Detection of Lightning Pattern Changes Using Machine Learning Algorithms

Aimée Booysens
School of Mathematics, Statistics
and Computer Science
University of KwaZulu-Natal
Durban, South Africa
210501411@stu.ukzn.ac.za

Serestina Viriri
School of Mathematics, Statistics
and Computer Science
University of KwaZulu-Natal
Durban, South Africa
viriris@ukzn.ac.za

***Abstract*—The distribution of lightning across the Earth's surface varies both with location and time. Seasonal changes in lightning activity recorded in Low Earth Orbit (LEO) satellite data have been studied by various authors, who used classical time series analysis techniques. This paper presents an alternative analysis based on automated pattern recognition, which identifies the changing state of lightning distributions using computer vision techniques. Due to the large quantity of data available, machine learning algorithms were used to predict the different lightning distribution patterns. The machine learning algorithm used to predict the seasonal patterns was K-Means which achieve a 64% global pattern detection rate. The Decision Tree machine learning algorithm was used to predict the spatial patterns this achieved 97% global pattern detection rate. The machine learning algorithm used to predict the time patterns were Naïve Bayes which achieved 31% global pattern detection rate. This system achieved a 63% global pattern detection rate for the larger dataset and for small dataset it achieved a 73% global pattern detection rate. This model not only has significant application in the analysis of historical lightning data but also helps in the forecasting of future lightning distributions.**

I. Introduction

Lightning is an important and critical weather phenomenon to study, as it affects people and the environment directly and indirectly. Lightning can occur in many different forms - these are inter-cloud, intra-cloud, cloud-to-air and cloud-to-ground flashes [1], [2]. Inter-cloud occurs when lightning flashes between two different clouds and intra-cloud occurs when lightning flashes inside the same cloud. Cloud-to-air occurs when lightning flashes are discharged into the sky and strike nothing around it and cloud-to-ground occurs when lightning flashes are discharged and the flash hits the earth's surface.

Quantification and mapping of lightning has become a very important area of research as it is beneficial to physicists trying to understand the phenomena of lightning. Along with meteorologists who can use lightning to understand climate change and severe weather storms. It is also of benefit to engineers and economists who use this information when planning buildings and infrastructure.

In this paper data from the Lightning Imaging Sensor (LIS) data aboard the Tropical Rainfall Measuring Mission (TRMM) [3], [4], [5], [6] is analysed with the aim of investigating global distribution patterns of lightning for the four year period from May 2002 to December 2006.

The rest of the paper consists of *Section 2* Related Works, *Section 3* Methodology and Design, *Section 4* Results and Discussion and lastly *Section 5* will be the conclusion.

II. Related Works

There is minimal information available on these types of systems but the following papers described their approaches thoroughly:

Finke and Hauf [7] used Lightning Position and Tracking System (LPATS). LPATS system is used to detect cloud-to-ground lightning discharges, by taking features from this type of lightning to reject or accept whether the lightning occurred - this system had 70% detection efficiency but this was only small datasets.

LPATS has a built in algorithm that estimates the peak current amplitude of the return stroke from the measured signal strength and calculated location. The algorithm used to calculate the peak current amplitude is defined in equation (1):

$$f(t,\vec{r}) = \sum_{i}^{N} \int (t - t_i) \int (\vec{r} - \vec{r_i}) \tag{1}$$

where the t_i and $\vec{r}$ denotes the time and location of the i^{th} detection. The Dirac function $\int (t - t_i)$ and $\int (\vec{r} - \vec{r_i})$ are used in their convectional meaning. By applying these integrals they used statistical analysis to see how the lightning changed over time and area.

One type of analysis done was the number of lightning events which occurred during the time interval $\triangle t$ around the central time t and inside the area $\triangle F$ around the location $\vec{r}$ were integrated, as defined in equation (2).

$$N(t, \vec{r}, \triangle t, \triangle F) = \int f(t + \tau, \vec{r} + \vec{p}) d\vec{p} d\tau \tag{2}$$

However, the results found by this system have been an estimate due to the fact that this type of detection is dependent on the location of the lightning to the location of the detection device.

