Predicting Wildfire using Data Mining

by

Richa Singh

A Project Report Submitted
in
Partial Fulfillment of the
Requirements for the Degree of
Master of Science
in
Computer Science

Supervised by

R. K. Raj

Department of Computer Science

B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York

May 2016

Abstract

Predicting Wildfire using Data Mining

Richa Singh

Supervising Professor: R. K. Raj

Every year United States of America spends millions in order to deal with the wildfire breakout. This has caused not only economic damage but also hampered ecological balance by destroying the vegetation and flora fauna. Wildfire is also responsible for pollution and changes in climatic condition over the period of time. Over the decade forest fire has become major concern as it has endangered lives of species. In spite of spending millions on trying to control the deceased fire, it still remains one of the prominent issue.

Fire fighters are aware about how forest fires can be unpredictable. But if they are alerted about the breakout well ahead in time, this phenomenon can be controlled and prevented. There are many traditional technologies that deal with wildfire threat analysis. In this project, we are aiming to solve the problem by analyzing the historical forest fire data along with weather data and predicting the size of fire. In our work, we are trying to explore data mining techniques which can help in predicting the intensity of forest fires. The intensity can be determined by the areas/hectares burned and how long did the fire last. Some of the data mining techniques used in our project are J48, Naive Bayes, Random forest, Support Vector Machine, Neural Network.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

The rise of temperature in all over the world has definitely put the ecosystem at risk by increasing the occurrences of wildfire. There can be many factors contributing to forest fire but those factors cannot be taken into account while predicting a fire. Amidst all these, weather conditions can play an important role in predicting a wildfire fire. These predictions can help fire man, to have a plan ahead of time and be prepared.

United States of America has witnessed some of worst wild fires of its time. Several reasons have contributed towards forest fire for most of the 20th century ranging from human caused fire, natural reason like lightening, weather conditions etc. Some of the deadliest fire that US has witnessed would be Peshtigo and Great fire 1910. Steadily the number of disastrous wildfire has been increasing. Over the years, US has spent more than a billion to subdue the forest fires. But wildfire across the western America has been getting bigger and worst. California is a diverse state and changes in climatic condition continue to bring a rise in temperature and similarly affect other weather conditions. The study says the wildfire of over 1000 acres in size have increased by 7 fire from year 1984 to 2011.

In this research, we are working with data set of historical forest fire of United States Of America from 1992 -2011. The initial statistical analysis says that year 2007 has had maximum occurrences of wildfire and California has seen severe fire in October 2007. Several study suggest that the climatic condition of California is major reason for frequent fire occurrences. Since California has been victim to maximum number of fire breakouts,it has become extremely important for the fire department of California to have a vision of wildfire, so that they can be prepared. This research would include studying the factors influencing the wildfire breakout. Data sets comprising of historical forest fire data and the meteorological data of year 2007 for California region is going to be considered. The two data sets would be clubbed together and analysis would be performed. The models thus built should be able to predict area burnt. Also using data mining algorithm, relationship can be established between different weather conditions as to how they influence the fire to spread.

1.2 Background

Every state has a forest department which collects data of the fire occurrences on daily basis. The abundant data has been used for visualization and statistical analysis a lot. But the humongous data can be used efficiently if it is to put it through some data mining algorithm to generate patterns and knowledge is extracted from it. This can help building predictive model which can be used for forecasting forest fire.

In past weather data has been used to prevent and help fire fighters department. When there were hardly any computer in 1970s, Canadian FWI index was designed using the weather data like temperature, relative humidity, rain and wind. Also using historical data, a lot of efforts have been put to do trend analysis and spatial-temporal forecasting which leads to predicting the hectare of burnt areas. Some researchers have put data mining techniques into analyzing the areas of higher risk.

1.3 Problem definition and motivation

Fire fighters are well aware of unpredictable nature of fire. But the early the fire is detected, the better for them. Also an old-style human analysis is costly and can easily be affected by various factors. This gives rise to the need of programmable solution. The satellite based solution which helps in predicting fire is an expensive operation. Whereas using historical

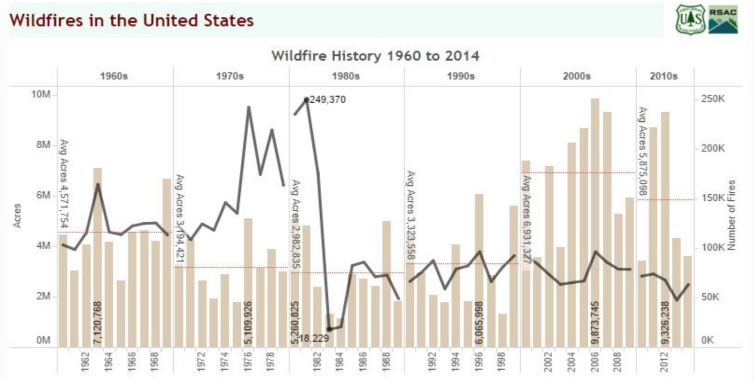

Figure 1.1: Fire History US (1960 - 2014)

data and analyzing it, is comparatively less expensive solution and does not incur cost of maintenance. Meteorological data like temperature, rain, humidity are recorded by weather stations on daily bases. If weather data is combined with historical data, new patterns can be generated.

