International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

Prediction of Sugarcane Yield using KNN and KNN Plus Clustering Algorithms

M. Naveen Kumar[1], Dr. M. Balakrishnan[2]

Research Scholar, R&D Centre, Bharathiar University

Principal Scientist, National Academy of Agricultural Research Management, ICAR, Rajendra Nagar, Hyderabad, India

Abstract: ***Traditionally, the crop analysis and agricultural production predictions were done based on statistical models. However, with the climate of the world changing to drastic degrees, these statistical models have become very ambiguous. Hence, it becomes prudent that we resort to other less vague methods. Through a traditional model, user interacts primarily with mathematical computations and its results and can help to solve well-defined and structured problems. Whereas, in a data driven model, user interacts primarily with the data and helps to solve mainly unstructured problems. At this point, enters the concept of Machine Learning. In this work we tried to find a new approach to reduce the input feature to reduce the processing power needed. In this work we have attempted at predicting the agricultural outputs of sugarcane production in Telangana Region area by implementing a KNN based machine learning model. Through this model, we tried to predict the approximate crop yield based on various parameters values analyzed for a particular season and area. Results from the simulated studies showed that the statistical models can roughly simulate pre harvest yield forecast of sugarcane under telangana region. This paper deals with study of KNN and clustered KNN algorithms which are suitable for the available predictors and predictions and to perform analysis, cleaning, pre processing and feature selection on the data and which will be very much useful to the researchers, policy and decision makers.***

Keywords: Sugarcane, Yield prediction, KNN, Clustered KNN

1. Introduction

Sugarcane is a most important cash crop of India. It involves less risk and farmers are assured up to some extent about return even in adverse condition. Sugarcane provides raw material for the second largest agro-based industry after textile. The sugar industry is an instrumental in generating the sizable employment in the rural sector directly and through its supplementary units. The Sugarcane plant offers a huge potential, not only as the sucrose of a very important food but also as a source of energy and valuable commercial products from fermentation and chemical synthesis. Sugarcane processing is focused on the production of cane sugar from sugarcane. Sugarcane is considered as one of the best converters of solar energy into biomass and Sugar.

Andhra Pradesh is a leading player in paddy, cotton, groundnut, sugarcane, maize, tobacco and chillies [1]. After bifurcation of state, there would be a stark contrast in the availability of fertile lands and water in the two regions. Telangana has emerged as a predominant player in cotton, paddy and maize, and Sugarcane with a total area of 81 lack hectares. Telangana farmers grow cotton in 14 lack hectares during a normal season. This shows how important the cotton crop would be for the new State. The other major crop where it virtually dominates is maize. Sugarcane is also is abundant in the region. In Telangana it is 26.00 Per Cent area and around 27.00 Per Cent of cane production and in Rayalaseema 15.00 Per Cent in area and cane production [2].

Yield prediction [3] is one of the most critical issues faced in the agricultural sector. Farmer's lack of knowledge about harvest glut, uncertainties in the weather conditions and seasonal rainfall policies, depletion of nutrition level of soils, fertilizer availability and cost, pest control, post harvest loss and other factors leads to decrease in the production of the crops. Regression Analysis can be defined as a structured approach which stresses on the analysis of data for the research purpose on decision making and problem solving. There are problems/situations that require simultaneous analysis of multiple variables or objects for efficient decision making.

Computer scientists and statisticians together brought many approaches and methodologies to improve the prediction power. It is mainly used by data scientists, data analysts and also for them who wants to use the raw data to predict or find trends in data. Regression analysis, one of the tools available in statistical analysis literature is the simple, common and important technique used to model the relationship between one or more independent or predictor variables and a dependent or response variable, which we want to predict. When all the predictor variables are continuous valued then the good selection of prediction technique is regression analysis.

It is evident from the results that to get highest accuracy value we must further improve on the accuracy of our sugarcane yield estimation model using data mining techniques of prediction. Thus in future we will try to effectively predict crop yields of given crops with high level of accuracy using efficient prediction models. The crop analysis and agricultural production predictions were done based on statistical models. However, with the climate of the world changing to drastic degrees, these statistical models have become very ambiguous. Hence, it becomes practically that we have to choose other less unclear methods. Through a traditional model, user interacts primarily with a mathematical computation and its results and can help to solve well-defined and structured problems. This study demonstrated that yield estimation can be done using input parameters for the sugarcane; Finding would benefit the farmers, sugarcane and industry by limiting the yield losses.

