Employing Data Mining Techniques to Predict Occurrence of Thunderstorm Using Hourly Weather Datasets: In The Case Of Gondar Control Zone

Abebe Mulu
Department of Information Technology
Faculty of Informatics
University of Gondar, Ethiopia

Belay Enyew
Department of Information Technology
Faculty of Informatics
University of Gondar, Ethiopia

Abstract—Thunderstorms has meaningfully effects on both terminal and en route flights and reduce airspace capacity that results delays and have increased substantially en-route congestion. Current technology cannot provide reliable long-term prediction of thunderstorms for aviation operation. The objective of this study was to apply the data mining techniques to predict the occurrence of thunderstorms using 10 years NMA's synoptic dataset of Gondar station using design science research method. From collected data sets seven important attributes (cloud amount, cloud type, temperature, pressure, wind speed, rain fall and humidity) was selected from other variables or attributes to build the model. The experiments have been conducted using the six-step hybrid process model using four selected modeling algorithms. After performing an experiment using classification algorithms decision tree and rule induction, the models is evaluated based on their prediction accuracy in classifying the instances of the data set into thundered and non-thundered situations. From those classifier PART is selected by having best classifying accuracy that can classify 10718 or 99.70% instances as correct out of 10750 instances which is processed from Gondar aeronautics and Synoptic station

Keywords—Thunderstorm, synoptic data, Data mining, PART classifier, predictive model.

I. INTRODUCTION

Thunderstorm is one of the most spectacular weather phenomena in the atmosphere. It is the towering cumulus or the cumulonimbus clouds of the convective origin and high vertical extent that are capable of producing lightning and thunder. Usually, these thunderstorms have the spatial extent of a few kilometers and life span less than an hour. However multi-cell thunderstorms developed due to organized intense convection may have a life span of several hours and may travel over a few hundreds of kilometers [2]. Rasika Kalbende and Nitin Shelke [12], define the development and occurrence of thunderstorm: when a layer of warm and moist air rises to a larger extent, and updrafts to the cooler regions of the atmosphere the updraft that contains moisture condenses in order to form massive cumulonimbus clouds and eventually leads to the development of precipitation. Almost all thunderstorms develop under atmospheric conditions of low static stability, with abundant heat and moisture at low levels. The formation begins when dense (sinking) cold air over lies less dense, warm and moist (rising) air. Development is greatly enhanced when a catalyst such as strong heating and/or a trough is present. Strong updraughts then gradually form and the heat energy in the air and water vapors gets converted to wind and electrical energy. When the atmosphere is sufficiently unstable and the immediate surrounding promotes continuous contribution of energy into a growing cloud, a severe thunderstorm then develops, Strong up- and down-draughts characterize a well-developed thunderstorm [3]. Mark Weber and Dimitris Bertsimas [14], describes Thunderstorms has significantly effects on both terminal and en route flights and reduce both terminal and en route airspace capacity that Results delays and have increased substantially in the past decade due to increase en route congestion. Current technology cannot provide reliable long-term forecasts of the aviation impact of thunderstorms. Even when good short-term forecasts are available, the current air traffic management system often cannot effectively exploit them to improve network flow because of workload and airspace management difficulties. Generally thunderstorm have the following risks such as severe turbulence, severe clear icing, large hail, heavy precipitation, low visibility, Gust Front, Downburst, Macro burst, Microburst and electrical discharges within and near the cell. Data mining (DM) techniques are very popular for solving various problems. As a brief description, data mining is a mechanism for obtaining patterns from an existing dataset. Those extracted patterns are used to interpret the new or existing data into useful information [5]. DM has a potential to identify hidden knowledge from huge datasets. Many researchers apply data mining to explore hidden pattern from met record data. This study will use data mining technique for developing best model for predicting the Thunderstorm using spatiotemporal and synoptic data to improve the prediction of thunder storms.

II. PROBLEM STATEMENT

Weather forecasting or predicting is one of the most technical and technological problems around the world [3]. Prediction of significant weather components such as tornado, thunderstorm and tunnel clouds has important effects on different economic and social activities of human being that helps to adjust him/herself with the event and to protect themselves from those weather effects. Especially forecasting thunderstorm is one of the most difficult tasks in weather prediction, due to their relatively small spatial and temporal extension and the inherent non-linearity of their dynamics and physics [2], as discussed in related works different Scholars studies about the
This study attempts to answer the following research questions:

- What are the most determinant attributes that uses for occurrence of Thunderstorm?
- Which mining algorithm produces best Thunderstorm prediction model?

