Middle-East Journal of Scientific Research
24 (Special Issue on Innovations in Information, Embedded and Communication Systems): 178-184, 2016
ISSN 1990-9233; © IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.IIECS.23158

Machine Learning Perspective for Predicting Agricultural Droughts Using Naive Bayes Algorithm

K. Sriram and K. Suresh

Department of Computer Science and Engineering, KCG College of Technology, Chennai, India

Abstract: This paper addresses the Supervised Learning mechanism and its one of the method Naive Bayesian Classification. The nature of Naive Bayesian Classification is to evaluate the parameters and it often performs better in many complex real-world situations. Drought affects a large number of people and cause more losses to society compared to other natural disasters. The frequent occurrence of drought poses an increasingly severe threat to the agricultural production. Drought has very complex phenomenon that is difficult to accurately quantify because it's immense spatial and temporal variability. Drought can also reduce water quality, [1, 2] because lower water flows reduce dilution of pollutants and increase contamination of remaining water sources. Existing system uses ISDI [3] model for evaluating the accuracy and the effectiveness. It assesses the performance of measuring the drought by using the Spatial and temporal characteristics of informations.

Key words: Droughts · Data Mining · Yield Production · Naive Bayes Classification

INTRODUCTION

Data mining (sometimes known as knowledge or data discovery) is the method of analyzing knowledge from completely different views and summarizing it into helpful data that reduces the revenue, cuts costs, or both. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome [4]. Data mining is the search for new, valuable and nontrivial information in large volumes of data. Data-mining is a process that can be applied to any type of data ranging from weather forecasting, electric load prediction, product design, among others.

The purpose of a data mining effort is normally either to create a descriptive model or a predictive model. A descriptive model describes the summary of the data points, making it possible to study important aspects of the data set. A predictive model is to allow the data miner to predict an unknown (often future) value of a specific variable.

Data mining consists of five major elements [2]:

- Extract, transform and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or Table.

The first developed indices are agricultural drought indices such as Standard Precipitation Index (SPI) and Palmer Drought Severity Index (PDSI). The SPI was one of the drought indices been widely used world wide, which is designed to be a spatially invariant indicator (spatially and temporally comparable) only based on in-situ precipitation data. The PDSI is calculated using historical temperature and precipitation and information of the available water content of the soil based on a soil moisture/water balance equation. Although the meteorological drought indices can get more accurate and spatially and temporally comparable drought conditions, their utilization is enslaved to the density and distribution of the station network. This type of indices also cannot reflect the vegetation condition induced by the water deficit.

In this paper the problem of predicting yield production is considered. Yield prediction is a very important agricultural problem that remains to be solved

Corresponding Author: K. Sriram, Department of Computer Science and Engineering, KCG College of Technology, Chennai, India

based on the available data. The problem of yield prediction can be solved by employing Data Mining techniques. This work aims at finding suitable data models that achieve a high accuracy and a high generality in terms of yield prediction capabilities. For this purpose, different types of Data Mining techniques were evaluated on different data sets. Some Data Mining techniques have not yet been applied to agricultural problems. As an example, Biclustering techniques may be employed for discovering important information from agricultural-related sets of data. The K-Means algorithm is able to partition the samples in clusters, but no considerations are made on the compounds that are responsible for this partition. In the Existing system, Using data-mining technology, this paper established a new method, named the Integrated Surface Drought Index (ISDI). ISDI integrates traditional meteorological data, remotely sensed indices and biophysical data and attempt to describe drought from a more comprehensive perspective. The evaluation results indicated that the construction models for three phases of growth season have very high regression accuracy. But it has large time taken for finding the results. It is more complicated to implement by using the spatial and temporal characteristics of data. For this apply the best machine learning mechanism for predicting the agricultural yields by using the time and spatial parameters.

The remainder of this paper is catalogued for further procedure. Section II is pursued with Related Work and to compare the performance and evaluate the level of quality. The Benchmark Analysis is tagged in Section III. Proposed System is prioritized in Section IV. The Results and its analysis are prompted in Section V. Conclusively Section VI, revealed the Conclusion for Bayesian classification.