The main motivation behind using all the historical data is that these data have hidden patterns and trends. These patterns can be used to make different observation, various prediction and recommendations. The Data mining algorithm can be used to recognize the hidden patterns and extract knowledge from it.

1.4 Related work

Several scientist have contributed to the field of data mining by analyzing forest fire data and in turn safeguarded the lives of human and helped in maintaining a balance in ecosystem. Data mining approach used by these researchers would include neural network along with sensor data to figure out areas which a risk of frequent fire. Paulo Cortes and Anibal Morais [3] used five different data mining algorithm to explore the historical forts fire data of Portugal. They worked with data of January 2003 to December 2003. They clubbed two datasets, the other one was weather data with factors like rain, wind, temperature and humidity. The five different algorithms used by them include multiple regression, Decision tree, random Forest, Neural network and Support Vector Machine. The approach as suggested in the papers could predict the burned area caused by fire of small sizes. Some researchers have used apriori algorithm to come up with patterns in weather condition that can suggest as to what factor contributes more towards a fire to spread. A lot of work has been done in spatial temporal data mining towards trying to predict forest fire. Also fire spread simulators are used along with cellular automata to predict forest fire

Not just USA, Canada and New Zealand are quite active in predicting wildfire using data mining techniques. Research is being done to find out all the reasons that play an important role in starting a fire and further. Several reasons that are responsible for forest fire to start and spread are human caused, lighting, insects, and vegetation. Scientists have achieved a 90 percent success rate in reducing occurrence of forest fire by building simulators which can predict fire using satellite images.

1.5 Hypothesis

The underlying problem that we are trying to identify and solve in our research is that earlier work has been done trying to extract hidden patterns from sensor data to predict forest fire. But reading the behavior of fire and forecasting the threat posed by it is not an easy task. Also using method like infrared and smoke scanner is very expensive. Whereas historical data of climatic condition which is known to influence the fire the most is readily available. Getting day to day weather data of any region will incur less cost. So using historical data and meteorological data of previously occurred fire, we try to predict the area burnt. It will help in preventing the outbreak of fire hours in advance. Some of the major agendas of the project would be studying various factors responsible for the wide spread forest fire along with forecasting the burned areas. This will help fire department to warn the public. Also fire department would have a schedule to fight the fire, they will be more alert, would be able to set their priorities right while dealing with fire and also will give them time to engage in various safety measures for dealing with huge fire

1.6 Approach and Solution Implementation

In our research, we aim to solve the problem of wildfire by predicting the burned areas of forest fire using different Data mining techniques. The paper shows two approach. One approach takes into consideration different classification algorithm that will help classify the wildfire into different classes 5 classes A, B, C, D, E based on size of area burnt in Acres. The other approach would use association rule mining to generate rules between weather data and wildfire. The preliminary task would be to collect historical forest fire data. Then collect weather data depending on the month and time of fire breakout. Consolidate all the data to be part of one data set so that it can be used for different analysis.

In our research, the data set we are considering consists of historical forest fire data of United States of America from year 1992 to 2011. The data set has more than 15 million records. Some of the important attributes to be considered for analysis while discarding others would be Latitude, longitude, region, area code, fire name, the size of fire, start date, end date, cause/reason of fire, state, county. Our study says California has witnessed some of the worst wildfires. The reasons that causes a forest fire to start in this region are plenty from human started fire to lighting. But the cause for fire to spread are dry torrid weather. In our study we are focusing on evaluating California State and study the weather factors that influence the fire here. Weather data consists of meteorological data with attributes like as rain, temperature, humidity, and wind.

Different classification algorithm would be used like Support vector machine, Nave Bayes to build a model which will be used for predicting the burned areas given a location and weather condition of the location. These prediction can be used by forest department in doing the math of number of fire fighters required to suppress the fire. Using feature selection some of the important factors affecting the model can be studied. The data would be divide into training and test data set. After using different classification techniques to build predictive model, it would be evaluated against the test data set. The performance of different model thus built would be analyzed to see which performs better against the other.

1.7 Roadmap

The report is presented in the following way: Section 2 talks about solution design of model. Section 3 contains information about data collection and pre-processing. Section 4 talks about implementation of four different data mining algorithms. Section 5 discusses about results and analysis of algorithms. Conclusion, future work and lessons learnt are presented in section 6 of the report.

