

Use of Artificial Intelligence

in statistical research

Ranjit Paul¹, Prawin Arya² and Shiv Kumar³

ICAR-Indian Agricultural Statistical Research Institute, New Delhi 110 012

Artificial intelligence can be defined as the intelligence exhibited by machines or software. To expand the idea further in simple terms one can, see this as art of making the computers think and act like humans. Activities that can be associated with human thinking, like decision making and learning can be programmed to machines so that they can act like human beings. The process of creating systems that think like humans include cognitive modeling, replicating human thought process and decision making using logical reasoning. Along with thinking, systems behave like humans using natural language processing for communication, storing information effectively and efficiently along with automated reasoning. Due to onset of growing population there arises need for fulfilling demands of billions of people in terms of goods across all the sectors of agriculture. This brings in, the evolving concept of artificial intelligence in use of agriculture to meet demands of future population.

Key words: Artificial Intelligence, Bayesian Model, Statistical Research

COMPUTATIONAL modeling uses simulations to study how human intelligence may be structure based on cognitive science. Cognitive science is the interdisciplinary study of cognition in humans, animals, and machines. It encompasses the traditional disciplines of psychology, computer science, neuroscience, anthropology, linguistics and philosophy. The goal of cognitive science is to understand the principles of intelligence with the hope that will lead to better comprehension of the mind and of learning and to develop intelligent devices. A central tenet of the cognitive science is that a complete understanding of mind/brain could be attained by studying at all possible levels. Cognitive science deals with intelligence and behavior, with a focus on how nervous systems represent, process, and transforms information. The fundamental concept is that “thinking can best be understood in terms of representational structures in the

mind and computational procedures that operate on these structures. Use of cognitive computing is one of the best solutions to cater problems of farmers, as a service, a chatbot platform provides help to farmers on different technological advancements which help them to increase their yield per hectare.

On daily basis huge amount of data is gathered for different crops, weather conditions, soil reports, rainfall, and images from drones. Using this data one can intelligently get insights to improve yield. Proximity Sensing and Remote Sensing techniques can also be used to gather high resolution data on soil testing. In case of remote sensing the sensors are built in the satellites or remote devices whereas proximity sensing requires sensors to be in close contact or range of work fields. This helps in characterization of soil in a particular place and desired crops can be planted accordingly.

Technology can also be used to identify optimal sowing period,

historic climate data, real time Moisture Adequacy Data (MAI) from daily rainfall and soil moisture to build predictability and provide inputs to farmers on ideal sowing time.

Image based farming is giving rise to Precision Farming. Drone based images are helpful in in-depth analysis, crop monitoring and so on. Using Drone data and images, farmers can get real time alerts on action needed in the fields, which can accelerate Precision Farming.

Computer vision technology and drone can be put to use in:

- *Disease detection* - Images captured in high quality from fields are segmented into non diseased and diseased part. Diseased parts is then cropped and send to remote labs for further diagnosis as it helps in pest detection and nutrient deficiency recognition so that necessary action can be taken.
- *Crop readiness identification* – Data on crop readiness is captured using images of different crops

under white, UV-A lights. This would help farmers to determine how ripe the crop is. Processing of the images using AI will help the farmer in taking appropriate decision and devising strategy to market the produce for availing of the maximum price in market to enhance their income.

- **Field management** - Using Drone Technology farmers can get an overview map of field, which can ease to find out areas where crops require water, fertilizer or pesticides. This helps in resource optimization to a huge extent.

Support of AI can help in identification of optimal mix for agronomic products. Based on multiple parameters like soil condition, weather forecast, types of seeds in a certain area, cognitive solutions provide recommendations to farmers on the best choice of crops and hybrid seeds. These recommendations can be further personalized based on the farmer's requirements like local conditions and data about successful farming in the past based on marketplace trends, prices or consumer needs. These factors may also enable farmers to take well informed decisions.

In terms of human intensive processes in farming, irrigation is one such process. Machines trained on historical weather pattern, soil quality and kind of crops to be grown can automate the irrigation process and increase overall yield.

Further, drone based AI solutions in agriculture have a lot of significance in terms of managing adverse weather conditions, productivity gains, Precision Farming and yield management.

Applications of AI

- **Machine learning** automates analytical model building. It uses methods from neural networks, statistics, operations research and physics to find hidden insights in data without being explicitly programmed where to look or what to conclude.
- **A neural network** is a kind of machine learning inspired by the workings of the human brain. Its a computing system made up of

interconnected units (like neurons) that processes information by responding to external inputs, relaying information between each unit. The process requires multiple passes at the data to find connections and derive meaning from undefined data.

