A Machine Learning Approach for Rainfall and Crop Prediction to Assist Farmers in Suitable Crop Production

Aditya Singh Shekhawat and Rishin Haldar*

School of Computer Science and Engineering, Vellore Institute of Technology,
Vellore, Tamil Nadu, India

6th Annual Agriculture to Nutrition Scientific Symposium, Kathmandu, 28th Nov, 2018

Outline

- Background and Motivation
- Objective
- Related Work
- Methodology
  - Data Preprocessing (Collection and Integration)
  - Classification
- Results & Discussion
- Conclusions & Limitations of the Study

Background and Motivation

- Agriculture sector contributes to more than 17% of the GDP of India. Therefore, the country is heavily dependent on the optimal availability of water for sufficient crop production to sustain the economy.
- The monsoon/rainy season (July to September) contributes to 70% of the annual rainfall, and Kharif crop production is almost entirely dependent on this rainy season.
- Most of the farmers in India depend only on the predictions given by the Indian Meteorological Department (IMD) to plan their agricultural activity, especially during this season.
- IMD predictions are based on current atmospheric conditions, and it is susceptible to rapid changes.

Objective

- Provide an alternate (or complementary) solution to the farmers by predicting the monthly rainfall after analyzing data archives.

- Provide recommendation to the farmer on the crop(s) that can be grown for that particular region by utilizing the
  - predicted rainfall
  - crops that grow in that region
  - Amount of rainfall that these crops require

Related Work

- Buishand et.al (1999), Cong et.al (2012), Betts et.al (2014) and others highlighted the correlation of factors like temperature, humidity, cloud cover etc. with precipitation.
  – In the absence of any of these attributes and /or the availability of other attributes, the effectiveness of the prediction may not be reliable.

- Kannan et.al (2010) used multivariate regression analysis on 5 years of precipitation data.
  – The predicted values were lower than the actual recorded values.

Related Work (Contd..)

- Kumar et.al (2016) compared popular data mining techniques and showed that Naïve Baye’s and kNN classifiers gave encouraging results with respect to rainfall prediction.
  - The highest accuracy was only 80%.
- Dabney et.al (2007) showed that nearest centroid classifiers were well suited for multidimensional applications.
  - The study was on genomic data.

Methodology

- **Data Collection and Integration**
  - Dataset A containing various attributes including the output attribute (rainfall amount)
  - Dataset B containing the district wise crop production
  - Dataset C containing the minimum and maximum rainfall range for crops
- **Classification/Prediction** (to predict the rainfall amount 'D', for a particular district)
  - i) Feature reduction on Dataset A
  - ii) Nearest Centroid Classification on reduced feature vector
  - iii) Find accuracy of the model by applying testing data
- **Map the dataset** B, along with dataset C with the predicted value of rainfall 'D' to generate 'E', the answer.
- Therefore, if the stakeholder (farmer) provides the input attributes, the **system will generate the crop(s) that can be grown**.

Data Collection and Integration :

- The data needed to predict the rainfall amount was gathered in .csv format from a government portal, http://www.indiawaterportal.org/met_data/
- By choosing the State, District, range of years and the attribute, one **.csv file** was generated which contained the monthly data for the specified attribute for all the chosen years.
- There were **11 attributes** : Precipitation, Minimum, Average, Maximum Temperature, Cloud Cover, Vapour Pressure, Wet Day Frequency, Diurnal temperature range, Ground frost frequency, Reference Evapotranspiration and Potential Evapotranspiration

Data Collection and Integration (Contd..)

www.indiawaterportal.org/met_data/

Python data visualiza | TCB Overview of Tutorials | Tutorial Program - In | Gromacs - Gromacs | ONCOMINE: A Cance | D