Chapter 2

Design

2.1 Solution Design

The basic idea of the project is to design a model which can be used to predict burned areas using predictive algorithm. Also to study the factors that influence the fire, rules would be generated using apriori Algorithm. The fire data set would be collected of U.S region. Along with fire data, weather data would be collected. The basic design rationale of project is that it will be divided into two phases. In the first phase a classifier would be built to predict the likelihood of a fire based on weather conditions like temperature, relative humidity, rain and wind. If it is concluded that a day has a likelihood of having at least one fire, the data could be given to second phase of classifier.

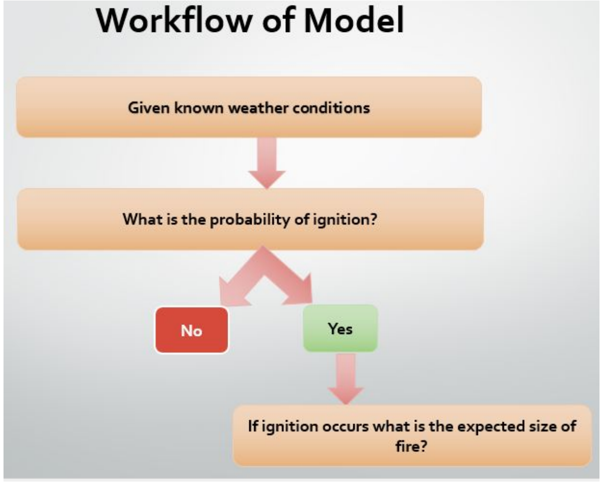
Workflow of Model
Given known weather conditions
What is the probability of ignition?
No
Yes
If ignition occurs what is the expected size of fire?

Figure 2.1: Workflow of model

Approach would involve the following steps:

1. Collecting historical forest fire data.
2. Collecting weather data depending on the month and time of the fire breakout.
3. Consolidating all the data to be part of one dataset.
4. Classification algorithm for predicting the fire breakout.
5. Algorithms like Decision Tree, Nave Bayes etc.
6. Association Mining for generating rules.
7. For Apriori Association rule mining we need prior information about the data.
8. Factors to be analyzed would be Rainfall, Moisture, Air temperature, Wind speed etc. as they play important role in determining if fire will break
9. Feature selection for determining the factors affecting the model.

Chapter 3

Data Processing

3.1 Data Collection

To construct the data set for analysis we needed a historical wildfire data along with weather data for a give region. After indulging in lot of research for getting the right set of data for our project, we finally found one such historical forest fire data on the official site of United States Department of Agriculture Forest Service [5] [4]. The data can be found at http://www.fs.usda.gov/rds/archive/Product/RDS-2013-0009.2/[5] and [4]. This data set had fire accidents recorded by government and local body. The data consists of spatial wildfire from year 1992 - 2011. It has more than 15 million records. The authors have already taken care of converting the data into readable format. Some of the important attributes of wildfire data includes Location(Latitude, Longitude), Discovery Date, Fire Size, Fire Name, Discovery Time, Fire Cause, Contain Date, Contain Time, State, County, Fire Type, Protection Type[6].

The weather data was obtained from University of California Agriculture and Natural Resources State Wide Integrated Pest Management Program [1]. Their official site maintains historical weather data for California Region. One gets to choose weather station to extract the historical data by supplying the date range. Daily weather data of Riverside was collected from Riverside Citrus Experiment Station. The attributes of the data were maximum temperature, minimum temperature, wind speed, wind direction, humidity [1].

Description and Statistical analysis of Weather Data

The fire data set had more than 15 million entries for the year 1992- 2012. It comprised of fire incidents for various parts of United States of America for those years. The data set included wildfire accidents of states like California, Arizona, Colorado, Connecticut, Oklahoma, Mississippi, Texas, Virginia, Washington, Georgia, Oregon etc. Each year saw an increase in the intensity and size of fire. The year 2007 especially saw the highest amount of fire accidents amongst all other years. The data set had records for reason of start of fire and most common ones were lighting, smoking, campfire, fire work, structure. Though these regions had most of the fire started by human, but reason for spread was different. On further statistical analysis, it was found that the state of California had witnessed the maximum number for fire accidents for that part of the year. The data for California was distributed over various counties which included Butte County, San Diego, Alameda, Fresno etc. Presence of different counties gave a lot of opportunity to study them. But on further probing, it was concluded collecting weather data of different region and merging it together would be difficult. So it was decided to work on analyzing the California state. So the data set was filtered to get the region with maximum number of fire accidents. Riverside County was selected as the area of study.

