Rainfall Forecast using an Effective Machine Learning Model

Numan Khan, Shubham Kumar

Computer Science and Engineering, Institute of Engineering and Management, Kolkata, West Bengal, India

ABSTRACT

Interfacing with the ever-increasing volumes of data in the sectors of technology, medicine, science, engineering, industry, and finance and transforming them into a comprehensible format. One of the primary criteria is a human user. Information-intensive applications will demand effective methods and the ability to learn from fresh data in order to swiftly identify and evaluate intricate patterns and requirements. Classification and clustering of widely accessible data is one way to address this. Within this research, we suggested a two - tier method for clustering big data sets for rain fall data prediction using SOM and SVM with ID3 in order to partially satisfy the market demand. This research examines a novel method for grouping the SOM and SVM with ID3. Specifically, application of agglomerative hierarchical clustering and ID3-participated clustering is examined. When compared to straight clustering of the data, the two-stage process, which first uses SOM to create the illustration and then examines SVM with ID3, performs well and cuts down on computation time.

KEYWORDS: *Rainfall prediction, SOM, SVM, ID3, ANN, Clustering, Forecasting, Entropy, Data mining, Weather data*

How to cite this paper: Numan Khan | Shubham Kumar "Rainfall Forecast using an Effective Machine Learning Model" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-3, June 2025, pp.436-441, URL: www.ijtsrd.com/papers/ijtsrd79960.pdf

IJTSRD79960

INTRODUCTION

Databases have expanded at unprecedented rates. Executing advanced perception out of large databases manually is difficult, time-consuming, and costly. There is no hope when data pass certain size and complexity limits. It is for this reason that scientific research has been focused on automatic analysis and presentation of large multi-dimensional data sets during the last decade. The main aim is to uncover pat terns and relationships in the data in order to uncover hidden but potentially valuable knowledge. One of the most promising areas in this vast field is artificial neural networks. Initiated by biomedical science advances, they form a family of algorithms that try to simulate the brain's neurological structure. One of the most widely used neural networks and one of the most in-demand unsupervised learning algorithms is the Self-Organizing Map (SOM).The maps have been used to great success in a great range of research areas, from financial analysis to speech recognition, and they map the natural sets and relations in the data to perfection. The Self-Organizing Map and SVM are two great tools in information visualization. Although there is still some mathematical analysis of the method, its basic implementation is simple, and the metaphor on the map is simple to understand. Moreover, the method scales very well, and the result is impressive. Visualizing and comprehending SOMs and SVMs has been an effective strategy in a broad spectrum of studies and 2 applications. Supervised learning models like SOM and SVM in machine learning that have identical learning methodologies for regression analysis and classification that consider data and determine patterns. A model that classifies new cases to either of the categories is defined by a SOM and SVM training algorithm, which generates a deterministic binary linear classifier from a set of training examples given that are various individually labelled as favorable and comprehending to among two classes. Visualizing SOMs and SVMs has been an effective strategy in a broad spectrum of studies and applications. In machine learning, the super vised learning models SOM and SVM have corresponding learning methodologies that learn data and determine patterns for regression analysis and classification. A SOM and SVM training methodology encapsulates a model that classifies new cases to either of the categories. It generates a non-probabilistic binary linear classifier from a given set of training examples that are each individually labelled as favorable to among two classes.

RELATED WORK

Many computer applications rely on data analysis, either during the design stage or during online operations. Procedures for data analysis can be divided into two catego ries: exploratory or confirmatory, depending on whether suitable models for the data source are available. However, a crucial component of both kinds of processes is the classification or grouping of computation according to either (i) the measurements' goodness-of-fit to a hypothesized model or (ii) naturally occurring assembling (clustering) that are discovered through analysis. Organizing a group of patterns into clusters according to similarity is known as cluster analysis.

The activity of searching and investigating the voluminous data to identify rational, unique, potentially beneficial, and eventually comprehensible patterns is referred to as data mining. Gradually Scanning and exploring databases whose complexity, dimensionality, and volume of data beyond the limit of manual scrutiny is now achievable due to powerful computer technology. The aim is to uncover trends, relationships, or patterns that allow us to formulate fresh insight and understanding of the data. Among the numerous algorithms used in this sporadic example are mining, SOM, and SVM. These are more specifically categorized as neural network algorithms. This is a class of algorithms that are based on analogies to the brain's neural structure. The SOM in ex clusivity was inspired by an interesting phenomenon: Doctors have found that certain brain tissue areas can be controlled in reaction to an input signal. Essentially, this ordering process is imitated by a computer program referred to as SOM and SVM.

