A Data Mining Approach for Parameter Optimization in Weather Prediction

Thushika N[a,1,*], Premaratne S.[b,2]

[a] Department of Mathematics, Eastern University, Sri Lanka
[b] Department of Information and Technology, University of Moratuwa, Sri Lanka
[1] thushikav@esn.ac.lk; [2] samindap@uom.lk
* corresponding author

ARTICLE INFO

Article history
Received August 08, 2019
Revised March 29, 2020
Accepted April 02, 2020

Keywords:
data mining
weather prediction
weka
classification
SSE values

ABSTRACT

More than two decades, there is a number of weather-related websites are available which approximately predict the weather and climate. By extracting essential data from the websites, a predictive data pattern can be produced to show the next day's weather is with rain or not. By applying different types of web mining and analyzing techniques those extracted weather-related data can be visualized to a typical pattern for weather forecasting with the main deciding factors of weather. With the use of these approaches, reasonably precise forecasts can be made up to about four to five days in advance. For the weather prediction analysis, we need to discover deciding factors of the next day's weather. Particularly, common weather dependent factors and the relationship of the prediction to the particular phenomenon. The solution proposed by this research can be used to analyze a large amount of weather data which are in different forms in each source. By using predictive mining task our solution allows us to make predictions for future instances according to the model what we have created. Evaluation measurements for the selected data mining technique such as accuracy percentage, TP & FP Rate, Precision, F-Measure, ROC area, SSE, and loglikelihood for classification and clustering leads to create a high-quality model of prediction. Knowledge flow interface provides the data flow to show the processing and analyzing data with precise association rules. In order to evaluate the model, SSE values and time to build the model, are considered in an effective manner.

1. Introduction

Extracting necessary and needed information from web pages are such an important task in this tech-covered world. However, the tools available for gathering, establishing, and distributing web content have not kept step with the rapid growth in information. So, the key analysis needed when web documents are in precisely content related.

Nowadays, there is a number of weather-related websites which approximately predict the climate and weather. Both seasonal and regional variability in weather directly influences in many fields like agriculture, tourism, disaster management, aircraft and shipping. By considering temperature, rainfall, evaporation, wind direction, wind speed, humidity, and air pressure most of the weather predictions are predicted and displayed. We have conducted research to offer in-depth analysis of weather prediction patterns using extensive data which are extracted from websites.

In this modern world, web development and its applications play an exciting role in everyone’s day to day life. So, the world wide web is moving into a suitable atmosphere, where all kind of users can able to acquire important information quickly and easily as they need. Web mining is primarily about using the web, the web structure, and exploring web content. Web Structure Exploration attempts to discover the underlying model of web link structures. This template can be used to categorize web pages and is useful for generating information such as the similarity and relationship between different websites. Web usage mining using web effects usage derived data detect user behavior patterns to automatically access web services. Web content mining is a kind of mining technique, which mines extracted information from selected web page elements.

Web content crawling is the analysis and exploration of text, images, and graphics on a web page to determine the importance of the content for the search query. This analysis is after the cluster in Web and the results with the huge amount of information available on the World Wide Web. Content mining provides search engine results lists in order of relevance to the keywords in the query [1]. Exploring web content with associated methodologies to extract information. These are the extraction of unstructured content, organized mining extraction, semi-organized content extraction and multimedia extraction.

The rapid development processes in web content mining have been initiated from the past few years. There are two approaches mainly considered within the web substance mining: Agent-based approach and Database approach [2]. The first approach is to improve the search and filtering of information the second approach to modeling data in the field of processing and analysis of data mining applications.

The extracted models are used to classify web objects, to recommend URLs or documents; to extract keywords for use in information retrieval and to infer the structure of objects semi-structured or unstructured [3]. Partitioned, Hierarchical, Graph-Based, Probabilistic algorithms are additional categorized accounting to the clustering method used into the other category [4]. Various algorithms and techniques have been represented for clustering web pages based on data content of which are Hard and fuzzy algorithm of clustering web pages document tarries based on key-words inside web pages and Cosine likeness measure [5] and clustering web pages based on the behavioral models of the users [6], clustering of web pages based on link structure between them which is based on the textual information in links [7], clustering web documentaries using neural networks based on key-words inside the documentaries [8].

In the field of document clustering, k-means is a simple unsupervised learning platform. K centers are defined for every cluster that is newly formed. While dealing with all substrings or phrases within a page Yamamoto and Church [4] provided us with a method of computing term frequencies (tf) and document frequencies for a set of documents by using the concept of suffix array. Later within the same platform K.M Hammoudoet.al. [9] came up with the idea of a phrase-based document indexing model which is later named as Document Index Graph (DIG). This method incorporates the construction of phrase-based indexes for the available document set but in an incremental manner. Indexing phrases repeatedly within the graph using the DIG model had a bottleneck of space complexity thus the STC algorithm introduced by Etzioni et-al. [10].

