|  |
| --- |
| [File:Santiago30std.jpg](https://upload.wikimedia.org/wikipedia/commons/0/04/Santiago30std.jpg) |
| Air Quality  RESEARCH ON |
| |  |  |  | | --- | --- | --- | | Sherin Sarah Varghese, Mohamed Rizwan and Jun Jie, Donald Leow | /18 | BIT | |

Contents

**Introduction2**

**Background knowledge of data set2**

**Summarized description of experimental procedures3**

Data Pre-processing3

Algorithms tested\_

Comparison of the algorithms\_

Justification of chosen algorithm\_

**Detailed analysis on the data set\_**

**Conclusion\_**

**References\_**

# Introduction

To replace the placeholder text on this page, you can just select it all and then start typing. But don’t do that just yet!

First check out a few tips to help you quickly format your report. You might be amazed at how easy it is.

* Need a heading? On the Home tab, in the Styles gallery, just click the heading style you want.
* Notice other styles in that gallery as well, such as for a quote, a numbered list, or a bulleted list like this one.
* For best results when selecting text to copy or edit, don’t include space to the left or right of the characters in your selection.

## Background knowledge on the data set

The Data set we chose was on Air quality and it has 9358 instances of metal oxide chemical sensors of hourly averaged responses detected in an Air quality chemical multisensory Device. There was a total of 15 attributes recorded which includes various contents of the air, date and time. The data was recorded at a seriously polluted area, at road level on a field, in an Italian city. The recorded data represents the longest openly available recordings of air quality were taken in a period of one year from March 2004 to February 2005.

Another co-located reference certified analyzer provided data such as hourly averaged concentrations of Carbon monoxide (CO), Nonmetallic Hydrocarbons, Benzene (C6H6), Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2). However, the air quality data set has Missing values, these values are tagged with -200 value.

This dataset can be used particularly for research purposes. Three research papers of De Vito et al., was used while collecting this data for evidences of cross-sensitivities as well as both concept and senor drifts are present in the papers.

## Summarized description of experimental procedures

## Data pre-processing

Before we start to use different data mining techniques, we cleaned the data by pre-processing it as quality decisions must be based on quality data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Before Pre-Processing | | | | | |
| Attribute Name | **Scale of the data** | **Type of data** | **Statistics** | **Range of data** | **Missing values** |
| Date | MM/DD/YYYY | Date |  | 10/03/2004 - 04/04/2005 | 0 |
| Time | HH:MM: SS | Time |  | 0:00:00 - 23:00:00 | 0 |
| hourly avg CO | mg/m3 | Numeric | Mean=2.153, StdDev = 1.453 | 0.1 - 11.9 | 1683 |
| hourly avg tin oxide | mg/m3 | Numeric | Mean=1099.833, StdDev = 217.08 | 647 - 2040 | 366 |
| hourly avg overall Non Metanic HydroCarbons | microg/m3 | Numeric | Mean=218.812, StdDev = 204.46 | 7 - 1189 | 8443 |
| hourly avg Benzene | microg/m3 | Numeric | Mean=10.083, StdDev = 7.45 | 0.1 - 63.7 | 366 |
| hourly avg titania | mg/ m3 | Numeric | Mean=939.153, StdDev = 266.831 | 383 - 2214 | 366 |
| hourly avg NOx | ppb | Numeric | Mean=246.897, StdDev = 212.979 | 2 - 1479 | 1639 |
| hourly avg tungsten oxide (nominally NOx targeted) | mg/m3 | Numeric | Mean=835.494, StdDev = 256.817 | 322 - 2683 | 366 |
| hourly avg NO2 | microg/m3 | Numeric | Mean=113.091, StdDev = 48.37 | 2 - 340 | 1642 |
| hourly avg tungsten oxide (nominally NO2 targeted) | mg/m3 | Numeric | Mean=1456.265, StdDev = 346.207 | 551 - 2775 | 366 |
| hourly avg indium oxide | mg/m3 | Numeric | Mean=1022.906, StdDev = 398.484 | 221 - 2523 | 366 |
| Temperature in Â°C | °C | Numeric | Mean=18.318, StdDev = 8.832 | -1.9 -- 44.6 | 366 |
| Relative Humidity (%) | % | Numeric | Mean=49.234, StdDev = 17.317 | 9.2 - 88.7 | 366 |
| AH Absolute Humidity | mg/m3 | Numeric | Mean=1.026, StdDev = 0.404 | 0.1847 - 2.231 | 366 |

