**Final Teleprompter Script**

**Sravan Kumar Borra**

**Contribution of Competitor’s Article**

There are so many research papers which were already published based on this topic. All those papers are being used for predicting climatic changes and weather conditions. But this paper particularly concentrates on the results and models which will help to build a smart city. All the predictions and predictive models will help to build a smart city which will run based on internet and connectivity. There are so many papers that used techniques for predicting weather conditions especially machine learning techniques such as SVM, linear regression, neural networks etc.

SVM works on statistical learning and it is also known as a kernel machine. It is a machine learning algorithm used for classification purpose and regression purpose. If we have n observations in a data set, Support vector machines can be explained as the simple co-ordinates which is related to each and every observation. Support vector machines segregates very two important parts of a plot. They are hyper plane and line.

In an article named “An efficient approach for Weather forecasting using Support Vector Machines” used support vector machines technique to predict maximum and minimum temperatures for a particular day. This article has been published in International conference in 2012. In this paper we are taking inspiration from this particular technique to perform the predictive model. This forecasting helps in building the smart city efficiently and helps the people who live in those cities. By taking inspiration from this paper we are also going to use decision tree model and neural network model to predict the temperatures. As mentioned above we are dividing the data into two sets one is for training and other is for testing the data. This approach will help us in predicting accurate maximum and minimum temperatures.

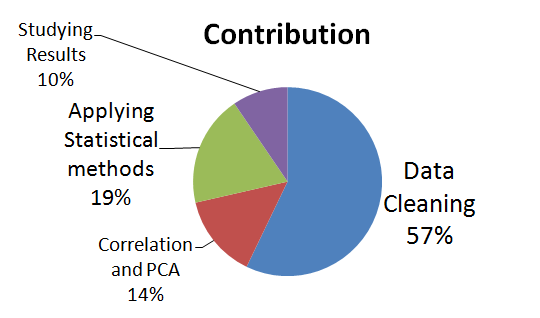
In competitor article they performed ARIMA and ANFIS statistical methods. We are using SVM and ANN. We out performed them as these statistical models gives better results. In the competitor article, they predicted only Maximum and minimum temperature. We are predicting Max and Min temperature, wind speed on a particular day, dew point at a particular time. The tool used in competitor article is SPSS software. We used Zeppelin in our project. We again outperformed in the selection of tools as Zeppelin if far more efficient than SPSS. For excluding null values their code is lengthy and not accurate.

**Description of Your Contribution**

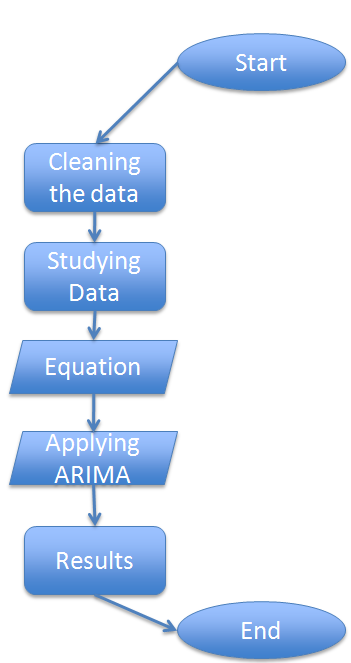
Smart city mostly runs on connectivity. This connectivity is established by using big data. We are using big data techniques, machine learning techniques, data mining techniques to build a model that will be used for the prediction of weather which helps in building a smart city. There are various machine learning techniques that are used for predicting the weather. We can use Support vector machines method; Decision trees techniques, neural networks.

All these methods are used in machine learning for predicting using the given data set. SVM works on statistical learning and it is also known as a kernel machine. It is a machine learning algorithm used for classification purpose and regression purpose. If we have n observations in a data set, Support vector machines can be explained as the simple co-ordinates which is related to each and every observation. Support vector machines segregates very two important parts of a plot. They are hyper plane and line. For this particular paper, we selected weather data of Aarhus city located in Denmark to predict the climatic changes. This prediction model helps in building a Smart city which runs on connectivity.

