Event Detection and Pattern Recognition using Support Vector Machine in Meteorology

Abhishek Jaswal
Department of Computer Science & Engineering
Jaypee University of Information Technology
Solan, India

Yashwant Singh
Department of Computer Science & Engineering
Jaypee University of Information Technology
Solan, India

ABSTRACT
Himachal Pradesh (India), having 85% of its population in rural areas produces 1 million tonnes of fruits yearly approximately, out of which 85-88% are Apples. But due to catastrophic nature of climate in the region, people across the state suffer severe damage to their crops causing huge financial loss every year. Anti Hail Guns firstly in India, were installed in Shimla district in particular apple growing areas, to prevent the damage to apples from Hailing. This Paper presents a credible Data Mining approach to discover the patterns and predict the formation of hails in clouds, so that guns can be fired accurately well in time reducing damage to fruit from hail and minimizing miss gun shots. Paper also presents the survey of work done in literature related to weather forecasting using data mining techniques. Support Vector Machine (SVM) a Data Mining algorithm have been implemented using different kernel functions upon 5 years weather data collected from the IMD (Indian Meteorological Department) Shimla to predict hailing. Performance and prediction accuracy of SVM with respect to all kernel functions and different size of training data have been analyzed to check the Sensitivity of the model.

Keywords
SVM, ANN, Hail, Anti Hail Gun, Prediction.

1. INTRODUCTION
Weather prediction has been one of the most fascinating and challenging domain since past. Weather is one of the most functional and powerful constraint, which affects our lives in many ways. Adverse effects of weather may cause a great loss to our lives, belongings and properties. During the last decade availability of climate data has increased tremendously due to enhanced technology. And it is important to find effective and accurate tools to analyze and withdraw hidden knowledge from this data, which can play a very crucial role in understanding the climate variability in future. As climate affects various sectors like agriculture, vegetation, water resources and tourism etc.

1.1 Anti Hail Gun
Frequent bad weather and Hailstorm events damaging a huge amount of crops particularly in Apple belts of Shimla district every year during last decade raised the alarm over the situation. Every year hailstorms destroys Apple crop worth crores of rupees. To address this alarming issue raised by farmers state government came up with *Anti Hail Gun* concept. *Anti Hail Gun* uses shockwaves to puncture the clouds formed over the region in order to prevent the formation of hails, and hence bring harmless rain instead of damaging hails. It can be used in all places where the hails can cause a severe damage. Ex-Automobile production unit.

1.2 Data Mining
The process of discovering interesting patterns from massive amounts of data. As a *knowledge discovery process*, it typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation, and knowledge presentation. The technique which one may use to get useful information from the available data depends directly to the domain of application, type of data and especially need of the customer/user. Data mining an application of machine learning, as Data Mining concept uses the machine learning algorithms in many different ways to tackle the problem of finding out hidden useful information from huge amount of raw meteorological data. [2, 18]

1.3 Machine Learning
Controls how computers can learn and boost their performance based on data. Main aim of a research aspect since a long time is to make computers such intelligent that they can make independent intelligent decisions in any situation without human intervention. They should implicitly learn to observe, identify complex patterns and make intelligent decisions based on the data of their own.

Types of Machine Learning:

1.3.1 Supervised Learning
Based upon the learning from available data and constructs a model, which classifies the new tuples to a particular class. Some examples of supervised learning techniques are SVM, decision tree [19] etc.

1.3.2 Unsupervised Learning
Implies clustering because initially no predefined classes are there in the data set. Clusters are built from the tuples which holds some similarity and after that user can map these clusters to a particular class. Commonly one uses clustering to identify the classes within the data. Unsupervised model built cannot tell us about the semantic meaning of the clusters identified, because training data is unlabeled.eg. KNN.

2. BACKGROUND STUDY
Available literature associated with data mining techniques applied in weather forecasting has been detailed technique wise below.

2.1 Support Vector Machine
SVM again a classification algorithm used for differentiating or classifying the data/tuples in to different classes based upon the maximum margin concept between the support vectors. Many kernel functions are used for

transforming the high dimensional data into different plane where it gets easy to partitions the data into two classes i.e. making data linear separable. Used for predicting max temperature and weather events coupled with regression technique aiming to enhance the performance or predicting power of the algorithm [21-23].A good choice of Kernel function is very important for effective SVM based classification. It transforms the non-linear separable data into high dimensional space where it becomes easy to separate the classes by maximum margin between support vectors. An appropriate Kernel function provides learning capability to SVM. Four types of Kernel Functions have been used for experimentation.