A. Objective of the study

General objectives: The general objective of this study developing Data Mining Predictive model that used to predict occurrence of thunderstorm in the case of Gondar Control Zone.

Specific objectives: The specific objective of the study are:

- To understand the problem domain by reviewing literatures on DM technology and their application in the prediction of thunderstorm.
- To identify the determinant attributes (features) that has great role for the formation of thunderstorm
- To prepare the data for analysis, to apply classification algorithms, to train, test and build the models using synoptic dataset.
- To compare the models based on their performance and select the best model.

B. Scope of study

The aim of this study was determining and predicting the occurrence of thunderstorm using of DM techniques on spatiotemporal datasets and ten years synoptic data from 2007 to 2016 that is stored in Ethiopian national meteorology agency data base having three observation time (06:00, 09:00 and 12:00). The model predicts the occurrence of thunderstorm before three hour of the event.

C. Research methodology

According to Smolander et al. (1990) [41], a method can be considered as a predefined and organized collection of techniques and a set of rules which state by whom, in what order, and in what way the techniques are used to achieve or maintain some objectives. For this study a design science research methodology which is developed by P. Ken, T. Tuure, A. R. Marcus and C. Samir (2007) [50], is used. DSRM provides a mechanism through which design, testing, and implementation of an IT artifact can be improved to the extent that it represents and the principle of what the artifact must to be [51].

Research design: This research was design to identify the determining factors for the formation of thunderstorm and predicting the occurrence in in Gondar control zone. To explore the application of data mining on this particular research, hybrid (Ciso.et al) data mining methodology was employed. Because this model is both academic and industrial. Ciso.et al involves sixth iterative process or steps including: Understanding the problem domain, Understanding of the data, Preparation of the data, Data mining, Evaluation of the discovered knowledge, and Use of the discovered knowledge steps [3].

Literature review: In this study Relevant literatures related to Thunder storms and its significant effect, the overview of data mining, knowledge discovery process model, data mining tasks, knowledge base system, and related works and Various books, journals, magazines, articles manuals, conference papers and related works and resources were reviewed.

Implementation Tools: The following tools and application software were used to accomplish the research process: WEKA 3.7.5 data mining tools: WEKA stands for Waikato Environment for Knowledge Learning. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Based on the prior knowledge of the researcher on the applicability of the tools for the purpose and freely availability of the software made the researcher to use the tools to build analysis and evaluate the model being developed for research goal. Ms-Excel: it will used for data preparation, pre-processing and analysis task because it has the capability of filtering attribute with different values. Besides, it is a very important application software to make ready the data and easily convert into of the file. Dataset preparation: method of Data collection & preprocessing.

Evaluation mechanisms and testing procedure: The data set collected from ENMA was preprocessed and divided into two data groups called training set and test set data. Using training set data the model was built using two decision tree and rule induction algorithms and the model tested using 10-fold cross validation test mode. These developed models in this research is compared using their classification accuracy and different confusion matrices (True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR)), precision, recall F-measure, Relative Operating Characteristics (ROC), the number of correctly classified instances, and number of leaves and the size of the trees, execution time.

D. Significance of the study

Weather forecasting is a vital application in meteorology and has been one of the most scientifically and technologically challenging problems around the world in the last century. Generally the study has the following major advantages in different fields of study.
ENMA: Ethiopian National Meteorology Agency (ENMA) forecast or predict the weather components using four categories. That is now casting- (current weather and forecasts up to a few hours ahead), Short range forecasts (1 to 3 days), Medium range forecasts (4 to 10 days) and Long range/Extended Range forecasts (more than 10 days to a season). All types of forecasting or prediction is depend on the previous data and present observation of observers which is exposed for human and technical errors. So the develop model help for NMA employee or professionals to easily and accurately Predict and identify special weather elements basically for thunderstorm which has short duration and great impacts on nature.

Aviation industries: aviation operations are constrained by different weather elements and all air navigation operators needs weather information for their safe operations. The result of these study helps for operators to design their routes and to reduce numbers of delays that is occurred due to thunderstorm and other related weather elements.

others: the study also have significant advantages for Air traffic controllers, maritime navigators and emergency and rescue handling organizations to make decisions during their day to day activities.

