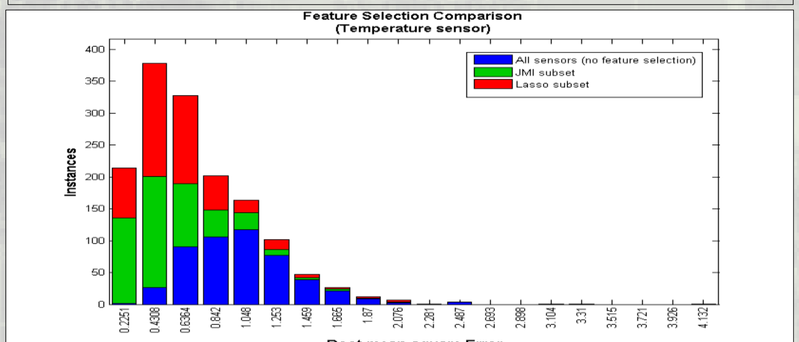
Sensor Validation with Machine Learning

Matt Smith (mksmith3@crimson.ua.edu), Charles Castello (castellocc@ornl.gov) , Joshua New (newjr@ornl.gov)

Since commercial and residential buildings account for 41% of the United States' energy consumption, making them more energy-efficient is a vital part of the nation's overall energy strategy. Sensors play an important role in this research by collecting data needed to analyze building performance. Given this reliance on sensors, ensuring that sensor data are valid is a crucial problem. In this research, we demonstrate the efficacy of machine learning techniques for this problem. We have looked at two such techniques: artificial neural networks and fuzzy clustering. We are trying to validate data for five sensors: outside temperature, outside humidity, refrigerator energy use, heat pump liquid line pressure, and heat pump liquid flow. Artificial neural networks have been able to predict data, and thus correct data, for three of the five sensors we are investigating. Our implementation of fuzzy clustering as a validation tool was not as successful. Our method was able to cluster data into "correct" and "errant" clusters reliably, but only when the points in the "errant" cluster were three to seven standard deviations away from their correct value.

1. Feature Selection

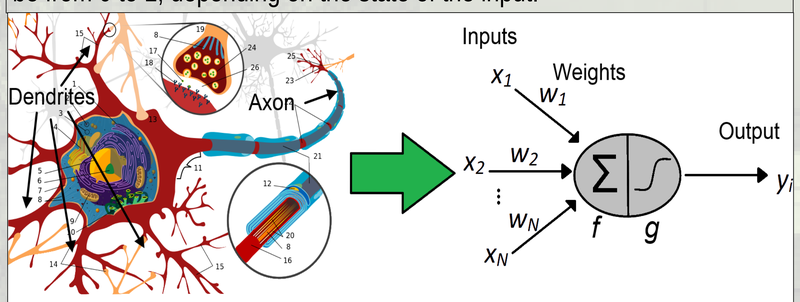
For neural networks (below), the goal is to train the network to be able to predict output given a set of inputs. For inputs, we used the data of other sensors. Thus, our choice of input sensors will affect the quality of the network's predictions. Feature selection is a collection of methods we use to determine which sensors will likely be good predictors of other sensors. We looked at two techniques for feature selection: Joint Mutual Information (JMI) and Lasso. The effect of feature selection can be seen in the figure below. We chose to use JMI for feature selection because it had slightly better accuracy and in general used fewer inputs.



Histogram comparing root-mean-square error (RMSE) values for JMI, Lasso, and control (no feature selection) with the temperature sensor. Note that the RMSEs for the control are centered around 1 °F, while those for JMI and Lasso are centered around 0.4 °F and 0.6 °F, respectively.

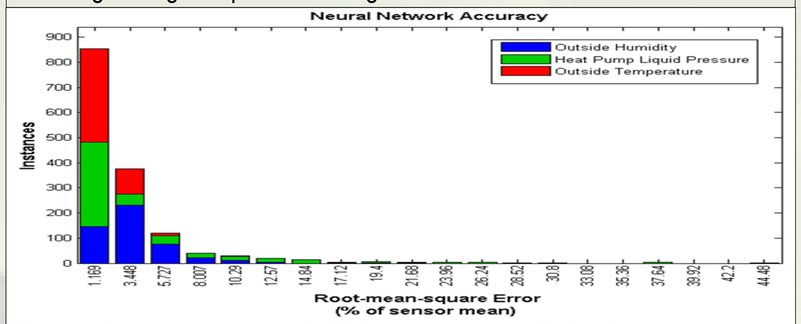
2. Neural Networks

Artificial neural networks are programs meant to mimic the activity of neurons in the brain. Neurons work by receiving input from other neurons through their dendrites. Based on the state of the input, they may fire an output through their axon. Similarly, artificial neurons receive input, perform a weighted sum of the inputs, and fire an output. This output may be from 0 to 1, depending on the state of the input.

