Using Naïve Bayes Algorithm to Classify Level of Drought

Mohamed Abdi1

1 Department of Computer Science and Software Engineering University of Washington, Bothell, WA USA

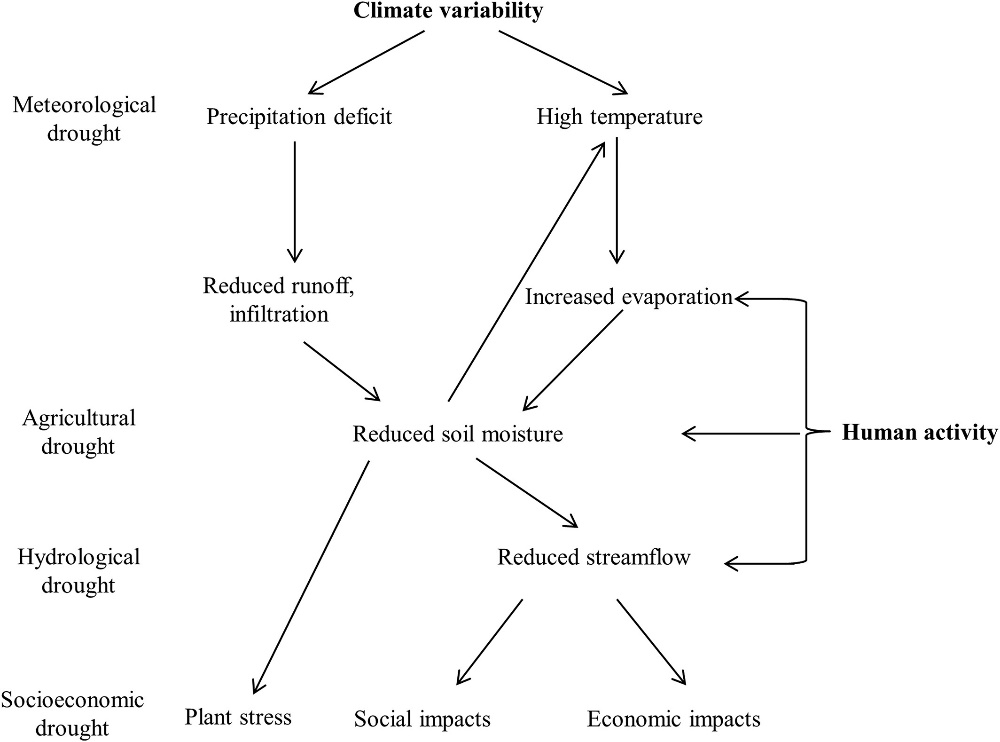
moabdi21@uw.edu

**Abstract.** Droughts are a natural disaster that affect large groups of people, which can lead to food shortages, wild fire, etc. This kind of natural disaster is still a common one, but it is becoming more severe as the years go by mostly due to global warming. This paper goes over a machine learning perspective that can be used to help classify droughts much quicker. Machine learning is a tool that’s very capable of producing much quicker and accurate calculations if implemented correctly which can also be used to predict, but without data, machine learning wouldn’t be possible. We currently live at time that we’re capable of collecting so much information on climate to take the necessary steps to accomplish identifying or predict certain possibilities. We are going to implement a Naïve Bayes Algorithm to classify the Standard Precipitation Index based on the average precipitation that occur by months.

**Keywords:** Naïve Bayes, Standard Precipitation Index (SPI), Palmer Drought Severity Index (PDSI) Famine, Droughts, Anthropogenic Drought, hydrological cycle, Available Water Content

1. INTRODUCTION

Natural Disasters have always been a danger to humanity, it does not matter how far we progress to the future they will always remain a danger to us today and the future. Droughts are among one of the most disastrous natural hazards [1]. There are several severe droughts that have occurred very recently such as in 2017 with a few countries in Africa [2], 2012-2016 California drought [3]. This drought affects several millions of people just like 2017 Somali’s drought which had about 6.2 million people at risk facing acute food insecurity [2]. We live at point to where humans are one of the few factors to causing such droughts today. When such drought is created by humans, they’re categized as anthropogenic drought. California sets an example of this kind of drought with its growing population and agriculture which uses a heavy consumption of water as it can lead to much quicker pace of droughts [4] (see fig 1 for other drought factors).