Term-based document model has the drawback as it ignores the relationships among individual words that it works on and thus not efficient nowadays. DIG has also become slow with the results with large space complexity. STC being a better option for clustering web documents still need the help of other techniques too to enhance its feasibility [11]. Beyond this, detailed forecasts are less useful, since some atmospheric conditions such as wind gust direction, temperature and wind direction are very difficult. The capabilities of retrieving and storing have increased, due to the latest industrial updates which resulting in the accessibility of enormous climatology dataset in various arrangements [12]. These data are generated both from the surface observation stations and aerial study stations. With the increase in the number of weather stations, a huge amount of data is available on daily, weekly, monthly and yearly basis and the data is stored exponentially [14].

The objective of this research is to achieve a concept-based term analysis regarding the weather on the sentence and document levels rather than a single-term analysis in the document set. The quality of web document clustering can be enhanced by dropping the noise in the data by pre-processing the structure of data representation and also by applying different clustering techniques.

Our aim is to achieve concept-based content analysis for web documents related to weather using web mining techniques and intend to afford an analysis study with our results of weather dataset. Weather prediction analysis is highly important in climatology and many another day to day life activities. So, this research has been conducted to check the most suitable model for the particular dataset.

2. Material and Methods

In this process of weather analyzing with data mining techniques all standard steps in the KDD process which contains data selection to evaluation are carried out. Through-out the process, the data set is cleaned, formatted and prepared for mining and interpretation.

2.1. Data Selection

The weather analyzing can be made by gathering data about the current state of the atmosphere and using considerate of atmospheric processes to predict how the atmosphere will change. Before any forecast can be made, first it is mandatory to understand what the current weather conditions are and what is producing them. This is done by examining a large quantity of observation data. The first set of data extracted from the weather content websites using python web scraping using BeautifulSoup.

Following steps followed to extract actual data from websites.

- Download the web page containing the weather details
- Create a BeautifulSoup class to parse the page
- Find the relevant style of HTML and assign to the index
- Inside the index, find each individual weather item.
- Extract and print the first weather item.
- Combine the data into Pandas DataFrame and analyze it.

All data converted to CSV format in order to use and make analysis inside the Weka tool for preprocessing process.

2.2. Data Preprocessing

Our weather dataset contains 6588 instances and 20 attributes, in order to predict a pattern to find out the next day is a rainy or light shower or no rainy day. Noise inside the dataset should be removed using various preprocessing techniques. Always noise reduces the quality of the data. Filters should be applied to the dataset in a proper manner to get effect quality. Therefore, our dataset preprocessed before further analysis.

2.3. Data Mining

This is the essential part of our research which used intelligent methods in order to determine exact data patterns. Not only that but also, discovering interesting knowledge associations, information gain, changes, anomalies and significant structures from our weather-related dataset. From association rules, the relationship between the selected attributes to the deciding factors can be determined.

In the classification part, different types of classification methods applied to find which outfit well for our model. Cluster analysis also was done with various techniques to visualize and get output with accuracy, true-false rate. Prediction and pattern analysis shows similar patterns which help to identify grouped attributes. According to the task we tried to achieve through our objectives.

2.4. Evaluation

Evaluation of the research presented through graphs and time period which are the main two factors of analyzing. According to the goal, evaluation should be precisely categorized to maximize efficiency. Many visualization packages and knowledge flow interfaces are available, including trees and distribute networks.

3. Implementation

Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, as well as graphical user interfaces for easy access to all functions. We have selected Weka since it includes many tools for preprocessing, classification, clustering, association ruled, regression and visualization as options.

The 'weka.filters' package is attached with sub-classes that transform our dataset by ignoring or adding attributes, resampling the dataset, and removing samples. This package offers important support for data preprocessing, which is a necessary step in each machine learning research.

3.1. Data Collection and Preprocessing

The dataset was exported into CSV format in order to use inside the Weka tool for preprocessing process. The Weka suite contains a set of visualization tools and algorithms for data analysis and predictive modeling, as well as graphical user interfaces to facilitate access to this feature.

Since, Weka is a platform-independent Visualization of all attributes in the database can be shown after loading the dataset. Preprocessing is one the important process in the research which reduces the inaccuracy of the dataset. When the particular dataset is an effective one, then only we can get the best results from classification.