Cannons and Kinsner [8] used an Axisymmetric Numerical Cloud Model (ANCM), this is an electrification model which represents a thundercloud in cylindrical axisymmetric form. This system uses equations to obtain temporal and spatial distributions. After calculating this, the Poisson's equation, as defined in equation (3), is used to calculate electrical capability, which is then used to model lightning discharges. A major problem is that the models created are large and complicated, and have difficulty in displaying the chaotic patterns of lightning discharges and this system had 60% detection efficiency.

$$f(x) = \frac{\lambda^x . e^{-\lambda}}{x!} \tag{3}$$

Dlamini [5] uses the Lightning Imaging Sensor (LIS) aboard the Tropical Rainfall Measuring Mission (TRMM) to obtain data over the area of Swaziland. Dlamini uses ArcGIS to ascertain if there is a link between varying elevations in Swaziland to the lightning that has occurred. This was done by using the machine learning algorithm of the Nearest Neighbor which is a ratio of the observed distance to the expected distance.The results that the Nearest Neighbor produced were 75% detection efficiency but this machine learning algorithm could only be used for small datasets.

III. Methodology and Design

The data processing system reads in various file types, of hdf4 to hd5. The data is then processed and analysed using machine learning algorithm, the resultant information enable the user to obtain a variety of graphs, data and maps, this is depicted in Figure 1.

A. Dataset

This paper uses data from the satellite- based Lightning Imaging Sensor (LIS) aboard the Tropical Rainfall Measuring Mission (TRMM). It was launched in November 1997 as a joint United States and Japanese mission aimed at understanding the global energy and water cycles by providing distributions of precipitation and the associated thermodynamics over the tropics[4], [5], [6]. The LIS TRMM measures all forms of lightning, inter-cloud, intra-cloud, cloud-to-air and cloud-to-ground flashes, with a high detection efficiency of every 90 seconds for both day and night conditions.

B. Data Manipulation

The data that was read into the system by the R Programming Language had to be manipulated as some variables such as location and time were in a different format. Which resulted in difficulty to produce any information.

The location values were separated by a comma to represent the Latitudinal and Longitudinal values. Due to this it had to be separated into different columns using the split function in the R Programming Language.

The time values were recorded in International Atomic Time (TAI) which is a high precision atomic coordinate time standard based on the movement of the Earth's geoid. Using the time package POSIXlt that is found in R packages it was able to convert the time to its correct Coordinated Universal Time (UTC). This was done in order to extract the day, month, year, hour, minute and second that the lightning strikes occurs.

C. Pattern

Once the data was correctly separated it was then used to plot the points of Longitude and Latitude with the corresponding Radiance using the R package ggmap [9]. By plotting all the points from each year it was evident that there was no clear pattern that emerged from the data, as depicted in Figure 2.

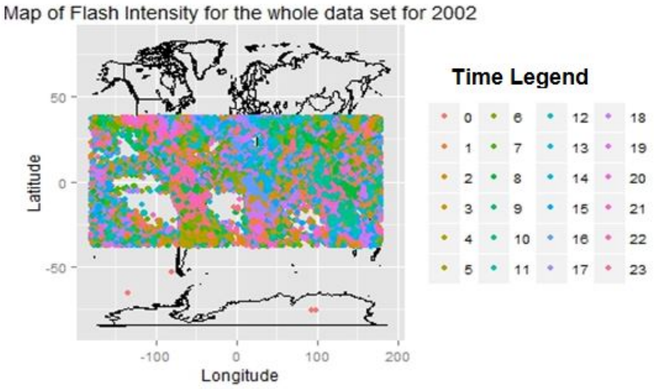

Fig. 2. The Whole Dataset

As no clear patterns were produced different maps of Radiance against variables such as the time and the month it occurred and the location were generated using the density function in ggmap. The density function uses a two-dimensional kernel density estimation with an axis-aligned bivariate normal kernel, as defined in equation (4). This equation takes a random sample of x values with a kernel (K) and a bandwidth (h) by which it tries to estimate the similarities in the x values and then links them together[9].

$$\hat{f}(x,h) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}) \tag{4}$$

Once these density maps were generated, clear correlations and patterns were observed. The main patterns that emerged were seasonal, time and spatial patterns.