- **Deep learning** uses huge neural networks with many layers of processing units, taking advantage of advances in computing power and improved training techniques to learn complex patterns in large amounts of data. Common applications include image and speech recognition.
- **Computer vision** relies on pattern recognition and deep learning to recognize what's in a picture or video. When machines can process, analyze and understand images, they can capture images or videos in real time and interpret their surroundings.
- **Natural language processing (NLP)** is the ability of computers to analyze, understand and generate human language, including speech. The next stage of NLP is natural language interaction, which allows humans to communicate with computers using normal, everyday language to perform tasks.

AI IN STATISTICAL ANALYSIS

Recently, Artificial Intelligence (AI) techniques have shown great ability in modeling and forecasting non-linear time series. AI techniques offer an effective approach for handling large amounts of dynamic, non-linear and noisy data, especially when the underlying physical relationships are not fully understood. This makes them well suited to time series modeling problems of a data-driven nature. In general, the application of AI technique does not require a prior knowledge of the process. Followings are the few types of machine learning techniques used for statistical analysis:

Machine learning in statistics

1. Support Vector Machine

2. Artificial Neural Network
3. Decision Tree and Random Forest
4. K- nearest neighbors
5. Bayesian Models

Support vector machine: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. In real world application, finding perfect class for millions of training data set takes lot of time. This is called regularization parameter. There are two terms: regularization parameter and gamma. These are tuning parameters in SVM classifier. Varying those we can achieve considerable non-linear classification line with more accuracy in reasonable amount of time.

Artificial neural network (ANN): Artificial neural networks (ANNs) are nonlinear data driven self-adaptive approach and are powerful tools for modeling, especially when the underlying data relationship is unknown. A very important feature of these networks is their adaptive nature, where "learning by example" replaces "programming" in solving problems. A neural network consists of a set of connected cells (neurons). Neurons receive impulses from input cells or other neurons, perform some kind of transformation of input, and transmit outcome to other neurons to output cells. The neural networks are built from layers of neurons connected so that one layer receives input from the preceding layer of neurons and passes the output on to the subsequent layer. Most commonly used ANN is multi-layer perceptron (MLP), a class of feed forward neural network. MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron, which distinguish

data that is not linearly separable.

Random forest: Decision tree learning is a method commonly used to create a model that predicts the value of target variable based on several input variables. Decision trees are of two types; (i) classification tree, and (ii) regression tree. Classification tree analysis is used when the response variable is a class label and the regression tree analysis is used when the response variable takes the values of real number. A classification tree is obtained by asking an ordered sequence of questions, where the type of questions asked at each step in the sequence depends upon the answers required for the previous questions of the sequence. The sequence always terminates in a prediction of the class label attached to the observation. The starting point of a classification tree is called the root node and consists of the whole data set at the top of the tree. A node in a tree can be a terminal or non-terminal node. A non-terminal (or parent) node is a node that split into daughter nodes. A node that do not split is called a terminal node and is assigned a class label. When an observation of unknown class is dropped down the tree and ends up at a terminal node, it is assigned to that class corresponding to the class label attached to that node. There may be more than one terminal node with the same class label. A single split tree with only two terminal nodes is called a stump. In case of binary splitted node, the split is determined by a Boolean condition on the value of a single variable, where the condition is either satisfied (“yes”) or not satisfied (“no”) by observed value of that variable. All the observations in the data set that have reached to a particular node and satisfy the condition for that variable drop down to one of the two daughter nodes and the remaining observations at that node that do not satisfy the condition drop down to the other daughter node.

k-nearest Neighbors: The model representation for k-nearest Neighbors (KNN) is the entire training dataset. KNN has no model other than storing the entire dataset,

so there is no learning required. Efficient implementations can store the data using complex data structures like k-d trees to make look-up and matching of new patterns during prediction efficient. Since, the entire training dataset is stored, we may want to think carefully about the consistency of the training data. It might be a good idea to curate it, update it often as new data becomes available and remove erroneous and outlier data.

Making predictions with KNN

KNN makes predictions using the training dataset directly. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. For regression this might be the mean output variable, in classification this might be the mode (or most common) class value. To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance.

