© IJCIRAS | ISSN (O) - 2581-5334
March 2020 | Vol. 2 Issue. 10

RICE CROP YIELD CLASSIFICATION BY USING MANHATTAN BASED KNN ALGORITHM

Htwe Htwe Pyone[1], Thin Thin Swe[2]

[1]Faculty of Computer Science, University of Computer Studies (Myitkyina), Myanmar
[2]Faculty of IT Support and Maintenance, University of Computer Studies (Myitkyina), Myanmar

Abstract

Rice crop production plays a vital role in our country. High crop production is dependent on suitable climate conditions. Detrimental seasonal climate conditions such as low rainfall or temperature extremes can dramatically reduce crop yield. To classify rice crop productivity in different climatic conditions, developing better techniques can assist farmer and other stakeholders in important decision making in terms of agronomy and crop choice. Data mining techniques such as K-nearest neighbor (KNN) classifier can be used to improve prediction of crop yield under different climatic scenarios. So, this system applies the Manhattan based KNN classifier to classify rice crop yield. In this system, the experimental results will show the performance of Manhattan based KNN classifier by using rice crop dataset.

Keyword: Rice Crop, Manhattan, KNN Classifier

1.INTRODUCTION

Today, agriculture forms the foundation of Myanmar economy. Due to both climatic and economic challenges, large areas of agricultural land are not achieving adequate crop production. Crop yield relies upon climate conditions. Farmers are faced with having to make difficult decisions as to how remain productive and sustainable with changing climates and market economic pressure. Water shortage, absence of high yielding varieties and poor technology transfer of best agronomic practices are considered to be the principal factors for low crop yields in Myanmar. To predict crop productivity in different climatic conditions, the use of technology and various computer science techniques can assist stakeholders and farmers in important decision making in terms of agronomy and crop choice. There are various computer science techniques such as data mining and machine learning which have been used to determine the influence of difference parameters and make predictions of the crop production. Data mining is the process of extracting important and useful information from large sets of data and is a relatively new inter-disciplinary concept involving data analysis and knowledge discovery from the database. It may provide crucial role in decision making for complex agricultural problems.

Due to the agriculture is the backbone of Myanmar economy, every farmer expects well their crop productivity. In this situation, this system will help farmers to have good yield. Data mining technique is useful for rice crop yield classification. So, this system uses the Manhattan based K-nearest neighbor (KNN) classifier to help farmers about rice crop yield. This system can also enhance the understanding of land-use change.

This paper is organized as follows: first, we introduce the rice crop yield classification system. In section 2, related works of the system are presented. Data mining and its classification methods are described in section 3 and 4. In section 5, the proposed system is presented with the system flow diagram. Then, the detail explanation of the system is described in section 6 and the experimental result of the system is expressed in section 7. Finally, the conclusion of the system is described in section 8.

2. RELATED WORK

In 2017, N. Gandhi and L. J. Armstrong [1] presented the data mining techniques that were applied to the historical agricultural dataset of semi-arid climatic zone of India to extract knowledge for predicting rice crop yield of kharif season. Free and open source software WEKA (Waikato Environment for Knowledge Analysis) was used to apply data mining techniques for the present agricultural dataset. They found that J48

classifier provided the best performance among the classifiers used for the semi-arid climatic zone of India data set.

In 2018, K. R. Akshatha and K. S. Shreedhara [2] implemented machine learning algorithm for crop recommendation using precision agriculture. The common problem existing among the Indian farmers are that they don't choose the right crop based on their soil requirements. In this system, this problem is solved by proposing a recommendation system through an ensemble model with majority voting technique using random tree, K-nearest neighbor and naive bayes as learners to recommend a crop for the site specific parameters with high accuracy and efficiency.

In 2019, M. P. Nanjesh Gowda and S. Ramya [3] presented the rice yield prediction system by using data mining technique. In this system, they took the soil from the farmers from their respective region and predicted the amount of rice that can be grown in these region by considering the various parameters. They considered a large number of dataset to predict the yield of the rice. They used various data mining technique to predict the yield of the rice. They pointed out the KNN algorithm it gives the accurate result to the farmers.

3. DATA MINING

Data mining is the discovery for knowledge of analyzing enormous set of data by extracting the meaningful data and thereby predicting the future trends with them. It deals with what kind of patterns can be mined. Based on the kind of data to be mined, there are two kinds of functions involved in data mining: descriptive model and predictive model. The former model deals with general properties of data in the database and the latter model predicts the class of objects whose class label is unknown.