II. RELATED WORKS (STUDIES)

Different scholars or authors were tries to develop different predictive models using different techniques and uses different data sets having different attributes to predict the occurrence of thunderstorm. Himadri Chakrabarty and Sonia Bhattacharya [1], tries to predict the occurrences of thunderstorms by K-Nearest Neighbor Technique using three types of weather variables such as moisture difference, adiabatic lapse rate (temperature), and wind-shear. The model that is developed by Himadri Chakrabarty and Sonia Bhattacharya [1], can classify with 82% accuracy of occurrences and non-occurrences of thunderstorm in 12 hour. Litta A. J et al [2], also develops Artificial Neural Network Model that used to predict the occurrences of thunderstorm using surface temperature. Their result clearly indicated that overall accuracy is 76%.

Himadri Chakrabarty et al [3], uses Artificial Neural Network to Predict Squall-Thunderstorms Using RAWIND Data and develop a model that used to forecast 98% ‘squall-storm days’ and ‘no storm days’. Himadri Chakrabarty and Sonia Bhattacharya [4], in 2014 also tries to Predict Severe Thunderstorms using artificial neural network technique. Multilayer Perceptron (MLP) has been applied on the weather parameters of moisture difference, adiabatic lapse rate and vertical wind shear which were recorded by the radiosonde rawind (RSRW) in the early morning at 06.00 am local time. MLP classified and predicted “severe storm” and “no storm” days nearly up to 70% having around 12 hours lead time.

Waylon Collins and Philippe Tissot [6], uses artificial neural network to forecast location of thunderstorm and develop a model having 50% accuracy.

III. DATA PREPROCESSING

Preprocessing the data includes multiple steps to assure the highest possible data quality, thus efforts are made to detect and remove errors, resolve data redundancies, and to handle errors. Different techniques were used in data preprocessing of this study.

A. Data Cleaning

Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data [14]. The data mining process get a confusion with the unclean data of real world database that contains incomplete, inconsistent and noisy data. Thus, data cleaning is mandatory in order to improve the quality of data that improves the performance of the data mining techniques.

Handling Missing Value: There are two methods to handle missing attribute values belong either to sequential methods (called also preprocessing methods) or to parallel methods (methods in which missing attribute values are taken into account during the main process of acquiring knowledge) [38]. This study used Sequential methods to handle missed values of the attributes from the collected data. For numeric attributes mean or average values was used and for nominal attributes highly frequent value or mode was used. The following table 1 below shows detail descriptions of attributes, total numbers of missed values, techniques that used to replaced missed values and the replaced values.
B. Data reduction
Data reduction is one of the tasks in data mining that needs to be done before the actual mining task is takes place. Although, 11400 instances are collected, after preprocessing the redundant instance that is measured in similar observation time as weather report and special weather report were selected one and rejected or deleted the other and finally the researcher re-sampled the dataset into 10750 instances for data mining. After formulating the required sample dataset, the data is converted in to comma delimited Excel file (CSV format), then to the arff file which is suitable for mining using WEKA 3.7

C. Data Integration
Even if most data set of the study was collected from on database some attributes such as Pressure and special weather components were collected from hourly registration log book which has different format and different data representation (i.e. symbolic and abbreviated forms of data). When matching those attributes from the database special attention was paid to integrate with similar event of other attributes structure of the data. This is to ensure that any attribute functional dependencies and referential constraints in the source system to match those in the target system.

D. Data Transformation
Among those techniques of transformation, data discretization / binning is selected for this study. Binning is used to reduce data size by dividing the range of a continuous attribute in to interval. Interval labels can then be used to replace the actual data values [30]. The collected data was transformed into a format appropriate for further Data Mining process by dimension reduction (such as feature selection and extraction, and record sampling), and attribute transformation (such as discretization of numerical attributes and functional transformation). The Divided range of continuous values was assigned into N intervals of equal sizes and labeled based on standards that is seated by WMO and discussion with domain expert.

IV. DATA MINING EXPERIMENTATION
Creating a Model is one of the major tasks which is undertaken under the phase of data mining in hybrid methodology. In this phase several data mining techniques are applied and their parameters are adjusted to optimal values. Some of the tasks include: - experimental setup or design, selecting the modeling technique, building a model and evaluating the model.

