ABSTRACT

Since the invention of computer, the information available in every field has been digitized to be accessible by people using computer resources. Hence, data are growing rapidly in gargantuan amount in every domain. One such domain of interest for researchers is agriculture field. To digitalize agriculture data, Government of Gujarat introduced Soil Health Card which contains macro and micronutrients records of soil samples taken from farms and examined in soil laboratory to record the details into database. It is a very large database containing information of all the farms in Gujarat. To find interesting patterns from the database, we applied and analyzed concepts of data mining on soil health card database in this thesis.

Data mining is the process of getting useful information by analyzing different kind of data. The object classification is an important area within the field of data mining and its application extends to various areas, whether or not in the branch of science. In this research, we have concentrated on k-Nearest Neighbor classification algorithm to classify soil sample instances into appropriate fertilizers deficiency category. Although k-Nearest Neighbor classification is simple and effective technique, having an extensive training set is an element of importance in order to obtain a high accuracy, on the other hand, it makes the classification of each object slower due to its lazy-learning algorithm nature. The k-Nearest Neighbour classifier has the large computational and storage requirements. In addition, the effectiveness of classification decreases because of uneven distribution of training data.

To overcome above mentioned limitations, we proposed fast k-Nearest Neighbor which generates training set prototype based on either Elbow method or Silhouette method, training set reduction k-Nearest Neighbor which reduces training set based on prototype selection and hybrid k-Nearest Neighbor classification methods which combines both prototype generation and prototype selection mechanism to generates prototype from original training set. These all methods aims at, how to decrease the requirement of time and space for classification task done by k-Nearest Neighbour classifier. We have applied our new approaches of classification on soil health card agriculture data set and our evaluation illustrates that these approaches can solve the mentioned problems effectively.