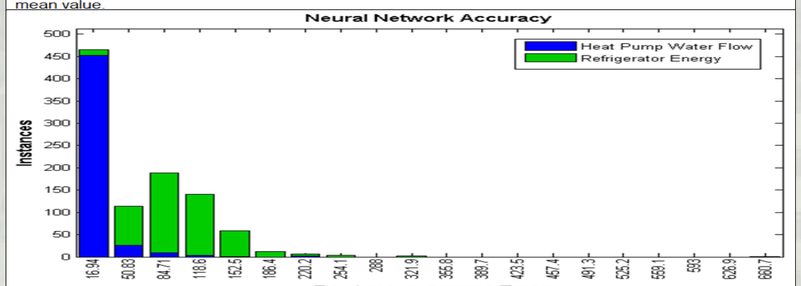


A diagram of a real neuron (left) and an artificial neuron (right). The neuron fires an output through its axon depending on the state of the inputs it receives from its dendrites. Similarly, the artificial neuron produces an output from 0 to 1 based on a linear combination of its inputs. A neural network is made of layers of individual neurons.

For each sensor, we trained 500 networks and calculated the root-mean-square-error between the network's output and the correct value. The following two figures present histograms of these results.



Histogram of root-mean-square errors (RMSE) for 500 networks each of the humidity, temperature, and pressure sensors. For scalability, each RMSE is presented as a percentage of its respective sensor's mean value.



Histogram of root-mean-square errors (RMSE) for 500 networks each of the water flow and refrigerator energy sensors. For scalability, each RMSE is presentad as a percentage of its respective sensor's mean value

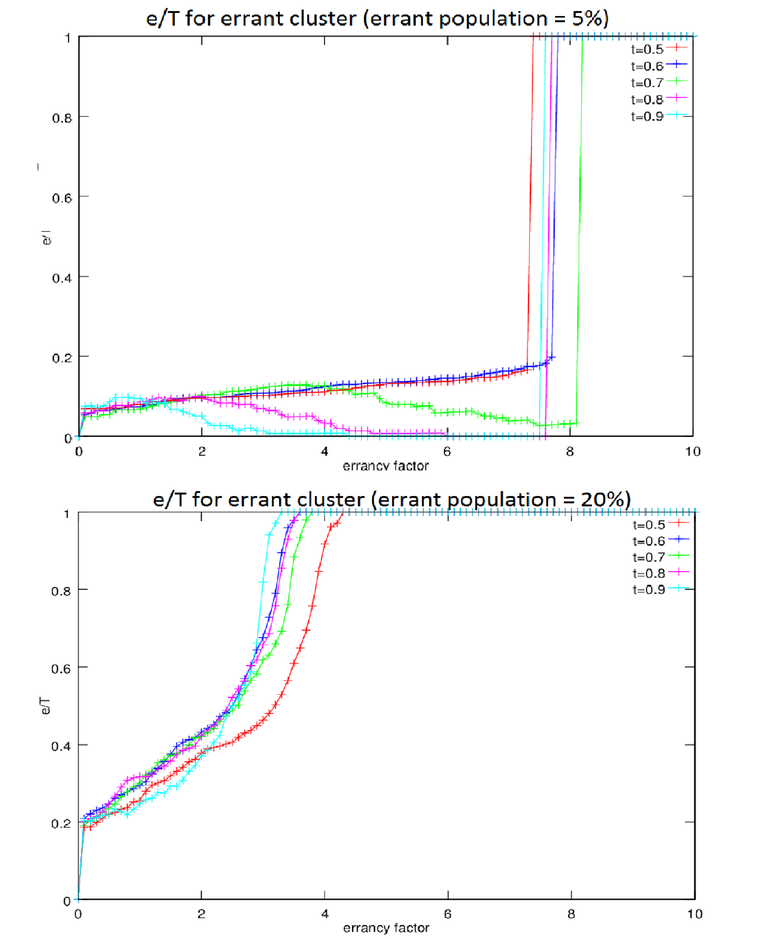
3. Fuzzy Clustering

Fuzzy clustering is a variant on classical set theory. Classical set theory divides the universe of discourse into crisp sets, where any point is in one set or no sets. Fuzzy clustering does away with this crispness. Instead, points may be a member of one set, or partially members of several sets.

To what degree a point is a member of a given cluster is determined by the points *membership function* for that cluster. The membership function ranges from 0 to 1, where 0 means the point is not at all a member of that cluster, and 1 means the point is completely a member of that cluster and no other.

We wanted to use fuzzy clustering as a validation tool by separating a sensor's data set into two cluster: "correct "and "errant". To make a point errant, we added *N* times the standard deviation of the data set to its original value, where *N* is called the *errancy factor*. For example, if *N*=3, then each errant data point is 3 standard deviations from its original value. We counted a point as "in" a cluster if its membership function exceeded a threshold value, *t*.

The results showed that this method doesn't really work. The errant points were not reliably grouped into their own cluster until they were off by 3 to 7 standard deviations, depending on how much of the population was made errant. The next figure shows what this looks like for the temperature sensor.



The graphs in this figure show the number of errant points *e* in the errant cluster divided by the total number of points *T* in the errant cluster. Whether a given point is "in" the cluster is determined by the threshold *t*. If the point's membership function for that cluster is greater than or equal to *t*, it is considered a member of that cluster. "Errant" points are made errant by adding to their original value the number $N\sigma$, where σ is the standard deviation of the data set and *N* is the "errancy factor". The "errant population" is the fraction of the total population to which this operation was applied.

If *e/T* = 1, then the errant cluster contains only errant points. The figure show that this does not happen until the data are made errant by ~3 standard deviations to ~7 standard deviations, depending on the errant population.

Acknowledgement: This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI).