Predict Acid Rain Based on Atmospheric Pollutant Composition Using Machine Learning

Data Mining Team Project Final Report

presented by
Team 11
Dietsche, Kang, Lei, Rao, Ruta

submitted to the
Data and Web Science Group
Prof. Dr. Heiko Paulheim
University of Mannheim

02.12.2018

Contents

1 Introduction

Acid rain is a major ecological problem around the globe. It is caused by emissions of sulphur dioxide (SO_2) and nitrogen oxides (NO_x), which react with the water molecules in the atmosphere to produce acids. It has been shown to have adverse impacts on forests, freshwater and soils, killing insect and aquatic life-forms, causing paint to peel, corrosion of steel structures such as bridges, and weathering of stone buildings and statues as well as having a variety of negative impacts on human health.[Dondapati u. a., 2014]

The goal of this project is to develop a machine learning classifier that predicts if the next rainfall event will be acidic rain or normal. The algorithm considers the Nitrogen oxides concentration, Sulphur dioxide concentration, rainfall amount and ambient temperature to classify the rain event. The economic loss due to acid rain is significant. As per the study conducted by National Environmental Monitoring

Network, China, the material loss caused by acid deposition in China was 32.165 billion Yuan (= 4 billion Euro) in 2013, accounting for 0.057% of GDP(Gross Domestic Product).[Zhang u. a., 2017] The prediction of acid rain, might have an impact on the regulations of pollutant level in the atmosphere.

2 Data set

This data mining project, is based on the environment pollution data of Taiwan recorded in the year 2015. The data originates from the Environmental Protection Administration (EPA) of Taiwan. EPA maintains the database of the air quality Pollution Standard Index (PSI) and related information.The database is updated hourly and is publicly accessible on the EPAs website.An English translation is available on the predictive modeling and analytic platform, Kaggle.[website:kaggle, 2016] The raw data set contains 23 columns (representing various environmental parameters) with 218641 rows of data. The data set consists hourly pollutant data from 25 stations: Banqiao, Cailiao, Datong, ..., Xinzhuang, Xizhi, Yangming, Yonghe, Zhongli, Zhongshan.[website:epa, 2015] All of these are located in the northern part of Taiwan and the cover mostly non-overlapping areas.The granularity of the data is 1 hour window per data point. The pollutant data includes Nitrogen oxides concentration, Sulphur dioxide concentration, rainfall quantity, pH level of the rainfall, ambient temperature, wind speed, carbon dioxide concentration, Ozone concentration and other values which may or maynot be linked to acid rain. Furthermore there are a variety of data points that need to be handled: # indicates invalid value of measurement by equipment inspection. * indicates invalid value of measurement by program inspection. x indicates invalid value of measurement by human inspection. NR indicates no rainfall and blank indicates no data available.
The pH scale measures how acidic a substance is. The scale ranges from zero (the most acidic) to 14 (the most basic). Pure water has a pH value of 7 and is considered neutral: neither acidic or basic. Normal, clean rain has a pH value of between 5.0 and 5.5. However, when rain combines with sulphur dioxide or nitrogen oxides, e.g. from automobiles,rain becomes much more acidic. The definition of acid rain varies across organizations. The United States Environment Protection Agency defines acid rain as rain water with pH level below 4.5. The Environment Protection Agency, Taiwan defines the acid rain by a pH level below 5.
Depending on the pH threshold, the data set is (almost) balanced for pH threshold of 4.5: 1740 acid rain events and 1597 regular rain events resulting in a ratio of 52% vs 48%. In case of a pH threshold of 5.0: 2685 acid rain events compared to 653 regular rain events is obtained, which results in an unbalanced set of 80% vs 20%.

3 Approach

The data mining project is implemented in Python using the Machine Learning library Scikit-learn for training and selecting of models.

3.1 Filtering of rain events

In the first step raw hourly data points are divided into rain and non-rain events. Based on a threshold of 0.05mm/hour, the data point is either above the threshold and therefore labeled as a rain event, or below/equal and therefore a non-rain event.