State RAJASTHAN

District BUNDI

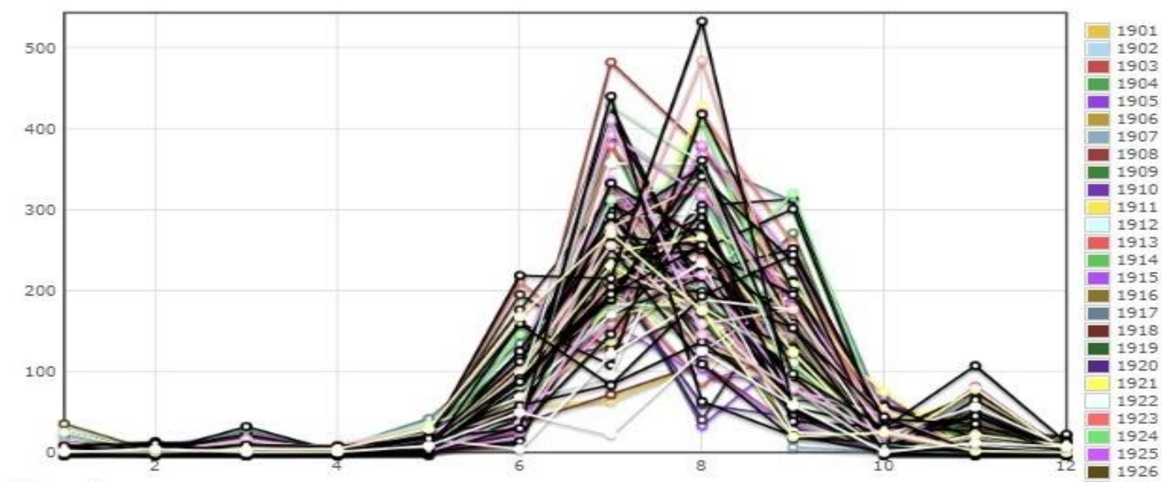
Date Type Precipitation

From 1901

To 2002

Generate Annual Mean

Generate Monthly Mean for each year | Show Annual Totals



Units: Integer / mm

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1901	11.621	0.94	0.681	0.484	5.349	33.241	245.404	208.835	22.062	24.815	0.18	5.126
1902	4.567	0	0	0	0.106	63.537	201.425	136.504	106.295	12.532	31.184	2.013
1903	0.833	0.187	0.256	0	7.806	11.044	332.421	268.145	138.167	6.488	0.18	0

Data Collection and Integration (Contd..) – Dataset A :

- **Bundi District** of Rajasthan State was chosen as a sample.
- Data from 1901 to 2002 were collected.
- Data from **11 .csv files** (for each of the attributes) were integrated. Thereafter, based on the **Kharif** crop season, data from **June to September** months were selected.
- There was no need to clean the data.
- Dataset A has been created

Year	Month	Min Temp	Max Temp	Cloud Cover	Other 7 features	Precipitation
1901	June					
1901	July					
1901	Aug					
1901	Sep					
Other years						
2002	June					
2002	Jul					
2002	Aug					
2002	Sep					

Data Collection and Integration (Contd..) – Dataset B :

- The district wise crop production data (32 districts in Rajasthan state) was gathered from https://www.rajras.in/index.php/rajasthan-agriculture-crops-snapshot/
- Dataset B

District	Crop
Ajmer	Jowar
Alwar	Bajra:Tur:Urad
Banswara	Tur:Sanhemp
Baran	Moth
Bharatpur	Caster Seed
Bhilwara	Maize:Urad
Bikaner	Groundnut:Guar
Bundi	Rice:Maize:Urad:Chowla:Sugarcane
Chittorgarh	Maize:Seasumum:Cotton:Sugarcane:Sanhemp
Churu	Moth:Millets
Dausa	Tur:Sugarcane

Data Collection and Integration (Contd..) – Dataset C :

- The minimum and maximum rainfall required for the crops were gathered from various agricultural web portals.
- Dataset C

Crop	Min Rainfall	Max Rainfall
Jowar	30	60
Bajra	50	100
Tur	45	65
Urad	50	65
Sanhemp	100	120
Moth	40	65
Maize	40	75
Groundnut	50	90
Rice	150	300
Sugarcane	100	140

Classification / Prediction : i) Feature Reduction

- Each row in the dataset A had 13 attributes, out of which the first two (Year and Month) did not have any relevance to prediction, thus they were removed.
- Out of the eleven attributes, there were ten input attributes/dimensions and one output attribute (Precipitation).

