RESEARCH ARTICLE

Asian Research Association

INTERNATIONAL RESEARCH JOURNAL OF
MULTIDISCIPLINARY TECHNOVATION

DOI: 10.54392/irjmt2437

Conglomerate Crop Recommendation by using Multi-label Learning via Ensemble Supervised Clustering Techniques

Surekha Janrao [a, *], Kamal Shah [b], Aruna Pavte [c], Rohini Patil [d], Sandeep Bankar [e], Anil Vasoya [f]

[a] Department of Computer Engineering, K.J. Somaiya Institute of Technology, Sion, Mumbai-400022, India
[b] Department of Information Technology, Thakur College of Engineering and Technology, Kandivali, Mumbai-400101, India
[c] Department of Cyber Security, Symbiosis Skills & Professional University, Pune-412101, India
[d] Department of Computer Engineering, Terna Engineering College, Nerul, Navi Mumbai-400706, India
[e] Department of Computer Engineering, NMIMS University Navi, Mumbai-410210, India
[f] Department of Information Technology, Thakur College of Engineering and Technology, Kandivali, Mumbai-400101, India
*Corresponding Author Email: surekha.janrao@somaiya.edu

DOI: https://doi.org/10.54392/irjmt2437
Received: 17-12-2023; Revised: 20-03-2024; Accepted: 10-04-2024; Published: 24-04-2024

Abstract: Existing crop recommender related to either binary classification or multiclass classification. This paper presents conglomerate crop recommendations which consist of a number of different and distinct crops that are grouped together. In this work we focus on transferring knowledge from single label output prediction to multiple label predicted output for a given input data instances. We proposed ESCT algorithm i.e. Ensemble Supervised Clustering Techniques in our research work. ESCT provides a combined approach of conventional clustering and enhanced supervised clustering methodology to optimize the conglomerate recommendation. We are focusing on K-mean clustering for conventional approach and ICCC i.e. Inter cluster correlation coefficient to achieve enhancement in supervised clustering. In conventional K-mean clustering there is a big challenge on how to optimize the k-value of clustering which directly affects the convergence of the clusters. To resolve this problem, we mainly apply function approximation on K-Value which provides us with better clustering and fast convergence. Existing methods for inter-clustering do not adequately address one of the key challenges i.e. exploiting correlations between labels and that is achieved by ICCC algorithm. This model provides learning and prediction of unknown observation by using Back propagation MLL algorithm which provides improved performance.

Keywords: Recommender, Supervised Clustering, ESCT, BP-MLL, ICCC, K-Mean Clustering

1. Introduction

All India is moving towards a more and more digitized world and now the future is here which can adapt to more advanced and evergreen technology like Artificial Intelligence, Machine Learning, Deep Learning, IoT etc. ML is the subset of Artificial Intelligence but AI is the discipline like physics which contains theory and methods. Our main research is working on one of the machine learning problems i.e recommendation systems in the agriculture sector. As per the world agriculture statistics, India ranked fifth largest producers with 80% of agricultural produce items [1]. In the agriculture sector, more research has been going on in Crop Yield Prediction, Crop Disease prediction, Fertilizer recommendation system, crop monitoring, Precision agriculture, crop recommendation system [2]. In this paper novel research is incorporated to provide Conglomerate Crop Recommendation System.

Existing crop recommender system is based on binary classification of single crop or multiclass classification in which single input vector is mapped with either binary predicted label for single crop i.e. "YES" or "NO" or "1" or "0" whereas in multiclass classification a problem is with more than two classes [3]. But we are dealing with the problem where an input vector is assigned to group of multiple classes or labels called as Multiple Label Classification (MLC).This type of classification is the addition of the traditional binary and multiclass classification machine learning models in which all independent variables of single input data instance is assigned to group of dependent output variables called as labels [4].

MLC is an abundant task in real world scenarios. So, to achieve the better performance of this type of classification particularly in the crop recommendation system we have proposed Ensemble Supervised Clustering Techniques (ESCT). Initially the k-mean algorithm is used as a traditional approach for unsupervised clustering irrespective of the number of clusters. As we are using labelled crop data set further k-mean algorithm is enhanced by using function approximation model which is used for supervised clustering [5]. K-mean method is extensively used in numerous clustering domains because of its simple complexity and better convergence characteristic in terms of identifying similarity between homogeneous clusters within the data. Main issue in this algorithm is that the value for a set of clusters should be given on a prior basis which directly affects the convergence result of homogeneous clusters [6]. So to tackle this issue function approximation algorithm is used which analyse purity and penalty after selecting the value of K. There are four different algorithms for deciding an accurate value for set of clusters namely Silhouette Coefficient, Gap Statistic, Elbow method and Canopy method [7]. Many researchers proved that every above-mentioned algorithm for K-value selection has its own characteristics. Out of those algorithms, the elbow method gives better performance by using Sum of Squared Error (SSE) as an evaluation metric which crosses the value of k and finds infection point. This method has a simple complexity so mostly preferred to select a number of which is further analysed and verified by function approximation algorithm for cluster analysis [8].

Furthermore, ESCT uses inter-clustering Correlation Coefficient Clustering to find out relations between inter-clusters formed by the K-mean clustering technique. Existing methods for inter-clustering do not adequately address one of the key challenges i.e. exploiting correlations between labels and that is achieved by ICCC algorithm derived from Inter clustering concept [9]. All the data observations are generated from k mean cluster analysis with function approximation algorithm which constructs the clusters with each one containing a composite set of closely located observations based on a Euclidean distance metric between data points [10]. Then ICCC algorithm is applied on this distribution of data clusters to find out correlation between two clusters to achieve Inter Clustering correlation. The main objective is to group these data clusters together if their principal correlation coefficient is satisfying the minimum distance [11].