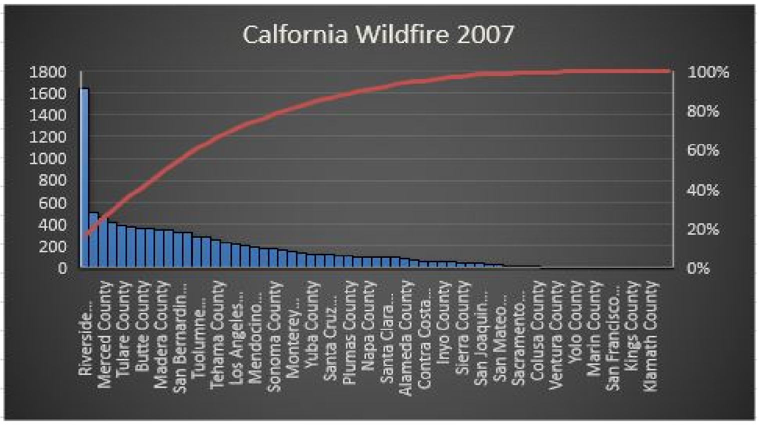

Figure 3.1: Graph for California Fire Breakout 2007

Description and Statistical Analysis of Weather Data:

Weather data set had attributes like maximum temperature, minimum temperature, wind speed, wind direction, humidity, precipitation[1]. Initial survey and research suggested the weather in California for the year was very bad as it saw rise in temperature. The weather highest recorded temperature was 111.8F on July 22nd, 2007 and lowest recorded temperature was 28F. Precipitation was recorded 72.9 mm. There were few rainfalls.

3.2 Data Cleaning and Preparation

Data Preparation for phase I

Wildfire data and weather data was in MS Access database. Wildfire data had total of more than 15 million instances which had seen forest fire from year 1992- 2012. The data set had lot of missing values. Cleaning started by getting rid of rows with most number of blank values. Skewed data set does not contribute much towards building a predictive model. Further duplicate values were also removed by querying the MS Access database. Some values were out of range, they were clearly the outlier and in process of cleaning even they were removed. At this point data set was free of complete blank rows, duplicate values and outlier.

Later we moved to filter the California state for year 2007 since it had seen the maximum number of fire. Total number of fire incident for California state for year 2007 was more than 10000. On filtering, it came to our notice that County variable had many missing values. We could not ave worked without county, as it was important to pull out the Weather data. Out of 10000 instances, more than 50 percent were missing from County field. We decided to fill the County column with the help of latitude and longitude available. So a restful service was written in java to consume and get the county based on coordinates. Google Map Geoencoding API was used for this purpose. Later it came to attention that even weather data set had some fields like precipitation, humidity, wind speed with blank values. To fill this up, Weather Underground API was used. If the API is supplied with

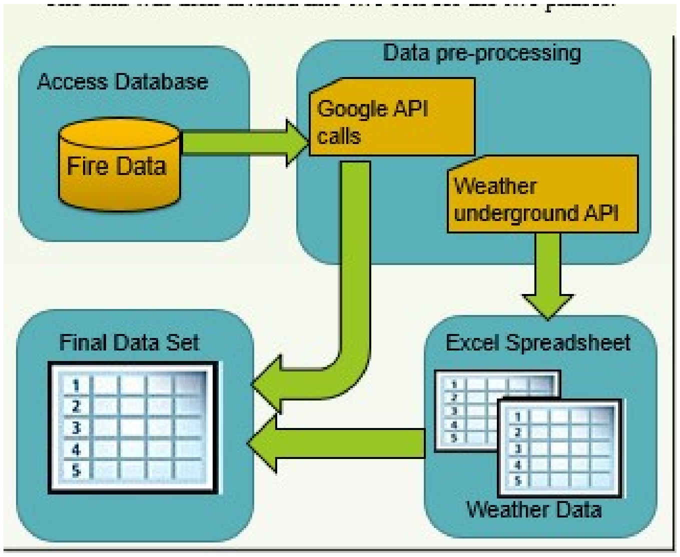

Figure 3.2: Data preparation

a date and region, it has provision of providing the environmental condition for the given region.

For further pre-processing, we used MS access queries to get rid of variables from both the dataset. Later the data was consolidated to form one single data set containing all fire characteristics and weather conditions. The start date of fire and the date when weather condition was recorded was considered as primary key in order to club the two data sets. The final data was extracted and kept in a excel sheet. Also for phase one, a new attribute was added which showed if a forest fire will breakout on a given day. So the first phase data set had all of the weather data and the binary field for fire yes/no [2]

Data Preparation for phase II

For second phase, first size was converted from hectares to acres (1 acre = 0.45 hectare).A new attribute names fire size class was added which was obtained from the previous attribute fire size [2]. The new attribute was added as fire size had fire sizes distributed over a wide range and classifications algorithm give better results when the number of resulting

Class	Class Description(in acres)
A	0 < A < =0.3
B	0.3 <B <=10
C	10 < C < =100
D	100 < D < =1000
E	1000 < E < =5000
F	5000 and up

Table 3.1: Fire Size Class

class are less. After creating the new attribute, it was added to the existing consolidated sheet. Data set for second phase had weather attributes and fire characteristics.

Final data set has total of 10 attributes and 1645 instances of Riverside county of California region.Table 3.2 has all the attribute for final data set.