A seasonal pattern was observed by plotting points where the lightning occurred for each hemisphere against the month it occurred. The observation made was that every three months the Radiance values changed from extremely high values to low values [10]. This showed that seasonal change correlated with that particular hemisphere's seasons. It was evident that when the Radiance and the number of lightning flashes where high, during that hemispheres summer season, shown in

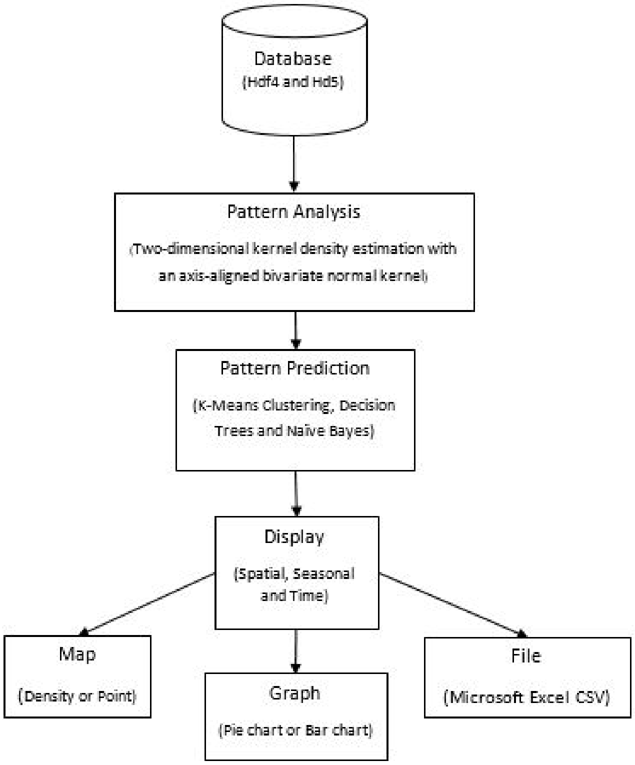
Database
(Hdf4 and Hd5)
Pattern Analysis
(Two-dimensional kernel density estimation with an axis-aligned bivariate normal kernel)
Pattern Prediction
(K-Means Clustering, Decision Trees and Naïve Bayes)
Display
(Spatial, Seasonal and Time)
Map
(Density or Point)
Graph
(Pie chart or Bar chart)
File
(Microsoft Excel CSV)

Fig. 1. The Subsystem Interaction

Figures 3 and 4. It was also evident that when the Radiance and the number of lightning flashes where low, during that hemispheres winter season [10]. When the Radiance and the number of lightning flashes were approximately the same in both hemispheres it was either autumn or spring season. Due to this autumn and spring seasons cannot be clearly distinguished between. From these results it has been established that there is lightning pattern that correlate to the hemispheres summer and winter seasons.

Time patterns were observed when plotting certain time values against the Radiance experienced at the time, as depicted in Figure 5. The results showed that globally certain areas experienced the same range of intensity values for that particular time. Hence showed evidence that there could be a time pattern for the dataset used.

Spatial patterns were observed by plotting latitudinal values with their corresponding Radiance value. What was evident was where the Radiance value was low the corresponding latitude was close to the equator and where the Radiance value was high, the latitude was further away. Due to this change in Radiance values as latitude increased, there is evidence that there could be spatial patterns, as depicted in Figure 6

D. Prediction

A number of different supervised and unsupervised machine learning algorithms were used in order to predict seasonal, spatial and time patterns. The machine learning algorithms used to test supervised learning were Naïve Bayes, Support Vector Machines (SVM), Random Forest, K-Nearest Neighbor Classification, K-Means and Decision Trees. The unsupervised learning was K-Means Clustering [11].