1) Replacing Missing Values: There were some tuples that have no recorded values for several attributes; then the missing values can be filled in for the attribute by different methods. Some of the null values were replaced with the maximum number of attribute option. The process was done with the use of filters.

2) Data Transformation: String values were changed to nominal featured attribute by applying the 'StingToBinary' filter to the dataset. By maintaining all values as nominal make the dataset more effective one according to the analysis.

3) Data Normalization: Normalization is one of the scaling techniques in the preprocessing stage of any problem statement. Where, we can find a new range from an existing range. This can be very useful for prediction or forecasting. So, the standardization technique is necessary to bring.

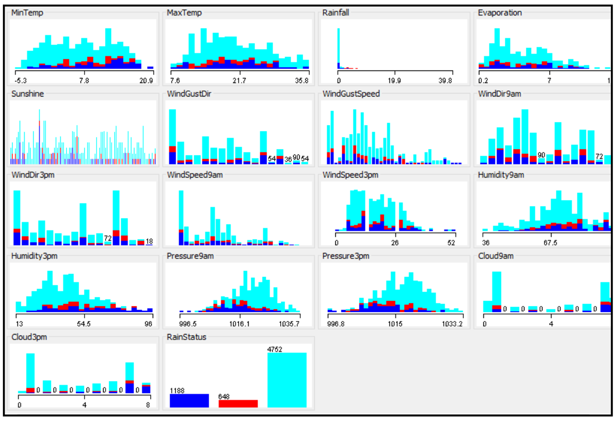


Fig 1. Visualization of attributes before preprocessing

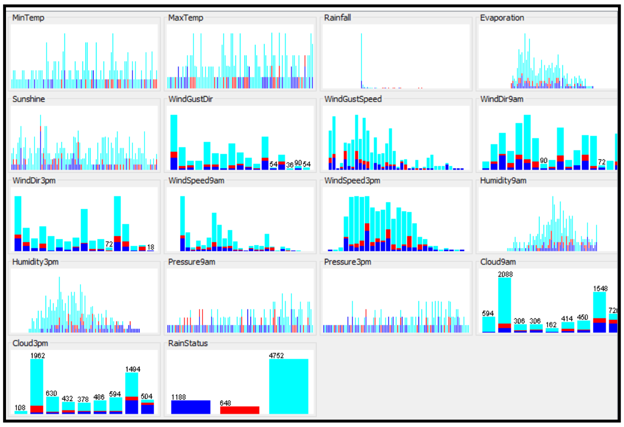


Fig 2. Visualization of attributes after preprocessing

Fig. 1 and Fig. 2 show the data distribution difference between the visualization of our selected attributes before and after the preprocessing respectively.

3.2. Weather Prediction Analysis

As the very first step, data exploration was done, where those 6588 instances dataset cleaned and transformed through several preprocessing techniques mentioned above. Significant factors and its dependent data were determined through the visualization patterns. Classification, Clustering, Association Rules, Decision Trees, Nearest Neighbour methods are some of the most important data mining techniques.

1) Using Suitable Classifier for Classification: From the literature review we came up to the conclusion that we can't define a particular classification algorithm is the best one to apply. The efficiency of the classification purely depends on the selected dataset. Naïve Bayes is a simple classifier based on the Bayes theorem. It is a statistical classifier which performs probabilistic prediction.

For Naïve Bayes classification, the following equation is used;

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)} \tag{1}$$

From the above equation (1), the idea of Bayes' rule is that the outcome of a hypothesis or an event (H) can be predicted based on same evidences (E) that can be observed from Bayes rule [13].Through many types of research, it was proven that Naive Bayesian classifier frequently used because of its simplicity which performs fantastically for very large datasets. Since our dataset is also considerably large, we went through the classification model of Naïve Bayesian.

J48 is one of the decision tree algorithm. Java implementation of the C4.5 algorithm creates a decision tree which can be used for classification based the value which are presented on dataset [13]. J48 also shown the more accurate result, but we cannot able to visualize the tree- since we dealt with 19 attributes. There are three classes, 'Yes': there will be rain, 'LightRainShower', 'NoRain'.

2) Applying Appropriate Clustering Method : Clustering in a sense grouping such objects in particular or similar group related to each other. Clustering methods can be seen as divided into two types: hierarchical methods and partitioning methods. Partitioning methods relocate instances by moving them one cluster to another, starting from an initial partitioning. Those methods generally require that the count of clusters should be pre-set by the user. For this research finally, we have compared two algorithms belonged to portioning methods which are more relevant and popular for this data set.