The techniques mentioned above are used in machine learning. For using those techniques we need big data set. To perform those machine learning techniques we divide the data set into two parts. One part is used for testing data where the machine learning technique is applied to the data set. Using the prediction results obtained from training set we will run a prediction on the remaining data set which is known as test data set. We already know the results for the test data set, so if the predicted results match with the original set of results then we can use our model to predict the weather for upcoming years.



We can predict weather or climatic changes for quarterly, yearly, seasonally, daily time periods for a year. We can compare our results after testing the data set with the original data set. This comparison is done by using a set of topics or metrics such as MAE and MSE. MAE is known as Mean absolute error and MSE is known as Mean square error. Among the machine learning techniques, the predictions done by using SVM (Support Vector Machines) technique turns out to be more accurate and reliable. We are comparing our predicted results with the results obtained from multiple linear regression models. It is a machine learning technique which is used for every type of prediction for comparing the results. Forecasting accurate weather conditions can also help in transportation. For flights and airplanes which are flying mostly in air perfect weather conditions are must and important. If we forecast the weather in advance it helps the aviation department to plan according and it also helps in damage control.



**Data Source and Content**

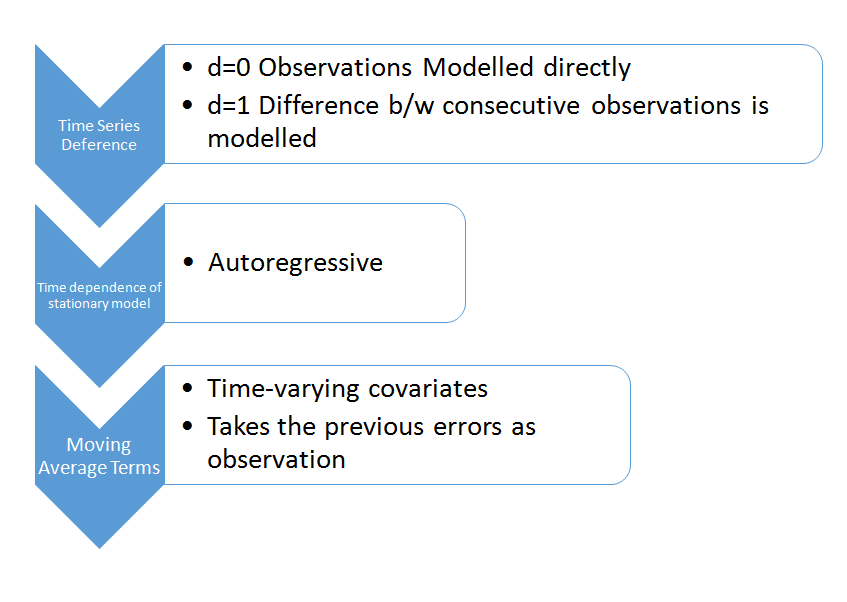
A data set with weather observations has been selected for this project. Data set observations are collected from the city of Aarhus in Denmark. This data set has been fetched form City pulse website, which is trying to provide models to build smart city. Data set has different types of fields such as Dew point, humidity, pressure, temperature, wind direction, wind speed. All these parameters are used for predicting weather conditions in different times of the year. We have two sets of data. One is from February-June 2014 and other set is from August-September 2014. These data sets were provided in JSON file format. We use Zeppelin platform to study the data. We initially load the data and study about the different fields in the data.

This connectivity is established by using big data. We are using big data techniques, machine learning techniques, data mining techniques to build a model that will be used for the prediction of weather which helps in building a smart city. There are various machine learning techniques that are used for predicting the weather.

**Your Method**

We are using big data techniques, machine learning techniques, data mining techniques to build a model that will be used for the prediction of weather which helps in building a smart city. There are various machine learning techniques that are used for predicting the weather. We can use Support vector machines method; Decision trees techniques, neural networks. All these methods are used in machine learning for predicting using the given data set. For this particular paper, we selected weather data of Aarhus city located in Denmark to predict the climatic changes. This prediction model helps in building a Smart city which runs on connectivity.