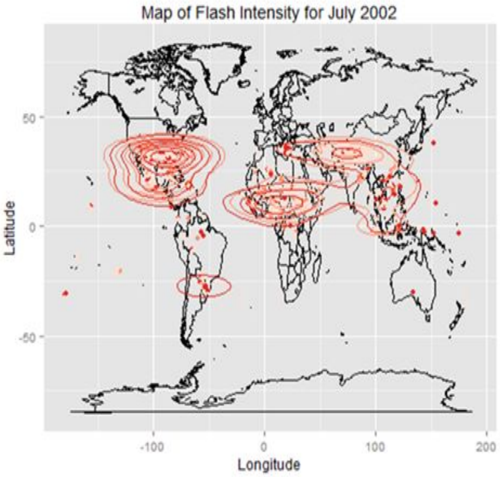

Fig. 3. Northern Hemisphere Summer Season Density Map

K-Means Clustering [11], [12] is a machine learning algo-

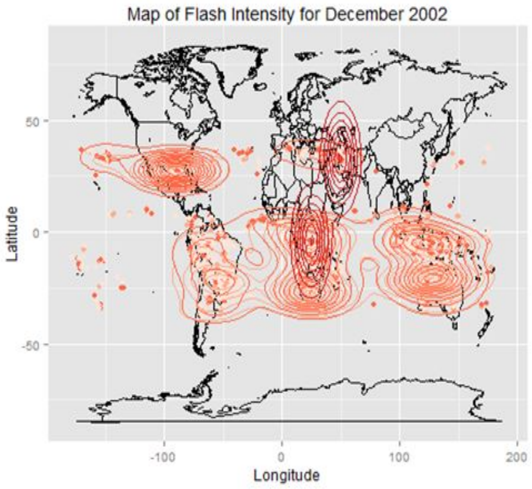

Fig. 4. Southern Hemisphere Summer Season Density Map

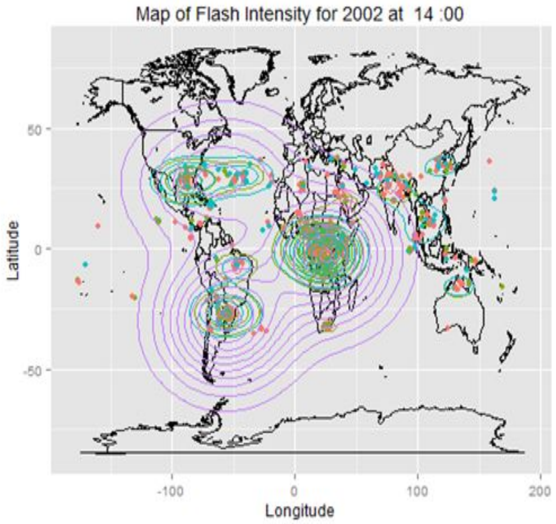

Fig. 5. Example of the Time Pattern Density Map

rithms used for seasonal pattern prediction. K-Means Clustering is calculated, by choosing a number of points c_j to be used as the initial centroid. Then subtracting all points $x_i^{(j)}$ from c_j then adding the subtracted answers together. Once the last point in the dataset has been reached and if the addition is not zero then it re-computes the centroid and restarts the calculation, as defined in the equation (5).

$$p = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - c_j\|^2 \tag{5}$$

Decision Tree[11], [13] algorithm is a suitable machine learning algorithms to use for spatial patterns. Decision Tree uses recursive partitioning to separate the dataset by finding the best variable and using the selected variable to split the data. Then using the entropy, defined in equations (6) and (7), to calculate the difference that variable would make on the results if it is chosen. If the entropy is 0 then that variable is prefect to use, else a new variable needs to be seleted.