A. Experimental setup
In any data mining process before building a model, we need to generate a procedure or mechanism to test the model's quality and validity. In this research 10750 datasets are used for training and testing the model. WEKA 3.6.9 software has used to set up and measure the quality, validity and test of the selected model. K-fold (10-folds) cross validation was used because of relatively its preference and low variations [30]

<table>
<thead>
<tr>
<th>s. no</th>
<th>Attribute name</th>
<th>Number of Missed values</th>
<th>Data type</th>
<th>Method or techniques</th>
<th>Replaced with</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Cloud amount</td>
<td>450</td>
<td>Numeric</td>
<td>MODE</td>
<td>CU</td>
</tr>
<tr>
<td>2.</td>
<td>Temperature</td>
<td>110</td>
<td>Numeric</td>
<td>AVERAGE</td>
<td>28.25</td>
</tr>
<tr>
<td>3.</td>
<td>Rain Fall</td>
<td>102</td>
<td>Numeric</td>
<td>AVERAGE</td>
<td>32.06</td>
</tr>
<tr>
<td>4.</td>
<td>Relative Humidity</td>
<td>430</td>
<td>Numeric</td>
<td>AVERAGE</td>
<td>68</td>
</tr>
<tr>
<td>5.</td>
<td>Wind Speed</td>
<td>420</td>
<td>Numeric</td>
<td>AVERAGE</td>
<td>4.37</td>
</tr>
<tr>
<td>6.</td>
<td>Wind Direction</td>
<td>328</td>
<td>Numeric</td>
<td>MODE</td>
<td>270</td>
</tr>
<tr>
<td>7.</td>
<td>Pressure</td>
<td>320</td>
<td>Numeric</td>
<td>AVERAGE</td>
<td>1023.67</td>
</tr>
<tr>
<td>8.</td>
<td>Observation Time</td>
<td>-</td>
<td>Numeric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Elevation</td>
<td>-</td>
<td>Numeric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>Latitude</td>
<td>-</td>
<td>Numeric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>Longitude</td>
<td>-</td>
<td>Numeric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>Sun shine</td>
<td>320</td>
<td>Numeric</td>
<td>MODE</td>
<td>9.8</td>
</tr>
</tbody>
</table>

Table 1 attribute missing values, replacement techniques and replaced values
B. Attribute selection

To select the best attribute subset selector, the investigator uses information gain attribute evaluator with ranker search method. Information gain attribute evaluator works by evaluating the worth of an attribute by measuring the information gain with respect to the class. As shown in figure 1 below, information gain attribute subset evaluator algorithm ranks the attributes based on the information gain with respect to class. The researcher selected 7 best attributes according to their rank from 12 independent attributes.

Figure 1: Attribute Selection table

Attribute Selection on all input data

Search Method: Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): # Thunder):
Information Gain Ranking Filter

Ranked attributes:
0.20792 7 Rainfall
0.22026 2 low_wid1_CRE
0.14459 1 cloud_amount
0.12146 3 tempmax
0.07483 4 RH
0.0097 6 windspeed
0.00202 8 pressure

Selected attributes: 7,2,1,3,4,6,5 : 7

C. Selecting modeling technique

Four predictive models involving J48, REPTree, PART and JRip classifier algorithms are constructed. J48, and REPTree are tree based classifiers in WEKA whereas PART and JRip are rule based classifiers. Four mining algorithms (J48, REPTree, PART and JRip) which were used to build the models yields for different models having different performance i.e. these algorithms yield a model having 99.53%, 99.25%, 99.70% and 99.57% accuracy respectively and all the modeling techniques have greater than 99% Recall, precision and F-measure. From all modeling techniques listed above table PART algorithm yield best performed model compared to all tested algorithms.

D. Evaluation of the developed models

In this study as discussed in table 2, the performance of the models is evaluated based on their prediction accuracy in classifying the instances of the data set into thundered and non-thundered situations. As we can see from the result of the experiments in the above scenarios, there is a slight difference between the results of each classifier. From those classifier PART is selected by having best classifying accuracy that can classify 10718 or 99.70% instances out of 10750 correctly. While other classifiers Results J48, JRip and REPTree show nearly equal number of incorrectly classified instances. The highest incorrect classification is scored by REPTree.

Comparison of the models using classifying accuracy, ROC, Precision and time execution

As we show from the following figure 2 below and table 2 from four experiments PART has best classifying accuracy, ROC and precision compared to other classifiers.