Related Work: A literature review includes the current knowledge including substantive findings, as well as theoretical and methodological contributions to a particular topic. The performance of measuring the drought by using the Spatial and temporal characteristics of informations is done by using different methods. The dataset was collected from different regions and also collect the time varying informations. Data mining techniques may be chiefly divided in 2 groups: classification and agglomeration techniques. Classification techniques square measure designed for classifying unknown samples mistreatment data provided by a collection of classified samples. This set is sometimes remarked as a coaching set, because, in general, it's wont to train the classification technique a way to perform its classification. Agglomeration techniques may be wont to split a collection of unknown samples into clusters. One in every of the foremost used agglomeration techniques is that the k-means technique. [5-7] The following four literature review attempts to demonstrate the fact and sight the survey parallel to the preceding survey.

The First Survey paper J.D. Bolten *et al*., on 2010 [8] attempts to define the contribution of integrating AMSR-E soil moisture retrievals into the drought detecting capability of the USDA IPAD soil moisture model. Given that the TRMM 3B40RT rainfall product (used to force the open loop case within the data denial experiment) accurately reflect the quality of real-time rainfall accumulation data available in data-poor areas, provide a credible estimate of the added utility provided by AMSR-E surface soil moisture retrievals for drought applications (like the USDA IPAD DSS) requiring near real-time root-zone soil moisture estimates within (potentially) data-poor land regions. Net improvement is noted in our ability to track root-zone soil moisture temporal dynamics and is observed for all non forested land cover types within the North American study domain most notably cropland areas of prime importance for the IPAD agricultural drought DSS.

The Second Survey paper J. Brown *et al*., on 2008 [9] defines VegDRI a new, national-level drought monitoring tool that provides near-real time, 1-km resolution maps that depict the geographic extent and severity of drought conditions on vegetation. The integration of 1-km vegetation condition observations, derived from AVHRR NDVI data, with climate-based drought index data and other biophysical information in the VegDRI model enables higher resolution drought monitoring information to be generated. The improved, 1-km resolution of VegDRI provides drought information at a more relevant spatial scale for local-scale planning, mitigation and response activities than the common drought indicators that are currently being used. As a result, the VegDRI could be used by a broad user community that includes agricultural producers, drought and natural resource specialists, policy makers and other stakeholders to make more informed decisions at national, regional, state and county levels VegDRI maps have been operationally produced and updated on a biweekly cycle during the growing season (May-October) for a 15-state region of the central United States for 2006 and 2007. Geographic expansion of VegDRI across the western United States is planned for the spring of 2008 and complete coverage of the

conterminous United States is targeted for 2009. In addition, the development of a retrospective time series of VegDRI maps dating back to 1989 for the conterminous United States is scheduled.

The Third Survey paper N. J. Doesken, J. Kleist and T. B. McKee *et al*., on 1993 [10] proposes a method called Standardized Precipitation Index (SPI) is calculated in the following sequence. A monthly precipitation data set is prepared for a period of m months, ideally a continuous period of at least 30 years. A set of averaging periods are selected to determine a set of time scales of period j months where j is 3, 6, 12, 24, or 48 months. These represent arbitrary but typical time scales for precipitation deficits to affect the five types of usable water sources. The data set is moving in the sense that each month a new value is determined from the previous i months. Each of the data sets are fitted to the Gamma function to define the relationship of probability to precipitation. Once the relationship of probability to precipitation is established from the historic records, the probability of any observed precipitation data point is calculated and used along with an estimate of the inverse normal to calculate the precipitation deviation for a normally distributed probability density with a mean of zero and standard deviation of unity. This value is the SPI for the particular precipitation data point. The SPI is uniquely related to probability. The precipitation used in SPI can be used to calculate the precipitation deficit for the current period. The precipitation used in SPI can be used to calculate the current percent of average precipitation for time period of i months. The SPI is normally distributed so it can be used to monitor wet as well as dry periods. where j starts with the first month of a drought and continues to increase until the end of the drought for any of the i time scales.

The Fourth Survey paper N.T. Son *et al.,* on 2012 [11] proposes a new multi-sensor microwave remote sensing drought index, the Microwave Integrated Drought Index (MIDI), for monitoring short-term drought, especially the meteorological drought over semi-arid regions, by integrating three variables: Tropical Rainfall Measuring Mission(TRMM) derived precipitation, Advanced Microwave Scanning Radiometer for EOS (AMSR-E) derived soil moisture and AMSR-E derived land surface temperature. Each variable was linearly scaled from 0 to 1 for each pixel based on absolute minimum and maximum values over time to relatively monitor drought. Pearson correlation analyses were performed between remote sensing drought indices and scale-dependent Standardized Precipitation Index (SPI) during the growing season (April to October) from 2003 to 2010 to assess the capability of remotely sensed drought indices over three bio climate regions in northern China. The results showed that MIDI with proper weights of three components outperformed individual remote sensing drought indices and other combined microwave drought indices in monitoring drought. It nearly possessed the best correlations with different time scale SPI; meanwhile it showed the highest correlation with 1-month SPI and then decreased as SPI time scale increased, suggesting that the MIDI was a very reliable index in monitoring meteorological drought. Furthermore, similar spatial patterns and temporal changes were found between MIDI and 1- or 3-month SPI in monitoring drought. Therefore, the MIDI was recommended to be the optimum drought index, in monitoring short-term drought, especially for meteorological drought over cropland and grassland across northern China or similar regions globally with the ability to work in all weather conditions.