Further studies are needed to investigate, between the sugarcane in Telangana Region, Also the likely incorrect interpretations of results using univariate analysis need to be investigated using historical data.

KNN:

The k-Nearest neighbourhood methodology is wide used adopted thanks to its potency [4][5]. The key plan of the algorithmic rule is to categorize a brand new sample within the most frequent class of its nearest neighbours within the coaching set. This is often the foremost selection formula on the category labels of its neighbours. The k-nearest neighbour classification algorithmic rule may be divided into 2 phases: coaching section and testing section. KNN is similar to kernel methods with a random and variable bandwidth. The idea is to base estimation on a x^{th} number of observations k which are closest to the desired point.

Suppose X 2 R^q and we have a sample $\{X_1, X_2 \ldots\ldots\ldots X_n\}$:

For any fixed point x 2 R^q; we can calculate how close each observation X_i is to x using the Euclidean distance $\| x \| = (x'x)^{1/2}$ this distance is

$$D_i = \| x - X_i \| = ((x - X_i)' (x - X_i))^{1/2}$$

This is just a simple calculation on the data set.

The order statistics for the distances D_i are $0 \leq D_{(1)} \leq D_{(2)} \leq \ldots \leq D_{(n)}$.

The observations corresponding to these order statistics are the "nearest neighbours" of x: The 1st nearest neighbour is the observation closest to x; the second nearest neighbour is the observation second closest, etc.

This ranks the data by how close they are to x: Imagine drawing a small ball about x and slowly initiating it. As the ball hits the …first observation X_i this is the "…first nearest neighbour" of x: As the ball further initiates and hits a second observation, this observation is the second nearest neighbour.

The observations ranked by the distances, or "nearest neighbours", are $\{X(1), X(2), X(3), \ldots X(n)\}$ The k^{th} nearest neighbour of x is X(k).

For a given k; let

$$R_x = \| X_{(k)} - x \| = D_{(k)}$$

denote the Euclidean distance between x and $X_{(k)}$: Rx is just the K^{th} order statistic on the distances D_i. When X is multivariate the nearest neighbour ordering is not invariant to data scaling. Before applying nearest neighbour methods, is therefore essential that the elements of X be scaled so that they are similar and comparable across elements.

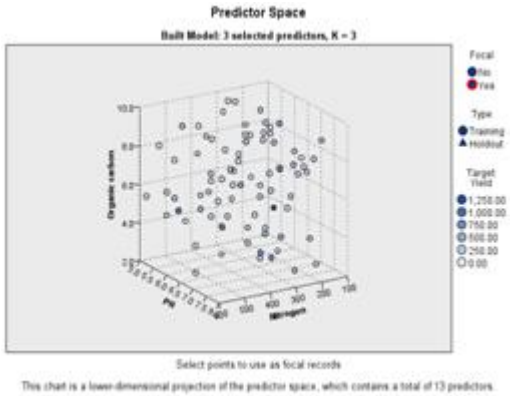

Figure 1: Lower Dimension Projection of Feature Subspace using KNN, when K=3 for Crop Yield Dataset

The 'k' value and KNN method to determine the minimum points and radius value automatically. Using these methods crop data set is analysed and determined the optimal parameters for the wheat crop production. Multiple linear regressions are used to find the significant attributes and form the equation for the yield prediction.

This model is simple, does not required any sophisticated statistical tools, required data for crop growing periods, yield data for past years and provides marginally good prediction. Therefore it can be used for district, agro climatic zone and state level prediction. After analyzing the results of statistical methods and KNN found that the results of KNN are less accurate than the statistical methods, but not getting the high level accuracy these methods through. Hence, it must be further improved for more accuracy and lower errors.

As the existing data mining algorithm KNN is not giving the satisfied accurate prediction results for the sugarcane crop yield. So, the designing and development of hybrid algorithm clustered KNN is done. This paper describes the design and development of hybrid algorithm - KNN+clustering, which was giving the best results.