3.2.1. Simple K-Mean Algorithm

In K-Mean clustering, k number of centers, one for each cluster. Each center should be as far as possible from each. Then each point belonging to a given dataset and associated it to the nearest center. Initial step completed when there are no pending points. After that, k new centroids are calculated again as center taken from the previous step. Now there are k new centroids, it has to be related between same dataset values and the nearest new center. Iteratively same procedures are carried out. The distance function between three points, $a = (x_1, y_1, z_1)$ and $b = (x_2, y_2, z_2)$ is defined as:

$$\rho(a, b) = |x_2 - x_1| + |y_2 - y_1| + |z_2 - z_1| \qquad (2)$$

Based on the distance function, as given in the equation (2), three clusters from the dataset obtained [16]. From the final result of these iteration, considerable change in k centers' location are observed and it will be continued until no more changes are done. When measuring cluster validity, several numerical measures are applied to keep track on various aspects of cluster validity. As an internal index, Sum of Squared Error (SSE) is displayed there. SSE is used to measure the effectiveness of a clustering structure without respect to external information. We choose three clusters where SSE is pretty high but by considering the sequence order of time taken and small changes in SSE value. We have identified three clusters, the first cluster contain 'yes' in a sense heavy rain is expected for the next day. Expecting 'light rain shower' belong to the second cluster and 'No rain' is considered to be the third cluster. With the visualization of the cluster, we can understand cluster assignments.

Table 1. Cluster instances in k-means clustering

Cluster Type	No. of Instances	Percentage
Cluster0: "No Rain"	3042	46%
Cluster1:"Light Rain Shower"	1854	28%
Cluster 2: "Yes"	1692	26%

3.2.2. Simple Expection-Maximization Algorithm

Expectation-Maximization (EM) algorithm is an iterative method for discovering maximum likelihood. According to the algorithm library, Gaussian distributions are usually modeled which are initialized randomly and whose parameters are iteratively optimized to fit better to the dataset.
This algorithm determines the maximum likelihood estimate of the marginal likelihood by the equation (3), where ϴ is the parameter vector, x and z represents observed and unobserved data values. Expected value of loglikelihood function, with respect to the conditional distribution of z given x under the current estimate of the parameters $\Theta^{(t)}$ [16]:

$$Q(\theta|\theta^{(t)}) = E_{Z|x,\theta^{(t)}}[\log(L(\theta; x, Z)] \qquad (3)$$

From the EM algorithm, we have identified three clusters. Through the likelihood values, there are significant values changes are not seen after cluster three. The first cluster contains 'yes' in a sense heavy rain is expected for the next day. Expecting 'light rain shower' belong to the second cluster and 'No rain' is considered to be the third cluster.

Table 2. Cluster instances in EM clustering

Cluster Type	No. of Instances	Percentage
Cluster0: “No Rain”	2790	42%
Cluster1:“Light Rain Shower”	1602	24%
Cluster 2: “Yes”	2196	33%

3.2.3. Clustering Using Density-Based Algorithm

The major importance of the density-based clustering algorithm is discovering nonlinear pattern structure based on the density. The influence of each data point can be formally modelled using a mathematical function, called an influence function. An influence function is a mathematical function typically used to model each datapoints which are comparatively influencing a lot. This influence function defines the neighbourhood data point impacts. The basic influence function of data object can be written as follows:

$$f_B^y\,(x) = f_B\,(x, y) \tag{4}$$

Equation (4) describes x and y be the points in a d-dimensional input parameters. Classifying the clusters and distribution patterns are that main goal of this particular method. This method can be used for finding clusters of arbitrary shape which are not necessarily convex.

Table 3. Cluster instances in density-based clustering

Cluster Type	No. of Instances	Percentage
Cluster0: “No Rain”	1980	30%
Cluster1:“Light Rain Shower”	2358	36%
Cluster 2: “Yes”	2250	34%

From the Density Based algorithm, we have identified three clusters. Through the likelihood values, there are significant values changes are not seen after cluster three. The first cluster contains ‘yes’ in a sense heavy rain is expected for the next day. Expecting ‘light rain shower’ belong to the second cluster where small amount of rain shower expected and ‘No rain’ is considered to be the third cluster which is not expected for rain on that day.

4. Results and Discussion

4.1. Data Collection and Preprocessing

One of the important outputs from the classifier is a confusion matrix. By going through the confusion matrix, we can find a number of evaluation factors such as precision accuracy, and recall to evaluate data mining classifiers. These measurements and their definition are given in the Table IV.

