Weather Analysis of Guntur District of Andhra Region using Hybrid SVM Data Mining Techniques

N.Rajasekhar, T.V.Rajini Kanth

Abstract*—In the recent years, weather prediction has drawn much attention for research community because it helps in safeguarding human life and their wealth. Apart from that, it is useful in effective prediction of natural calamities, agricultural yield growth, air traffic control, marine navigation, forests growth & military purposes. Literature studies shows that Machine Learning Algorithms proved to be good than the existing techniques / methodologies/traditional statistical methods. Hence development of new Hybrid SVM (Support vector machines) model is required for effective weather prediction by analyzing the given weather data and to recognize the patterns existing in it. SVM comes under the set of supervised learning methods for classifications & regression. It will be yielding good results in predicting the weather than the existing machine learning programming techniques. In this paper, Guntur district weather data sets were considered for analysis using the hybrid SVM data mining techniques.

Index Terms*— weather prediction, Machine learning, Data mining techniques, Hybrid SVM.

I. INTRODUCTION

Estimation of Weather is used in order to calculate the atmospheric pattern and also it is highly useful in predicting the weather status at desired locations for a stated future time period. Guntur district is located in Andhra Pradesh along the east coast of the Bay of Bengal. The district's coastline is approximately 100 kilometers. Guntur City is the largest city and is the administrative center of the Guntur District. It has 57 mandals. All the numerical data about the previous and current status of the rainfall is collected for the estimation of weather. It is
useful to study trends and patterns of weather in a scientific way for effective prediction. Appropriate existing weather prediction models are used to estimate weather based on the climatic conditions. The Automated expert systems are not available to pick up the best possible predictive models in order to forecast the atmosphere. It involves design of new mathematical models for addressing huge computational complexity which occurs due to uncertainty of atmospheric conditions apart from using the existing methods. These new models will help us in enhancing accuracy levels of estimation or prediction.

Manuscript received on April, 2014.

N.Raja Sekhar, Assistant professor, Department of Computer Science & Engineering, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad, India.

Dr. T.V. Rajini Kanth, Professor, Department of Computer Science & Engineering, Sri Nidhi Institute of Science & Technology, Hyderabad, India.

Effective Weather forecasting has lot of uses like providing early warnings about cyclones, tsunami, earthquake etc. in order to make public to save their lives and reduce the loss of property. These weather predictions also make farmers and traders to plan their respective crops and yields. The weather prediction is categorized based on time in to four types' namely Short range- forecast up to 48 hours, Extended-forecast from 3 – 5 days, Medium range–3 to 7 days and Long-range forecasts– more than 7 days [1]. We consider the rainfall data of three districts of Andhra Region. First we consider Guntur district data for study and analysis for effective weather forecasting.

II. LITERATURE SURVEY

K- means clustering algorithm [6, 7, 8] is certainly the most widely used clustering algorithm in practice. K-means clustering has been successfully applied in domains such as relational databases, gene expression data and decision support. The drawbacks of K-means includes choice of locations of centroids which is at random at the start of the algorithm, variables treatment as numbers and the clusters number K is unknown. Impact of the first drawback can be accessed through multiple runs or specific initialization methods. However, the specific initialization methods are not better than random centroids. There is a possibility to overcome the drawback of handling the categorical parameters by using the measure of matching dissimilarity. Concerning the third point, the number of clusters is an input parameter that is fixed as priori in the standard K-means algorithm. One way to address this challenge is through the use of cluster validity indices. As many other data mining algorithms, K-means has reduced reliability when treating high-dimensional data because data sets are nearly always too sparse. This is because the use of the Euclidean distance becomes meaningless in high dimensional sparse spaces. A solution involves combining K-means with feature extraction methods such as Principal Component Analysis (PCA) and Self Organizing Maps (SOM). But this clustering has taken place on the class attribute Station.

Support vector machines (SVMs) [2, 5] are a set of related supervised learning methods that create a decision-maker system which tries to predict new values. In a simple way, select a collection of training examples that belongs to either of the category, and this SVM algorithm develop a model in order to predict the given example to which category it falls. SVM model represents the examples as points in space, and mapped such a way that the different category examples are divided by a clear gap i.e. as distant as possible. The given examples to be predicted are then mapped into the category

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

that they belongs to based on which side of the gap they fall on, in the same space.