Artificial neural networks are used in investigative data mining to provide another dimension to data mining, specifically methods focused on classifying and clustering data. We employ neural networks as a 3 practical data mining method that may provide models and statistical insights from big data sets, as well as how Self-Organizing Only data clustering applications may effectively use, an unsupervised learning neural network paradigm and Kohonen Maps, for data mining. We demonstrate that it is possible to estimate high-dimensional data to a lower dimension and to cluster data together while maintaining the essential information. Kohonen Map-based clustering algorithms were used in 1991 by the World Bank to display two-dimensional maps from multidimensional data sets as well as group of related data items. Due to its significant visualization qualities, the SOM is only appropriate for data surveys. It creates a collection of prototype vectors that show the information. establish and complete a topology-protected projection of the proto types onto a low-dimensional grid from the d-dimensional input space. This well-organized grid can serve as a practical surface for visualizing various SOM (and thus data) properties.

Self-organizing map's components: Sample data is the initial component of SOM, and weight vectors are the second.

Data Sample: Initial section of SOM manages with the data, like three-dimensional data, which are frequently employed in SOM experiments. Here are the three colors red, blue, and green are represented by dimensions. SOM's primary goal is to give n-dimensional data the greatest possible display.

Weights: There are two components in every weight vector. Its data is represented in the first section, while the natural location is represented in the second.

Speech Recognition: Creating objective metrics to improve voice quality is the primary goal of SOM in speech-recognition. Apply Teuvo Kohonen's groundbreaking work on Self-Organizing Maps in speech recognition.

Control Engineering: In engineering applications, SOM serves as a look-up table. Similar rows in a specified input pattern are returned by the look-up table. We identify the best matching unit in a map by employing the learning phase of the algorithm.

Applications for Self-Organizing Maps: The Self-Organization Map made use of a number of Listed below are a few application areas.

Biomedical Sciences: Self-Organizing Maps have made a significant contribution to the emerging field of bioinformatics study.

Financial Analysis: A number of authors, including Kaski et al., have utilized the SOM to look up business financial accounts and reveal connections between a company's type and insolvency risk.

Being a statistical method that solves the problem in the same way as ANN,SVM are linear machines. Structural risk minimization, or SRM, is used to model well being on unobservable data. Structural Risk Minimization is exploit in the same manner by means of SVM. On the positive side, it has all the advantages of ANN, as well as on some of the inherent issues in ANN implementation outlined (ASCE Task Committee, 2000a, b). The present paper merely asserts the advantages of SVM over ANN. Re searchers are now interested in SVM classifiers and their uses; more recently, regression analysis and time series forecasting. SVM has been applied by Mukherjee et al. (1997) for non-linear prediction of chaotic time series like the Mackey Glass, Ikeda, and Lorenz time series, and compare results applying a variety of similar methods . Finally, SVM was superior to the other methods in chaotic time series.

According to Dibike (2000), SVM is superior to ANN and genetic programming in rainfall runoff modeling. SVM performed better than ANN in Venice city for forecasting water level for a 12-period horizon, concluded Babovic et al. (2000). Siva Pragasm & Liong (2000) and Liong & Siva Pragasam (2000) concluded.

that SVM performs better in flood fore casting and rainfall-runoff modeling.

Classification of text (and hypertext): We transform natural text into a predetermined, fixed number of categories based on its content. Topic-based document sorting, email filtering, online search, etc. This can be seen as a sequence of binary classification problems, one for each category, as a document can belong to more than one category.

- Hand-written character recognition
- Image classification
- Bioinformatics (Protein classification, Cancer classification)

Basic clustering methods are ineffective for large, multi-level datasets. To handle this, we use SOM and SVM with ID3, along with a statistical model that supports multi level structures and distribution mixtures. Hierarchical SOM, with its structure defined through unsupervised training, is central to our approach. It is well-suited for tasks like weather forecasting that require hierarchical clustering.

SYSTEM STRUCTURE

Clustering algorithms are a key aspect of data mining and represent a major unsupervised learning challenge. A cluster groups similar items together while distinguishing them from other groups. In this study, we illustrate a real-world prediction method used in India, as shown in Fig. 1. It has two parts: one predicts the city's weather for the next five days (temperature, humidity, wind, rain, etc.), and the other provides warnings about extreme events like cyclones or thunderstorms. The dataset was collected and pre-processed from various meteorological sources in India and regions like Indonesia.