KNN+Clustering

K-Nearest Neighbor rule (KNN) has been one of the most well-known supervised learning algorithms in pattern classification, as it was first introduced. This rule simply retains the entire training set during learning and assigns to each query a class represented by the majority label of its k-Nearest Neighbours in the training set [6]. The Nearest Neighbor rule (NN) is the simplest form of KNN when k = 1. KNN has several main advantages: simplicity, effectiveness, intuitiveness and competitive classification performance in many domains. It has been found that the asymptotic error rate of KNN approaches the optimal Bayes error rate $R*$ when the number of the samples N and the number of neighbours k tend to infinity and $k=N \rightarrow 0$, and the error rate of NN is bounded above by twice the optimal Bayes error rate $2R*$. Furthermore, the appeal of KNN stems from only a single integer parameter k, and the high classification performance with increasing the amount of training samples. As an improvement to KNN with the basic idea of weighting close neighbours more heavily, according to their distances to the query [7]. However, the number of available samples in real applications is usually too small to

obtain a good asymptotic performance, which often leads to dramatic degradation of the classification accuracy, especially in the small sample size cases with the curse dimensionality and existing outliers.

Nowadays, one major challenging problem, yet to be resolved for KNN, is the selection of the neighbourhood size k, which can have a significant impact on the performance of KNN-based classifiers. It has been found that the classification performance of KNN intrinsically results in the estimate of the conditional class probabilities from training set in a local region of data space, which contains k nearest neighbours of the query. The estimate is affected by the sensitivity of the selection of the neighbourhood size k, because the radius of the local region is determined by the distance of the k-th nearest neighbour to the query and different k yields different conditional class probabilities. If k is very small, the local estimate tends to be very poor owing to the data sparseness and the noisy, ambiguous or mislabelled points. In order to further smooth the estimate, we can increase k and take into account a large region around the query. Unfortunately, a large value of k easily makes the estimate over smoothing and the classification performance degrades with the introduction of the outliers from other classes.

To deal with the problem, the related research works have been done to improve the classification performance of KNN. Our work is inspired by the sensitivity issue of different choices of the neighbourhood size k in KNN-based classifiers. We propose a new improved Clustering k-Nearest Neighbour rule using the dual distance-weighted function. In this new rule, we employ the dual distance-weights of k-Nearest Neighbours to determine the class of the query by majority weighted voting. The experimental results suggest the superiority of our proposed classifier in the crop yield estimation of sugarcane crop.

Based on the common nearest neighbour technique for classification we develop a much more flexible tool that extends the basic method in two directions. First we introduce a weighting scheme for the nearest neighbours according to their similarity to a new observation that has to be classified. Based on the fact, that the voting of nearest neighbours is equivalent to the mode of the class probability distribution, the second extension uses the median or the mean of that distribution, if the target variable shows an ordinal or even higher scale level. The nearest neighbour method represents one of the simplest and most intuitive techniques in the field of statistical discrimination. It is a nonparametric method, where a new observation is placed into the class of the observation from the learning set that is closest to the new observation, with respect to the covariates used. The determination of this similarity is based on distance measures.

Formally this simple fact can be described as follows: Let

$$L = \{(yi, xi), i = 1,..., nL\}$$

be a training or learning set of observed data, where yi ∈ {1, . . . , c} denotes class membership and the vector $x_i = (x_{i1}, . . . , x_{ip})$ represents the predictor values. The determination of the nearest neighbours is based on an arbitrary distance function d(., .). Then for a new observation (y, x) the nearest neighbour (y(1), x(1)) within the learning set is determined by

$$d(x, x_{(1)}) = \min_i(d(x, x_i))$$

5.1.1 Improved Distance based Clustering scheme for neighbors

This extension is based on the idea, that such observation switch in the learning set, which are particularly close to the new observation (y, x), should get a higher weight in the decision than such neighbours that are far away from (y, x) [8].

This is not the case with KNN: Indeed, only the k-nearest-neighbours influence the prediction; however, this influence is the same for each of these neighbours, although the individual similarity to (y, x) might be widely different. To reach this aim, the distances, on which the search for the nearest neighbours is based in the first step, have to be transformed into similarity measures, which can be used as weights.

Thus again in the first step the k nearest neighbours are selected according to the Minkowski distance [9]. As before, for that purpose one needs two parameters: The number of neighbours k and the Minkowski parameter q for selection of the distance measure.

To put equal weight on each covariate in computing the distances, one has to standardize the values. In the case of ratio or difference scale level, this aim is reached simply by dividing the variables by their standard deviation. Subtraction of the mean is not necessary, as this operation has no influence on the distances between observations.