Attributes	Description
Latitude	Latitude in degrees
Longitude	Longitude in degrees
Date	day, month and year of fire occurrence
Max Temperature	Temperature in Fahrenheit
Min Temperature	Temperature in Fahrenheit
Humidity	Average humidity in percentage
Precipitation	Amount of snowfall/ran recorded in mm
Wind Speed	Speed measured in mph
Wind Direction	Direction of wind in degrees
Fire Size Class	Size of fie in acres

Table 3.2: Final Data Set

Chapter 4

Implementation

4.1 Software details

Details of the software used for the project are as follows:

1. Tools: Weka, Microsoft Excel Sheet, Tableau
2. Languages :Java, SQL
3. Database : Microsoft Access DB
4. API calls : Weather underground API, Google Maps Geo coding API
5. Algorithms : J48, Naive Bayes, Support Vector Machine, Artificial Neural Network

4.2 Implementation for phase I

1. **Naive Bayes**

   Naive Bayes is a classification algorithm and belongs to supervised learning category. In case of supervised learning, the label / classes for the data is already known. It is probabilistic algorithm which is used to classify / assign class label for new instances of a data set based on prior probability of already seen instances. Nave Bayes calculates the prior probability of each class independent of other variables. Then it calculates the likelihood of a new instance belonging to a particular class. Finally posterior probability is calculated by taking product of prior probability and Likelihood[5]. The class with maximum posterior probability is assigned to the unseen instance.

In our case, the first phase of implementation had task of predicting whether a fire would breakout or not based on the weather condition for a given day. The class label was Yes and No. To perform the analysis task, Weka a Data mining and analysis tool was used. The prepared weather data set of Riverside County for year 2007 was loaded in Weka. The visualization of raw data showed that data values were distributed over a large number. So the data was discretized and all the attributed were put into two bins. After pre-processing, Nave Bayes was selected from the several data mining algorithm available. The default setting was chosen and instead of percentage split, 10 fold cross validation was used for building the classifier. Naive Bayes calculates the prior probability of known classes in as given: Prior probability of class Yes = Number of instance of Yes/ Total number of weather data Prior probability of Class No = Number of instance of No/ Total number of weather data. Later Nave Bayes calculates the Likelihood of a new instance to be part of Class Yes and No. Both the posterior probabilities are compares and the unseen instance is assigned a class with maximum posterior probability.

2. **J48 Decision Tree**

J48 is also a classification algorithm and hence belongs to supervised learning category. The algorithm takes a decision of assigning label to unseen data by using a top down tree construction. It analysis each and every attributes and stores its value on go. It goes on to constructing tree until it reaches it has exhausted all the attribute value and made a final conclusion. The result is usually stored in leaf node. The label assignment is dependent on value of other attribute and hence it is a dependent algorithm. The splitting of branches is calculated using entropy and information gain.

For implementation, the data that was preprocessed for Nave Bayes was used for applying J-48. The default setting were selected and using 10 fold cross validation, the algorithm was run. On completion, the decision tree was visualized. It had several branches and visualization was not very clear. Still the results of tree definitely indicated that temperature and humidity are the most important features while analyzing weather a fire would break out on a particular day.

4.3 Implementation for phase II

1. **Support Vector Machine**

   Support vector machine is a machine learning algorithm. It performs better than many other supervised learning. The reason behind using machine learning algorithms performing slightly better than other data mining algorithms is that it uses iterative approach in process of finding hidden patterns. This benefits the classification of new instance as more the classifier is learned, the better classification result it gives. As the name goes, Support Vector machine involves mapping the input data to a higher dimensional vector space model. It uses kernel function which is nothing but similarity function to find the optimal edge hyperplane. This makes sure that the error rate is minimal. It works well with a normalized data set. There are several kernel functions available and is chosen as per the suitability of data set. We have linear, sigmoid, Radial Basis function and Polynomial kernel function.

   The data set prepared for phase II was loaded to Weka. But it was noticed that data was not evenly distributed and hence all the attributed were normalized to a scale of (0,1). Since data set was nonlinear , all the three nonlinear kernel functions like Sigmoid, Polynomial and RBF were applied. Cost function was selected as 1. Gamma was kept at 0.25 and SVM type was selected as configuration for building the classifier. 10 fold cross validation method to iteratively build the model which would predict the size of fire by tagging it class of A,B,C,D,E,F.