Crop recommender model has been provided which considers the crop data set having soil and whether parameters of the historical data by using Ensemble Supervised Clustering Technique (ESCT) instead of using only conventional approach. Later ESCT model has been enhanced with multi-label learning for predictive analysis of multiple crops by using Back propagation algorithm neural network classifier which gives better performance as compared to other MLL algorithms [12]. This model recommends multiple crops to the farmers which can help to increase the crop yield instead of taking the same and same crop from many years which further leads to soil erosion problem [13].

The paper proposes a multi-label prediction method for Conglomerate crop recommendation which consists of Ensemble Supervised Clustering Techniques (ESCT) and Inter cluster correlation coefficient (ICCC) using K-means clustering algorithm. A main challenge in conventional K-means clustering is the determination of the optimal value of K which directly affects the clustering results. To resolve this issue, this study mainly applies a function approximation method.

Our research contribution towards paper is a Scalable, incremental, ensemble supervised clustering multi-label classification algorithm, i.e. ESCT-BPMLL (i.e. Ensemble Supervised Clustering technique for Back Propagation Multi-label Learning) has been developed. This system has been successfully designed and implemented with an evaluation of different performance metrics such as hamming loss decreased, one error, ranking loss and average precision increased by 0.072%. These results have been validated by experimentation and same presented in the result section.

2. Related work

Literature review on the crop recommender model defines the diverse methods such as collaborative recommender system, content and fusion based hybrid recommender system by means of issues found in the working of the mentioned systems in the reference [14]. Some recommender system has been proposed for precision-based agriculture to predict a output in terms of crop by compelling the data set of soil input features by using general machine learning techniques as SVM and ANN explained in detail in the provided reference [15]. Crop prediction system based on the demand or request is developed the crop yield and crop price by using nonlinear regression model and sliding window technique on historical data [16]. Many researchers has done research on crop recommendation system by using various ML techniques which recommends the right crop to produce in the farm based on research data of various soil and whether parameters [17].

In literature review many researchers proved that ensemble clustering achieves better than a single approach in most of the applications as it provides a diverse approach between various clustering techniques [18]. Ensemble approach can excellently increase the efficiency through assimilating numerous techniques compared with a single technique, which has been proved in most of the existing research [19, 20]. Fundamentally ensemble approach provides two major techniques as ensemble classifiers and ensemble clustering [21]. In this paper, our research focuses on the ensemble clustering technique, and the main goal is to combine multiple bases clustering to obtain a possibly improved and more robust performance. The superiority of the base clustering plays an essential role in the entire ensemble process. Ensemble approach is very popular as various base clustering can apply combined to make the clustering more efficient and better.

K-mean clustering gives superior performance and due to its simplicity and fast convergence we use this type of clustering as base clustering for Ensemble technique [22]. Cluster analysis by using k-mean technique need to be use optimistic value for K i.e. number of clusters but in reality it is mostly difficult to define. K. Hansen and P. Salamon solved this difficulty by summarizing the various techniques to optimize the value for k through their experimental demonstration and analysis with definite representation [23]. Kittler *et al,* proved that execution of k-mean clustering is faster than the other conventional clustering methods due to use of euclidean distance as a similarity measure for cluster analysis. This clustering model reduces the iteration for comparing different clusters and saves the cost for computations [24].

After applying the traditional clustering approach, to optimize the value of K, a function approximation algorithm is used which comes under the supervised clustering category. Numerous approaches for supervised clustering is used to intensify the accurateness of the particular clustering model namely Clustering Function Approximation (CFA), Conditional Fuzzy Clustering (CFC), Alternation Fuzzy Clustering (AFC) algorithm explained precisely in the reference [25, 26].These algorithms has been applied in different domains as per specified in existing work [27].From above mentioned research study it has been concluded that function approximation supervised clustering is extra effective with respect to both computation time and performance as compare to fuzzy clustering and alternative fuzzy clustering algorithms. CFA algorithm is used as a complete model to approximate the function used for supervised clustering to provide optimized results for cluster analysis [28].

Existing work presents various algorithms which are used to find the distance between inter clusters like Distance based algorithms, Hierarchical based algorithms in which clusters are grouped again in the later phase [29]. But these algorithms did not address the correlation between inter clusters to analyze the relationship between the multiple class labels. In our proposed work this problem is addressed and resolved by a novel view of ICCC algorithm i.e. Inter Clustering Correlation Coefficient.

For developing the predictive modelling multi-label version back propagation neural network algorithm is used analysis of this algorithm is experimentally demonstrated in the reference paper [30]. In existing research work this algorithm is evaluated with other multi-label learning techniques namely BOOSTTEXTER boosting style algorithm, ADABOOST. MH 5 decision tree-based algorithm, RANK-SVM ranking and kernel-based support vector machine algorithm practically explained in the reference [31]. All above mentioned algorithms are general purpose program transformation multi label algorithms. From above related work researchers conclude that global error function used by the Back propagation neural network optimize the performance as compared to normal error function [32]. In our proposed work ESCT-BP-MLL is compared by Basic Back Propagation, FW, MAIAC, PLST, RAKEL. Whereas FW is the Four Class Pairwise (FW) method trains a multi-label base classifier for every pair of classes. MAIAC is a multi-label classification algorithm using auto-encoders which transforms the labels using layers of auto encoders. PLST is Principle Label Space Transformation it uses Singular Value Decomposition method (SVD) to generate a matrix that transforms the label space. RAKEL is a multi-label classification with Random K-label sets which defines a feature set.

3. Ensemble Supervised Clustering Techniques Algorithms

In this section, first we provide the problem definition of our conglomerate crop recommendation scenario. Second, we present the initial approach of Ensemble clustering i.e. k-mean clustering algorithm with function approximation for validating the number of clusters. Third, we extend ensemble clustering by using ICCC algorithm i.e. Inter correlation coefficient algorithm. In the fourth section we analyzed prediction of conglomerate crop by using Backpropogation learning algorithm for Multi-label classification.