In a high dimensional space the SVM constructs a hyper plane or set of hyper-planes that can be used for data mining techniques like classification, or regression. The hyper plane with largest distance has a good separation even to the closest training data points that may belong to whatever class. Normally there will be smaller generalization of classifier error whenever there is higher margin. This technique can be understood as a combination of two steps:

Learning: It consists of examples with SVM training.

Prediction: It is required when ever results are not known by insertion of new samples.

Every example is represented as a pair of (input, output) in which data set is the input and how to categorize is denoted by output. In Mathematical format each example is represented as a pair (x, y), in this real numbers vector is denoted by x and Boolean value is denoted by y (i.e. y/n, 1/-1, T/F), or a real number. Out of two cases first one, is discussed as a problem of classification, where as the second one is a problem of regression. Moreover, the learning results are compared to the points denoted by pairs (x, y) representing a real function in the space where the interpolated function is generated by SVM interpolates these series of points to the best.

The "Training Set" consists of group of pairs that make up the SVM, where as in "Test Set" it contains the group of pairs for prediction. The data set is derived after clustering the data set

III. PROPOSED APPROACH

The realistic data sets that are collected are in raw format. In first step they have to be converted in to experimental suitable data[3] sets format. K-means clustering technique will be applied on the data sets and successively apply SVM classification technique over that clustered data set. This SVM hybrid technique [4] is suitable for effective analysis of weather data.

IV. IMPLEMENTATION OF THE PROPOSED APPROACH

K-means clustering was applied on the data set for about 102 years namely Monthly Mean for each year Average Temperature (1901-2002). K-means clustering algorithm was applied on the data set [3] Monthly mean for each year Average Temperature and the Clustering model is on full training set. There are 13 attributes namely year, Jan, Feb, Mar, April, May, June, July, August, Sep, Oct, Nov and Dec. Euclidean distance was applied for making clusters. The number of iterations is 8 and within cluster sum of squared errors is 34.48144163072698. Time taken to build model (full training data) is 0.02 seconds. In this, missing values are globally replaced with mean/mode. Table 1 is representing the cluster centroids for about 5 clusters. Total number of Instances is 102. Cluster 0 has 32 (31%), Cluster 1 has 14(14%), Cluster 2 has 19(19%), Cluster 3 has 18 (18%) and Cluster 4 has 19(19%) number of instances. Fig.1 shows cluster graph of Monthly Mean for each year Average temperature with Instance number along x-axis and year along y-axis.

SVM Classifier for Monthly Mean for each year Average Temperature: SVM classification was done with 15 attributes namely Instance_number, year, names of 12 months and Clusters. The total number of Instances is 102. It was evaluated on total training set. LibSVM wrapper, actual code by Yasser EL-Manzalawy (= WLSVM). The time taken to build the model is 0.25 seconds and 0.02 seconds to test the model on training data set. Agreement between the classifications and the true classes is represented by Kappa Statistic. It is a Chance-corrected measure. It is measured by considering expected agreement by chance away from observed agreement and divides it by highest possible agreement. If the value is more than 0 it means that the classifier is doing better. The Kappa statistic is near to 1 so it is near to perfect agreement. The kappa statistic measures the agreement of prediction with the true class 0.9368 signifies almost there is complete agreement. The correctly classified instances are 97 and incorrectly classified instances are 5. That means it has classified perfectly. There is no considerable error most of the errors are near to zero except the root relative squared error. Fig.2 shows SVM classifier graph of monthly mean for each year Average temperature. In this graph Clusters were taken along x-axis and predicted cluster along y-axis. The Annual Mean values of Minimum, Average and Maximum temperatures for all the years 1901-2002 are shown in Fig.3 along with their trend line Polynomial equations. In this month's are taken along x-axis and temperature along y-axis. The equation for Mean Minimum Temperature is given by

$$y = 0.000x^6 - 0.023x^5 + 0.389x^4 - 3.170x^3 + 12.58x^2 - 19.80x + 28.19 \text{ ---- (1)}$$

