin, a logistic regression model will be developed to understand the predictability of the data. Additionally, varying methods of decision tree models will be created to further explore model validity and increase performance. Further modeling will be discussed and implemented as appropriate as knowledge and curiosities are enhanced. There shouldn’t be much customization to the model development. The data appears to be structured well and compiled in such a way that it is ready to be modeled and explored.

**What software will you use to implement the network? Why?**

Pycharm will be the main software used in this project as it is most familiar and functional in regard to model development. Additional software will be used in relation to creating the graphic interface for algorithm deployment.

**What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?**

During the creation of the learning software, we will use scholarly articles to gain further insight into occupancy detection. Sources about the effectiveness of occupancy detectors will be referenced to determine the effectiveness of the learning model. We will also use multiple sources to determine what features of these detectors are the most and least important. The greater impact that this type of research will have will be presented and expanded research will be suggested.

**How will you judge the performance of your results? What metrics will you use?**

The classification report and confusion matrix will be a good place to start in evaluating the performance of the models. Beyond, the Receiver Operating Characteristic (ROC) curve will also be examined to understand the balance of the models predictability and accuracy.

**Provide a rough schedule for completing the project.**

**27 Oct - 02 Nov**

Data exploration and scientific research. Develop the narrative for the project and perform data pre-processing. Discuss goals and distribute personal tasks.

**03 Nov - 09 Nov**

Start developing models, test validity and optimize results. Continue to develop additional data mining techniques to be included in the project.

**10 Nov - 16 Nov**

Continue to develop models and apply new concepts to the data. Begin writing the report to express the results.

**17 Nov - 23 Nov**

Begin building GUI. Ensure model performance and clarity of methods.

**24 Nov - 30 Nov**

Finish building GUI and test functionality. Finalize report and author individual contribution reports. Create presentation.

**01 Dec - 04 Dec**

Test GUI and ensure performance of model. Make final edits to report and practice for presentation.