3.1 Problem Definition

In the auxiliary Data Matrix there are 11 input features present, some features are related to soil properties and some are related to weather properties. Mapping of Auxiliary with target variables are defined as,$|D|_{X,Y} \leftrightarrow |XY|_{i,j}$ where X=Input feature matrix and Y={1, 2, 3, 4, 5….k} be the finite set of labels. In the target data there are 9 dependent variables as Y= $\{Y_1, Y_2, Y_3 \ldots\ldots Y_k\}$ where k=1 to 9.Training data set given as X =$\{X_1, X_2, X_3 \ldots\ldots X_m\}$ where (m=No of input features), for Row1 (R_1 to R_i) and $R_i = \{X_{11}, X_{12}, X_{13} \ldots . X_{ij}\}$ .In first pass cluster analysis has been applied on Intra Clusters by using k-mean clustering with a function approximation algorithm that is an ensemble approach. The resulting data matrix is defined as, $X_{ensemble}$ = $|X|_{9*11}$ from $X_{original}$ = $|X|_{2500*11}$ .Then For further optimization Dimensionality reduction technique is apply to get relevancy of attribute on $X_{ensemble}$ data matrix. We will get further reduced data matrix as $X_{Reduced}$=$|X|_{9*9}$ from $X_{ensemble} = |X|_{9*11}$ and target vector matrix=$|Y|_{1*9}$. Final hypothesis is h: $X \rightarrow Y^2$ from transformation of matrix $X \rightarrow Y^T$ which has been achieved in the second pass by using Inter correlation coefficient technique. This technique is applied on a reduced matrix to get the distance between inter clusters. For this we need to convert $X \rightarrow Y$ to $X \rightarrow Y^T$ where $Y^T = I$ (I=identity matrix).Once the hypothesis has been achieved we will get the multi-label data which is further used to build the prediction model by Back Propagation Multi-label Learning Algorithm (BP-MLL) which works more efficiently on the resultant Multi-label data set.

3.2 K-mean Clustering with function approximation (K-CFA)

In this section, we briefly describe the K-mean clustering with FA (Function Approximation) algorithm. Given a set of n I/O data vectors X → f(x) representing an unknown mapping function of f(x). The main goal is to find this unknown function which can act as a fitness function for cluster analysis and mainly for validation of the number of clusters. As this is a function approximation problem, it is necessary to incorporate the output variable into the clustering algorithm. So, we are using a traditional approach first that is k-mean clustering. The K-mean algorithm is a simple iterative clustering algorithm. For a given auxiliary data set X containing n multidimensional data points and the target data Y is to be divided, the Euclidean distance is selected as the similarity index as shown in below equation 1.

$$d(X_, Y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \qquad (1)$$

Where, n is the number of dimensions, x_k and $y_k$are the k-th attribute values of X and Y.

However, the K-value of clustering needs to be given in advance and the choice of K-value directly affects the convergence result. To solve this problem, we use the elbow method for the K-value selection algorithm. The basic idea of the elbow rule is to use a square of the distance between the sample points in each cluster and the centroid of the cluster to give a series of K values. The target of the clustering is to minimize the sum of the squares as given in following equation 2.

$$d(x_{i,} c_k) = \sum_{k=1}^{k}\sum_{i=1}^{n}\|x_i - c_k\|^2 \qquad (2)$$

$$\frac{\partial}{\partial c_k} = \frac{\partial}{\partial c_k}\sum_{k=1}^{k}\sum_{i=1}^{n}(x_i - c_k)^2$$

$$=\sum_{k=1}^{k}\sum_{i=1}^{n}\frac{\partial}{\partial c_{k_i}}(x_i - c_k)^2$$

$$=\sum_{i=1}^{n}2(x_i - c_k)$$

Let Equation (2) be zero, then $c_k = \frac{1}{n}\sum_{i=1}^{n}x_i$

The sum of squared errors (SSE) is used as a performance indicator. Iterate over the K-value and calculate the SSE. Smaller values indicate that each cluster is more convergent. When the number of clusters is set to approach the number of real clusters, SSE shows a rapid decline. When the number of clusters exceeds the number of real clusters, SSE will continue to decline but it will quickly become slower as shown in the experimental results and discussion section.

K-CFA algorithm further demonstrated by using fitness function to approximate the function f(x). We used the following fitness function which validates the number of clusters analyzed by elbow method of K-mean clustering as shown in following equation 3.

O.F. =min (f(x))

F(x) =Impurity(X) +β*Penalty (k)

Where $Impurity(X) = \frac{No\ of\ Infrequent\ observations}{n}$, and

$$Penalty(k) = \{\sqrt{\frac{k-c}{n}},\ k \geq c\ 0,\ k < c \qquad (3)$$

Where n being the total number of observations and c being the number of classes. The parameter β (0< β ≤2.0) determines the penalty that is associated with the numbers of clusters k, in a clustering. Higher values for β imply larger penalties for a higher number of clusters. The main goal of fitness function is to minimize the value of f(x). So minimum value of fitness function will approximate the results of clustering and that clustering is considered as better clustering. This fitness value is compared with the sum of squared errors (SSE) calculated by equation 3. We found the following relationship between No of clusters (k), Value of SSE, and value of f(x) as shown in following expression 4.

$$f(x) \propto K \propto \frac{1}{SSE} \qquad (4)$$

From the above equation it has been observed that no of clusters k is inversely proportional to SSE and directly proportional to f(x). So, fitness function gives us optimal value for number of clusters which can be demonstrated in the experimental results and discussion section. Hence, we have solved the big challenge of K-mean clustering that is how to optimize the k-value of clustering which directly affects the convergence by using K-CFA algorithm. Above equations are demonstrated in experimental results and discussion sections.