Equation for Mean Average Temperature is given by

$$y = 0.000x^6 - 0.027x^5 + 0.454x^4 - 3.668x^3 + 14.26x^2 - 21.85x + 34.70 \text{ --- (2)}$$

Equation for Mean Maximum Temperature is given by

$$y = 0.000x^6 - 0.031x^5 + 0.519x^4 - 4.177x^3 + 16.00x^2 - 24.05x + 41.38 \text{ --- (3)}$$

The regression equation [9] for Annual Mean values is given by

Mean Avg. temp = Mean cloud cover * 0.01581 + Mean Diurnal temp.range * -0.6906 + Mean Ground frost Frequency * -2.025 + Mean Precipitation * 0.01491 +Mean Vapour Pressure * 0.07776 +Mean Wet day frequency * -0.4255 +Mean poential Evapotranspiration * 1.991 +Mean reference crop evapotranspiration * 1.502 + 13.83 --- (4)

V. CONCLUSION

In the K-means cluster analysis given by Table.1 cluster 0 with highest number of instances has 1925 year as the centroid with low temperature 31.04860c in the month of April. Cluster3 has highest temperature values in all months except Apr, May and Jun with 18 instances. Cluster2 with number of instances 19 has highest temperature values in the three months Apr, May and Jun. Detailed accuracies are given by Table.2 indicates that cluster 0 has TP rate as 1 and F-measure as 0.941. Cluster2 and cluster4 has same TP rate value as 0.947 and F-measure values are 0.947 and 0.973 respectively. It tells that cluster 0 was classified accurately followed by cluster2 and cluster4. In the confusion matrix represented by the Table.3 each column represents instances in the predicted class and each row indicates instances in actual class. The classifier has classified accurately for the

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

cluster0, cluster2, cluster3 and cluster4 but less accurate in case of cluster1 indicated by confusion matrix. In cluster 1 out of 14 instances 2 were classified under cluster0. Only cluster 0 has classified accurately. Fig.4. gives the following conclusions or inferences in which months are taken along x-axis and different weather parameter values are taken along y-axis. There is a correlation between mean precipitation and mean cloud cover its value is 0.754384589. Likewise there is a correlation between average temperature and mean vapour pressure and is 0.780409787. There is a correlation between mean diurnal temperature and mean potential evapotranspiration and is 0.686833118. There is negative correlation between Mean ground frost frequency and mean wet day frequency and its value is -0.840915846. Meaan average temperature has correlation with all the parameters namely mean cloud cover, diurnal temperature range, ground frost frequency, precipitation, vapour pressure, wet day frequency, potential evapotranspiration and reference crop evapotranspiration. It has highest correlation with reference crop evapotranspiration and negative correlation with ground frost frequency. The future scope is this Hybrid SVM technique can be extended to any weather data for further analysis based on various weather parameters.

Cluster Centroids:

Table 1: Cluster Centroids of Monthly mean for each year Average temperature.

	Cluster #					
Attribute	**Full Data (102)**	**0 (32)**	**1 (14)**	**2 (19)**	**3 (18)**	**4 (19)**
Year	1951.5	1925.375	1933.071	1970.737	1990.389	1953
Jan	23.7332	23.5988	23.046	24.1585	24.251	23.5504
Feb	25.7257	25.3925	24.9439	26.3476	26.5412	25.4684
Mar	28.4904	27.8745	27.8522	29.1598	29.3087	28.5533
Apr	31.3154	31.0486	30.4034	31.851	31.8018	31.4402
May	33.0835	33.0241	32.3386	33.7081	33.0784	33.1124
Jun	31.3508	31.5223	30.7781	31.7839	31.4556	30.9515
Jul	29.1686	29.1733	28.9411	29.3551	29.7928	28.5503
Aug	28.6422	28.7749	28.2217	28.8192	28.9369	28.2722
Sep	28.4556	28.2975	28.3962	28.6719	29.2188	27.8264
Oct	27.3615	27.2112	27.5579	27.3945	27.9018	26.9252
Nov	24.9237	24.6118	25.2622	24.9729	25.9337	24.1935
Dec	23.3658	23.1157	23.5789	23.6323	24.0607	22.7049