Table1. SVM Kernels

Types of kernels	**Full Name**	**Functions**
NP	Normalized Polynomial Kernel	K(x,y)=<x,y>/sqrt(<x,x><y,y>) Where<x,y>=Poly kernel<x,y>
PK	Poly Kernel	K(x,y)=<x,y>^p or (<x,y>+1)^p
PUK	Pearson VII Function Based Universal Kernel	$f(x) = H/[1 + \left(2(x - x_{\circ})\sqrt{2^{\frac{1}{\omega}}} - \frac{1}{\sigma}\right)^2]^{\omega}$
RBF	Radial Basis Function Kernel	K(x,y)=e^-(gamma*<x-y><x-y>^2)

2.2 Decision Tree

A supervised classification algorithm[19] which have different varieties in internal node splitting algorithm like ID3,CART,C4.5,Gini Index have been implemented by authors in predicting temperature, rainfall, evaporation, wind speed and weather events [3,8,9,11]. As it is supervised algorithm, previous years or day's data is used for training the classifier and then used for prediction purpose of unlabeled data.

2.3 K-Nearest Neighbor

KNN as name implies consider the K nearest numbers of already classified tuples in order to assign the label to newer one. Nearest measured by Euclidean distance. Other distance measures like Manhattan, Minkowski can also be used, but as errors are normally distributed thus solution by least square i.e. Euclidean distance gives appropriate results. Temperature, humidity, weather events have been predicted individually or by mixing with clustering technique. [4, 6]

2.4 Artificial Neural Network

ANN having some input layers, hidden layers and output layers is used for predicting the weather events by using the previous data(supervised) in order to adjust its weights associated with each of the neurons/nodes in hidden layers nodes, and running the model till the errors gets minimized to the already classified instances, termed as back propagation neural network and as Multi Layer Perceptron in Waikato Environment of knowledge Analysis (*WEKA)* [20] used for temperature, rainfall, wind speed prediction[5,7,8,13].

2.5 Clustering

Unsupervised technique used for grouping the data or instances based upon some similarity measures. As one is not aware in the start about the label for the tuples, so data is grouped by some similar feature and given a class name, and thus followed by adding the more tuples having same characteristics. In practical way can be understood as when a new species of animal or reptiles is found which have different characteristics to all others already known to human kind, a new name is given to that one and all other found in future having similar features are assigned to this new category. Used for temperature, humidity, cloudburst prediction [6, 10, 16] also exploited by mixing with KNN too.

3. METHODOLOGY

Steps followed initially from raw data to the evaluation of SVM classifier has been listed below

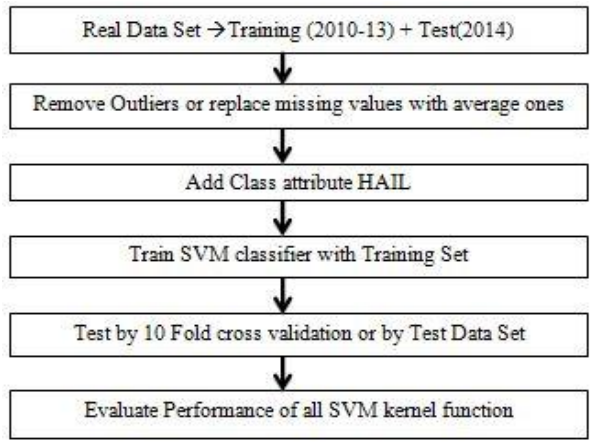


Figure1. Steps followed in retrieving knowledge from data set.

3.1 Creation of Training and Test Data Sets

5 years Data collected from IMD is partitioned into two sets for training and testing of classifier. 2010-2013(4 years) data set is clustered season wise i.e. Winter (December to March), Summer (March to June) and Rainy or Monsoon (June to September).2014 data cluster used for testing purpose. Outliers like vacant fields or unexpected values have been filled with average ones or removed.

3.2 Adding Class Attribute HAIL

Threshold values for all attributes of 3 clusters have been set after analyzing the various hail affecting parameters, past years frequency of hails in each month and how other parameters affects each other contributing to formation of hails [24].

Threshold limits for Class attribute are set as:

Winter Season

IF (MaxTemp<=17, MinTemp>=3.5, SealevelPressure>1610, Relative Humidity>=50 OR Wind Speed>0) then Hail=Yes

Summer Season

IF (MaxTemp<24, MinTemp>10, SealevelPressure>1560, RelativeHumidity>=55 OR Wind Speed>0) then Hail=Yes

Monsoon Season

IF (MaxTemp<23, MinTemp>14, SealevelPressure>1580, RelativeHumidity>=65 OR Wind Speed>0) then Hail=Yes

As Rainfall and Hail are only different form of precipitates which condenses at different temperature and altitude. Wind direction also not considered due to similar characteristics as of wind speed in the data set.

3.3 Parameters Observed

WEKA is used for implementing SVM and Results and Performances metrics like Accuracy, Root Mean Square Error (RMSE), Area under Receiver Operating Characteristics Curve (ROC), Precision, Recall, True Positive Rate (TP), False Positive Rate (FP) and False Negative (FN) have been observed for all kernel functions.