3.3 Inter Correlation Coefficient Algorithm for Inter Clustering (ICCC)

Cluster analysis and validation of K-CFA algorithm further extended to find out distance between inter clusters. Correlation-based distance considers two objects to be similar if their features are highly correlated, even though the observed values may be far apart in terms of geometrical distance. Pearson correlation distance is used to find out the relationship between two inter clusters. This measures the degree of a linear relationship between two clusters that is already defined by our previous algorithm K-CFA. So this type of Inter clustering is known as supervised clustering. The correlation coefficient ‘r’ takes values from –1 (large, negative correlation) to +1 (large, positive correlation).This value has been calculated by using following equation 5 and we use Absolute Pearson Correlation distance the absolute value of the Pearson correlation coefficient hence the corresponding distance lies between 0 and 1. After the initial k points are determined, we need to fill the clusters by means of defining an optimization problem by finding out the distance between two clusters k and j. $x_i^{\rightarrow j}$ is the input vector which is related to cluster j and $x_i^{\rightarrow k}$is the input vector related to cluster k. Thus, we calculate the correlation coefficient of the corresponding integer vectors for these two clusters j and k, denoted as corr $(x_i^{\rightarrow j}, x_i^{\rightarrow k})$ as shown in below equations (5, 6, 7, 8).

$$d_{corr}\left(x_i^{\rightarrow j}, x_i^{\rightarrow k}\right) = 1 - \frac{\sum_{i=1}^{n}\left(x_i^{\rightarrow j} - \underline{x}_i^{\rightarrow j}\right)\left(x_i^{\rightarrow k} - \underline{x}_i^{\rightarrow k}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i^{\rightarrow j} - \underline{x}_i^{\rightarrow j}\right)^2 \sum_{i=1}^{n}\left(x_i^{\rightarrow k} - \underline{x}_i^{\rightarrow k}\right)^2}} \quad (5)$$

For cluster j,

$$x_i^{\rightarrow j} = \left(x_{i,1}^{j}, x_{i,2}^{j}, \ldots.., x_{i,n_i}^{j}\right) \quad (5.1)$$

For cluster k,

$$x_i^{\rightarrow k} = \left(x_{i,1}^{k}, x_{i,2}^{k}, \ldots.., x_{i,n_i}^{k}\right) \quad (5.2)$$

$$d\left(x_i^{\rightarrow j}, x_i^{\rightarrow k}\right) = d\left(x_i^{\rightarrow j}, x_i^{\rightarrow k}\right) \quad (6)$$

For representing the correlation between inter clusters consider a matrix model M_{nxk} is a Boolean matrix, whose elements depict the belonging inter clusters. Each row of the array represent an $x_i(i = 1,2,\ldots n)$ objects) and each column is a cluster $Y_j(1,2,\ldots k; k = number\ of\ clusters)$.If an element of this Boolean matrix x_{ij} takes a value equal to 1, it implies that the object i belongs to cluster j, otherwise, when x_{ij} takes a value equal to 0, it means that object i does not belong to cluster j as shown in the following matrix [48].

$$M_{n\times k} = X_1\, X_2 \,:\, X_k\ [Y_1\, Y_2 \,..\, Y_k\ ] \quad (7)$$

The sum of elements of each row shows number of labels that an object can belong to a single group of clusters having multiple labels or classes:$\sum_{j=1}^{k} m_{ij} = 1; \forall_i = 1,2,3 \ldots.., n$. This sum predicts the group of conglomerate recommendations of crops used as a predicted dependent variable in our data set. The sum of each column shows the size of each cluster, $\sum_{i=1}^{n} m_{ij1} = e_j; \forall_j = 1,2,3 \ldots.., k$.in this way inter clustering has been achieved and we have reached to the final hypothesis which is h: $X \rightarrow 2^y$ defined in above correlation matrix M_{nxk} from transformation of matrix.

$X \rightarrow$Y. Following the above hypothesis proved by correlation matrix, we pose this problem as a linear correlation problem, the objective function (O.F.) is given by the expression. This will be experimentally demonstrated in the Experimental results and discussion section.

$$O.F = min(\sum_{j=1}^{k} \sum_{i=1}^{n} d\left(x_i^{\rightarrow j}, x_i^{\rightarrow k}\right), X_{i,j} \quad (8)$$

3.4 Backpropogation Neural Network Multi-label learning (BP-MLL)

An intuitive approach to solving a Multi-label problem is to decompose it into multiple independent binary classification problems. This kind of method does not consider the correlations between the different labels of each instance and the expressive power of such a system can be weak which has been proved by researchers in the reference. In this paper, a neural network algorithm named BP-MLL, i.e., Back propagation for Multi-label Learning, is proposed, which is the first Multi-label neural network algorithm defined as below. The basic architecture of BP-MLL is described in the reference.

Let $\{X_i\}_{i=1}^{i} \subset R^n$ denote the auxiliary domain of instances and let $Y \subset R^d$ be the output data set. y= {1, 2, 3 …k} be the finite set of labels. Given a training set$Q = \{(x_1, Y_1), (x_2, Y_1), \ldots, (x_m, Y_m)\}$

$(x_i \in X, Y_i \subseteq y)$ drawn from unknown distribution D, the objective of learning system is to output a multi-label classifier as defined in the matrix of above section and in the hypothesis as h: $X \rightarrow Y^2$ which optimizes some specific evaluation metric. The goal of Multi-label learning is to predict the label sets of unseen instances, an intuitive way to define the global error of the network on the training set could be BP-MLL is derived from the popular Back propagation algorithm through replacing its error function with a new global error function defined to capture the characteristics of Multi-label learning as shown in the following equation (9, 10, 11).

$$E = \sum_{i=1}^{m} E_i \quad (9)$$

$$E_i = \sum_{j=1}^{k}\left(c_j^i - d_j^i\right)^2 \quad (10)$$

Where $c_j^i = c_j(x_i)$ is the actual output of network on x_i on the jth class and d_j^i is the desired output of the x_i on the jth class, which takes the value of either 1 (j $\in Y_i$) or 0(j $\notin Y_i$).