Other popular distance measures include:

- **Hamming Distance:** Calculate the distance between binary vectors.
- **Manhattan Distance:** Calculate the distance between real vectors using the sum of their absolute difference. Also called City Block Distance.
- **Minkowski Distance:** Generalization of Euclidean and Manhattan distance.

The computational complexity of KNN increases with the size of the training dataset. For very large training sets, KNN can be made stochastic by taking a sample from the training dataset from which to calculate the K-most similar instances.

Bayesian models: Bayesian Model statistics is a branch of statistics where quantities of interest (such as parameters of a statistical model) are treated as random variables, and one draws conclusions by analyzing the posterior distribution over these quantities given the observed data.

While the core ideas are decades or even centuries old, Bayesian ideas have had a big impact in machine learning in the past 20 years or so because of the flexibility they provide in building structured models of real-world phenomena. Algorithmic advances and increasing computational resources have made it possible to fit rich, highly structured models which were previously considered intractable.

Bayesian methods are trying to solve one of the following problems:

- **Parameter estimation:** Suppose we have a statistical model of some domain, and we want to use it to make predictions or maybe we think the parameters of the model are meaningful, and we want to fit them in order to learn something about the world. The Bayesian approach is to compute or approximate the posterior distribution over the parameters given the observed data.
- **Model comparison:** We may have several different models under consideration, and we want to know which the best match to the data is. A common case is that we have several models of the same form of differing complexities, and we want to trade off the complexity with the degree of fit.

Different types of Bayesian models

- **Logistic Regression:** A discriminative model for predicting binary targets for given input features.
- **Bayesian networks (Bayes nets):** Bayes nets are directed graphs which encode patterns of probabilistic dependencies between different random variables, and are typically chosen to represent the causal relationships between the variables. A statistical model represents a set of variables and their conditional dependencies via graph. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. While Bayes nets can be learned in a non-Bayesian way, Bayesian techniques can be used to learn both the parameters and structure (the

set of edges) of the network.

- *Latent Dirichlet Allocation:* a “Topic Model,” where a set of documents (e.g. web pages) are each assumed to be composed of some number of topics, such as computers or sports. Treats data as observations which contains hidden variables. These hidden variables reflect thematic structure of the collection of data. The approach would be to infer hidden structure using posterior inference, discovering topics in the collection using Bayesian inference, placing new data into the estimated model etc. Related models include nonnegative matrix factorization and probabilistic latent semantic analysis.
- *Linear Dynamical Systems:* A time series model where a low-dimensional gaussian latent state evolves over time, and the observations are noisy linear

functions of the latent states. This can be thought of as a continuous version of the Hidden Markov Models (HMM). Inference in this model can be performed exactly using the Kalman filter and smoother.

- *Sparse Coding:* A model where each data point is modeled as a linear combination of a small number of elements drawn from a larger dictionary. When applied to natural image patches, the learned dictionary resembles the receptive fields of neurons in the primary visual cortex.

Limitation of AI

Though Artificial Intelligence offers vast opportunities for application in agriculture, there still exists a lack of familiarity with high tech machine learning solutions in farms. Exposure of farming to external factors like weather conditions, soil conditions and

presence of pests is quite a lot. So what might look like a good solution while planning during the start of harvesting, may not be an optimal one because of changes in external parameters results in change in average crop yield per hectare.

AI systems also need a lot of data to train machines and to make precise predictions. In case of vast agricultural land, though spatial data can be gathered easily, temporal data is hard to get. For example, most of the crop-specific data can be obtained only once in a year when the crops are growing. Since the data infrastructure takes time to mature, it requires a significant amount of time to build a robust machine learning model.

¹Scientist, ²Principal Scientist, ³Principal Scientist, ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi. Corresponding authors' e mail: ranjitstat@gmail.com

Ant Colony Optimization

Nature has gifted, extraordinary capability to ants to perform complex problem solving through their ability to act in coordinated manner. One particularly example of their skill is to find shortest path between food and their store house. Researchers have been motivated by this particular skill of ants. In computer programming field, researchers have been formulating various ways to get computers to find shortest path between two given points for finding solutions to various real life problems. Problems include finding shortest distance between two cities connected by road, finding nearest service centre to service a client, etc. Ant colony optimization method is a meta-heuristic to mimic in a computer algorithm, how in a given population, individuals can cooperate by sharing information as a matter of habit for problem solving.



Compiled by: Dr Himanshu Varshney, Scientist, ICAR-DKMA, himanshu@icar.org.in