CS506 Progress Report

Namir Fawaz, Sarthak Jagetia, Kasim Patel

Introduction:

As stated in our Project Proposal, the objective of our project is to work with SICK in order to analyze their sensor data and return meaningful results. They allowed us to develop our own questions and hypotheses instead of telling us exactly what to look for. We have decided to look into three questions:

1. **If voltage is spiking or or lowering, how is this affecting the operational data?**
2. **If the conveyor belt is speeding up, how well is the barcode information being captured? (At what speed does the sensor fail to read the barcode)**
3. **How does the sensor temperature affect the capturing of barcode information.**

Data Retrieval:

All of our data has been organized into two different data sets, Operational Data and Sensor Diagnostics. We have already been provided with all of the data so there is no reason to have to do any type of data retrieval. The data is organized into XML files for the Operational Data and TXT files for the Sensor Diagnostics Data. Luckily, in one of our first CS506 homework assignments we learned how to parse through XML data, making this portion of the project rather straightforward. The TXT files are more tricky to parse, because they are formatted as if they are in tables but are only tab-delimited, making the data a bit harder to parse through. In any case we will read and write both data sets into Pandas Dataframes in order to make manipulating and interpreting the data much more straightforward.

Data Cleaning:

As stated above, the majority of the “cleaning” necessary to work with our data will simply be to parse through the large data sets and retrieve the relevant categories of data. For example, we will retrieve the sensor temperature or sensor voltage from the Sensor Diagnostics data, and we will retrieve the barcode capture result from the Operational Data, and we will store all of this information into a Pandas Dataframe. From there we will easily be able to check for any type of Regression, most likely Linear, in order to see if the data correlates together in a meaningful way. With the data stored in dataframes, we will also be able to potentially use forms of clustering in order to detect any patterns, and additionally search for any potential outliers that may skew our data.

Methodologies/Experiments:

This section has been organized to reflect each of our hypotheses/questions

1. We will load the data relating to sensor, retrieving the given voltage level for each time the sensor was used the from the Sensor Diagnostics data. We will then use the Operational Data, namely the Heartbeat Data to find all the errors corresponding to the sensors, and combine these data sets to create a dataframe containing each sensor's id, the voltage level for each instance of use, and and a boolean flag (1/0) to mark if an error occurred, which we will save as a CSV file. We will most likely need to use the timestamp in order to connect each sensor use with the resulting error. We will then run a Linear Regression test and maybe some other testing such as outlier detection or clustering to detect any anomalies.
2. We will use the Operational Data to find the speed of the conveyor belt during each reading, and then use the Heartbeat Data to find any possible error that could have occurred. We will then organize this data into a dataframe, and save it to a CSV file. We will then run a Linear Regression test and maybe some other testing such as outlier detection or clustering to detect any anomalies. We will draw a conclusion of what speed threshold substantially affects the results of the barcode reading.
3. We will retrieve the sensor temperature for each given sensor from the Sensor Diagnostics Data. We will then retrieve the Heartbeat Data to take note of any errors that may have occurred for a given sensor. We will combine these data sets to create a dataframe containing each sensor's id, the temperature for each instance of use, and and a boolean flag (1/0) to mark if an error occurred, which we will save as a CSV file. We will most likely need to use the timestamp in order to connect each sensor use with the resulting error. We will then run a Linear Regression test and maybe some other testing such as outlier detection or clustering to detect any anomalies.

Timeline:

- 11/27: Finished retrieving and organizing all data based on Sensor ID for each question, saved all of these dataframes into CSV files.
- 12/1: Finished writing algorithms to check for Linear Regressions
- 12/8: Finished writing any other algorithms or methods to be used, begun writing writing conclusions of analysis
- 12/12: All work has been reviewed, confirmed, and analysis has been written/poster has been made