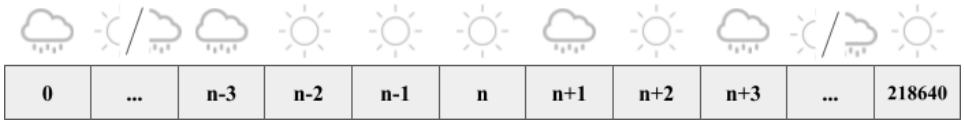

Figure 1: hourly data, marked with rain (cloud) or no rain (sun)

3.2 Marking of acid rain

The second step is only applied to events labeled as rain event. The pH levels of rain-labeled data points are compared with the pH threshold. As a result every rain event is either labeled as acidic or non-acidic.

3.3 pH value

Since there is no standard pH range for acidic rain, it's necessary to adapt the approach depending on the definition applied. As it is not a parameter which need to be optimized, we do not consider it as a variable.

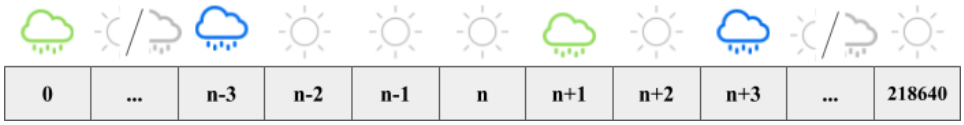

Figure 2: hourly data, acid rain (green cloud) and regular rain(blue cloud)

3.4 Variables

Parameters which can be selected by us, independent of any definition, are described in this section. The goal is to find the best setup of parameters, to maximize the F1 Score.

Windowing

The raw data is split into 1 hour long events. Yet there is no particular reason to assume 1 hour is the right time frame to decide for an acidic rain event. In the Windowing process, consecutive rain data points are accumulated into Rain Windows. A Rain Window is considered as acidic when at least one of it's data points is labeled as acidic.

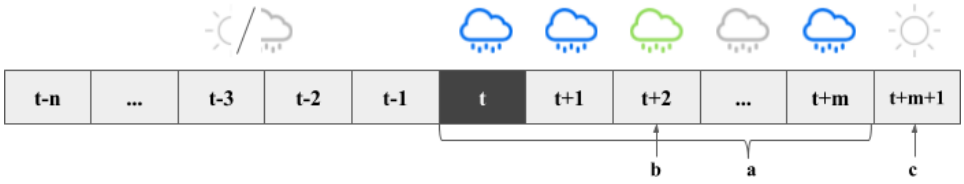

Figure 3: a Rain Window of length m hours (a) is considered acidic in case at least one hour is marked as acid rain (b). The end of the window reached, when one hour has no rain event (c)

The next possible window could start in t+m+2, if it contains rain labeled event. This leads to a variation of the research question. From originally "How likely is it that the rain in the next hour will be acid rain" to "How likely is it that the rain starting next hour will be acidic at some point?"

Rainfall

Unlike NO_x and SO_2, the previous rainfall is not as directly linked to acid rain. Yet,it is worth investigating if the previous rain, which alters the NO_x and SO_2 concentration, might impact the prediction.

Ambient temperature

Chemical processes occur at different speed, depending on the temperature of the substances involved. The temperature of the environment might have an influences on the creation of acid rain.

Previous data points

Another unknown variable is how many previous data points are relevant for the prediction. As a hard requirement, we defined it a necessity that the previous data point of a rain event will have valid NO_x and SO_2 values. If either one is not valid the case the rain event will be ignored. Yet, similar to the line of thinking behind the Windowing process, there is no reason to assume that 1 hour is the perfect knowledge basis to make a prediction.

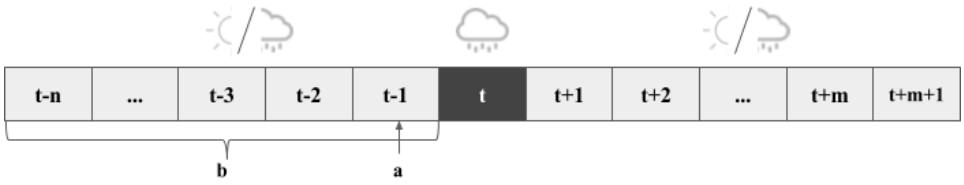

Figure 4: n previous data points (b) before the rain event, the predecessor (a) is required to contain valid NO_x and SO_2 values.