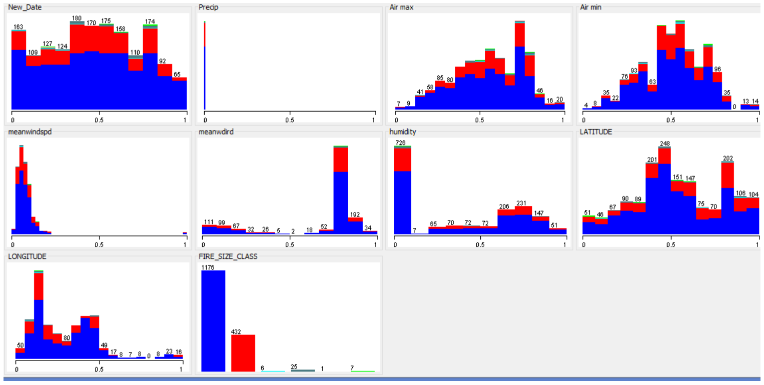

Figure 4.1: Visualization of raw data

2. **Articial Neural Network**

Artificial network comprises of many hidden layer and neurons. It uses these hidden layer and neurons for training a data set. The algorithm makes the model learn like human. It uses the feed forwarded back propagation method. So the input nodes are trained iteratively till it finds the minimum error. Weight is assigned to each node randomly. But the output node informs the hidden node about the inaccuracy and sends it back to calculate new weight. This is done to reduce the error rate.

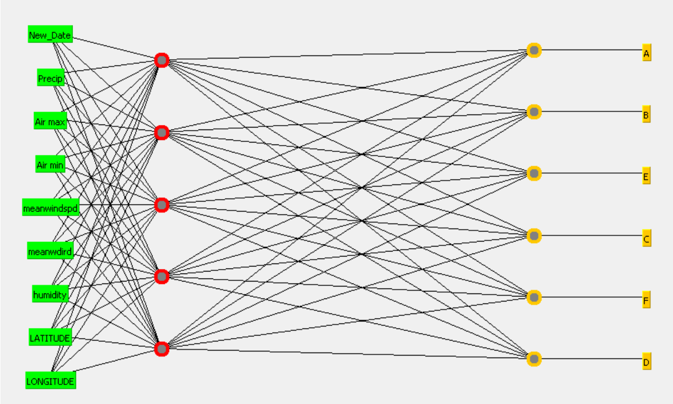

Figure 4.2: ANN with 5 nodes

It is said the neurons in a hidden layer should always be more than input nodes for better learning process of model. For our purpose, Multilayer perceptron algorithm was used for building the classifier.

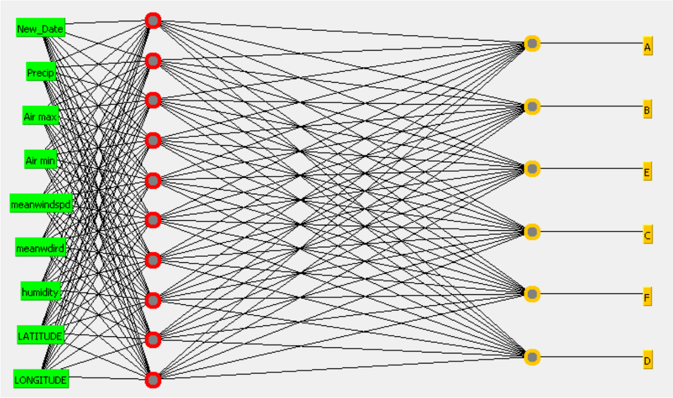

Figure 4.3: ANN with 10 nodes

The data that was normalized for running SVM, remained unchanged and was used for running ANN. ANN is dependent largely on values of number of neurons, hidden layer, learning rate and momentum. In our case, after making several variations, learning rate was chosen as 0.3 and momentum 0.2. Using 1 hidden layer gave best results. The data set was run by changing the number of neurons and keeping the hidden layer as constant. The values for neurons were 5, 10, 12 and so on.

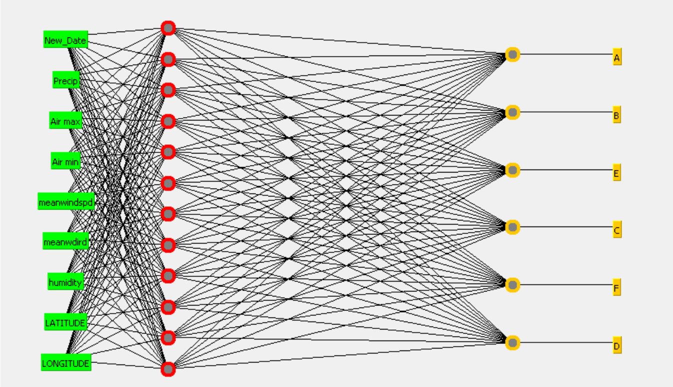

Figure 4.4: ANN with 12 nodes

Chapter 5

Results and Analysis

Performance measurement:

The performance of the classifier was measured using confusion matrix obtained after running the algorithms for 10 fold cross validation. For all the algorithm implemented, the fold count was kept constant so that it becomes easier to compare the result. It consists of correctly classified instances, and wrongly classified instances. Mainly it is made up of four variable, they are called false positive, false negative, true positive and true negative. These four variables play a really important role in determining the other performance evaluation variables which are accuracy, precision and recall. For analyzing the model, we also take into consideration kappa statistic and Area under Roc curve. Accuracy is nothing but correctly predicted instances. Precision is descried as instances that were supposed to be correctly classified and are correctly classified. Recall is instances that were correctly classified taking into considered TP and TN. Kappa is a measurement metric which simply tries to link together the correctly classified instances to the one that should have been the real output. Roc curve is used to plot the false positive against true positive.