Artificial Neural Networks and Support vector Machines are the methods used in Data mining and Machine learning. We are going to use those techniques to predict the maximum and minimum temperatures for a particular day, for a particular hour a day. We are also going to predict the humidity scale for a particular day. We are going to predict the wind speed and direction of a particular day.

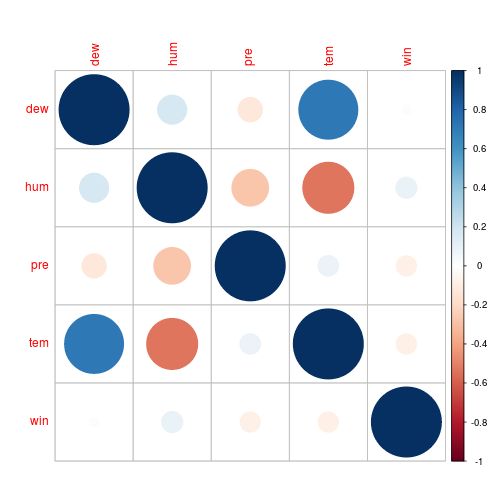


* **Method**: Time series Analysis Using ARIMA.
* Parameters used for comparing ARIMA and ANFIS
  + Mean Absolute Error (MAE)
  + Root Mean Square Error (RMSE)
  + R-square error and the Sum of Square Error (SSE).
* Parameters used for estimating ARIMA
  + Estimated error and Standard Error
  + ARIMA(2,0,1)
  + AR1, AR2, MA

We have two sets of data. We use one of the data set to train the model and use another set to test the model and we can also test the model by predicting for upcoming days. Data Cleaning, implementation of methods on data set are going to be performed in Zeppelin platform. We are going to use Python language in Zeppelin to clean the data and run the methods on data set.

The techniques mentioned above are used in machine learning. For using those techniques we need big data set. To perform those machine learning techniques we divide the data set into two parts. One part is used for testing data where the machine learning technique is applied to the data set. Using the prediction results obtained from training set we will run a prediction on the remaining data set which is known as test data set. We already know the results for the test data set, so if the predicted results match with the original set of results then we can use our model to predict the weather for upcoming years.

**Quantitative Results**

****

We had the correlation results based on the experiments done by using zeppelin. I performed precision correlation method to find the correlation between the variables. It shows that Temperature and density highest percentage of correlation. Our aim in this project is to provide an efficient model which helps to build a Smart city. Smart city mostly runs on connectivity. This connectivity is established by using big data. We are using big data techniques, machine learning techniques, data mining techniques to build a model that will be used for the prediction of weather which helps in building a smart city. There are various machine learning techniques that are used for predicting the weather. We can use Support vector machines method; Decision trees techniques, neural networks.

Correlation Table:

|  |  |  |
| --- | --- | --- |
| **Temperature** | **Dew Point** | **71%** |
| **Temperature** | **Humidity** | **49.7%** |
| **Pressure** | **Humidity** | **36.2%** |

Above results explains the comparison of results between the two methods we have applied on our data set. We have applied Support vector machines and ARIMA on our data set. We have compared the results for SSE, R^2, RMSE and MAE. By comparing the table we can observe that the value of SSE is better when compared to the SSE value of ARIMA method. This explains that the method we chose is better than the method applied by the competitor. When we compare the R^2 values of SVM and ARIMA the results are better with the SVM and here also the method applied by us is better than the competitors method. If we compare RMSE and MAE values we can observe that the values obtained by SVM method are better than the values obtained by ARIMA. By this we can conclude that SVM method results are much better than the results obtained by SVM.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **ar1** | **ar2** | **ma1** |
| **error** | **1.5** | **-0.5** | **0.86** |
| **Std. error** | **0.026** | **0.86** | **0.01** |

