An Improved Feature Selection and Classification using Decision Tree for Crop Datasets

Surabhi Chouhan
Department of Computer Science
University Institute of Technology,
Barkatullah University, Bhopal

Divakar Singh
Department of Computer Science
University Institute of Technology,
Barkatullah University, Bhopal

Anju Singh
Department of Computer Science
University Teaching Department,
Barkatullah University, Bhopal

ABSTRACT

In this paper a more improved Feature Selection and Classification technique is implanted on Benchmark Datasets such as Mushroom and Soyabean. The Proposed Methodology implemented is based on the Hybrid Combinatorial method of Applying PSO-SVM for the selection of Features from the Dataset and Then Classification is done using Fuzzy Based Decision Tree. Experimental results when performed on Various Datasets prove that the proposed methodology extracts more features as well as provides more accuracy as compared to existing methodologies.

Keywords

PCA, GA, PSO, SVM, Decision Tree, Naïve Bayes, CART, J48.

1. INTRODUCTION

The process of extracting significant and constructive information from large sets of data is called Data Mining. Various data mining techniques which are applied to agriculture and their applications to agriculture related areas are described. In many developing countries, hunger is forcing people to cultivate land that is unsuitable for agriculture and which can only be converted to agricultural use through enormous efforts and costs, such as those involved in the construction of terraces. Each country is known for its core competence. India's is agriculture. Yet, it only accounts for 17 per cent of the total Gross Domestic Product. With the pressure of urbanization, it is going to be a challenge to produce food for more people with less land and water. The current research mainly concerns with the impact of climate change over agricultural landscapes based on the impact on crop productivity as well as impact on soil organic carbon due to combined effect of soil erosion and climate change. To execute such a study, agricultural area with spatial variability in soil type, topography and weather should be a must.

Agriculture plays a key role in overall economic and social well being of India. Agriculture or farming forms the backbone of any country economy, since a large population lives in rural areas and is directly or indirectly dependent on agriculture for a living. Income from farming forms the main source for the farming community. The essential requirements for crop harvesting are water resources and capital to buy seeds, fertilizers, pesticides, labor etc. Most farmers raise the required capital by compromising on other necessary expenditures, and when it is still insufficient they resort to credit from sources like banks and private financial institutions. In such a situation, the repayment is dependent on the success of the crop. If the crop fails due to several factors, like bad weather pattern, soil type, improper, excessive and untimely application of both fertilizers, pesticides, adulterated seeds and pesticides etc. then is pushed into an acute crisis causing severe stress [1]. In addition, the plant growth depends on multiple factors such as soil type, crop type, and weather. Due to lack of plant growth information and expert advice, most of the farmers fail to get a good yield. Most knowledge of soil in nature comes from soil survey efforts. Soil survey, or soil mapping, is the process of determining the soil types or other properties of the soil cover over a landscape, and mapping them for others to understand and use. Primary data for the soil survey are acquired by field sampling and supported by remote sensing.

Techniques to improve the productivity and sustainability of agricultural production systems are one strategy to tackle the challenges of climate change and available agricultural land. Frequent research experiments and field trials are examine the influence of method of land use, soils, climate and agronomic practices on farm production systems [2] [3]. However, the use of data mining and other analytical techniques has become increasingly important in crop prediction and decision making [4]. This is especially crucial in relation to farmers dealing with short-term seasonal variability. The study is significant for the agricultural industry in general and for farmers specifically. The quality and relevance of agricultural information is crucial for farmers who need accurate predictions of crop yield to help make strategic decisions. An efficient data mining mechanism based on the combination of Principal Component Analysis (PCA) as a preprocessing method and a modified Genetic Algorithm (GA) is used to get crop yield implement, in order to reduce computational cost and time by keeping a number of features as discriminating and small as possible [5], [6]. In so doing, generating agricultural crops classification models is efficient and characterization is improved. The PCA-GA data mining mechanism will be implemented for agricultural crops dataset to identify key attribute combinations and characteristics that determine crop performance. To overcome this heterogeneity and complexity in climate and topography, high resolution spatial simulations have been performed by incorporating high resolution datasets with agro ecosystem models. This generates a quantitative and visual idea of spatial impact of climate change over a complex landscape. The present study considered mainly two aspects of impact of climate change over an agricultural landscape which includes: (1) Impact on crop productivity and (2) Impact on Soil organic carbon due to combined effect of soil erosion and climate change.