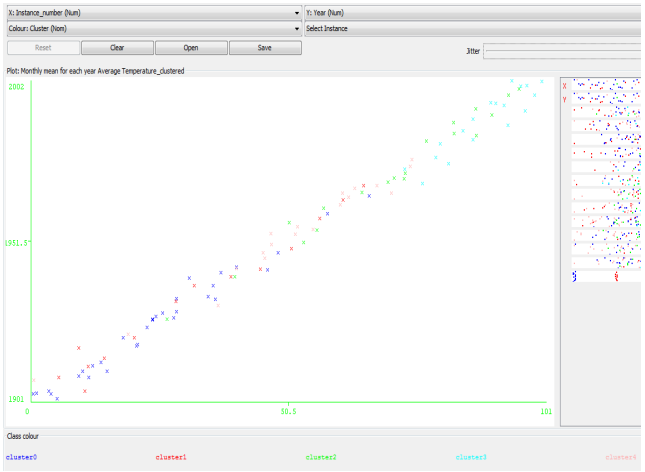

Fig.1: Cluster graph for Monthly Mean for each year Average Temperature

Table2: Detailed accuracy of the clusters by class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.057	0.889	1.000	0.941	0.915	0.971	0.889	Cluster0
0.857	0.000	1.000	0.857	0.923	0.915	0.929	0.877	Cluster1
0.947	0.012	0.947	0.947	0.947	0.935	0.968	0.907	Cluster2
0,944	0.000	1.000	0.944	0.971	0.966	0.972	0.954	Cluster3
0.947	0.000	1.000	0.947	0.973	0.968	0.974	0.957	Cluster4
0.951	0.020	0.955	0.951	0.951	0.938	0.965	0.915	

Table 3: Confusion Matrix

a	*b*	*c*	*d*	*e*	<-- *classified as*
32	*0*	*0*	*0*	*0*	| *a = cluster0*
2	*12*	*0*	*0*	*0*	| *b = cluster1*
1	*0*	*18*	*0*	*0*	| *c = cluster2*
0	0	1	17	0	| d = cluster3
1	0	0	0	18	| e = cluster4

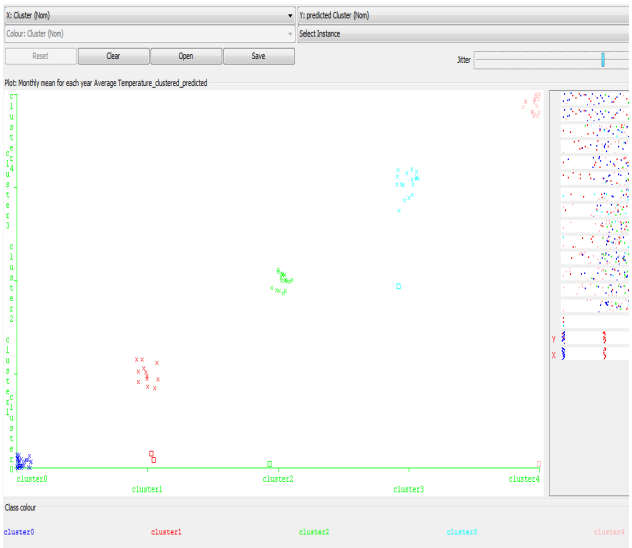

Fig.2: SVM classifier graph of monthly mean for each year Average temperature

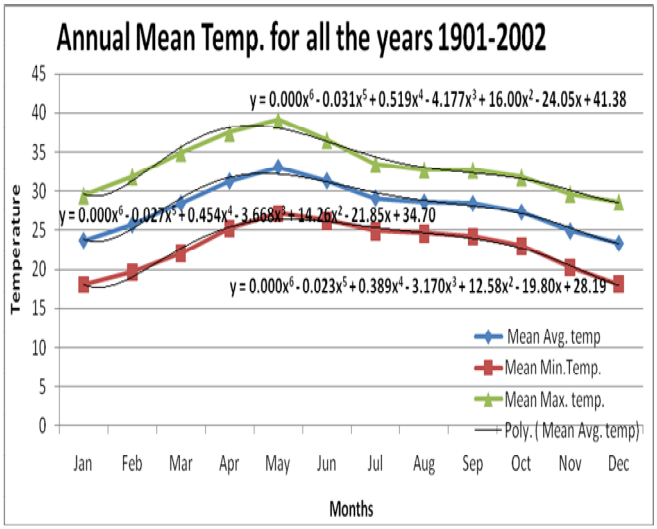

Fig.3: Graph of Annual Mean Temperatures for all the years 1901-2002 along with Trend line equations