Functions

When a time series of data points is considered, the data points need to be accumulated. In our approach we will compare: summarizing, averaging and the calculating of the median value. Average and median have the inherent benefit of normalizing itself by division by the amount of data points considered. In case of sum there is no such normalization. The average and median approach are supposed to answer if a certain (e.g. NO_x, SO_2 or ambient temperature) threshold needs to be passed for a extended period of time. The summary approach on the other hand is supposed to be more effective when certain spikes in values cause an acid rain event.

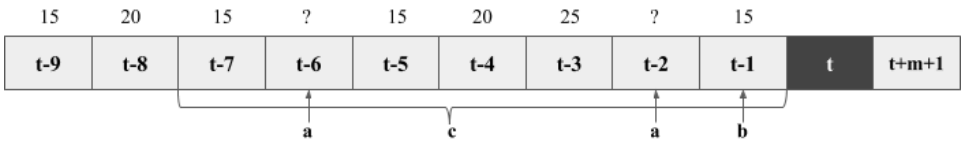

Figure 5: Example NO_x data: the previous 7 hours will be considered (c), for t-1 NO_x and SO_2 must exist, in this example t-6 and t-2 have no valid NO_x values (a). In either case the NO_x values 15, 15, 20, 25 and 15 will be used: the average of it therefore is 18.0, the median 15.0 and the sum 90.

4 Evaluation

The aim of this project is to predict whether the rain event will be acid rain or not. We have implemented two classifiers with different parameters on binary prediction - KNN, normal Naive Bayes and Naive Bayes with 5 and 10 stratified k folds, all of these preprocessed with and without time windows. We split our data set into 2 parts - 20% test data and 80% training data. For imbalanced data and considering some classifier ,like KNN, need to compute the distance between samples, we have used Min-Max Normalization to scale both training and test data between 0 and 1 for better data structure and higher accuracy.[Shalabi und Shaaban 2014] The main goal is to optimize the F1-score. The analysis is structured as follows: We first set up a basic baseline model to have a reference for different methods and parameter settings. Next, we implement two consecutive iterations of our model and one iteration that contains more than 1 hour time period.

4.1 Methods

Brief introduction to applied classifiers

KNN classifier predicts the label of point i by the label of k nearest neighbors.[D.Coomans und D.L.Massart 1982] KNN classifier calculates the distance between the data point to each of the centroid point. Therefore, KNN is more efficient when applied on comparatively smaller dataset. After filtering rain events, there are 3337 rows of data remained. Naive Bayes classifier is supervised learning algorithm based on applying Bayes theorem with the naive assumption of conditional independence between every pair of features.[Lorenzo Trippa und Parmigiani 2015] The data set is split into a 80% training set and 20% test set. To avoid bias, K-Fold Cross Validation is used to validate through generating different combinations of the dataset.The Naive Bayes 5-fold and 10-fold is implemented as our 3^{rd} and 4^{th} classifiers.

4.2 Baseline

Scientific research showed that acid rain is mainly caused by SO_2 and NO_x, this was combined with considering only at the values at t-1. For this the KNN classifier showed the highest F1 score = 0.57 (see details in Table 1).

4.3 Variables

Evaluation setup

To find the best configuration brute force was used on all variable parameters. The main distinction, which affects the acid rain labeling process, is the pH threshold of either 4.5 and 5.0. For both pH values the Windowing approach is compared to hourly rain data points. These 4 possible scenarios are executed for 1 to 12 previous data points, which results in 96 scenarios. For the scenarios which involve more than 1 data point, the sum, avg and median function are used to aggregate the previous data point. Therefor the initial run compared 284 scenarios Since this first iteration indicated a saturation of F1 score quality, a rerun with a more specified configuration: pH 4.5 and 5.0, no Windowing, always include ambient temperature, with and without previous rainfall, sum+avg+median functions was run for 13 to 20, 25, 30, 50, 100, 150, 300 and 600 data points. This results in a total of 440 configurations which are compared in this section.

pH Values

Due to the varying definition of the acid rain across global organizations, the evaluation is done independently, considering pH level of 4.5 and 5 as the threshold, for the classification of rain event. Based on the threshold, the rain event is classified as acidic or normal rainfall event. The basic aim of the algorithm is to predict if the rain event shall be acidic or neutral.