Data mining techniques are divided into two groups: classification and clustering techniques. Classification techniques are essential for classifying unknown samples using information provided by a set of classified samples. To train the classification, the training set is used. If a training set is not available, clustering technique is used to split a set of unknown samples into clusters. [4]

4. CLASSIFICATION

Classification is a technique that finds the rules from the large database. The main task of classification is to recognize the similar observation from large dataset and arrange them into a set. Classification techniques help to find different patterns among large dataset. The advantages of classification (classifier) are as follows:

- Efficiency is good
- Handles the noisy data
- Well suited for multimodal classes
- Requires short computational times. [5]

Classification produces a function that maps a data item into one of several predefined classes, by inputting a training data set and building a model of the class attribute based on the rest of the attributes. There are various classification methods that are K-nearest neighbor, naive bayesian, support vector machine, decision tree classifier and so on. [6]

4.1. KNN Classifier

K-nearest neighbor (KNN) classifier is the most commonly used classification method. KNN is very easy to implement and give fairly good results. It does not require any prior knowledge regarding data set for classification. It performs classification purely on similarity basis. [7]

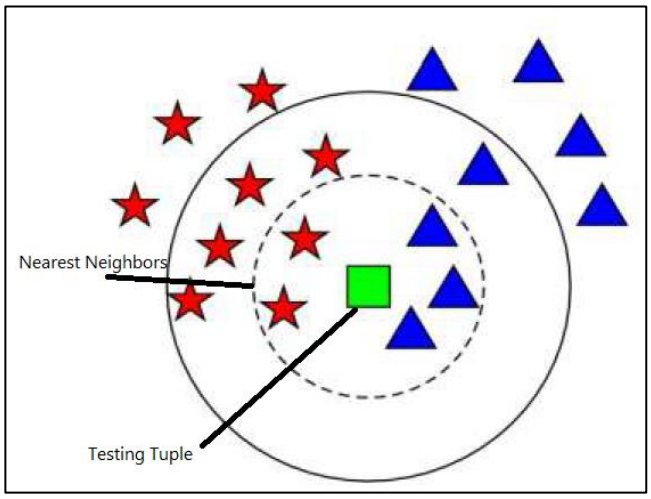

Figure 1 KNN Classifier [3]

KNN classifier is shown in Figure 1. The procedure of classification in KNN starts with a data set. The data set is constituted of certain number of attributes that define a data set. Data set is divided into two sets: training set and test set. In KNN classifier, the training set can be

viewed as a set of data points in an n-dimensional space, where n dimensions are the set of n attributes describing the data set. When an unknown tuple comes for classification, it is needed to find out the k nearest data points to it in the n-dimensional space. To find the k nearest data points to the unknown tuple, various distance metrics are used. Euclidean distance, Minkowski distance and Manhattan distance method are popular methods. [7]

4.2. K-Nearest Neighbors (KNN) Algorithm

The k-nearest neighbors (KNN) algorithm is as follows:
Input parameters: Data set, k
Output: Classified test tuples

- Step 1: Store all the training tuples.
- Step 2: For each unseen tuple which is to be classified, KNN classifier computes the distance of unseen tuple with all the training tuples by using Manhattan distance method. Manhattan distance between data tuples X and Y is computed as:

$$\sum_{1 \le i \le n} |x_i - y_i| \quad (1)$$

- Step 3: Find the k nearest training tuples to the unseen tuple.
- Step 4: Assign the class which is most common in the k nearest training tuples to unseen tuple. [7]

4.3. Hold Out Validation Method

In the hold out validation method, data are divided into two separated sets, i.e. training data and testing data. The process of hold out method is shown in Figure 2.

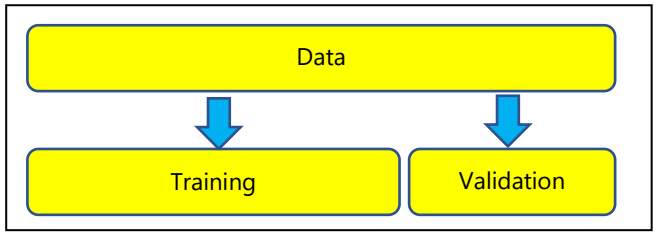

Figure 2 Hold Out Method [8]

The proportion between the training data and testing data is not binding but to ensure that the variant in the model is not too wide, usually 2/3 of the data are used as the training and the other 1/3 used as the testing data, while data is divided into 70% for training data and 30% for testing data. [8]

5. PROPOSED SYSTEM DESIGN

For rice crop yield classification, the system flow diagram is shown in Figure 3.