Table 4. Evaluation measurements for classifier

Measure	Meaning	Relevant Formula
Precision	Percentage of positive predictions which are correct	TP/ (TP+FP)
Recall	Percentage of positive labeled instances	TP/ (TP+FN)

	that are predicted as positive	
Accuracy	Percentage of predictions which are correct	(TP+TN)/ (TP+TN+FP+FN)

Classifiers provide few sets of measures named, TP rate, FP rate, ROC area, and F-measure. TP rate is equal to sensitivity, while FP rate equal to one minus specificity. F-measure calculated by precision and recall. The overall ability of the test to distinguish between usefulness and uselessness can be quantified by the ROC curve area. A truly useless test has an area of 0.5. A perfect test has an area of 1.00. Usually better models are having higher TP rate, lower FP rate and ROC space close to 1.00. Comparison of the confusion matrixes and weighted averages in the classification model used for a rainy day: “Yes” are given in the following Table V.

Table 5. Comparission of different classification for rainy day

Technique	NavieBayes	J48
TP Rate	0.723	0.987
FP Rate	0.034	0.000
Precision	0.823	1.000
F-Measure	0.777	0.994
ROC Area	0.956	0.999
Recall	0.723	0.987

Table VI shows the comparison of two classifiers for a light rain shower day. Here we can able to see that, J48 shows the TP rate is almost 1, where the model classifies almost more effective. But the drawback here is we can’t be able to visualize the tree as it is a big data set with more branches and associations.

Table 6. Comparission of different classification for light rain shower day

Technique	NavieBayes	J48
TP Rate	0.935	1.000
FP Rate	0.021	0.000
Precision	0.832	1.000
F-Measure	0.981	1.000
ROC Area	0.994	0.999
Recall	0.935	1.000

Table VII describes the comparison of different methods of classification for no rainy day. Here also, the J48 FP rate shows 0.008, which means a very less false rate leads to an effective classification method. But comparatively Naïve Bayes classification also high in TP rate.

Table 7. Comparission of different classification for no rainy day

Technique	NavieBayes	J48
TP Rate	0.956	1.000

FP Rate	0.150	0.008
Precision	0.943	0.997
F-Measure	0.949	0.998
ROC Area	0.972	0.999
Recall	0.956	1.000

4.2. Evaluation for Clustering

In order to evaluate the model, SSE values and time to build the model, are considered in the clustering. To evaluate the accuracy of the data model, datasets deployed in Weka tool and clustering algorithms are applied to the dataset with classes to cluster evaluation option. Table VIII shows the within cluster SSE values for a different number of clusters used in KMean algorithm with Euclidean Distance function. Moreover, Fig. 3 shows the decrement values SSE for various clusters. We have selected the seed value as -2 which made comparatively low SSE values among others.

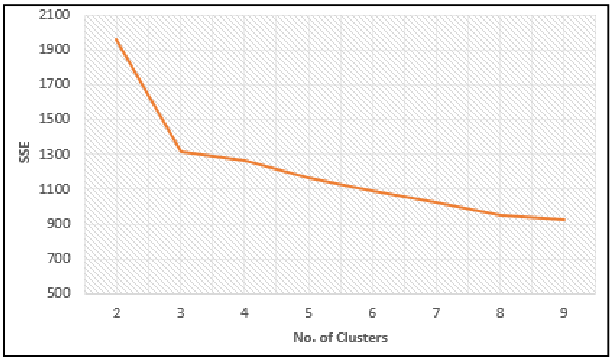


Fig 3. Graph of SSE within clusters versus different clusters

Intended for the evaluation of EM clustering, a number of clusters analyzed to find loglikelihood value-effective time period. If loglikelihood of sample is greater under one model that other, we tend to infer that the former model is more likely than later. Table IX shows the variations of loglikelihood with a number of clusters. And Fig. 4 shows graph flow among these two factors.

Table X indicates the time taken for K-Means, EM, Density-Based algorithms. From this analysis it is obvious that K-Means algorithm takes minimum time to make clusters in comparison with other clustering algorithms. Hence, time is more effective for K-means clustering algorithm.