Figure 2: Comparison of the Algorithms Using Classifying Accuracy, ROC, and Precision.
to evaluate the performance of the classifiers used in this study

<table>
<thead>
<tr>
<th>Classifier</th>
<th>TP RATE</th>
<th>FP RATE</th>
<th>INSTANCES</th>
<th>CCI</th>
<th>ICI</th>
<th>Precision</th>
<th>RECALL</th>
<th>F-Measure</th>
<th>ROC</th>
<th>Accuracy</th>
<th>T-Execution</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>0.996</td>
<td>0.005</td>
<td>10699</td>
<td>99.53</td>
<td>51</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
<td>0.999</td>
<td>99.53</td>
<td>0.39</td>
<td>No-thunder</td>
</tr>
<tr>
<td></td>
<td>0.995</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Thundered</td>
</tr>
<tr>
<td>REP Tree</td>
<td>0.993</td>
<td>0.008</td>
<td>10669</td>
<td>99.25</td>
<td>81</td>
<td>0.994</td>
<td>0.993</td>
<td>0.994</td>
<td>0.998</td>
<td>99.25</td>
<td>0.06</td>
<td>No-thunder</td>
</tr>
<tr>
<td></td>
<td>0.992</td>
<td>0.007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Thundered</td>
</tr>
<tr>
<td>PART</td>
<td>0.997</td>
<td>0.003</td>
<td>10718</td>
<td>99.70</td>
<td>32</td>
<td>0.998</td>
<td>0.997</td>
<td>0.997</td>
<td>0.999</td>
<td>99.70</td>
<td>1.98</td>
<td>No-thunder</td>
</tr>
<tr>
<td></td>
<td>0.995</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Thundered</td>
</tr>
<tr>
<td>JRip</td>
<td>0.994</td>
<td>0.002</td>
<td>10704</td>
<td>99.57</td>
<td>46</td>
<td>0.998</td>
<td>0.994</td>
<td>0.996</td>
<td>0.998</td>
<td>99.57</td>
<td>15.74</td>
<td>No-thunder</td>
</tr>
<tr>
<td></td>
<td>0.998</td>
<td>0.006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Thundered</td>
</tr>
</tbody>
</table>

Table 2 TPR, Precision, Recall, F-Measure, ROC, Accuracy and time execution of the algorithms

In the other word the performance of the classifier is also evaluated in terms of different confusion matrices (True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR), precision, recall F-measure, Relative.

Operating Characteristics (ROC), the number of correctly classified instances, and number of leaves and the size of the trees and execution time. Based on the above result in experiment 3 (PART rule induction classifier list) is selected. Because from all four experiments which was conducted using 10-fold cross validation as we show from table 2 Experiment 3 (PART rule induction classifier list) is best based on accuracy which registered 99.7% and correctly classified instance which is accounts 10718 out of 107500 instances.

V. CONCLUSION AND RECOMMENDATION

Thunderstorms has significantly effects on both terminal and enroute fights and reduce airspace capacity that Results delays and aviation impact of thunderstorms. The goal of this research was to apply the data mining techniques to predict the occurrence of thunderstorms using synoptic dataset. The research uses design science research methodology and follow the six-step Cios et al. (2000) hybrid data mining process model and 10750 datasets are used for training and testing the model using WEKA 3.7 Four mining algorithms (J48, REPTree, PART and JRip) were used to build the models get for different models having different performance i.e. these algorithms yield a model having 99.53%, 99.25%, 99.70% and 99.57% accuracy respectively and all the modeling techniques have greater than 99% Recall, precision and F-measure. From all modeling techniques PART algorithm yield best performed model compared to all tested
algorithms. The rule induced from PART algorithm is used to develop the knowledgebase system using Swing prolog that is used to assist air traffic controllers to make decision. The results obtained from this study are very promising and the model also have better performance related to previous studies. The researchers thought that considering the special attributes such as cloud type, measured rainfall, cloud amount and current temperature improves the performance of the model. The results obtained from this research indicate that data mining is useful in bringing relevant information to the service providers (NMA) as well as decision makers (duy air traffic controllers). This research is mainly conducted for an academic purpose. However, the results of this study are found to be hopeful to be applied to address practical problems that is observed real life activity in aviation industries. This research work can contribute a lot towards a comprehensive study in this area in the future, in the context of our country. The results of this study have also shown that the uses of DM technologies with knowledge base system are well applicable in other weather element prediction. Hence, based on the result of this study the researcher believes further researches have to be done to increase the benefits of the developed model the following are recommended for future study

1. Using Satellite data addition to synoptic datasets to build the model: The Agency have geospatial database which recorded events throughout the day using satellite. Constructing the model using those images and satellite data is complete and more reliable than those lease quality data in that synoptic database.

2. Building a model by considering Topography, plants Distribution (vegetation index) and water surface of the earth which has significant role for development of thunder bring clouds. So considering these variables may give significant change on rules gain from the model and gives better prediction model

3. Integrating Data mining model with the Knowledge Based System.

REFERENCES