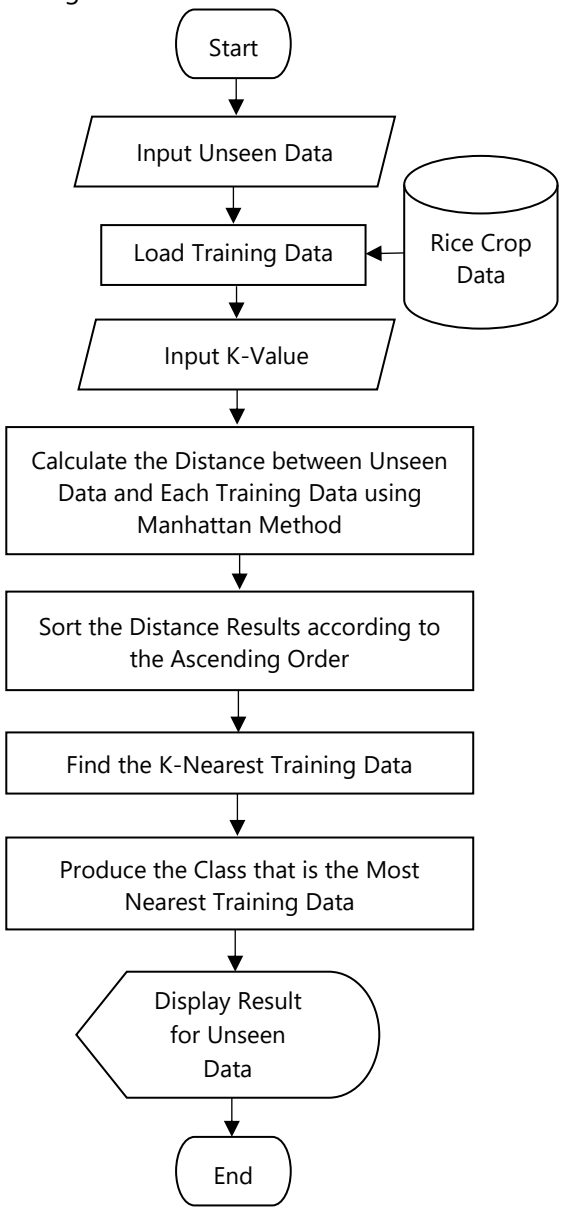

Figure 3 System Flow Diagram

The proposed system classifies the rice crop yield by using Manhattan based KNN classifier. Firstly, the user must input the unseen data that has unknown class label. For classification, this system loads the training data from the rice crop database. According to the K-value, this system classifies the k-nearest neighbor training data for the unseen data. To find the nearest neighbor, this system uses the Manhattan distance method that calculates the distance (or similar) between each training data and unseen data.

After calculating the distance, this system sorts the distance results according to the ascending order. From this order, this system produces the class that has the most similar result (minimum distance) between training and unseen data. Finally, this system displays the result (class) to the user.

6. EXPLANATION OF THE SYSTEM

For classification, this system uses the rice crop yield dataset from Maharashtra state in India.

Table 1. Training Rice Crop DataSet

District	Record No	Year	Precipitation	Min-Temperature	Avg-Temperature	Max-Temperature	Reference Crop Evapotranspiration	Area	Production	Yield
Ahmednagar	1	1998	153	27	26	31	4.62	9000	5600	High
	2	1999	128	27	25	32	4.50	6000	5700	High
	3	2000	93	27	25	31	4.57	4000	5500	Moderate
	4	2001	96	27	25	31	4.54	1200	7300	Low
	5	2002	87	25	25	33	4.57	1200	7400	Low
Amravati	6	1998	143	26	28	32	4.99	1523	8700	Moderate
	7	1999	159	26	26	32	4.83	2633	8800	Moderate
	8	2000	97	26	27	31	4.93	1523	8900	Low
	9	2001	81	28	27	30	4.89	2633	5000	Moderate
	10	2002	118	27	27	30	4.86	1600	1100	Moderate

This dataset consists of ten attributes that are district, year, precipitation (In mm), reference crop evapotranspiration, area (in hectare), minimum temperature, average temperature, maximum temperature, production (in tonnes) and yield (tonnes/ hectare). Among them, this system uses the eight attributes for rice crop yield estimation. As a sample, the training rice crop data is shown in Table 1. The user inputted unseen (unknown) data is as follows:

- Precipitation (In mm): 126
- Minimum Temperature: 27
- Average Temperature: 26
- Maximum Temperature: 30
- Rice crop evapotranspiration: 4.96
- Area (In Hectare): 600
- Production (In Tonnes): 1200
- Yield: **????**

Then, the user input the "K" value. In this sample, the K=1 is used for classification. For rice crop production (yield) classification, this system uses the Manhattan based KNN algorithm. This system calculates the distance between the user inputted unseen data and each training data. The distance results are shown in Table 2.