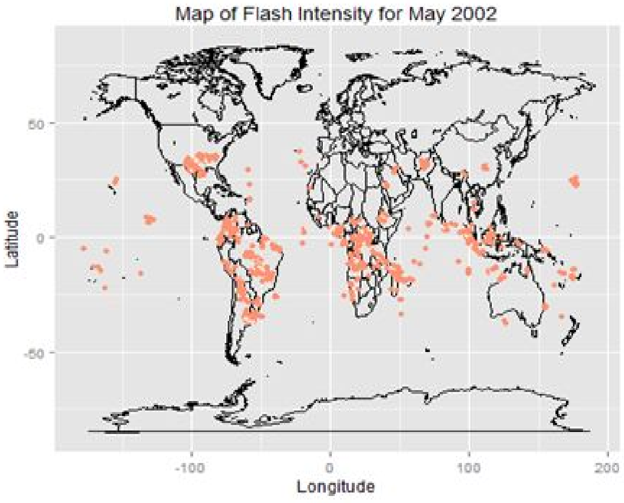

Fig. 6. Example of Spatial Pattern Density Map

$$H(D) = -\sum_{i=1}^{k} P(C_i|D)\log_k(P(C_i|D)) \tag{6}$$

$$P(C_i|D) = \frac{number of correct observation for that class}{total observation for that class} \tag{7}$$

Naïve Bayes [11], [14] is used for the prediction of the time pattern. Naïve Bayes classifies an instance by assuming the presence or absence of a particular feature and sees if it is unrelated to the presence or absence of another feature, given in the class variable. This is done by calculating the probability for which it occurred, as defined in equation (8).

$$P(x_1, ..., x_n|y) = \frac{P(y)P(x_1, ..., x_n)}{P(x_1, ..., x_n)} \tag{8}$$

IV. Results and Discussion

The system has been coded using the R Programming Language, the reason for this is because it has the ability to open and read hdf4 and hd5 files by downloading an extension package. The Graphical User Interface (GUI) was also coded in R Programming Language using the packages RGtk2, gWidgetsRGtk2, cairoDevice and plotrix as R Programming Language does not have a one set GUI package as in Java [15].

The dataset utilized was for the period from May 2002 to December 2006 and was acquired from the Space Science Research Institute at the University of KwaZulu-Natal. This data consisted of information on the date, time, location, number of events, number of groups and radiance. From the above variables this system used the time, the location and the radiance. These variables were then imported and saved as a Microsoft Excel CSV (Comma Delimited) file for spatial-temporal analysis.

The datasets used had between 900 000 - 1 050 000 lightning strikes per year. This system had a high complexity time to load data and produce the maps or graphs this could be a flaw of using R programming language. Once the dataset was loaded it was easier and quicker to work with the data.

A number of tests were done to investigate which machine learning algorithm would be suitable to determine the best results for seasonal, spatial and time patterns. The dataset sizes ranging from 0.01% to 50% of the original dataset were tested against the different machine learning algorithms. These testing datasets were filled with randomly chosen points from the original dataset.

After testing the dataset it was found that K-Means Clustering was suited for seasonal patterns, Decision Trees was suited for spatial patterns and Naïve Bayes were best suited for time patterns. The dataset sizes ranging from 0.01% to 50% of the original dataset were tested against these machine learning algorithms. These new testing datasets were filled with randomly chosen points from the original dataset.

A. Seasonal Pattern Prediction Results

Figure 7. showed the that for small amounts of data points all the machine learning algorithms have the ability to produce a result. Figure 7. also showed that the moment the dataset become too large Random Forest, Support Vector Machines and K-Nearest Neighbor could no longer produce results.

Decision Tree did produce the best results but it had high time complexity of over 3.5 minutes to compute a result for the whole dataset. Naïve Bayes produced average results but here too had a high time complexity. K-Means was found to be the best machine learning algorithm as on average had a 63% success rate for all dataset sizes. Also the time take to produce these results were less than a minute.