5.1 Analysis for phase I

For phase I, classifier models were built by implementing J48 and Naive Bayes on data set that had just the weather variables like max temperature, minimum temperature, humidity, precipitation, wind, wind speed and a new binary class which had information if or not a fire broke on a particular day. Both J48 and Naive Bayes performed really well by outputting an accuracy of almost 90 percent. Such a high accuracy is difficult to achieve. On further analysis it was found that the weather data had more than 75 % of the instances belonging to class yes . So this resulted in data set being biased and we cannot establish the fact that the predicted class are being rightly assigned. Also precision and recall was recorded for both the algorithm. It is said that higher the precision better the model is. In our case precision was higher than recall for both J48 and Naive Bayes. For my first phase analysis both the classifier performed more or less the same with Naive Bayes being still better than J48. Table below gives an overview of how both the algorithms performed in terms of accuracy, precision and recall.May be along with weather data, if we had vegetation and soil data of forest, it would have produced better result.

Algorithm	Accuracy	Precision	Recall
J48	90.01 %	0.923	0.763
Naive Bayes	91.03 %	0.945	0.663

Table 5.1: Result for Phase I

5.2 Analysis for phase II

For the second phase of project Support vector machine and Artificial neural network was implemented. The aim was to predict the size of fire given the weather attributes and fire characteristics. Support Vector machine was run for three different kernel functions which included Sigmoid, Radial basis function and polynomial. The performance was measured in terms of accuracy, mean absolute error(MAE) and Root mean square error(RMSE). Sigmoid and polynomial resulted in accuracy of 71.5 % and 71.04 % respectively. The accuracy of radial basis function was better with 73 %. The results were almost similar to those of the authors Cortez and Morris[1].

Multilayer Perceptron is an Artificial Neural Network algorithm which was also implemented for second phase. It uses the concept of hidden layer and neuron. Different neuron and hidden layers were used to get better results. The result started getting better with just one hidden layer. So sticking to one hidden layer and increasing the number of neuron,

Kernel Function	RMSE	MAE	Accuracy
Polynomial	0.3087	0.1	71.4026 %
RBF	0.2	0.05	73.03 %
Sigmoid	0.30	0.009	71.5 %

Table 5.2: Result for Support Vector Machine(LibSVM)

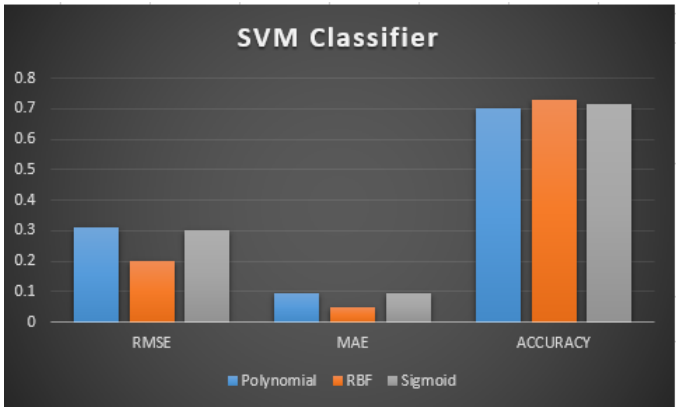

Figure 5.1: Support Vector Machine

various observations were made. Observations include accuracy measurement, RMSE and MAE With just five neurons, the classifier resulted in outcome of 69.5 % accuracy. With 10 the accuracy was 69.9 %. By increasing the count to 14, the accuracy decreased to 69.2 % and kept decreasing as the number of neurons was increased. So it was concluded that the algorithm performed best with just 10 neurons and one hidden layer. This time the input nodes were more than the attributes and algorithm performed better just in line with the study.

Hidden Nodes	RMSE	MAE	Accuracy
5	0.2624	0.13	69.52 %
10	0.2645	0.1273	69.9454 %
14	0.2667	0.128	69.16 %

Table 5.3: Result for ANN with different nodes

Figure 5.1 has performance comparison of SVM using three different kernel function. Figure 5.2 has comparison results for ANN run for different number of neuron. Fig 5.3