3.3.1 Accuracy:

Number of instances correctly classified /total no. of instances. Along with true positive rate of the data we have keep an eye on false positive rate also which will cause a false alarm and miss gunshots.

3.3.2 ROC Curve:

A plot that explains the performance of a classifier. Plotting True Positive Rate along y-axis and False Positive Rate along x-axis creates ROC curve. More the area under the curve more will be its performance.

3.3.3 Precision:

Fraction of instances correctly classified as +ve out of all instances the algorithm classified as +ve i.e. TP/ (TP+FN)

3.3.4 Recall:

Fraction of instances correctly classified as +ve out of all +ve instances i.e. TP/ (TP+FN).

4. RESULTS ANALYSIS

SVM is applied on two different sizes of data sets in order to check the *Sensitivity* of the SVM model. *Sensitivity* of the system refers how the output, results of system/model changes when corresponding input data is changed. Crucial method to analyze performance and scalability of the model built.

4.1 Training Data Set = 734 instances (March-August of 4years), Test Data Set=184 instances of 2014.

Table 2: Confusion Matrix

NP Kernel		Poly Kernel		PUK		RBF Kernel		Class HAIL
a	b	a	b	a	b	a	b	
16	3	16	3	16	3	0	19	a=Yes
3	162	2	163	1	164	0	165	b=No

Table 3: Accuracy, RMSE and ROC Curve area

	NP Kernel	Poly Kernel	PUK	RBF Kernel
Accuracy	96.7391	97.2826	97.8261	89.6739
RMSE	0.1806	0.1648	0.1474	0.3213
ROC Curve area	0.912	0.915	0.918	0.5

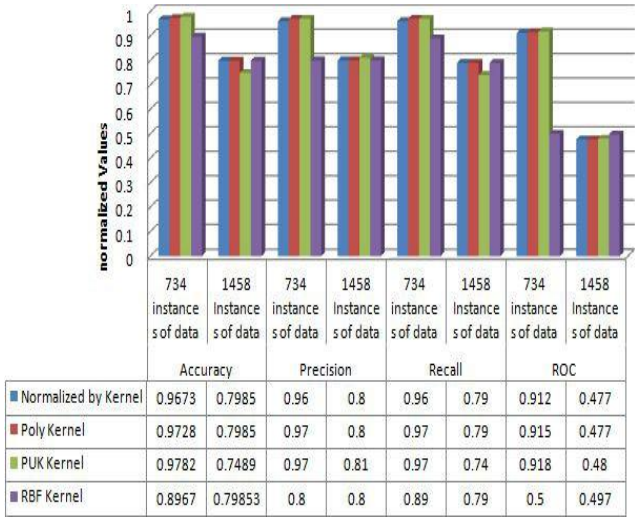
4.2 Training Data Set = 1458 instances (Seasonal wise of 4years) Test Data Set=273 instances of 2014.

Table 4: Confusion Matrix

NP Kernel		Poly Kernel		PUK		RBF Kernel		Class HAIL
a	b	a	b	a	b	a	b	
2	26	2	26	4	24	2	26	a=Yes
29	216	29	216	45	200	29	216	b=No

Table 5: Accuracy, RMSE and ROC Curve area

	NP Kernel	Poly Kernel	PUK	RBF Kernel
Accuracy	79.8535	79.8535	74.8938	79.8535
RMSE	0.4488	0.4488	0.5027	0.4488
ROC Curve area	0.477	0.477	0.48	0.47



	Accuracy 734	Accuracy 1458	Precision 734	Precision 1458	Recall 734	Recall 1458	ROC 734	ROC 1458
Normalized by Kernel	0.9673	0.7985	0.96	0.8	0.96	0.79	0.912	0.477
Poly Kernel	0.9728	0.7985	0.97	0.8	0.97	0.79	0.915	0.477
PUK Kernel	0.9782	0.7489	0.97	0.81	0.97	0.74	0.918	0.48
RBF Kernel	0.8967	0.79853	0.8	0.8	0.89	0.79	0.5	0.497

Figure 2.Comaprison of normalized values of Precision, Accuracy, Recall and ROC area for both different sizes of Training Data set.

5. CONCLUSION AND FUTURE WORK

After observing the simulated results and various performance metrics it is clear that variation in the size of input data affects the outputs and performance of classifier significantly. In small data size simulation PUK Kernel comes out to be best among others having large areas under ROC curve and all other parameters and when size of input data size gets large, PUK performance decreases while others gives better results. SVM gives better results with all its different kernel functions and can be deployed in real scenario and in future Wireless Sensor Network (WSN) can be deployed in the apple belt region to collect the weather attributes periodically in combination with SVM to predict the hailing and making installation of anti hail guns more worthy.