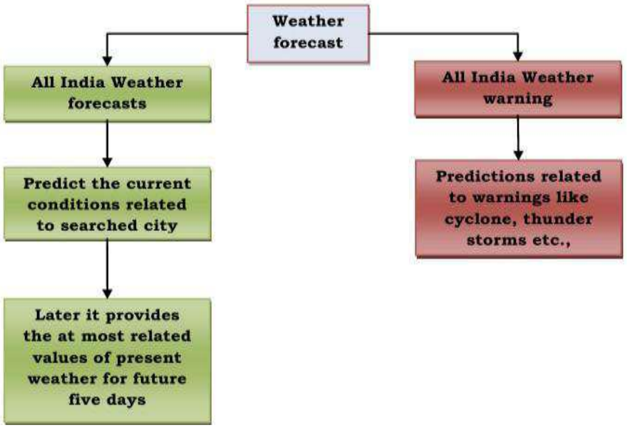
Weather forecast
All India Weather forecasts
Predict the current conditions related to searched city
Later it provides the at most related values of present weather for future five days
All India Weather warning
Predictions related to warnings like cyclone, thunder storms etc.,

Fig. 1 Weather prediction actual approach

Our research employs a two-level clustering method: Self-Organizing Maps (SOM) for first-level unsupervised learning and an SVM algorithm with ID3, as presented in Figure. 1. We also use Artificial Neural Networks (ANNs), which can handle noisy and imprecise data. ANNs have future potential for data mining exploration, but we utilize them together with SOM and SVM as exploratory data analysis methods. SOM and SVM with ID3 assist in reducing high-dimensional data to a 2D map, in which clusters can be visualized and comprehended more easily. For accurate predictions, we utilize the Global Framework Model with SVRK (Support Vector Regression Kernel Agent).

SOM Algorithm -:

The algorithm begins the map's weights using different methods. The first uses random values unrelated to the training data, making convergence slow. The second uses random samples from the training data, offering better initial alignment and faster training. However, due to randomness and fewer map nodes, it might still miss important data patterns. There's also a risk of selecting outliers or overlooking them entirely.

SVM Algorithm -:

Similar to the Direct SVM method, the algorithm begins with a close pair of points from opposite classes. It then greedily adds a violating point from the dataset to the candidate set. If existing Support Vectors block the addition of this point, it is removed. To ensure linear separability in kernel space, a quadratic penalty is applied. Finding the closest pair requires n2n^2n2 kernel com- putations, but with kernel the exponential kernel which is a distance-preserving, this step can be skipped. The goal is to add a new Support Vector, ccc, to the existing set SSS of Support Vectors.

ID3 Algorithm -:

ID3 makes use of a fixed number of examples to induce a decision tree. The induced tree will categorize future samples. This includes numerous feature and is an example of a class (e.g., yes/no). The name of the class is stored in the leaf nodes of the decision tree, while a decision node is a non-leaf node. Each of the branches relates to a possible value for the attribute, and the decision node is an feature test. Learning a decision tree has the advantage that a computer may learn from an expert instead of a knowledge engineer. Fig.3 depicts this basic structure. The central ideas of the ID3_ algorithm are:

- Every non-internal node in a decision tree is the feature, and the value of an attribute is given by its edge. When the path originates from root node to the leaf node defines the input attributes, the leaf node is the desired value of the output attribute.

- Every non-leaf node of a "good" decision tree is an attribute which, out of all input attributes not used so far, contains most information regarding the output attribute.

Training Set and Data: ID3 is a supervised learning algorithm that requires training data to make decisions. It works with categorical data, so continuous values must be discretized. The dataset includes five attributes: outlook, temperature, humidity, wind, and rain. Using these, we've developed a weather forecasting mechanism, as shown in Tables 1 and 2.

Table 1: Data attributes

	VALUES
Temperature	85
Outlook	sunny
Humidity	85
Rainy	NO
Windy crop	False

Table 2: Actual data

Weather	Temp.	Humidity	Wind	Crop
Rainy	70	96	False	Yes
Sunny	85	85	False	No
Overcast	64	65	True	Yes
Rainy	68	80	False	Yes
Sunny	80	90	True	No
Overcast	83	86	False	No
Rainy	65	70	True	No

Entropy -:

Starting with a root node, a decision tree is constructed top-down by dividing the data into subsets that comprise homogeneous (similarly valued) occurrences. The ID3 algorithm determines a sample's homogeneity using entropy. When the sample is evenly distributed, its entropy is one, and when it is entirely homogenous, it is zero.

$$E(S) = -(p+) * \log 2\,(p+) - (p-) * \log 2(p-) \ldots\ldots.(1)$$

"S" stands for the data, where the numbers "p+" and "p-" indicate how many elements in the set "S" have positive and negative values, respectively. The ID3 algorithm's objective is to use decision trees for data classification in a way that produces homogenous, zero-entropy leaf nodes.