The error function defined in equation (10) concentrates only on single label discrimination, i.e., whether a particular label belongs to the instance x_i or not; it does not consider the correlations between the different labels in Y_i should be ranked higher than those not in Y_i. These characteristics of multi-label learning are appropriately addressed by rewriting the global error function as follows:

$$E = \sum_{j=1}^{n} E_j = \sum_{j=1}^{n} \frac{1}{|Y_j|} \sum_{(k.m)\in Y_j \times Y_j}^{n}) \quad (11)$$

The right hand side of the equation (11) defines the error of the network on the i^th^ multilable training example(x_i, Y_j).Hence Y_j is the complementary set of $Y_j$in y.This term $c_k^j - c_m^j$ measures the difference

between the outputs of the network on one label belonging to x_i(k $\in Y_j$) and one label not belonging to it (m ∈ Y_j). If this difference is bigger it gives better performance. The negation of this difference is fed to the exponential function in order to severely penalize the i^{th} error term if c_k^j is much smaller than c_m^j .The summation in the i^{th} error term takes account of the accumulated difference between the outputs of any pair of labels with one belonging to x_i and another not belonging to x_i. In this way, the correlations between different labels of x_i i.e., labels in Y should get larger network outputs than those in Y_j , are appropriately addressed. BP-MLL consumes much more time in the training phase than all the other algorithms. So we are using a reduced data matrix resulting from an ensemble process to create a BP-MLL model for predicting the multiple crops or Conglomerate crops as shown in the following figure 1.

4. Experimental Results and Discussion

In this section, we first present the description of the datasets then in the second section we define an evaluation metric. In the last section we present the evaluation of results of all the algorithms used in ensemble process and for multi-label learning with the comparison of other methods.

4.1 Description of Dataset

Our Proposed algorithm on crop dataset has been collected from annual reports of the year 2021-22, 2019-20 and 2017-18, from Maharashtra http://agricoop.nic.in/ (Ministry of Agriculture and Farmers welfare). Total 11 features have been used in the data set including one class label attribute. This dataset contains total $2.5 * 10^3$ Data Instance detailed Statistical analysis of data is described in the following table1.

4.2. Performance Evaluation Parameters

Evaluation parameters for clustering are introduced in the following section.

Sum of Squared Error (SSE) is the most common measure, for each point, the error is the distance to the nearest cluster. To get SSE, we square these errors and sum them as shown in equation 12.

$$SSE = \sum_{i=1}^{K} \sum_{x\in C_i}^{k} dist^2(m_i, x) \qquad (12)$$

Class impurity, Impurity(X) measured by the percentage of minority examples in the different clusters of a clustering X A minority example is an example that belongs to a class different from the most frequent class in its cluster.

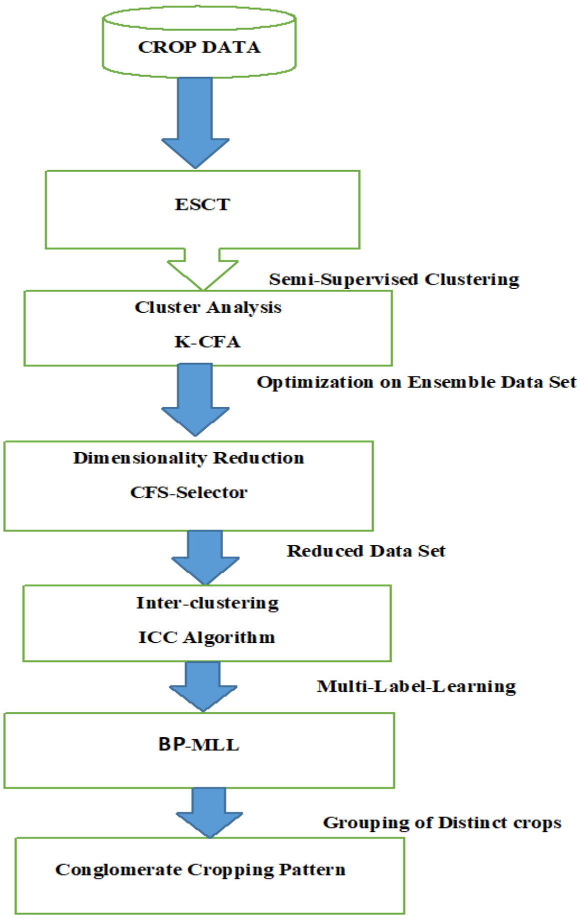

Figure 1. Ensemble Supervised Clustering Technique for Conglomerate Crop Recommendation System

Table1. Statistical data for features

Input Features	Data Type	Statistics Values for Features			
		in	**Max**	**Mean**	**Standard Deviation**
Soil_Ph	Numeric	5.4	8.6	7.31	0.67
N	Numeric	25	175	86.107	43.997
P	Numeric	25	100	49.92	22.09
K	Numeric	10	100	43.86	24.91
Soil_Depth	Numeric	5	59	23.56	12.09
Temperature	Numeric	14	49	26.44	6.11
Rainfall	Numeric	50	1395	765.24	314.22
Humidity	Numeric	20	28	26.08	2.692
Soil type	Alluvial soil, Red Soil, Black soil, Arid soil				
Water storage	Yes, No(can be taken as 1,0)				
Target Outcomes(**DataType :Categorical)**	CORN, COTTON, SOYABEAN, BAJRA, SUGARCANE, RICE, GROUNDNUT, WHEAT, JOWAR (multiple crops labels)				

A Fitness Function for Supervised Clustering In particular, we used the following fitness function in our experimental work (lower values for q(X) indicate a 'better' solution) already discussed in the proposed algorithm section.

$$q_X = Impurity(X) + \beta * Penalty(k) \qquad (13)$$

4.3 Evaluation of results of proposed Algorithms.

From below results it has been observed that no of clusters k is inversely proportional to SSE and directly proportional to f(x) expression of equation 4 has been proved over here. The Elbow Method algorithm uses

SSE as a performance metric, traverses the K value, finds the inflection point, and has a simple complexity. The inadequacy is that the inflection point depends on the relationship between the K value and the distance value. Following figure 2 shows the inflection point at two values of k i.e. k=7 and k=9 and if the inflection point is not obvious, the K value cannot be determined accurately.