Benchmarking: A particular data mining algorithm is usually an instantiation of the model/preference/search components.

The more common model functions in current data mining practice include [4]:

- Classification: classifies a data item into one of several predefined categorical classes.
- Regression: maps a data item to a real valued prediction variable.
- Clustering: maps a data item into one of several clusters,where clusters are natural groupings of data items based on similarity metrics or probability density models.
- Rule generation: extracts classification rules from the data.
- Discovering association rules: describes association relationship among different attributes.
- Summarization: provides a compact description for a subset of data.
- Dependency modeling: describes significant dependencies among variables.
- Sequence analysis: models sequential patterns, like time-series analysis. The goal is to model the states of the process generating the sequence or to extract and report deviation and trends over time. To proceed further, four techniques has been taken to compare the efficiency.

Naive Bayesian Classification: The Concept of Naive Bayesian classifier is based on Bayes theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$ and $P(x|c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \quad (1)$$

Where $P(c \mid x) = P(x_1 \mid c) \times P(x_2 \mid c) \times ... \times P(x_n \mid c) \times P(c)$
P(c|x) is the posterior probability of class (target) given predictor (attribute), P(c) is the prior probability of class, P(x|c) is the likelihood which is the probability of predictor given class, P(x) is the prior probability of predictor. Naive Bayesian network is a powerful tool for dealing uncertainties and widely used in agriculture datasets. Bayesian network is a graphical model which encodes probabilistic relationship among variable of interest when it is used with statistical technique, the graphical model has several advantages for data analysis [12, 13].

Neural Networks: Neural Network or an artificial neural network is a biological system that detects patterns and makes predictions. The greatest breakthroughs in neural network in recent years are in their application to real world problems like customer response prediction, fraud detection etc. Data mining techniques such as neural networks are able to model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business applications [14]. This powerful predictive modelling technique creates very complex models that are really difficult to understand by even experts. Neural Networks are used in a variety of applications.

K Nearest Neighbor: K nearest neighbor techniques is one of the classification techniques in data mining. It does not have any learning phase because it uses the training set every time a classification performed. Nearest Neighbor search (NN) also known as proximity search, similarity search or closest point search is an optimization problem for finding closest points in metric spaces. K nearest neighbor is applied for simulating daily precipitation and other weather variables.

Decision Trees: The decision tree is one of the popular classification algorithms in current use in Data Mining and Machine Learning. Decision tree is a new field of machine learning which is involving the algorithmic acquisition of structured knowledge in forms such as concepts, decision trees and discrimination nets or production rules. Application of data mining techniques on drought related data for drought risk management shows the success on Advanced Geospatial Decision Support System (GDSS).

Comparison of Techniques: The conclusion can be drawn from the comparison of the above stated algorithms. Every single method plays its vital role in different categories of Data Mining Techniques, which issues its drawback at last. In order to acknowledge Basically, there are two approaches used for prediction. They are Empirical method and dynamical methods. The empirical approach is based on the study of historical data of the rainfall and its relationship to a variety of atmospheric and oceanic variables over different parts of the world. In dynamical approach, predictions are generated by

Table 1: Comparison of Techniques

Data Mining Techniques	Advantages	Disadvantages
Naïve Bayesian Classification	Fast to train and classify Not sensitive to irrelevant features.	Assumes independence of features
Neural Network	Neural networks realized as specialized hardware systems. Useful for network learning.	Too much of a black box. Neural networks are not probabilistic. Neural networks are not a substitute for understanding the problem deeply.
K-Nearest Neighbor	Robust to noisy training data. Effective if the training data is large.	Need to determine the value of K. Computation cost is quite high.
Decision Trees	Are simple to understand and interpret. Can be combined with other decision techniques.	Calculations can get very complex. Information gain in decision trees are biased.