Table 2. Distance Result

Record No	Training Data & Unseen Data	Distance Result
1	Training Data 1 & Unseen Data	12827.66
2	Training Data 2 & Unseen Data	9902.54
3	Training Data 3 & Unseen Data	7666.61
4	Training Data 4 & Unseen Data	6669.58
5	Training Data 5 & Unseen Data	6760.61
6	Training Data 6 & Unseen Data	8443.03
7	Training Data 7 & Unseen Data	9666.87
8	Training Data 8 & Unseen Data	8594.97
9	Training Data 9 & Unseen Data	5789.93
10	Training Data 10 & Unseen Data	892.9

After calculating distance between each training data and unseen data, this system produces the class name (Yield name) that is the most similar among all training data. In this sample, training record 10 is the most similar the unseen data. This system produces the "**Moderate**" yield for the user inputted unseen data.

7. EXPERIMENTAL RESULT OF THE SYSTEM

To show the performance of the system, this system is tested by using 500 data records from 27 districts. These districts are namely, Ahmednagar, Amravati,

Aurangabad, Beed/Bid, Bhandara, Buldhana, Chandrapur, Dhule, Gadchiroli, Gondia, Hingoli, Jalana, Jalgaon, Kolhapur, Latur, Nagpur, Nanded, Nasik, Osmanabad, Parbhani, Pune, Sangli, Satara, Solapur, Wardha, Washim and Yavatmal. For accuracy of the system, this system uses the hold out validation method. The accuracy results of the system are shown in Table 3.

Table 3. Accuracy Result

Test No	Number of Rice Crop Yield	Accuracy Result
1	100	95%
2	200	91%
3	300	85%
4	400	87%
5	500	90%

The experimental results of the system are shown in Figure 4.

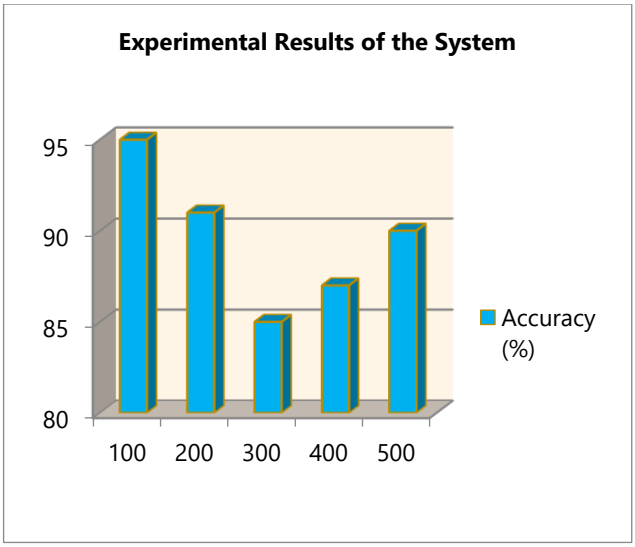

Figure 4 Experimental Results of the System

8. CONCLUSION

Agriculture is the most significant application area particularly in the developing countries like Myanmar. Farmers are being encouraged to continue to plant rice and continue to improve crop productivity. In this situation, this system enables to help farmer and stakeholders to improve crop production. This system demonstrates the potential of using classifier to improve the decision support system about the prediction of crop yield productivity. This system applies Manhattan based KNN classifier to predict rice crop production.

REFERENCES

[1] N. Gandhi and L. J. Armstrong, "Application of Data Mining Techniques for Predicting Rice Crop Yield in Semi-Arid Climate Zone of India", IEEE, 2017.
[2] K. R. Akshatha and K. S. Shreedhara, "Implementation of Machine Learning Algorithms for Crop Recommendation using Precision Agriculture", International Journal of Research in Engineering, Science and Management (IJRESM), vol. 1, no. 6, pp. 58-60, 2018.
[3] M. P. Nanjesh Gowda and S. Ramya, "Rice Yield Prediction using Data Mining Technique", International Research Journal of Engineering and Technology (IRJET), vol. 6, no. 4, 2019.
[4] A. Mistry and V. Shah, "Brief Survey of Data Mining Techniques Applied to Applications of Agriculture", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE),
vol. 5, no. 2, pp. 301-304, 2016.
[5] P. Akulwar, S. Pardeshi and A. Kamble, "Survey on Different Data Mining Techniques for Prediction", IEEE, 2018.
[6] N.Midha and V. Singh, "A Survey on Classification Techniques in Data Mining", International Journal of Computer Science & Management Studies (IJCSMS), vol. 6, no. 1, 2015.
[7] A. Lamba and D. Kumar, "Survey on KNN and Its variants", Interantional Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), vol. 5, no. 5, pp. 430-435, 2016.
[8] I. K. Hadihardaja and Nurhayati, "A Study of Hold-Out and K-Fold Cross Validation for Accuracy of Groundwater Modeling in Tidal Lowland Reclamation using Extreme Learning Machine", IEEE, 2014.