Following table2 evaluate the results of very first proposed algorithm i.e. K-CFA. The main goal of this algorithm is to minimize the value of f(x) and compare with the result of SSE i.e. sum of squared error which is evaluated by using elbow method in K-mean clustering algorithm. But this method does not give a definite result for k-value selection. So function approximation function is used to solve this issue as the minimum value of fitness function will approximate the results of clustering and that clustering is considered as better clustering challenge to optimize the k-value of clustering which directly affects the convergence. So by using the fitness function i.e. q(x) of K-CFA algorithm k value has been validated by calculating class impurity and penalty for the value of k. From above results given in table2 it has been analysed that for k= 9 gives better clustering, minimum error and less impurity. Cluster analysis and validation of K-CFA algorithm further extended to find out distance between inter clusters by applying ICCC (Inter Correlation Coefficient algorithm) based on the following condition for correlation.

$$x_i^{\to j} = \{ 1,\ r_{corr}(x_i^{\to j}, x_i^{\to k}) \geq 0.9\ 0,\ r_{corr}(x_i^{\to j}, x_i^{\to k}) < 0.9 \qquad (14)$$

Above results interpret the multi-label dataset now multi-label algorithm has been applied to train the model and give the conglomerate crop recommendation. ESCT-BP-MLL algorithm is compared with Basic Back Propagation, four class Pairwise method FW, Maniac -Multi-label classification using Auto encoders, PLST is Principle Label Space Transformation, RAKEL is a multi-label classification with Random k Label sets. All these algorithms are related to adaptation methods for multi-label learning instead of problem transformation methods.

Following figure 3 (a,b,c,d) illustrates and validates how various metric values of ESCT-BP-MLL are comparable to other values in terms of Hamming Loss, One-Error, Ranking Loss and average precision.

This result concludes that ESCT-BP-MLL gives better performance as compare to other multi-label learning.

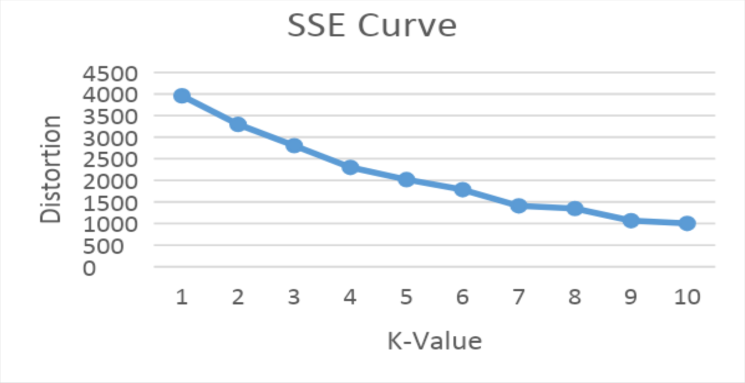

Figure 2. The elbow method showing the optimal value of k

Table 2. Evaluation of Crop Data set results for K-CFA algorithm

K value	SSE	Class Impurity	β Constant	Penalty	q(x)	Execution Time(sec)	Iteration
K=1	3.9×10^3	0.2	0.4	0	0.2	0	1
K=2	3.2×10^3	0.2	0.4	0	0.2	0.11	5
K=3	2.8×10^3	0.2	0.4	0	0.2	0.06	11
K=4	2.3×10^3	0.2	0.4	0	0.2	0.11	18
K=5	2.0×10^3	0.2	0.4	0	0.2	0.13	21
K=5	1.7×10^3	0.2	0.4	0	0.2	0.14	8
K=7	1.4×10^3	0.2	0.4	0	0.2	0.09	13
K=8	1.3×10^3	0.2	0.4	0	0.2	0.16	13
K=9	1.0×10^3	0.2	0.4	0	0.2	0.14	7
K=10	1.0×10^3	0.2	0.4	0.02	0.40	0.14	7