Rainfall

The acidity of the rainfall depends on the quantity of rainfall. The more the quantity of rain, more is the neutralizing effect of the pollutants.The algorithm was developed considering the NO_x, SO_2 concentration and rainfall one hour prior to rainfall event as the key parameters and the pH level as the affected parameter.
The performance of the model improved with this measure. For baseline model K Nearest Neighbour classifier, Naive Bayes classifier and Naive Bayes with k stratified fold(k=5 and 10) was used. In the end, KNN classifier showed the highest F1 score(F1 score = 0.62). Detail about the performance of the methods is shown in Table 1.

Windowing

The Windowing approach turned out to be a dead end. Clustering into rain windows blurred the the outcome of the predict. Furthermore it reduced the amount of data for evaluating itself by almost 2/3, compared to the no window approach. The highest no-Windowing F1 score is 0.85 compared to 0.69 for Windowing approach (Table 1).

Ambient temperature

A scatter plot of the acid rain event and non acid rain event w.r.t. ambient temperature was plotted (see Figure 6). From the figure it can be concluded that the probability of acid rain is higher when the temperature is lower. The algorithm was developed considering the NO_x,SO_2 concentration,rainfall and ambient temperature one hour prior to rainfall event as the key parameters and the pH level as the affected parameter.

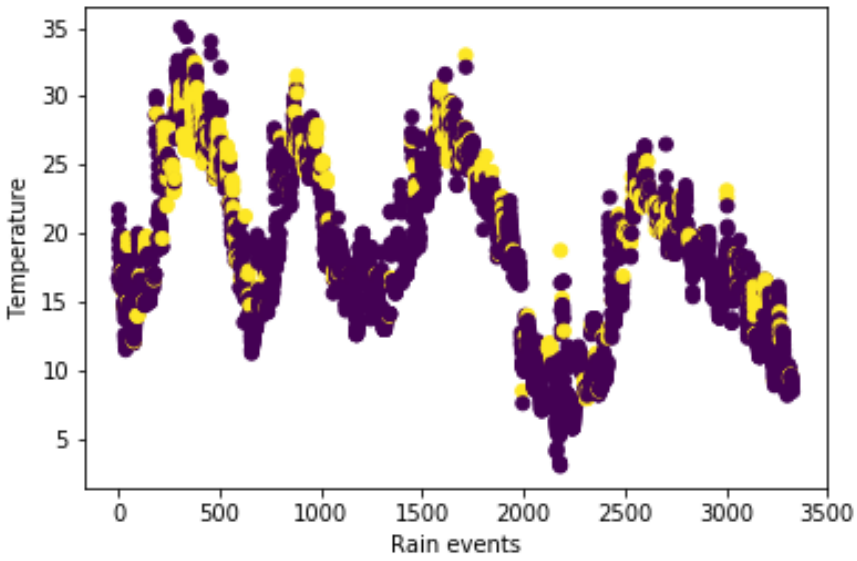

Figure 6: Acidic rain events and temperature

The performance of the model improved with this measure. For baseline model K Nearest Neighbour classifier, Naive Bayes classifier and Naive Bayes with k stratified fold(k=5 and 10) was used. In the end, KNN classifier showed the highest

F1 score(F1 score = 0.66). Detail about the performance of the methods is shown in Table 1.

Number of previous data points and Functions

Till this point, the data of key parameters one hour prior to rain event was considered. But what if the acidity of the rain event depends not only on key parameter measurement one hour prior but over 6 hours or 12 hours or 24 hours? To analyze the same, model was modified to consider the previous data of the key parameters.The performance of the model greatly improved with the number of data points as shown in the figure 7 below.