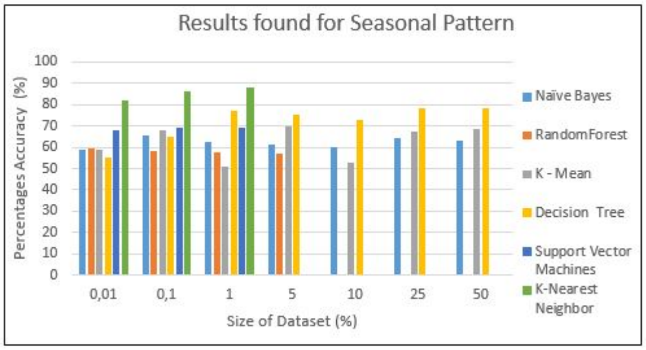

Fig. 7. Graph for percentages achieved for machine learning algorithm of the Seasonal Pattern

Hence, K-Means was used to predict the seasonal pattern for each year from 2002-2006, this produced 63%. After testing it was found that when the dataset was small it produced average results due to the fact that the machine learning algorithm did not have enough data to produce the pattern. The large datasets yielded good results, which is needed as the size of the data is extremely large. The computational time for K-Means was short for both large and small datasets, as depicted in Figure 8.

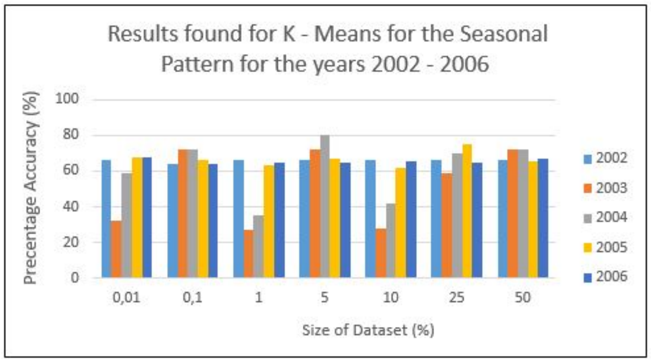

Fig. 8. Graph for percentages achieve for K-Means machine learning algorithm

B. Spatial Pattern Prediction Results

The observations made with spatial patterns and the machine learning algorithms on average produced results of about 90%.This showed that only a few points per dataset were close to the Equator and had low Radiance values hence making it easier to predict and obtain the correct results as most of the points were for the area far from the Equator and had higher Radiance. The moment the dataset became too large, once again Random Forest, Support Vector Machines and K-Nearest Neighbor could no longer produce results.

Due to the fact that the results produced were so high it was best to use the time it took to produce a result as the main factor. Naïve Bayes and K-Means both took on average 1.5 minutes for small datasets but for large datasets took 4-5 minutes to compute. Whereas with Decision Tree for both large and small datasets took about a minute to compute the result, as depicted in Figure 9.

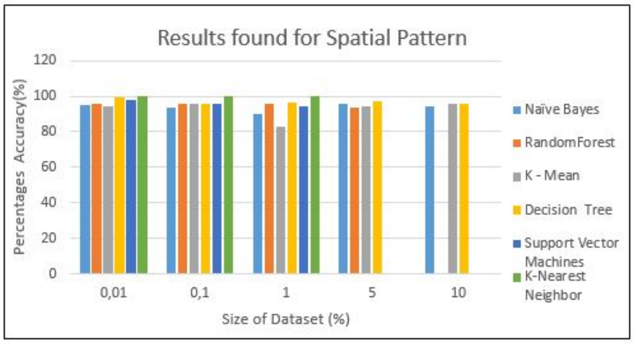

Fig. 9. Graph for percentages achieved for machine learning algorithm of the Spatial Pattern

Therefore the Decision Tree, which produced a 97% global pattern detection rate, was used to predict the Spatial Pattern for the years 2002-2006. It was found that when the size of the dataset was 0.01% for all years produced the same result of 99%. As the dataset got larger the results dropped by 1-2% but still maintained the above 90% average for all years, with some years even obtaining 100%, as shown in Figure 10 .