Gain -:

The reduction in entropy following the splitting of a dataset according to an attribute serves as the basis for the information gain. Finding the feature that yields the greatest information gain that is, the most homogenous branches is the key to building a decision tree. It is calculated as

$$\text{Gain}(S,A) = \text{Entropy}\,(S) - S((|Sv| / |S|)\ * \text{Entropy}(Sv)) \ldots\ldots(2)$$

where 'A' is the attribute and 'S' is the set. The subset of S' having attribute A' equal to v' is denoted as SV. The cardinality of set 'S' is denoted as |S|', and the number elements of the subset "Sv". ID3 chooses the attribute with the largest gain to con- struct nodes in the decision tree. ID3 chooses one of the other attributes to create more nodes until all of the subsets become homogeneous if the created subsets are not of entropy zero or equal to zero.

Table 3: Data

If temp. = 71 then Prediction = No
If temp. = 85 then Prediction = Yes
If temp. = 81 then Prediction = No
If temp. = 80 then Prediction = No
If temp. = 64 then Prediction = Yes
If temp. = 70 then Prediction = Yes
If temp. = 83 then Prediction = Yes
If temp. = 65 then Prediction = No
If temp. = 68 then Prediction = Yes

RESULT

This work demonstrates the effectiveness of SOM and SVM by providing a process for automatically learning a recognition model. There are numerous possible uses for gesture recognition. This technique is used on high-dimensional rainfall data. Here, we examine the various predictions for whether it will rain or not, which are displayed in Fig. 5, using the SOM and SVM with ID3 algorithms. Tables 3 and 4 display the attributes that are taken into consideration in this approach. We use rainfall data for experiments. The input data set consists of numerical data values with variables like temperature, humidity, dew point, sea level pressure, and rainfall at that hour. Forecasting rainfall happens at case, yes case, and if so, at what level (in millimeters)? Table 5 displays this, and the graph in Figure 2 illustrates how the prediction is only dependent on temperature and breezy %.

Table 4: Results

TEMPERATURE	PREDICTION
72	No
80	No
69	Yes
85	No
83	Yes
70	Yes

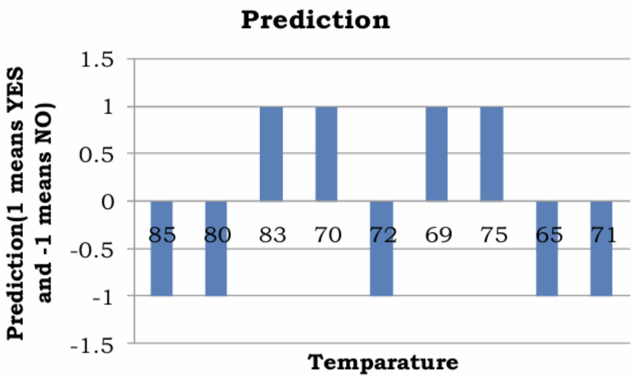

Fig 2. Prediction results

Table 5: Actual data set

Day	Weather	Temp.	Humidity	Wind	Crop
1.	Cloudy	83	78	False	Yes
2.	Cloudy	72	90	True	Yes
3.	Rainy	71	85	True	No
4.	Sunny	85	85	False	No
5.	Cloudy	81	75	False	Yes
6.	Sunny	80	90	True	No
7.	Rainy	70	96	False	Yes
8.	Sunny	75	70	True	Yes
9.	Rainy	68	80	False	Yes
10.	Rainy	75	80	False	Yes
11.	Rainy	65	70	True	No
12.	Sunny	69	70	False	Yes
13.	Cloudy	64	65	True	Yes
14.	Sunny	72	95	False	No

The real suggested prediction is obtained after using Table 4 as the input data set and our suggested method (SOM, SVM with ID3). We can plainly see that the accuracy is around 82% when we compare Table 5 with the actual data set and

Table 6 with the forecast values. In actuality, this is forecasted based on the strongest winds, temperature, and humidity. According to the research and application done here, the algorithm ID3 of decision tree performs admirably in any category task including datasets having discrete values. According to the study, the classification tree constructed with the ID3 algorithm works well as a result. To examine the output of result analysis, the categorized datasets with framed design soft margins are modularized to have independent data.