Algorithm: Clustering based k-Nearest-Neighbor classification (cKNN)

1) Let $L = \{(yi, xi), i = 1, . . . , nL\}$ be a learning set of observations *xi* with given class membership *yi* and let *x* be a new observation, whose class label *y* has to be predicted.
2) Find the $k + 1$ nearest neighbours to *x* according to a distance function $d(x, xi)$.
3) The $(k + 1)^{th}$ neighbour is used for standardization of the *k* smallest distances via

$$D_{(i)} = D(x, x_{(i)}) = \frac{d(x, x_{(i)})}{d(x, x_{(k+1)})}$$

4) Transform the normalized distances $D(i)$ with any kernel function $K(.)$ into weights $w(i) = K(D(i))$.
5) As prediction for the class membership *y* of observation *x* choose the class, which shows a weighted majority of the *k* nearest neighbours

$$\hat{y} = \max_r \left(\sum_{i=1}^{k} w_{(i)} I(y_{(i)} = r) \right)$$

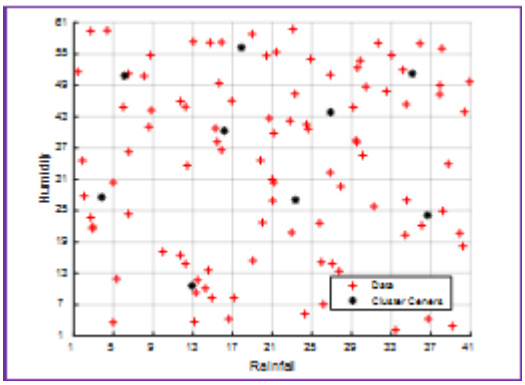

Figure 3: KNN Clusters for the Crop Yield Dataset

2. Conclusion

The present study concluded that a hybrid intelligence based approach, clustered KNN, to facilitate Agriculture Managers and Governments in sugarcane crop yield estimation through improvement of the KNN+Clustering is developed by a fusion of KNN and clustering. KNN was utilized to discover the underlying mapping between influencing factors. Clustering is applied as the best optimizer to search for KNN optimum parameters. By this mechanism, the proposed system can operate autonomously because it eliminates the trial-and-error process in parameter setting. After being trained, KNN can act as a causal prediction model to make inference some parameters whenever an input pattern is provided. Result comparisons have demonstrated that the prediction accuracy of the proposed system is far superior to that of statistical methods and KNN, achieves the lowest Error and accurate accuracy for both training and testing data sets. These facts prove the strong potential of the new hybrid system is a useful tool for sugarcane yield prediction. Thus, making long-term forecasts of crop yields using hybrid intelligence approaches can be promising future research directions in the field of agriculture and allied sectors.

References

[1] Chand, R., Raju, S. S. (2008). Instability in Andhra Pradesh Agriculture - A Disaggregate Analysis, *Agricultural Economics Research Review*, Vol. 21(2), PP: 283-288.

[2] Vakulabharanam, V. (2004). Agricultural Growth and Irrigation in Telangana: A Review of Evidence. *Economic and Political Weekly*, Vol. 39(13), PP: 1421-1426.

[3] Singh, K. K., Reddy, D. R., Kaushik, S., Rathore, L. S., Hansen, J., Sreenivas, G. (2007). Application of seasonal climate forecasts for sustainable agricultural production in Telangana subdivision of Andhra Pradesh, India, *Climate Prediction and Agriculture - Springer, Berlin, Heidelberg*, ISBN: 978-3-540-44650-7_12, PP: 111-127.

[4] Aiken, L. S., West, S. G., Pitts, S. C. (2003). Multiple Linear Regression, *Handbook of Psychology*, https://doi.org/10.1002/0471264385.wei0219 Cited by: 23.

[5] Peterson, L. E. (2009). K-Nearest Neighbor, *Scholarpedia*, Vol. 4(2), PP: 1883.

[6] Kramer, O. (2013). K-nearest neighbors, *Dimensionality reduction with unsupervised nearest neighbors, Springer, Berlin, Heidelberg*, PP. 13-23, ISBN: 978642386527.

[7] Tapas Kanungo, David M. Mount, Member, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu. (2002). An efficient k-means clustering algorithm: analysis and implementation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24(7), PP: 881–892.

[8] Chen Zhang, Shixiong Xia. (2009). K-means Clustering Algorithm with Improved Initial Center, *Knowledge Discovery and Data Mining 2009. WKDD 2009. Second International Workshop on*, PP. 790-792.

[9] Jianpeng Qi, Yanwei Yu, Lihong Wang, Jinglei Liu, Yingjie Wang, (2017), An Effective and Efficient Hierarchical K-Means Clustering Algorithm, *International Journal of Distributed Sensor Networks*, Vol. 13(8), PP: 1-17.