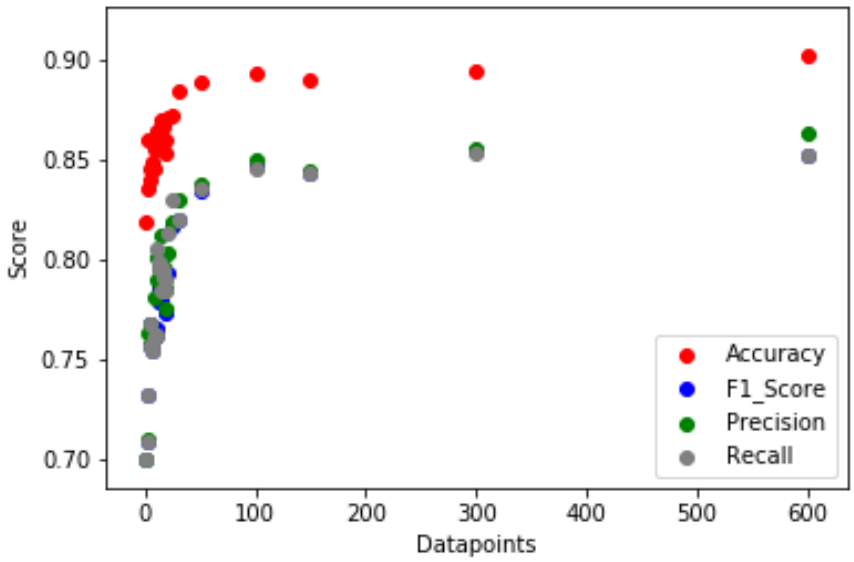

Figure 7: Variation in score vs number of previous data points considered.

5 Results

The result of the performance after each improvement in algorithm is shown in the below table.

Baseline	Accuracy	Precision	Recall	F1-score
KNN	0.784	0.615	0.566	0.573
Naive bayes	0.786	0.573	0.523	0.511
Naive bayes with 5 folds	0.790	0.589	0.529	0.518
Naive bayes with 10 folds	0.789	0.583	0.526	0.515
Considering rainfall	Accuracy	Precision	Recall	F1-score
KNN	0.629	0.628	0.628	0.628
Naive bayes	0.581	0.623	0.572	0.528
Naive bayes with 5 folds	0.791	0.620	0.556	0.560
Naive bayes with 10 folds	0.791	0.620	0.556	0.560
Considering temperature	Accuracy	Precision	Recall	F1-score
KNN	0.701	0.700	0.700	0.700
Naive bayes	0.784	0.606	0.554	0.558
Naive bayes with 5 folds	0.792	0.630	0.569	0.577
Naive bayes with 10 folds	0.791	0.628	0.568	0.576
Previous 50 data points	Accuracy	Precision	Recall	F1-score
KNN (avg)	0.835	0.836	0.834	0.835
Naive bayes (sum)	0.624	0.652	0.618	0.599
Naive bayes with 5 folds (sum)	0.626	0.656	0.615	0.594
Naive bayes with 10 folds (sum)	0.624	0.653	0.613	0.592
Previous 300 data points	Accuracy	Precision	Recall	F1-score
KNN (sum)	0.855	0.856	0.854	0.854
Naive bayes (sum)	0.766	0.616	0.604	0.609
Naive bayes with 5 folds (sum)	0.780	0.631	0.606	0.614
Naive bayes with 10 folds (sum)	0.782	0.636	0.612	0.621

Table 1: Evaluation table of best result of each classifier with acid rain defined as pH < 5