Table 6: Prediction Dataset

Day	Weather	Temp.	Humidity	Wind	Play
1.	Sunny	69	70	False	Yes
2.	Sunny	85	85	True	Yes
3.	Cloudy	83	78	False	Yes
4.	Rainy	65	70	True	No
5.	Sunny	80	90	True	Yes
6.	Sunny	72	95	False	Yes
7.	Rainy	70	96	False	Yes
8.	Rainy	68	80	False	Yes
9.	Rainy	65	70	True	No
10.	Cloudy	64	65	True	Yes
11.	Sunny	72	95	False	Yes
12.	Sunny	69	70	False	Yes
13.	Rainy	75	80	False	Yes
14.	Sunny	75	70	True	Yes
15.	Cloudy	72	90	True	Yes
16.	Cloudy	81	75	False	Yes
17.	Rainy	71	85	True	No
18.	Sunny	80	90	True	Yes
19.	Rainy	70	96	False	Yes
20.	Rainy	68	80	False	Yes

CONCLUSION

The finest machine learning methods for predicting and visualizing massive data sets are SOM and SVM. We use a two-level architecture in our suggested work for both prediction and visualization. The data set or record first separated within small several of vectors, which are subsequently grouped using clustering techniques. Because lowering the computational cost for each cluster is the primary goal of hierarchical algorithms. Secondly, it provides a rough visual representation of every cluster. As a result, it eventually offers benefits over earlier methods in terms of computation, cost, prediction, and reliability. Better visualization and predictions can also be obtained through the use of soft computing techniques.

REFERENCES

[1] In 2011, Daoud BA, Sauquet E, Lang M, Bontron G, and Obled C A com- parative analysis of precipitation predictions using an analog sorting method. Advances in Geoscience 29:103–107.

[2] Ratna SB, Dodla VBR (2010) Using a high resolution mesoscale, mesoscale characteristics and forecasting of an uncommon severe heavy precipitation event across India doi:10.1016/j.atmosres.2009.10.004 model. Atmos Res 95:255–269.

[3] Lee J-W, Hong S-Y (2009) Evaluation of the WRF model's ability to replicate a flash flood with intense rainfall over Korea. doi:10.1016/j.at- mosres.2009.03.015 Atmos Res 93:818–831.

[4] Mahajan PN, Khaladkar RM, and Narkhedkar SG (2007). A research con- ducted in July 2004 examined the accuracy of NCMRWF models in forecasting periods of heavy rainfall during the SW monsoon season. RR-116, Re- search Report No.

[5] Mitra A, Sinha SK, and Narkhedkar SG (2008) A mesoscale analysis of daily rainfall over the Indian summer monsoon region using satellite and traditional data 159–177 in Geofizika 25.

[6] Bhowmik RS, Durai V (2010) Use of multi-model ensemble methods at the district level in real time rainfall predictions for the Indian region over a short time frame. Atmospheric Meteorology 106:19–35.

[7] Predictability of flood events in light of present hydrology and meteorology in the Czech Republic was examined by Březková L, Alek M, and Soukalová EC (2010). 156–168 in Soil Water Res 2(4) .

[8] Extreme precipitation patterns linked to tropical cyclones in the central region of the North American monsoon, Cavazos T, Turrent C, Lettenmaier DP (2008). Lett Geophys Res 35:L21703.

[9] Jason Ong, School of Computer Science, "Data Mining Using Self Organizing Kohonen Maps: A Technique for Effective Data Clustering and Visualization" of Malasia for Sciences .

[10] Vipin Kumar, Eui-Hong (Sam) Han, and George Karpis Chameleon: A Dynamic Modeling-Based Hierarchical Clustering Algorithm.

[11] Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82, Espoo 1997, 57 pages. Kaski, S., "Data exploration using self-organizing maps." The Finnish Academy of Technology is the publisher.

[12] Subimal Ghosh, Munir Ahmad Nayak, Theoretical and Applied Climatology, Predicting Extreme Rainfall Events using Weather Pattern Recognition and Support Vector Machine Classifier November 2013, Issue 3-4, Volume 114, pages 583-603, March 6, 2013.