Table 8. Variation of SSE within cluster for different clusters

No. of clusters	SSE
2	1958.2636
3	1311.10083
4	1269.08451
5	1167.59455
6	1092.65681
7	1021.33567
8	953.76428
9	923.93444

Table 9. Variation of SSE within cluster for different clusters

No. of clusters	Log-likelihood
2	-55.90600
3	-51.49529
4	-52.49095
5	-50.38657
6	-51.61624
7	-51.21999
8	-48.56916
9	-48.86482

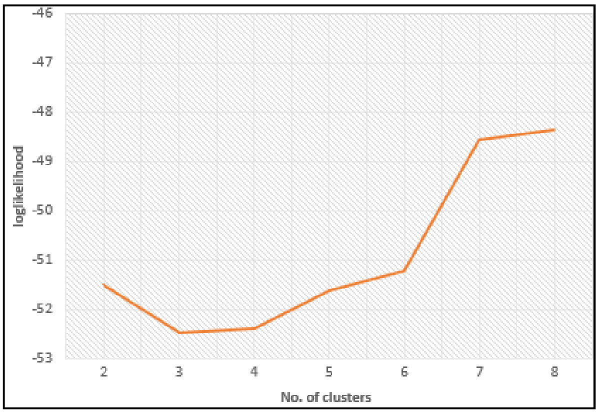


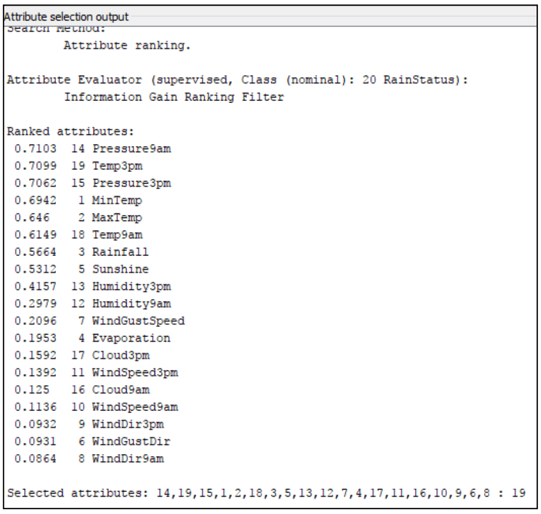
Fig 4. Graph of log-likelihood versus different clusters

Table 10. Time taken for clustering by different clustering algorithm

Clustering Algorithm	Time (sec.)
K-means	0.19
EM	4.21
Density-Based	0.29

4.3. Evaluation for Clustering

Most necessary part of attribute evaluator is information gain. Through this evaluation how much attribute information gives about the class can be evaluated. Perfectly partition always produces a higher rate of information. Information gain for attributes from weather data can be obtained from the information gained before the split and after the split. InfoGainAttributeEval attribute evaluator is used with full training set. Information gain for our dataset is shown in the above Fig. 5.

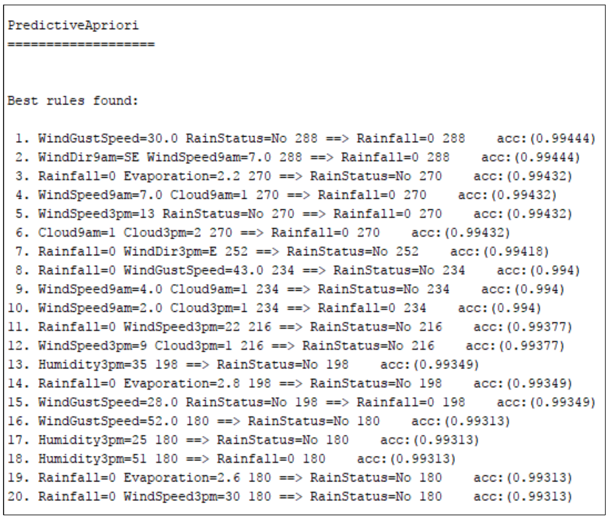


```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 20 RainStatus):
        Information Gain Ranking Filter

Ranked attributes:
 0.7103  14 Pressure9am
 0.7099  19 Temp3pm
 0.7062  15 Pressure3pm
 0.6942   1 MinTemp
 0.646    2 MaxTemp
 0.6149  18 Temp9am
 0.5664   3 Rainfall
 0.5312   5 Sunshine
 0.4157  13 Humidity3pm
 0.2979  12 Humidity9am
 0.2096   7 WindGustSpeed
 0.1953   4 Evaporation
 0.1592  17 Cloud3pm
 0.1392  11 WindSpeed3pm
 0.125   16 Cloud9am
 0.1136  10 WindSpeed9am
 0.0932   9 WindDir3pm
 0.0931   6 WindGustDir
 0.0864   8 WindDir9am

Selected attributes: 14,19,15,1,2,18,3,5,13,12,7,4,17,11,16,10,9,6,8 : 19
```