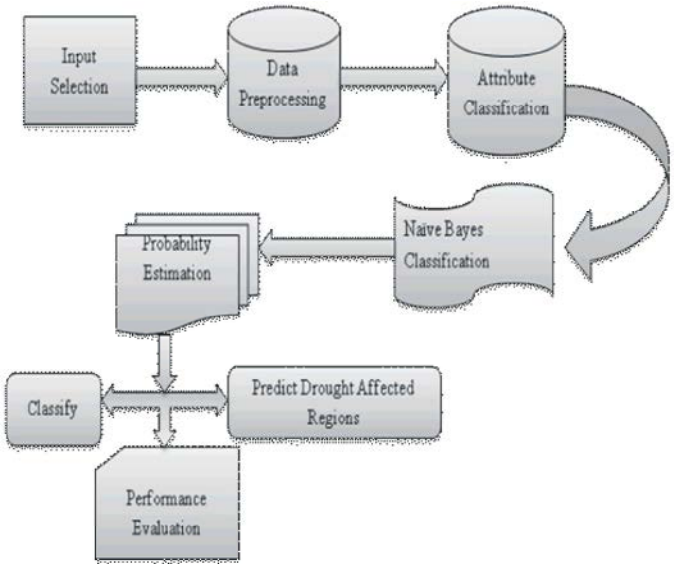

Fig. 1: Architecture of Proposed System

physical models based on systems of equations that predict the evolution of the global climate system in response to initial atmospheric conditions are generated by physical models based on systems of equations that predict the evolution of the global climate system in response to initial atmospheric conditions.

Proposed System: The system architecture shown in Fig. 1, explains the working of the proposed system. In this paper, multi-source information is integrated to achieve the purpose of accurately monitoring agricultural drought. Agricultural drought generally occurred after a period of time. since the cessation of rain, when the available stored water will support the actual evapotranspiration. It involves a variety of meteorological drought characteristics (e.g., the scarcity of precipitation, actual evapotranspiration at only a small fraction of the potential evapotranspiration rate and the shortage of soil moisture) impact on agriculture (e.g., yield reduction). Therefore, the agricultural drought condition is affected by many factors, such as precipitation, soil moisture, temperature, vegetation type, soil type and phenology. Based on the defined drought criteria, the intensity, temporal and spatial distribution of agricultural drought can be monitored. By using this type of attributes, drought conditions by using the supervised machine learning methods were predicted.

Input Selection is the first process. In this, first have to browse and select the input for the process. Input of the process is dataset. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. Normally Dataset pre-processing is the method for cleaning the dataset. Datas may be incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data noisy: containing errors or outliers inconsistent: containing discrepancies in codes or names. In this elimination of this type of occurring in the dataset is going to be performed. Eliminating the unwanted value or symbols or characters in the dataset.

Classification is a way of categorising the data (records) for an attribute. Attributes can use different classifications for the same data to change the nature of the display; this can be achieved based on the attributes in the dataset. Attributes in the dataset are the scarcity of precipitation, actual evapotranspiration at only a small fraction of the potential evapotranspiration rate and the shortage of soil moisture) impact on agriculture (e.g., yield reduction). In the implementation process, implementation of the drought research by using the supervised classification algorithm called Naive Bayes Classification algorithm. A Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood. Predicting the drought affected region by using the naïve bayes algorithm and evaluate the performance by using the parameters of the process. Then, evaluation of the graph based on the extreme, mid, low, normal droughts etc., is done like that. These are estimated by using the probability values in the dataset. The graph can be evaluated by using this type of parameters form probability estimations.

RESULT AND DISCUSSION

Craven and Shavlik *et al.,* defines in there paper [3] listed five criteria for rule extraction and they are as follows:

Comprehensibility: The extent to which extracted representations are humanly comprehensible.

Fidelity: The extent to which extracted representations accurately model the networks from which they were extracted.

Accuracy: The ability of extracted representations to make accurate predictions on previously unseen cases.

Scalability: The ability of the method to scale to networks with large input spaces and large numbers of weighted connections.

Generality: The extent to which the method requires special training.