Min Temp	Avg Temp	Max Temp	Cloud Cover	Vapour Pressure	Wet Day Freq	Diurnal Temp Range	Ground Frost Freq	Ref Evapo Transp	Potential Evapo Transp	**Precip-itation**

- For larger dimensions, a simple Decision Support Tree (ID3 algorithm, using Entropy) can be used to determine the most significant causal/independent attributes.
  - The nodes of the tree represents the input attributes. The root of the tree signifies the most important input attribute. The importance reduces as one traverses down the tree.
- However, since the dimension value was low, correlation was tried out for each of the input attributes with respect to the Precipitation.
- Top four attributes were chosen: Max Temperature, Cloud Cover, Vapour Pressure and Wet Day Frequency.

Wet Day Freq	Max Temp	Cloud Cover	Vapour Pressure	**Precipitation**

Classification /Prediction: ii)Nearest Centroid Classification

- Training sample / object is represented by a vector having input attributes and an output attribute (classifier)
  - Vector having Four input attributes and One output (precipitation)
  - Major percentage of the vectors were put in the training set, while the rest were placed in the testing set
- Choose a value of k, signifying the number of groups/classes that should be there in the dataset.
  - Chose five, where 1 represents Very Low and 5 represents Excessive Rainfall
- Choose k random objects as the centroid for each of the k classes.
  - Chose five vectors randomly from 300 odd vectors. These would be the initial five centroids , representing each class.
- Calculate the distance between each of the samples/vectors to the k centroid vectors, and assign the samples to the group(centroid) which is nearest to the sample.
  - Chose Euclidean distance to measure the closeness of a vector to the centroid vector of a class

Classification / Prediction :
ii) Nearest Centroid Classification (Contd..)

- Iterate till the centroid value does not change for all the k classes. This signifies that the data elements have been allocated to their corresponding groups.
- The centroid represents the mean of all the vectors for that group. The classifier value is the mean of the output attribute value of all the vectors in that group.
- Each group is now populated with vectors who are closer to their group members than the other groups.
- The output value (Rainfall) would also fall in line !!!

Suggestion of Crop(s) :

- Once the rainfall is predicted, the value is mapped with Dataset B and Dataset C
- The Dataset B gave us the crops grown in a particular district
- The Dataset C gave us the minimum and maximum rainfall for the chosen crops.
- The predicted rainfall was compared with the min and max rainfall boundaries of the chosen crops. This led to the crop(s) that can be suggested to the farmer.

Evaluation of Classification :

- Vectors from the testing data, **without the output variable**, are now compared with each of the k centroid values.

Wet Day Freq	Max Temp	Cloud Cover	Vapour Pressure	**Acrual Precipitation**

Wet Day Freq	Max Temp	Cloud Cover	Vapour Pressure	

- The output attribute value of the nearest centroid is copied to the vector from the testing data. This is the predicted value.

Wet Day Freq	Max Temp	Cloud Cover	Vapour Pressure	**Predicted Precip**

- The actual value of the output variable (of the vector in testing data) is now compared with the predicted value.
  - If the actual value is beyond the range of the minimum and maximum precipitation of that class, then the model has failed in the prediction.

Results & Discussion

- The Dataset A (over 400 entries) was split into training data and testing data by random holdout method.
- By applying Nearest Centroid Classifier, the following accuracy values were obtained.

Training data / Testing Data	Accuracy
50% / 50%	79%
60% / 40%	84%
70% / 30%	81%
80% / 20%	81%
90% / 10%	83%

- The results were marginally better than the Naïve Baye’s and the kNN classifiers, as mentioned in the literature.
- A k fold cross validation method could have been tried also to generate a richer training set, resulting in better accuracy.

Conclusions

- Nearest Centroid Classifiers performed well for multidimensional applications.
- If archived data is available for all countries/regions, machine learning can be applied for the benefit of the people.

<u>Limitation of the study</u> :

- This work only provides the farmers with a set of suitable crops, it does not address the profitability issue.
- In addition, if the nature of the soil, soil reusability, soil fertility information would have been available , then this effort can be scaled up to predict an optimal rotation of crops along with profitability.

Thank You

Questions ?