2. LITERATURE SURVEY

Geraldin et al. has [7] proposed an efficient data mining methodology based on PCA-GA, it is implemented to describe agricultural crops. The technique illustrates enhancements to classification difficulties by using Principal Components Analysis (PCA) as a pre processing scheme and a modified Genetic Algorithm (GA) as the function optimizer. The fitness function in GA is modified for that reason using well-organized distance measures. The advance is to asses, the PCA-GA hybrid data mining technique, using various agricultural field data sets, generate data mining classification models and establish meaningful relationships. The experimental outcomes give you an idea about improved classification rates and generated characterization representations for agricultural crops. The domain model outcome may have advantages to agricultural researchers and farmers. These generated classification models can also be exploited and with pleasure have as a featured into a decision support system. Based on the effects of the experiment, the implementation of the algorithm based on PCA-GA is proficient in optimizing the data mining process, generating classification models and rules for agricultural crops characterization. This may be attributed to the optimization characteristics of the GA in the data mining process.

Raoranne and Kulkarni [8] has used the effectiveness of data mining as a tool, use of data mining to crop yield estimation is discussed. The study assessed new data mining techniques and was applied to various variables to establish if meaningful relationships can be found. It was observed that efficient techniques can be developed and analyzed using appropriate data to solve complex agricultural problems using data mining techniques.

One of the key steps in data mining is finding ways to reduce dimensionality without sacrificing correctness. PCA is applied and found that it handles sparse data and generated fewer and improved association rules. PCA is a multivariate technique, that analyzes a data table in which observations are described by several inter correlated quantitative dependent variables. Its goal is to transform the data, represent it as a set of new orthogonal variables called principal components [9].

Farina et al., [10] applied EPIC model to simulate the interactive effect of climate change, CO_2 enrichment, soil management and two crop rotations on crop yield and SOC (0-100 cm depth). The current study used GCM (general Circulation Model) derived future climatic conditions to study the climate change impact on crop productivity. Outcomes of the study concluded that conventional tillage practices led to massive C loss rate.

Kumar and Kannathasan [11] used data mining and pattern recognition techniques for soil data mining which is being used in the agricultural soil science and its allied area. The proposals arising from this research survey are: A comparison of different data mining methods could produce an efficient algorithm for soil classification for multiple classes. The advantages of a greater understanding of soils could improve productivity in farming, maintain biodiversity, decrease reliance on fertilizers and create a better integrated soil management system for both the private and public sectors. Here author has using efficient techniques can be extended and tailored for solving complex soil datasets using data mining.

S.Veenadhari, et al., [12] observed the research studies on application of data mining methods in the field of agriculture. Some of the methods, such as ANN, ID3, the k-means, and the k-NN and support vector machines applied in the field of agriculture were offered. Data mining in application in agriculture is a comparatively novel approach for forecasting or predicting of agricultural crop or animal management. This item investigates the applications of data mining methods in the field of agriculture and allied sciences. The supply chain operation of companies employed in industries that use agricultural produce as raw material is important for Historical crop yield information. Animal feed, seed, chemical, poultry, fertilizer pesticides, seed, paper and many other industries use agricultural products as intergradient in their production processes. A correct estimate of crop size and risk facilitates these companies in planning supply chain decision like production scheduling. Business such as seed, fertilizer, agrochemical and agricultural machinery industries plan production and marketing activities based on crop production estimates.

3. PROPOSED METHODOLOGY

Take an input Training Dataset such as Mushroom or Soyabean.

Apply PSO-SVM Feature Selection Algorithm for the Selection of Important Features from the Dataset on the basis of which classification can be done accurately.

Apply Fuzzy Decision Tree for the Generation of Decision Tree and generate rules using Fuzzy Decision Tree.Store the Generated rules.

4. RESULT ANALYSIS

The Table shown below is the analysis and comparison of various Techniques implemented for the classification of Soyabean and mushroom datasets. The Existing Methodology implemented uses PCA-GA based Feature Selection and classifier is used for the classification of final Decision. But here a more improved PSO-SVM based Feature Selection method if used for the selection of features from the dataset and then Fuzzy Based Decision Tree is implemented as classifier for the classification of Dataset.

The Methodology when applied on Various Datasets on Various Techniques it is concluded that the proposed methodology outperforms more accuracy as compared to other existing methodologies.