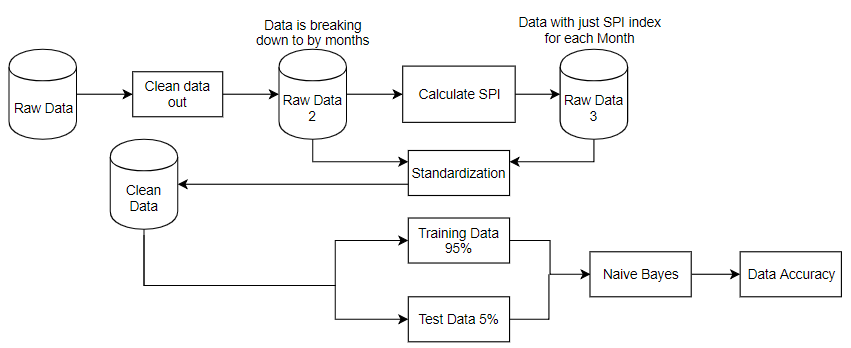


**Fig. 1.** A figure from taken from Hao article Represents the interaction of different factors in the hydrological cycle in droughts [1]

1. METHODS

This section gives an overview of how droughts are measured, describes and analyses the data we used, and how machine learning was implemented to achieve a program that can classify droughts.

This is a high-level flowchart of how the experiment was performed:



**Fig. 2.** A strategy of how this experiment was broken down. The raw/clean data represents single city data set, not all four cities.

* 1. Measuring Droughts

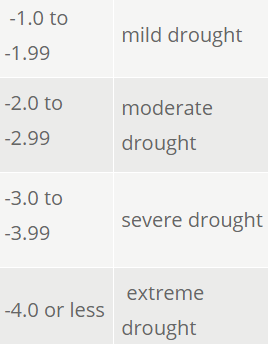
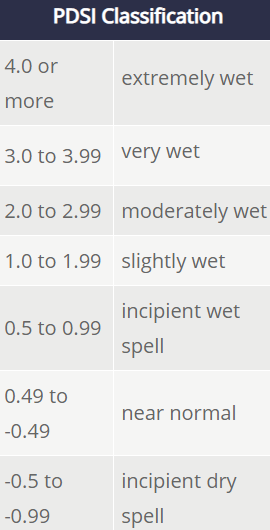
There are a few ways to measure droughts these measurements are used a lot now a days such as the Standard Precipitation Index (SPI), and Palmer Drought Severity Index (PDSI). This is tools developed by research to help define onset, severity, and end of droughts. These measurements are possible by taking the data of rainfall, snowpack, streamflow, etc. [5].

The SPI is index only based on the amount of precipitation for a various of time frames [5]. This drought measurement is widely used around the world, since it can be used on a variety of time scales this method of measurement can be used for short-term agricultural and long-term hydrological applications [5]. SPI is classified into 7 different groups going from extremely wet to extremely dry (see figure 3). In this paper we are going to be using the SPI, due to its simplicity of needing just one attribute & issue encountered with the dataset (talked about in 2.2). The way SPI is calculated by inadequacy of the gamma distribution by just using precipitation but won’t be explained on how its calculated since we used SPI program from National Drought Mitigation Center (check out reference 6 for how SPI is calculated).



**Fig. 3.** Different classified groups for SPI values.

The PDSI is bit more difficult to measure due to having more indices required by it. In order to identify its level of drought we need to know the amount of precipitation, temperature, and Available Water Content (AWC) of the soil [5]. PDSI can give a more accurate drought classification as it has 11 different groups going from extremely wet to extreme drought (shown in figure 4), but its more complex drought measurement that also has unspecified, build in time scale that can be a bit misleading [5].



**Fig. 4.** Different classified groups for PDSI values.

* 1. Dataset and Data processing

First step in getting a dataset was to first choose a geographic location, somewhere that has weather centers with consistent and complete attributes. East Africa was first choice of location, but due to missing sets of data which would’ve led to inaccurate classifications. We moved to California and targeted four different cities which are Monterey, Redding, Riverside, and Visalia. Each city had daily climate data and the attributes they had are precipitation, temperature average, snow, & evaporation rate as shown in table 1.