6. REFERENCES

[1] Fayyad, U., Shapiro, G., and Smyth, P. 1996. From data mining to knowledge discovery in databases. AI magazine 17, no. 3, pp. 37-54.

[2] Han, J., Kamber, M. And Pei, J. 2011. Data mining: concepts and techniques. Elsevier.

[3] Petre, E. 2009. A Decision Tree for Weather Prediction. Buletinul, Vol. LXI No. 1, 2009 pp.77-82.

[4] Jan, Z., Abrar, M., Bashir, S. and Mirza, A. 2008. Seasonal to inter-annual climate prediction using data mining KNN technique. In Wireless networks, information processing and systems, Springer Berlin Heidelberg, 2008, pp. 40-51.

[5] Kohail, S. And Halees, A. 2011. Implementation of Data Mining Techniques for Meteorological Data Analysis.IJICT Journal Volume 1 No. 3, 2011.

[6] S. Badhiye, S., Chatur, P.N. And Wakode, B. 2012. Temperature and Humidity Data Analysis for Future Value Prediction using Clustering Technique: An Approach. International Journal of Emerging Technology and Advanced Engineering, 2250-2459, Volume 2, Issue 1, January 2012.

[7] Baboo, S. and I Shereef,I . "An Efficient Temperature Prediction System using BPN Neural Network." International Journal of Environmental Science and Development 2, no.1, 2011, pp.49-54.

[8] Olaiya, F. and Adeyemo, A. 2012. Application of Data Mining Techniques in Weather Prediction and Climate Change Studies. I.J. Information Engineering and Electronic Business, 2012, pp. 51-59.

[9] Yeon, S., Sharma,S.,Yu, B., and Jeong,D. 2012.Designing a Rule-Based Hourly Rainfall Prediction Model. IEEE IRI 2012, August 2012

[10] Pabreja,K. 2012 .Clustering technique to interpret Numerical Weather Prediction output products for forecast of Cloudburst. International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 3 (1), 2996 - 2999, 2012.

[11] Prasad, N., Kumar, P. and Naidu, M. 2013. An Approach to Prediction of Precipitation Using Gini Index in SLIQ Decision Tree. 4th International Conference on Intelligent Systems Modeling & Simulation (ISMS), IEEE, January 2013 pp. 56-60.

[12] Saxena, A., Verma, N. and Tripathi, K. 2013.A review study of weather forecasting using artificial neural network approach. International Journal of Engineering Research and Technology, Vol. 2. No. 11 ESRSA Publications, 2013.

[13] Naik, A. And Pathan, S. 2012. Weather Classification and forecasting using Back Propagation Feed-forward Neural Network. International Journal of Scientific and Research Publicationsvol2 no.12, 2012.

[14] Kaur, G. 2012. Meteorological Data Mining Techniques: A Survey. International Journal of Emerging Technology and Advanced Engineering, Volume-2, Issue-8, August 2012,pp. 325-327,

[15] Ghosh, S., Nag, A., Biswas,D., Singh, J., Biswas, S., Sarkar, D. and Sarkar, P. 2011. Weather data mining using artificial neural network. In Recent Advances in Intelligent Computational Systems. IEEE 2011 pp. 192-195

[16] Kalyankar, M. and Alaspurkar, S. 2013. Data Mining Technique to Analyze the Metrological Data. International Journal of Advanced Research in Computer Science and Software Engineering 3(2), February –2013, pp. 114-118.

[17] Ganguly, A. and Steinhaeuser, K. 2008. Data mining for climate change and impacts. IEEE International Conference on Data Mining Workshops, 2008,pp. 385-394

[18] Japkowicz, N. and Shah, M. 2011. Evaluating Learning Algorithms: A Classification Perspective. First Edition, Cambridge University Press, 2011.

[19] Rokach, L. And Maimon, O. 2014. Data mining with decision trees: theory and applications" World scientific, 2014.

[20] Holmes, G., Donkin, A. And Witten, I. 1994. Weka: A machine learning workbench", Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, IEEE1994.

[21] Radhika, Y. and Shashi, M. 2009. Atmospheric temperature prediction using support vector machines. International Journal of Computer Theory and Engineering, 2009, 1(1), p.55.

[22] Rani, R. And Rao, T. 2013. An Enhanced Support Vector Regression Model for Weather Forecasting. IOSR Journal of Computer Engineering (IOSR-JCE), 2013, pp. 2278-0661.

[23] Rao, T., Rajasekhar, N. and Rajinikanth, D. An efficient approach for Weather forecasting using Support Vector Machines" In Proceedings International Conference Computer Technology Science (ICCTS), Vol.47, pp. 208-212.

[24] www.amssdelhi.gov.in/news events/ Shimla climate.pdf