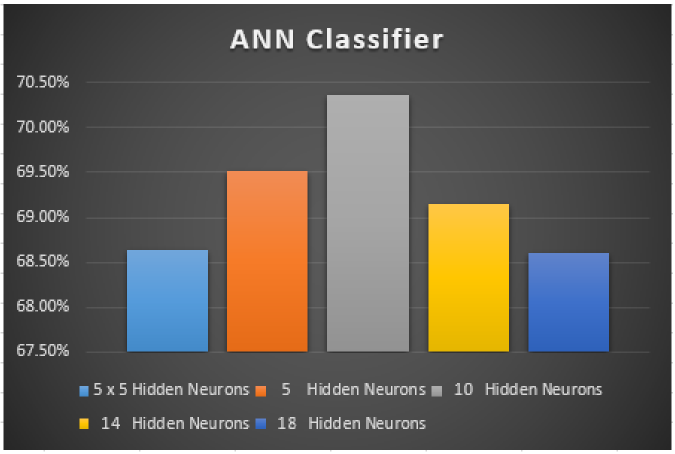

Figure 5.2: Artificial Neural Network

gives comparison graph for all the different algorithms that were run

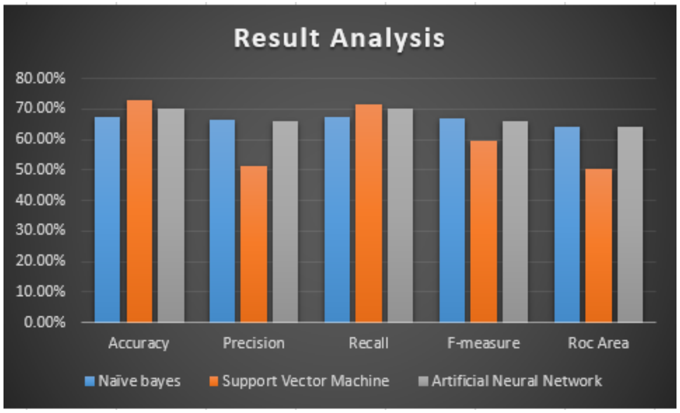

Figure 5.3: Comparison of Algorithms

Chapter 6

Conclusions

6.1 Current Status

In this project, five data mining algorithms were applied to predict the size of the fire for Riverside county of California state in the year 2007. Performance of a classifier is based on the dataset used to train the model. The first classifier was built to predict if a fire will break out given the weather conditions of a day. The model gave an outcome of 90 % accuracy which can be misunderstood as a good result initially. Later it was concluded that the model built after phase I was biased as most of the instances in the training set belonged to class yes and very less instances belonged to class no.

The classifier built post phase II implementation was used to evaluate the risk of fire based on the weather condition and fire characteristics of a given day. Model built using support vector machine gave an accuracy of 73 % which was the highest accuracy amongst all other algorithms employed. The results were analogous with the observation of authors Coretz and Morais [3]. Overall performance was moderate as the data set used to train the model was of California region and the weather of this region has hardly been exposed to any extreme winter days. So the model is able to predict the size of fire less accurately, if the given instance is of an extreme winter day. Clearly, the data set was not very versatile consisting of fire accidents of all the season. Weather conditions recorded from different parts of United States can be included for better results. Also it was observed that humidity and temperature were dominating feature I determining the size of fire. Also creating classes based on fire size can played important role in improving the performance as the machine learning algorithms work well in presence of less number of classes.

6.2 Future Work

This project can further extended to perform better by including weather conditions and fire instances of different parts of America. This way the model will be better trained and results will be better. Also we can have UI developed for the application to do some real time results. The workflow of the UI model can be, user enters the local and zip code may be. Using the zip code, we fetch latitude and longitude using some API and consuming those locations coordinates as parameters, get the weather condition like max temperature, min temperature, humidity, wind speed etc. for of the particular day. After we get that information, behind the scene the model built using the historical weather rand fire data can be utilized to test if a fire will breakout for the region one is trying to look up. And if so we can go ahead and predict the risk of breakout. This application will come handy for fire departments across the fire department as it could be made available just for them with some credentials. This will sure help fire fighters help curb the breakout and save them some time in controlling the fire spread.

6.3 Lessons Learned

This project has been a complete exponential learning graph. It started with learning about data collection. Since the data was not so easily available, the data had to gathered, a lot of research had to be put in. Also data was not in a readable format, since most of fire data was in GIS format. The historical weather data was not available either. It required consuming some APIs to get the data together. To club he historical fire data and weather data there was an issue of primary key initially. So finally the start date of fire was chosen as one to the club it. Learning involved dealing with different data formats, writing web service calls, using excel sheet, learning about data integrity like primary and foreign key.

Bibliography

[1] http://www.ipm.ucdavis.edu//.

[2] http://www.cs.wcupa.edu/rburns/DataMining/hw/example-project-S13.pdf//.

[3] P. Cortez and A. Morais. A data mining approach to predict forest fires using meteorological data. pages 512–523, 2007.

[4] Lisa M.; Miller Carol; Nelson Cara Parks, Sean A; Holsinger. Wildland fire as a self-regulating mechanism: the role of previous burns and weather in limiting fire progression. 2015.

[5] Lisa M.; Miller Carol; Nelson Cara R Parks, Sean A.; Holsinger. Fire atlas for the crown of the continent ecosystem (glacier national park, great bear wilderness, bob marshall wilderness, and scapegoat wilderness). 2015.

[6] K. C. Short. A spatial database of wildfires in the united states, 1992-2011. 2014.