**Table 1.** Sample Raw dataset of daily weather at Monterey, CA. Raw data for all cities were collected from National Oceanic And Atmospheric Administration (NOAA)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **STATION** | **NAME** | **DATE** | **PRCP** | **TAVG** | **SNOW** | **EVAP** |
| USW00023259 | MONTEREY PENINSUL AIRPORT, CA US | 4/1/1998 | 0.49 | 51 |  |  |
| USW00023259 | MONTEREY PENINSUL AIRPORT, CA US | 4/2/1998 | 0.12 | 54 |  |  |
| USW00023259 | MONTEREY PENINSUL AIRPORT, CA US | 4/3/1998 | 0.6 | 52 |  |  |
| USW00023259 | MONTEREY PENINSUL AIRPORT, CA US | 4/4/1998 | 0.14 | 52 |  |  |
| …. | …. | …. | …. | …. | …. | … |

Due to some attribute such as evaporation rate (EVAP), and PRCP being inconsistent for some of the cities. It’s impossible to calculate the PDSI without these attributes. We limited to just using SPI as our drought classification.

Since we are focused on classifying on SPI, there must be a certain time frame whether it’s a month, 3 months, or year. We decided to break it down for each city to have a time scale of one month. Which meant for each year is going to have a total of 12 months. Each following month would have the attribute average precipitation, SPI index, and its classification as shown in table 2.

**Table 2.** Clean Data for Monterey, CA for 1995

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Average PRCP** | **SPI** | **SPI Classification** |
| 1/1/1995 | 0.342258 | 1.23 | Moderately wet |
| 2/1/1995 | 0.026071 | -0.84 | Near normal |
| 3/1/1995 | 0.234194 | 1.14 | Moderately wet |
| 4/1/1995 | 0.074667 | 0.57 | Near normal |
| 5/1/1995 | 0.01871 | 0.18 | Near normal |
| 6/1/1995 | 0.046667 | 1.25 | Moderately wet |
| 7/1/1995 | 0.000645 | -0.26 | Near normal |
| 8/1/1995 | 0.000968 | -0.1 | Near normal |
| 9/1/1995 | 0 | -0.32 | Near normal |
| 10/1/1995 | 0.000968 | -1.15 | Moderately dry |
| 11/1/1995 | 0.007333 | -1.19 | Moderately dry |
| 12/1/1995 | 0.075484 | -0.12 | Near normal |
| … | … | … | …. |

Once the clean data is obtained, we get a step closer to implementing the Machine learning Technique, but first we must divide our data to training and test data set. The training dataset was group up with the years 1995-2017 and test dataset with only year 2018. Both training and test datasets had the same structure. Instead of going by years, this dataset went by months and each month would hold possible attributes with its classification for that single month. For example, the training dataset had years 1995-2017 and let say for the month of January would hold all the possible attributes and classifications for the years 1995-2017 for that single month. Then we can begin implementing with the machine learning technique.

* 1. Machine Learning Technique

The Machine learning Technique chosen for this experiment is Naïve Bayes Classifier Algorithms. This is an easy to build, with no complicated iterative parameter estimation which it very useful for very large datasets. The Naïve Bayesian often commonly used due to how well it performs with classification problems, as Naïve Bayesian is all about probability of what might occur. As by using you using data about the past it can provide a way of calculating the posterior probability, P (c | x), from P(c), P(x) and P (x | c) [8].

*P(c|x) = P(x|c)P(c) /* P(x) (1)

Where P(c | x) = P(x1 | c) x P(x2 | c) x ….. x P(xn | c) x P(c), P(c | x) is the posterior probability of targeted class given its feature, P(c) is the prior probability of class, P(x | c) is the likelihood which is the probability of predictor given class, P(x) is the prior probability of predictor [8].

1. EXPERMENTS & RESULTS

The first plan in mind was to take use of both SPI and PDSI drought measurements and have total of just two classifications which would have been a yes for a drought is going to occur and no for the opposite result. Due to having inconsistence with attributes such as average temperature, and evaporation rate also not having Available water content it became impossible to calculate PDSI. The project got simplified to just identifying the level of SPI it at.

Since changing to classifying SPI, the experiment would change to problem of identifying a total of 7 different classes (see fig. 2 in section 2.1). The clean data was set up by months and for every city we made a prediction of what level of SPI it is at for each following month (a total of 12 month for each city). Due to the test dataset being 2018 the experiment is only capable of predicting 10 months (January – October) as shown in table 3.

**Table 2.** Monterey, CA Naïve Bayes Classification Predication for 2018

|  |  |
| --- | --- |
| **Month** | **SPI Classification** |
| January | Near normal |
| February | Near normal |
| March | Near normal |
| April | Near normal |
| May | Near normal |
| June | Near normal |
| July | Near normal |
| August | Near normal |
| September | Near normal |
| October | Near normal |

The SPI Classification for Monterey, CA 2018 is shown that from January – October has been neither a drought or wet, but Near normal. Which means that it stayed at its average amount of precipitation in that region. Now how accurate is this prediction (see fig. 5).