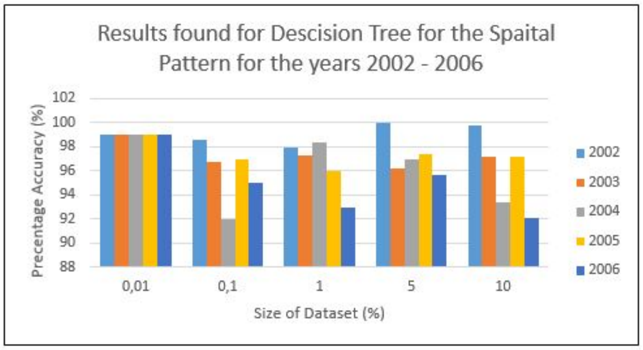

Fig. 10. Graph for percentages achieve Decision Tree machine learning algorithm

C. Time Pattern Prediction Results

Machine learning algorithms used to predict the time patterns on average produced poor results compared to the other two types of patterns as none of the percentages were above 50%. A reason that the percentage was so poor could be the fact that time patterns had 24 different categories from which it could choose from, where as seasonal and time patterns only had 2 different categories to choose from. Due to this all machine learning algorithms were finding it difficult to correctly classify the time patterns.This also showed that the time patterns were not that well defined in the dataset used. It was found that the moment the dataset became too large Random Forest, Support Vector Machines and K-Nearest Neighbor could no longer produce results, as shown in Figure 11.

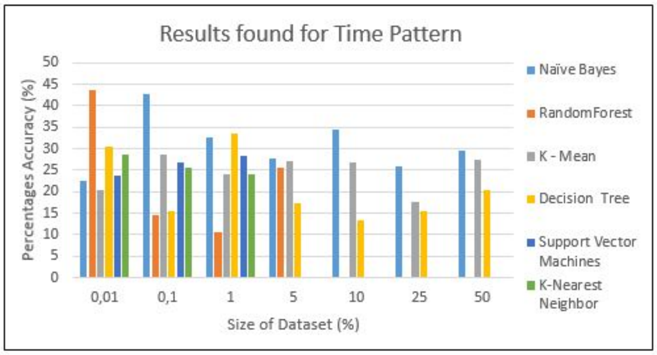

Fig. 11. Graph for percentages achieve for machine learning algorithm for the Time Pattern

The machine learning algorithm that produced better results was with Naïve Bayes with 31%. This was used to predict the time pattern for the years 2002-2006. It was found that when the dataset was small it produced results lower than normal. The large datasets had a 2-4 % higher result, with a few datasets going above 40%. The computational time for Naïve Bayes was longer due to the fact that it had many different categories to choose from, this was found for both large and small datasets, as depicted in Figure 12.

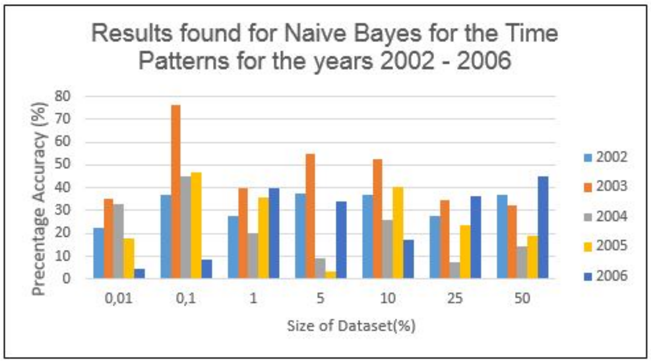

Fig. 12. Graph for percentages achieve for Naïve Bayes machine learning algorithm