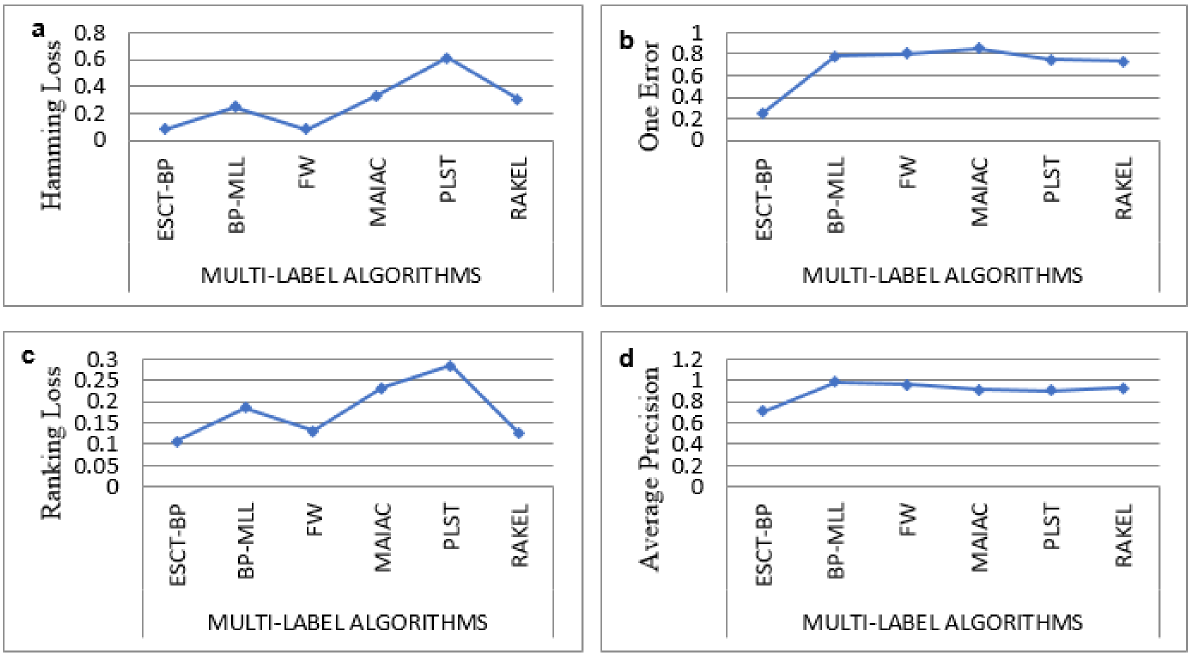

Figure 3. The performance of different multi-label Algorithms for different evaluation metrics (a) Hamming loss. (b) One-error. (c) Ranking loss. (d) Average precision.

Table 3. Comparison Between Existing and Proposed Multi-label algorithms in terms of various performance metrics like (Hamming Loss, One Error, Ranking Loss, Average Precision)

Evaluation Criteria	Multi-label Algorithms					
	ESCT-BP	BP-MLL	FW	MAIAC	PLST	RAKEL
HAMMING LOSS	0.08	0.25	0.08	0.333	0.611	0.306
ONE ERROR	0.25	0.78	0.80	0.85	0.75	0.73
RANKING-LOSS	0.107	0.185	0.130	0.232	0.285	0.125
AVERAGE PRECISION	0.706	0.983	0.958	0.911	0.9	0.93

At the end after validating the results of ESCT-BP-MLL, the multi-label learning algorithm gives minimum value for Hamming Loss, One Error and Ranking loss and Maximum value for average precision which gives the better performance for multi-label classification as shown in the following table 3.

From the above results it has been observed that our research contribution is a Scalable, incremental, ensemble supervised clustering multi-label classification algorithm, i.e. ESCT-BPMLL (i.e. Ensemble Supervised Clustering technique for Back Propagation Multi-label Learning) has been developed. This system has been successfully designed and implemented with an evaluation of different performance metrics such as hamming loss decreased by 0.2 %, one error decreased by 0.5%, ranking loss decreased by 0.1%, and average precision increased by 0.2%. These results have been validated by experimentation and same presented in the table 3.

At the end after applying grouping of inter-clusters on results of Table3 we got conglomerate grouping of crops which is much more helpful for the farmers and also other agriculture experts in terms of crop management and yield improvement. And they would be properly educated regarding which crop to grow, and which not to grow.

5. Conclusion

In this paper, we presented a novel Ensemble approach of supervised clustering by using different techniques on crop data sets. This approach is further enhanced to get the relationship of multiple labels by applying correlation models on clustered data. This is further validated by applying multiple label learning. Results of this interpret BP-MLL gives superior performance as compared to other multi-label learning algorithms in terms of different evaluation criteria. Then inter-clustering- of multiple labels of crops provides the conglomerate crop recommendation rather than single label crop recommendation. This recommender system helps the farmers or any other agriculture experts in terms of crop management and yield improvement. And they would be properly educated regarding which crop to grow, and which not to grow.

References

[1] K.K. Jha, A. Doshi, A comprehensive review on automation in agriculture using artificial intelligence. Artificial Intelligence in Agriculture, 2, (2019) 1-12. https://doi.org/10.1016/j.aiia.2019.05.004

[2] S.K. Prasad, B.S. Sreedharan, S. Jaishanth, (2017) Crop monitoring and recommendation system using machine learning techniques. Madras Institute of Technology, Chennai.

[3] N. Hegde, S. Sannidhi, R. Navada, S. Jambarmath, R. Madhavi, Survey paper on Agriculture Yield Prediction Tool using Machine Learning. International Journal of Advance Research in Computer Science and Management Studies, 5(11), (2017) 36-39.

[4] G. Tsoumakas, I. Katakis, I. Vlahavas, Effective and Efficient Multi-label Classification in Domains with Large Number of Labels. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD '08), 21, (2008) 53-59.

[5] E. Gibaja, S. Ventura. A tutorial on multi-label learning. ACM Computing Surveys, 47(3), (2015) 1–38. https://doi.org/10.1145/2716262

[6] T. Kanungo, D.M. Mount, N. Netanyahu, An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), (2002), 881-892. https://doi.org/10.1109/TPAMI.2002.1017616

[7] M.H. Almannaa, M. Elhenawy, H.A. Rakha, A novel supervised clustering algorithm for transportation system applications. IEEE transactions on intelligent transportation systems, 21(1), (2020) 222-232. https://doi.org/10.1109/TITS.2018.2890588

[8] Z. H. A. I. Dong-hai, Y. Jiang, G. Fei, Y.U. Lei, D.I.N.G. Feng, K-means text clustering algorithm based on initial cluster centers selection according to maximum distance. Application Research of Computers/Jisuanji Yingyong Yanjiu, 31(3), (2014) 713–719.