We downloaded the adult data set from UCI Machine Learning website. Missing data may cause incorrect or even misleading statistics. During the beginning stage of pre-processing the missing values are cleaned from the data. There was a total of 16701 missing values which is 12% of the data set. We first replaced the missing value with ‘?’ in the Microsoft excel file we downloaded. It was easy to replace the missing values as Microsoft excel has the find and replace function. Then we converted the Microsoft excel file into CSV file and further amendments were made on the notepad (@relation @attribute @data) to make sure that the file can be processed by weka. This file was later converted to arff so that it can be open the file in weka. Weka was able to process the data and identify the missing values in the data set and gave us the visualization of the data set. Weka also show the mean value and standard deviations of each attribute.

We used the ReplaceMissingValues’ filter in weka which is a rudimentary filter that fills in missing values; numeric values are replaced with the sample mean and nominal values are replaced with the sample mode. In this case it was numeric values. After using this filter, we came across an issue of one of the attributes named AH still having missing values and it was not able to be replaced using the filter ‘ReplaceMissingValues’. We discovered that the filter doesn't work on the last attribute. Therefore, we added another attribute called ‘extra’ to the data set and used the filter ‘ReplaceMissingValues’. Then it worked on AH attribute this time and therefore we removed the attribute called ‘extra’.

Subsequently, we realized that there were outliers in the data set. Outliers are extreme values in a set of data which is much higher or lower than the other numbers. They can cause serious problems in statistical analyses. Hence, we worked on the data to remove the outliers and we used the filter ‘InterquartileRange’ to identify the outliers in all the attributes. Then we used the filter ‘RemoveWithValues’ to remove the outliers from all the attributes. After cleaning the data, weka recalculated the minimum value, maximum value, mean and standard deviations for each of the attribute as the missing values were replaced and outliers were removed. The difference in the data set is recorded in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| After Pre-Processing | | | | | |
| Attribute Name | **Scale of the data** | **Type of data** | **Statistics** | **Range of data** | **Missing value** |
| Date | MM/DD/YYYY | Date |  | 10/03/2004 - 04/04/2005 | 0 |
| Time | HH:MM: SS | Time |  | 0:00:00 - 23:00:00 | 0 |
| hourly avg CO | mg/m3 | Numeric | Mean=2.053, StdDev = 1.142 | 0.1 - 6.8 | 0 |
| hourly avg tin oxide | mg/m3 | Numeric | Mean=1089.445, StdDev = 200.167 | 667 - 1975 | 0 |
| hourly avg overall Non Metanic HydroCarbons | microg/m3 | Numeric | Mean=218.579, StdDev = 62.666 | 9 - 1084 | 0 |
| hourly avg Benzene | microg/m3 | Numeric | Mean=9.67, StdDev = 6.666 | 0.1 - 40.5 | 0 |
| hourly avg titania | mg/ m3 | Numeric | Mean=927.061, StdDev = 247.363 | 383 - 1782 | 0 |
| hourly avg NOx | ppb | Numeric | Mean=227.911, StdDev = 154.592 | 2 - 800 | 0 |
| hourly avg tungsten oxide (nominally NOx targeted) | mg/m3 | Numeric | Mean=839.497, StdDev = 236.566 | 345 - 1826 | 0 |
| hourly avg NO2 | microg/m3 | Numeric | Mean=110.622, StdDev = 40.737 | 2 - 274 | 0 |
| hourly avg tungsten oxide (nominally NO2 targeted) | mg/m3 | Numeric | Mean=1449.101, StdDev = 333.855 | 551 - 2746 | 0 |
| hourly avg indium oxide | mg/m3 | Numeric | Mean=1022.661, StdDev = 365.973 | 221 - 2519 | 0 |
| Temperature in Â°C | °C | Numeric | Mean=18.497, StdDev = 8.683 | -1.9 - 44.6 | 0 |
| Relative Humidity (%) | % | Numeric | Mean=48.947, StdDev = 16.998 | 9.2 - 88.7 | 0 |
| AH Absolute Humidity | mg/m3 | Numeric | Mean=1.03, StdDev = 0.397 | 0.185 - 2.231 | 0 |