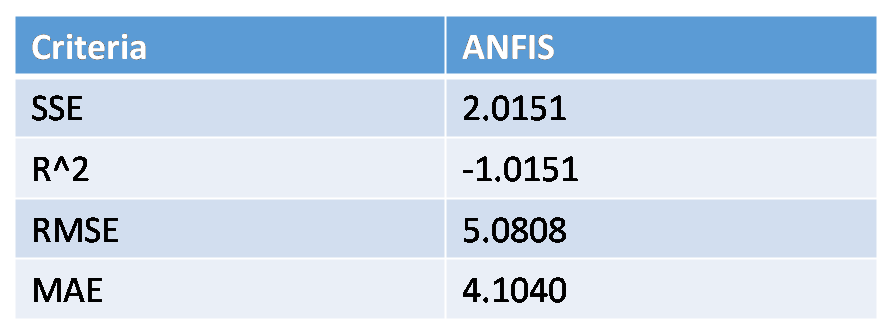
* + **Moving Average Terms -> ACF**
  + **Auto regressive terms -> PACF**
  + **SSE --- 0.213**
  + **RMSE --- 1.69**
  + **MAE --- 1.2**
  + **R^2 --- 0.78**

Above are the results obtained after applying time series analysis on the weather data set to obtain maximum and minimum temperatures. In below sections we are going to compare our results with the competitor’s article.

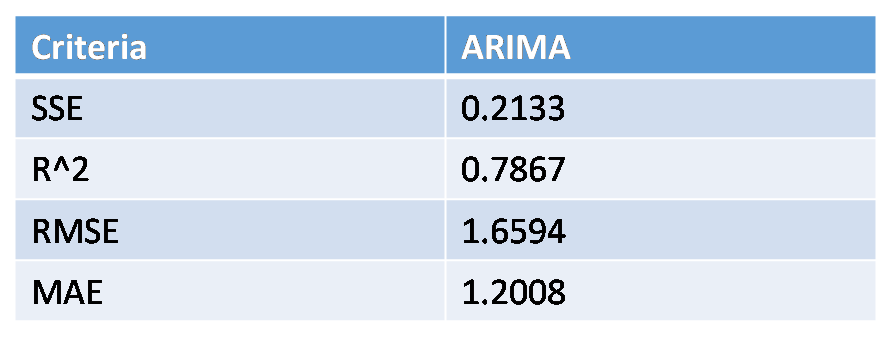
**Discussion: Comparison With Your Competitor**

We out performed them as these statistical models gives better results. In the competitor article, they predicted only Maximum and minimum temperature. We are predicting Max and Min temperature, wind speed on a particular day, dew point at a particular time. The tool used in competitor article is SPSS software. We used Zeppelin in our project. We again outperformed in the selection of tools as Zeppelin if far more efficient than SPSS. For excluding null values their code is lengthy and not accurate. It consumes more time. In our article, it took almost less than 10 lines of code to exclude null values as we are using spark. Pearson Correlation for temperature and dew point- more than 70%. Our data has less outlier as we excluded the null values before performing the statistical analysis. The regression value of temperature and humidity grouped by pressure is -7.500000 and the Intercept is 123.500. We still do not have the R^2 value for our data as we did not performed SVM on our data. The work is still going on. In the competitor article, people have not performed correlation on their data. So they don’t know the relation between their data variables. We performed correlation on our data. We have a 70% relation between temperature and dew point. With this result we can predict the temperature based on dew point variable.

Competitor results

****

My results

****

**Performance on Big Data: Time Measurements**

|  |  |  |
| --- | --- | --- |
| Process | Time Measurement | Work Done |
| Data Loading | 75 sec | Loading six data files into zeppelin |
| Data Cleaning | 60 sec | Converting JSON format files into DATETIME format files and removing outliers. |
| Applying Correlation and PCA | 30 sec | Finding correlation between the variables and dimensionality reduction using PCA. |
| ARIMA | 45 sec | Running time series analysis. ACF, PACF. |

**Conclusion**

As we can see, the result obtained by ARIMA has less error than ANFIS which has been applied by my competitor. The results are more than thirty percent accurate when compared to competitor article. Achieved two important points:

* I was able to prove ARIMA is better statistical method than ANFIS for this kind of weather forecasting.
* Tried to outperform competitor and achieved it.