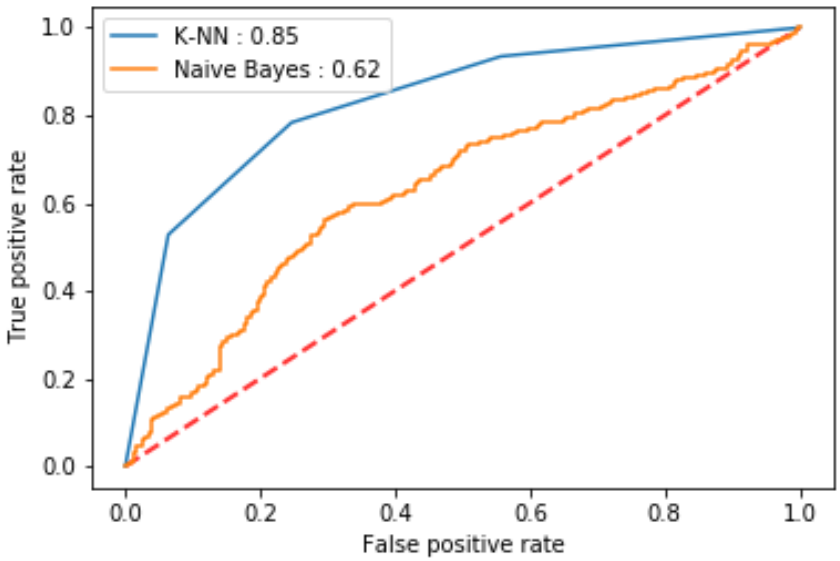

Figure 8: Receiver Operating Characteristic

The ROC curve for the fine tuned parameter is as shown in above figure 8.[website:roc] The area under the curve for the KNN algorithm and Naive Bayes classifier for the fine tuned parameter is 0.85 and 0.63 respectively. The ROC Curve for KNN is closer to the upper left corner, which indicates a superior performance.

6 Conclusion

In this report we have shown that Machine Learning methods can be used to predict if the rainfall event will be acidic or neutral with a good precision. Acidity of rainfall can be determined by considering the previous data points of NO_x concentration, SO_2 concentration, rainfall quantity and ambient temperature. We observed that the inclusion of more previous data points increase the quality of the prediction. Yet at 50 hours before the rain event the F1 score saturates at 0.83 and the further quality increase is marginal.
As further step the previous pH value of the rainfall can be used to determine the acidity of the rain event. As in case of continuous rainfall, the acidity of the rainfall tends to reduce. This may further improve the performance of the algorithm.

Bibliography

[website:roc] *Detector Performance Analysis Using ROC Curves*. – url = "http://ww2.mathworks.cn/help/phased/examples/detector-performance-analysis-using-roc-curves.html"

[website:epa 2015] *Air quality in northern Taiwan*. 2015. – url = "https://taqm.epa.gov.tw/taqm/en/YearlyDataDownload.aspx"

[website:kaggle 2016] *Air quality in northern Taiwan*. 2016. – url = "https://www.kaggle.com/nelsonchu/air-quality-in-northern-taiwan"

[D.Coomans und D.L.Massart 1982] D.COOMANS ; D.L.MASSART: *Alternative k-nearest neighbour rules in supervised pattern recognition Part 1. k-Nearest neighbour classification by using alternative voting rules*. Bd. 163. S. 15–27, Elsevier, 1982

[Dondapati u. a. 2014] DONDAPATI, Naveen ; POLEPELLI, Buddha R. ; AHMED, Imtiyaz ; MATHSA, Chandrakanth: Computational approach to identify the acid rain patterns by adopting satellite imagery data mining technique. In: *International Conference on Computing and Communication Technologies* (2014)

[Lorenzo Trippa und Parmigiani 2015] LORENZO TRIPPA, Curtis H. ; PARMIGIANI, Giovanni: BAYESIAN NONPARAMETRIC CROSS-STUDY VALIDATION OF PREDICTION METHODS. (2015)

[Shalabi und Shaaban 2014] SHALABI, Luai A. ; SHAABAN, Zyad: Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix. In: *International Conference on Computing and Communication Technologies* (2014)

[Zhang u. a. 2017] ZHANG, Yinjun ; LI, Qian ; ZHANG, Fengying ; XIE, Gaodi: Estimates of Economic Loss of Materials Caused by Acid Deposition in China. (2017)