V. Conclusion

This paper presented an alternative analysis based on automated pattern recognition, which identifies the changing state of lightning distributions using machine learning algorithm. The data file in the format hdf4 to hd5 is processed and analysed using particular machine learning algorithm, the resultant information enable the user to obtain a variety of graphs, maps and data. The machine learning algorithm used to predict the seasonal patterns was K-Means which achieve a 64% global pattern detection rate. The Decision Tree machine learning algorithm was used to predict the spatial patterns this achieved 97% global pattern detection rate. The machine learning algorithm used to predict the time patterns were Naïve Bayes which achieved 31% global pattern detection rate. This research work achieved a 63% global pattern detection rate for the larger datasets and for small datasets it achieved a 73% global pattern detection rate. Future work for this work is to optimize the time complexity through parallel computing.

Acknowledgment

The authors would like to thank the National Research Foundation (NRF) and the South Africa National Space Agency (SANSA) for all forms of assistance during this research work.

References

[1] C. Neuwirth, W. Spitzer, and T. Prinz, “Lightning density distribution and hazard in an alpine region,” *Journal of Lightning Research*, vol. 4, pp. 166 – 172, 2012.

[2] J. W. van Wagtendonk and D. R. Cayan, “Temporal and spatial distribution of lightning strikes in california in relation to large - scale weather patterns,” *Fire Ecology*, vol. 4, no. 1, pp. 34 – 56, 2008.

[3] H. R. Christian, D. Blakeslee, W. Boccippio, D. Boeck, K. Buechler, S. Driscoll, J. Goodman, W. Hall, D. Koshak, D. M. Mach, and M. Stewart, “Global frequency and distribution of lightning as observed from space by the optical transient detector,” *Journal of Geophysical Research*, vol. 108, pp. 1 – 15, 2003.

[4] A. B. Collier, S. Bremner, J. Lichtenburger, C. J. Downs, J. R.and Rodger, P. Steinbach, and G. McDowel, “Global lightning distribution and whistlers observed at dunedin, new zealand,” *Annales Geophysicae*, vol. 28, pp. 499 – 513, 2010.

[5] A. Dlamini, “Integrating satellite data and gis to map lightning distribution,” *PositionIT*, vol. 12, no. 1, pp. 1 –4, 2007.

[6] R. Marshall, U. S. Inan, T. Neubert, A. Hughes, G. Sátori, J. Bórl, A. Collier, and T. H. Allinl, “Optical observation geomagnetically conjugate to sprite - producing lightning discharges,” *Annales Geophysicae*, vol. 23, pp. 2231 – 2237, 2005.

[7] U. Finke and T. Hauf, “The characteristics of lightning occurrence in southern germany,” *Beitr. Phys. Atmosph.*, vol. 69, no. 3, pp. 361 – 374, 1996.

[8] J. Cannons and W. Kinsner, *Modelling of Lightning Discharge Patterns as Observed from Space*. Manitoba: Department of Electrical & Computer Engineering Signal & Data Compression Laboratory, 2000.

[9] H. Kahle, D.and Wickham, “ggmap: Spatial visualization with ggplot2,” *The R Journal*, vol. X/Y, pp. 1 – 12, 2012.

[10] A. Sugita and M. Matsui, “Lightning distributions in winter observed by the jldn,” in *21st International Lightning Detection Conference*, Orlando, 2010.

[11] P. Domingos, *A Few Useful Things to Know about Machine Learning*. Washington: University of Washington, 2010.

[12] T. Kanungo, M. Mount, N. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k - means clustering algorithm: Analysis and implementation,” *Transations on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 881 – 892, 2002.

[13] L. Rokach and O. Maimon, *Data Mining and knowledge discovery Handbook for Beginners*. France: Tel-Aviv, 2010.

[14] D. Lowd and P. Domingos, *Naive Bayes Models for Probability Estimation*. Washington: University of Washington, 2003.

[15] M. F. Lawerence and J. Verzani, *Programming Graphical User Interfaces in R*. London: CRC Press Taylor and Francis Group, 2012.