[9] H. Pomares, I. Rojas, M. Awada, O. Valenzuela, An enhanced clustering function approximation technique for a radial basis function neural network. Journal of Science Direct on Mathematical and Computer Modelling, 55(3-4), (2012) 286-302. https://doi.org/10.1016/j.mcm.2011.07.010

[10] K.O. McGraw, S.P. Wong, Forming inferences about some intraclass correlation coefficients. Psychological methods, 1(1), (1996) 30-46. https://psycnet.apa.org/doi/10.1037/1082-989X.1.1.30

[11] X. Li, N. Ye, Grid- and dummy-cluster-based learning of normal and intrusive clusters for computer intrusion detection. Quality and Reliability Engineering International,18(3), (2002) 231–242. https://doi.org/10.1002/qre.477

[12] M.L. Zhang, Z.H. Zhou, Multilabel neural networks with applications to functional genomics and text categorization. IEEE transactions on Knowledge and Data Engineering, 18(10), (2006) 1338-1351. https://doi.org/10.1109/TKDE.2006.162

[13] A. Manjula, G.Narsimha, (2015) XCYPF: A Flexible and Extensible Framework for Agricultural Crop Yield Prediction. 9th International Conference on Intelligent Systems and Control (ISCO), IEEE, India. https://doi.org/10.1109/ISCO.2015.7282311

[14] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the art and possible extensions. IEEE Transactions on Knowledge & Data Engineering, 17(6), (2005) 734-749. https://doi.org/10.1109/TKDE.2005.99

[15] N. Dumbre, O. Chikane, G. More, System for Agriculture Recommendation Using Data Mining. International Education & Research Journal (IERJ), 1(5), (2015) 2454-9916.

[16] K.A. Reddy, R.K. Kumar, Recommendation System: A Collaborative Model for Agriculture. International Journal of Computer Sciences and Engineering, 6(1), (2018) 120-123. https://doi.org/10.26438/ijcse/v6i1.120123

[17] R. Rajak, A. Pawar, M. Pendke, P. Shinde, S. Rathod, A. Devre, Crop recommendation system to maximize crop yield using machine learning technique. International Research Journal of Engineering and Technology (IRJET), 4(12), (2017) 950-953.

[18] S. Raja, K.S. Rishi, R. Sundaresan, E. Srijit, (2017) Demand based crop recommender system for farmers. IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), IEEE, India. https://doi.org/10.1109/TIAR.2017.8273714

[19] S. Pudumalar, E. Ramanujam, R. Rajashree, H. Kavya, C. Kiruthika, J. Nisha, (2017) Crop recommendation system for precision agriculture. 2016 Eighth International Conference on Advanced Computing (ICoAC), IEEE, India. https://doi.org/10.1109/ICoAC.2017.7951740

[20] M.J. Mokarama, M.S. Arefin, (2017) RSF: A recommendation system for farmers. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), IEEE, Bangladesh. https://doi.org/10.1109/R10-HTC.2017.8289086

[21] S. Mao, W. Lin, L. Jiao, S. Gou, J. Chen, End-to-End Ensemble Learning by Exploiting the Correlation Between Individuals and Weights. IEEE Transactions on Cybernetics, 51(5), (2021) 2835-2846. https://doi.org/10.1109/TCYB.2019.2931071

[22] Z.H. Zhou, (2015) Ensemble learning. Encyclopedia of Biometrics. Boston, Springer, USA. https://doi.org/10.1007/978-1-4899-7488-4_293

[23] L.K. Hansen, P. Salamon, Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(10), (1990) 993–1001. https://doi.org/10.1109/34.58871

[24] J. Kittler, M. Hatef, R.P. W. Duin, J. Matas, On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), (1998) 226–239. https://doi.org/10.1109/34.667881

[25] S. Vega-Pons, J. Ruiz-Shulcloper, A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence, 25(3), (2011) 337–372. https://doi.org/10.1142/S0218001411008683

[26] A.K. Jain, Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), (2010) 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[27] L.B. Han, Q. Wang, Z.F. Jiang, Z.Q. Hao, Improved k-means initial clustering center selection algorithm. Jisuanji Gongcheng yu Yingyong (Computer Engineering and Applications), 46(17), (2010)150–152.

[28] UCI Machine learning repository. Available, http://archive.ics.uci.edu/ml

[29] M. Ali, X. Li, Z. Yang Dong, (2005) Efficient Spatial Clustering Algorithm Using Binary Tree. Intelligent Data Engineering and Automated Learning - IDEAL 2005. Lecture Notes in

Computer Science, Springer, Berlin. https://doi.org/10.1007/11508069_39

[30] J. Gonzalez, I. Rojas, H. Pomares, J. Ortega, A. Prieto, A new clustering technique for function approximation. IEEE Transactions on Neural Networks, 13(1), (2002) 132–142.

[31] Zhang, M. L., & Zhou, Z. H. (2013). A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering, 26(8), 1819-1837. https://doi.org/10.1109/TKDE.2013.39

[32] S. Ji, L. Sun, R. Jin, J. Ye, Multi-label multiple kernel learning. Advances in Neural Information Processing Systems, 21 (2008).

Acknowledgement

The authors thanks to all co-authors for their research contribution for writing this article through different levels.

Authors Contribution Statement

Dr. Surekha Janrao: Conceptualization, Methodology, Data collection, Formal analysis, Writing - Original Draft, Writing - Review & Editing; Kamal Shah - Formal analysis, Writing - Original Draft, Writing - Review & Editing; Aruna Pavte- Writing - Original Draft; Rohini Pati – Methodology, Software validation, Writing - Review & Editing; Dr. Sandip Bankar - Writing - Review & Editing; Anil Vasoya - Writing - Review & Editing. All the authors read approved the final version of the manuscript.

Funding

The authors declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors declares that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

Data will be provided upon request.

Has this article screened for similarity?

Yes

About the License