Fig 5. Information gain evaluation

4.4. Evaluation for Clustering

Association rules look for the whole sets of items that have support larger than the minimum support, and then use huge sets of items to generate the desired rules that have confidence greater than minimum confidence. To find association rules, we applied predictive apriori algorithm.



```
PredictiveApriori
===================


Best rules found:

 1. WindGustSpeed=30.0 RainStatus=No 288 ==> Rainfall=0 288    acc:(0.99444)
 2. WindDir9am=SE WindSpeed9am=7.0 288 ==> Rainfall=0 288    acc:(0.99444)
 3. Rainfall=0 Evaporation=2.2 270 ==> RainStatus=No 270    acc:(0.99432)
 4. WindSpeed9am=7.0 Cloud9am=1 270 ==> Rainfall=0 270    acc:(0.99432)
 5. WindSpeed3pm=13 RainStatus=No 270 ==> Rainfall=0 270    acc:(0.99432)
 6. Cloud9am=1 Cloud3pm=2 270 ==> Rainfall=0 270    acc:(0.99432)
 7. Rainfall=0 WindDir3pm=E 252 ==> RainStatus=No 252    acc:(0.99418)
 8. Rainfall=0 WindGustSpeed=43.0 234 ==> RainStatus=No 234    acc:(0.994)
 9. WindSpeed9am=4.0 Cloud9am=1 234 ==> RainStatus=No 234    acc:(0.994)
10. WindSpeed9am=2.0 Cloud3pm=1 234 ==> Rainfall=0 234    acc:(0.994)
11. Rainfall=0 WindSpeed3pm=22 216 ==> RainStatus=No 216    acc:(0.99377)
12. WindSpeed3pm=9 Cloud3pm=1 216 ==> RainStatus=No 216    acc:(0.99377)
13. Humidity3pm=35 198 ==> RainStatus=No 198    acc:(0.99349)
14. Rainfall=0 Evaporation=2.8 198 ==> RainStatus=No 198    acc:(0.99349)
15. WindGustSpeed=28.0 RainStatus=No 198 ==> Rainfall=0 198    acc:(0.99349)
16. WindGustSpeed=52.0 180 ==> RainStatus=No 180    acc:(0.99313)
17. Humidity3pm=25 180 ==> RainStatus=No 180    acc:(0.99313)
18. Humidity3pm=51 180 ==> Rainfall=0 180    acc:(0.99313)
19. Rainfall=0 Evaporation=2.6 180 ==> RainStatus=No 180    acc:(0.99313)
20. Rainfall=0 WindSpeed3pm=30 180 ==> RainStatus=No 180    acc:(0.99313)
```

Fig 6. Rules generated from Predictive Apriori algorithm

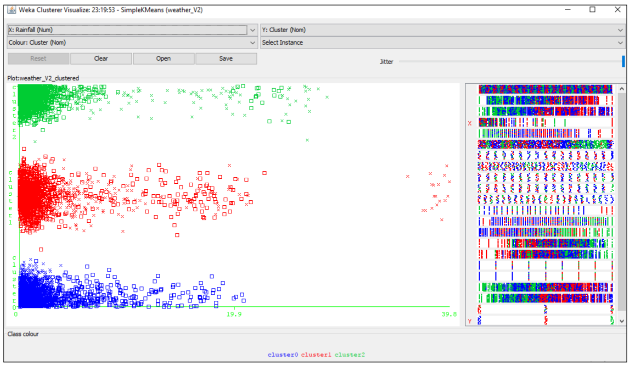
Because, comparison of apriori & predictive apriori associators study gives that, predictive apriori is ended with higher accuracy[15]. Fig. 6 shows 20 associate rules with accuracy values, which are generated from our dataset.

5. Conclusion

From the analysis of weather predictions, we found the major deciding factors of the rain prediction. Since there are many factors need to be considered for analyzing, according to the dataset extracted and generated we ignored some attributes to reduce the sum of squared errors. By ignoring several attributes incorrectly clustered attributes' count reduced tremendously. Around the overall world climate may differ with its geographical factors. Therefore common factors are considered and clustered. One sample attribute visualization with cluster variation is given in Fig 7.

Major deciding factors of the weather prediction are listed down by ignoring attributes sequentially compared with SSE values. Fig. 8 shows the reduction of SSE values with the selected attributes.

The overall least SSE error we got is considerably little high, it means incorrectly clustered rate is high. If it can be reduced the value of SSE, then the output will be more accurate to visualize the pattern. Accuracy of the tool highly depends on the dataset.