The table aims to come out of the techniques being used in the agricultural soil science and its allied area. Table 2, analyzes some data mining techniques and its accuracy to predict the rainfall [14]. From this table, we understand that the Naive Bayesian has better accuracy in result than other data mining techniques. Accuracy is determined by the formula 100-RMSE. RMSE (Root Mean Square Error) is one of the measuring

Table 2: Accuracy of Data Mining Techniques

Data Mining Techniques	Accuracy
Naïve Bayes	85.77%
KNN (k=30)	81.81%
Decision Trees	81.40%
Neural Networks	82.81%

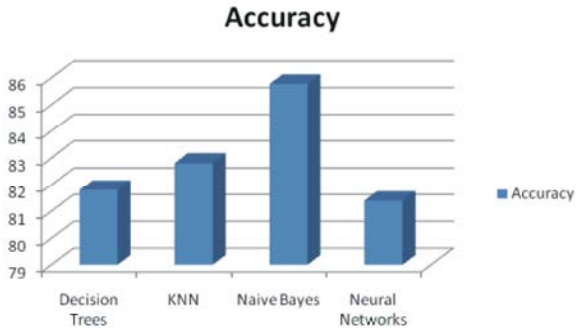

Fig. 2: Graphical representation of Data mining techniques and its accuracy

techniques for predicting rainfall. It is measured by the differences between values predicted by a model and the values actually observed from the model [12].

CONCLUSION

The Bayesian model attempts to characterize drought in a more accurate and comprehensive way. Cross-validation analysis proved that the regression accuracy is very high and the drought condition can be accurately portrayed using the input variables. This model application using a variety of methods and data, there is still some work to be Done in the future research because of the complex spatial and temporal characteristics of drought. To overcome imitation, it assesses performance of measuring the drought by using the Spatial and temporal characteristics of informations. Collecting the dataset from different regions and also collect the time varying information's like that.Prediction of the drought conditions can be done by using the supervised learning mechanism. It can be implemented by using the Bayesian supervised machine learning algorithm. Through this the performance and effectiveness will be improved.

REFERENCES

1. Aldridge, K.T., T.M. Heneker, M.R. Hipsey, L.M. Mosley, E. Leyden, D. Skinner and B. Zammit, 2012. The Impact of Extreme Low Flows on the Water Quality of the Lower Murray River and Lakes (South Australia). Water Resources Management, 26: 3923-3946.
2. Mosley, L.M., 2014. Drought impacts on the water quality of freshwater systems; review and integration. Earth Science Reviews. DOI: 10.1016/j.earscirev.2014.11.010.

3. Fengying Zhang, Jianjun Wu, Jianhui Zhang, Jie Zhang, Lei Zhou, Lin Zhao, Ming Liu, Song Leng, and Yu Shi, 2013. The Integrated Surface Drought Index (ISDI) as an Indicator for Agricultural DroughtMonitoring: Theory, Validation and Application in Mid-Eastern China, 6(3).
4. Ashok Kumar, D. and N. Kannathasan, 2011. 8(3): 1, May 2011., A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining.
5. Agnes Begue, Dino Ienco, Elodie Vintrou and Maguelonne Teisseire, 2013. Data Mining, A Promising Tool for Large-Area Cropland Mapping, IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing, 6(5), October 2013.
6. Dean Diepeveen, Leisa J. Armstrong and Rowan Maddern, The application of data mining techniques to characterize agricultural soil profiles, International Journal of Computer Science and Technology IJCST.
7. Latika Sharma and Nitu Mehta, 2012. Data Mining Techniques: A Tool For Knowledge Management System In Agriculture, International Journal Of Scientific and Technology Research, 1(5), June 2012.
8. Bolten, J.D., *et al.*, 2010. Evaluating the utility of remotely sensed soil moisture retrievals for operational agricultural drought monitoring, IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., 3: 57-66.
9. Brown, J., *et al.*, 2008. The Vegetation Drought Response Index (VegDRI): A new integrated approach for monitoring drought stress in vegetation, GIS Remote Sens., 45: 16-46.
10. Doesken, N.J., J. Kleist and T.B. McKee, 1993. The relationship of drought frequency and duration to time scales, in Proc. 8th Conf. Applied Climatology, Anaheim, CA, USA, pp: 179-184.
11. Son, N.T., *et al.,* 2012. Monitoring agricultural drought in the Lower Mekong Basin using MODIS NDVI and land surface temperature data, Int. J. Appl. Earth Observ. Geoinf., 18: 417-427.
12. Bhargavi, P. and S. Jyothi, 2009. Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils IJCSNS International Journal of Computer Science and Network Security, 9(8): 117, August 2009.
13. Quinlan, J.R., 1986. Induction of decision trees. Machine Learning, 1(1): 81-106, March 1986.
14. Andrews, R., J. Diederich and A.B. Tickle, 1995. A survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-Based Systems, 8(6): 378-389.