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

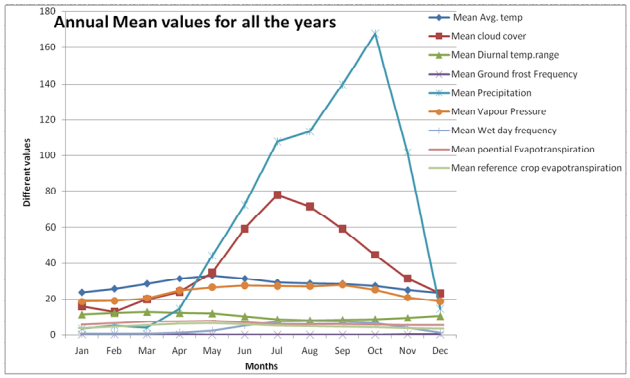

Fig.4: Graph of Annual mean values of weather parameters for all the years

REFERENCES

[1] http://www.timeanddate.com/weather/forecast-accuracy-time.html

[2] http://en.wikipedia.org/wiki/Weather_forecasting

[3] Folorunsho Olaiya, Application of Data Mining Techniques inWeather Prediction and Climate Change Studies, I.J. InformationEngineering and Electronic Business, 2012, 1, 51-59, Published OnlineFebruary 2012 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijieeb.2012.01.07

[4] Dr. M.H.Dunham, Companion slides for the text, "Data Mining,Introductory and Advanced Topics", Prentice Hall, 2002.

[5] Y.Radhika and M.Shashi, "Atmospheric Temperature Predictionusing Support Vector Machines" International Journal of ComputerTheory and Engineering, Vol. 1, No. 1, April 2009 1793-8201

[6] Dr.T. V. Rajini Kanth, Ananthoju Vijay Kumar, Estimation of theInfluence of Fertilizer Nutrients Consumption on the Wheat Crop yield in India- a Data mining Approach, 30 Dec 2013, Volume 3, Issue 2, Pg.No:316-320, ISSN: 2249-8958 (Online).

[7] Dr.T. V. Rajini Kanth, Ananthoju Vijay Kumar, A Data Mining Approach for the Estimation of Climate Change on the Jowar Crop Yield in India, 25Dec2013,Volume 2 Issue 2, Pg.No:16-20, ISSN: 2319-6378 (Online).

[8] A. Vijay Kumar, Dr. T. V. Rajini Kanth "Estimation of the Influential Factors of rice yield in India" 2nd International Conference on Advanced Computing methodologies ICACM-2013, 02-03 Aug 2013, Elsevier Publications, Pg. No: 459-465, ISBN No: 978-93-35107-14-95.

[9] Dr.David, B.Stephenson, Data analysis methods in weather and Limate research Department of Meteorology University of Reading, July,20, 2005 ,http://www.met.rdg.ac.uk /cag/courses/

N.Rajasekhar did B.Tech in Mechanical Engineering from NBKRIST, Vidyanagar, Nellore and M.Tech in Computer Science & Engineering from Bharath University, Chennai, INDIA, in 2006. Currently pursuing Ph.D from Acharya Nagarjuna University, Guntur. He is presently working as Assistant Professor in the Department of Computer Science & Engineering VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad, India. He is a professional member of ISTE.

Dr.T.V.K.Rajinikanth received M.Tech from Osmania University, Hyderabad in 2001, PhD from Osmania University, and Hyderabad in 2007. He is currently working as a Professor, Department of Computer Science & Engineering, Sri Nidhi Institute Of Science & Technology, Hyderabad, India. He has written and attended several National & International publications and conferences. His research interests include Data warehouse & mining, Semantic Web, Spatial Data mining, ANN. He is a professional member of CSI and ISTE.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.