**Fig. 5.** Accuracy of the Naïve Bayes prediction to the expected result for all four cities.

Figure 5 shows Monterey received an 100% accuracy, with the other cities at 90% accuracy. This experiment only had one feature which is the average precipitation for every month and that’s the only necessary attributed need to identify the level of SPI.

1. DISCUSSION

Based on what we’ve seen from the results, using Naïve Bayes algorithm showed a high accurate rate, which is quite surprising especially since the average precipitation data who had the same month, but different years showed a difference from each other and SPI is calculated by the probability of precipitation for various time frames [5]. If there wasn’t any issue with the other attributes in the raw data (average temperature, evaporation rate) and was implemented into the clean data as two more features would the accuracy dropped? Or increased? But still having inconsistent and incomplete data shows that’s there’s progress need to be done in collecting data.

This experiment is just idea and start of implementing for machine to classify a level of drought. The Original goal was to experiment on producing a program that’s capable of forecasting an upcoming drought in the future, but it a project that is going to take much longer time then expected to be. For future work, we would like to continue working on this project. When better quality data are collected, we then can take the step of using a more advanced machine learning techniques like Artificial Neural Networks. Drought prediction is quite important for the future especially with the situation we’re at with our environment. By creating this program, we can help people to be more prepared and readier for when such harmful disasters are on their way.

References

1. Hao, Z., Singh, V. P., & Xia, Y. “Seasonal Drought Prediction: Advances, Challenges, and Future Prospects”. Reviews of Geophysics, 5 Jan.2018. https://doi.org/10.1002/2016RG000549
2. Anyadike, Obi. “Drought in Africa 2017.” IRIN, 19 Dec. 2017, [www.irinnews.org/feature/2017/03/17/drought-africa-2017](http://www.irinnews.org/feature/2017/03/17/drought-africa-2017).
3. “2012-2016 California Drought: Historical Perspective.” California Droughts Compared | USGS California Water Science Center, 3 Apr. 2018, ca.water.usgs.gov/california-drought/california-drought-comparisons.html.
4. AghaKouchak, Amir. “Water and Climate: Recognize Anthropogenic Drought.” Nature News, Nature Publishing Group, 25 Aug. 2015, [www.nature.com/news/water-and-climate-recognize-anthropogenic-drought-1.18220](http://www.nature.com/news/water-and-climate-recognize-anthropogenic-drought-1.18220).
5. “MEASURING DROUGHT.” Home, National Drought Mitigation Center, drought.unl.edu/ranchplan/DroughtBasics/WeatherandDrought/MeasuringDrought.aspx.
6. Blain, Gabriel C., & Meschiatti, Monica C.. (2015). Inadequacy of the gamma distribution to calculate the Standardized Precipitation Index. Revista Brasileira de Engenharia Agrícola e Ambiental, 19(12), 1129-1135. https://dx.doi.org/10.1590/1807-1929/agriambi.v19n12p1129-1135
7. SPI Program, <https://drought.unl.edu/droughtmonitoring/SPI/SPIProgram.aspx>
8. “Naive Bayesian.” Model Deployment, www.saedsayad.com/naive\_bayesian.htm.
9. Sriram K., Suresh K., “Machine Learning Perspective for Predicting Agricultural Droughts Using Naive Bayes Algorithm”. Middle-East Journal of Scientific Research,
10. Park, Seonyoung, et al. “Drought Assessment and Monitoring through Blending of Multi-Sensor Indices Using Machine Learning Approaches for Different Climate Regions.” NeuroImage, Academic Press, 11 Nov. 2015, www.sciencedirect.com/science/article/pii/S0168192315007467.
11. Wood, Eric F., et al. “Prospects for Advancing Drought Understanding, Monitoring, and Prediction.” Journal of Hydrometeorology, vol. 16, no. 4, 29 July 2015, pp. 1636–1657., doi:10.1175/jhm-d-14-0164.1. “https://journals.ametsoc.org/doi/full/10.1175/JHM-D-14-0164.1”
12. “Drought Basics.” National Drought Mitigation Center,
13. “drought.unl.edu/ranchplan/DroughtBasics/WeatherandDrought/MeasuringDrought.aspx”
14. Belayneh A. Adamowski J. “Standard Precipitation Index Drought Forecasting Using Neural Networks, Wavelet Neural Networks, and Support Vector Regression”. 18 July 2012