Weka Clusterer Visualize: 23:19:53 - SimpleKMeans (weather_V2)

X: Rainfall (Num)

Y: Cluster (Nom)

Colour: Cluster (Nom)

Select Instance

Plot:weather_V2_clustered

Fig 7. Cluster visualization with Rainfall factor

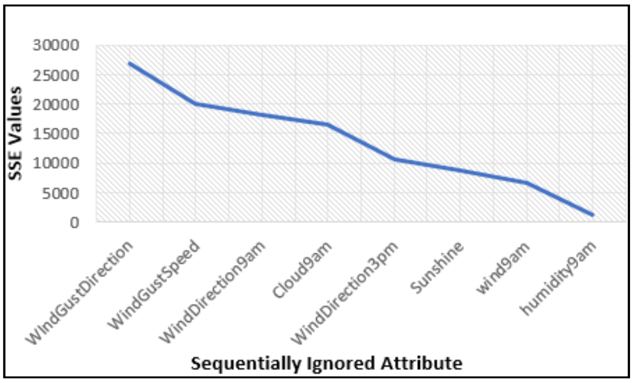


Fig 8. Graph of SSE values vs Sequentially ignored attributes

Our analysis reveals the significance of temperature and moisture among the influences such as Rain Fall, Evaporation, Pressure, Cloud status, and late-night Humidity. And also, Wind Gust Direction, Wind Direction, Minimum Temperature and Wind Gust Speed are low dependent factors in the rain prediction for the next day. These important patterns recognized from our research that can be used to model a climate pattern in a particular area.

References

[1] B. Rajdeepa and Dr. P. Sumathi, “An Analysis of Web Mining and its types besides Comparison of Link Mining Algorithms in addition to its specifications,” International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 3, issue 1, Jan. 2014.

[2] Kamlesh Patidar, Preetesh Purohit and Kapil Sharma, “Web Content Mining Using Database Approach and Multilevel Data Tracking Methodology for Digital Library,” Medicaps Institute of Technology & Management IJCST, Vol. 2, Issue 1, March 2011.

[3] M. Baglioni; U. Ferrara; A. Romei; S. Ruggieri; F. Turini, “Preprocessing and Mining Web Log Data for Web Personalization,” LNCS, vol. 2829, pp. 237–249. Springer, Heidelberg (2003).

[4] Michael Azmy, “Web Content Mining Research: A Survey”, DRAFT Version 1, - Nov. 2005.

[5] Menahem Friedman, Mark Last, Yaniv Makover and Abraham Kandel, “Anomaly Detection in web documents using Crisp and fuzzy-based cosine clustering methodology,” Information sciences 177, pp. 467-475 (2007).

[6] Qinbao Song and Martin Shepperd, “Mining web browsing patterns for E-commerce,”, Computers in Industry 57, pp. 622-630 (2006).

[7] Xiaofeng He, Hongyuan Zha, Chris H.Q. Ding and Horst D. Simon, “Web Document Clustering Using Hyperlink Structures,” Computational Statistics & Data Analysis 41, pp. 19-45 (2002).

[8] M. Shamim Khan and Sebastian W. Khor, “Web clustering Using a hybrid neural network,” Applied Soft Computing 4, pp.423-432 (2004).

[9] K.M. Hammouda and M.S. Kamel, “Effective Pharse-Based Document Indexing for Web Document Clustering,” IEEE Trans.Knowledge and Data Eng., vol. 16, no. 10, pp. 1279-1296, Oct. 2004.

[10] O.Zamir and O.Etziono, “Web Document Clustering: A feasibility Demonstration,” Proc. Third Int’l Conf. Research and Development in Information Retrieval (SIGIR),1998.

[11] S. K. Sahu and S. Srivastava, “Review of Web Document Clustering Algorithms,” Third Int’l Conf. IEEE Computing for Sustainable Global Development (INDIACom), 2016.

[12] Y. W. Dou, L. Lu, X. Liu and Daiping Zhang, “Meteorological Data Storage and Management System”, Computer Systems & Applications, vol. 20, no.7, (2011) July, pp. 116-12.

[13] Seif, H. (2016). Naïve Bayes and J48 Classification Algorithms on Swahili Tweets: Perfomance Evaluation. International Journal of Computer Science and Information Security (IJCSIS), 14(1).

[14] Kamber, M. and Pei, J. (2006). Data Mining:concepts and techniques. 2nd ed. heidellberg london: Morgan Kaufmann.

[15] Bharati, M. and Ramageri, A. (2013). Data Mining techniques and applications. Indian Journal of Computer Science and Engineering, 1(4), pp.301-305.

[16] Vanitha, K. and Roch Libia Rani, G. (2010). Analysis of Classification and Clustering Algorithms using Weka For Banking Data. International Journal of Advanced Research in Computer Science, 1(4).