Table 1. Comparison of Accuracy of Various Techniques

Classifier	Soyabean	Mushroom
k-NN	99.85	99.85
J4.8	98.68	100
Naïve Bayes	92.53	97.22
MLP	98.83	99.96
Proposed	100	100

The Table shown below is the analysis and comparison of various Techniques implemented for the classification of Soyabean and mushroom datasets. The Existing Methodology implemented uses PCA-GA based Feature Selection and classifier is used for the classification of final Decision. But here a more improved PSO-SVM based Feature Selection method if used for the selection of features from the dataset and then Fuzzy Based Decision Tree is implemented as classifier for the classification of Dataset.

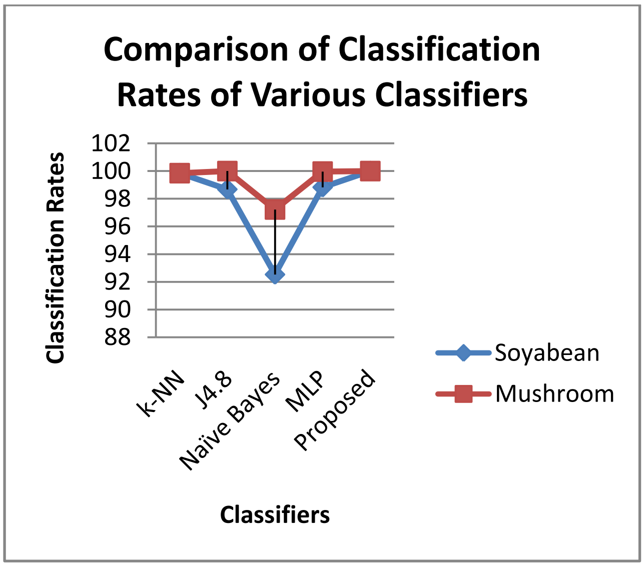

Figure 1. Analysis of Accuracy on Various Dataset

The table shows the comparison of number of reduced features and hence the accuracy of the proposed methodology and SVM without PSO. The proposed methodology optimizes the selection of features from SVM and hence provides better accuracy as compared to the existing technique of features selection using PSO.

Table 2. Comparison of accuracy with and without PSO - SVM approach for Soyabean Dataset

	Original set of features	Reduced feature subset	Accuracy (%)
Without PSO-SVM	35	-	80.5%
With PSO-SVM	35	5	100.0%

Table 3. Comparison of accuracy with and without PSO - SVM approach for Mushroom Dataset

	Original set of features	Reduced feature subset	Accuracy (%)
Without PSO-SVM	22	-	82.45%
With PSO-SVM	22	4	100.0%

Table 4. Comparison of Various Feature Selection approach for Dataset

Datasets	PCA	k-NN-GA	J4.8-GA	Naïve bayes-GA	Proposed Work
Mushroom	58	2	21	17	2
Soyabean	41	2	18	26	1

5. CONCLUSION

The result analysis shows the performance of the proposed methodology. The proposed methodology implemented here provides more accuracy for the classification of Datasets such as Soyabean or Mushroom Dataset. Various Experimental Results when performed on these dataset provides more generated rules and high selection of features using PSO-SVM algorithm and Fuzzy Decision Tree. Hence provides high Accuracy as compared to the existing methodology and less Error Rate and High Positive Rate.

Although the methodology applied here provides efficient results as compared to the other existing techniques, but further enhancements can be done related to the execution time of the methodology as well as reducing the rules generated.

6. REFERENCES

[1.] Sudarshan Reddy S, Vedantha S, Venkateshwar Rao B, Sundar Ram Reddy and Venkat Reddy. Gathering Agrarian Crisis Farmers Suicides in Warangal district. Citizens Report, 1998.

[2.] Anderson, W. K. (2010). Closing the gap between actual and potential yield of rainfed wheat. The impacts of environment, management and cultivar. . Field Crops Research, 116(1), 14-22.

[3.] Asseng, S., & Pannell, D. J. (Adapting dryland agriculture to climate change: Farming implications and research and development needs in Western Australia. Climatic Change, 1-15, 2012.

[4.] Drew, J. Operating In a Change Environment – NEAR/Drought Reform. Retrieved from (2010).

[5.] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," IEEE Intelligent Systems and Their Applications, vol. 13, no. 3, pp. 44-49, March/April 1998.

[6.] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Reading Menlo Park: Addison-Wesley, vol. 412, 1989.

[7.] Geraldin B. Dela Cruz, Bobby D. Gerardo, and Bartolome T. Tanguilig, "Agricultural Crops Classification Models Based on PCA-GA Implementation in Data Mining" International Journal of Modeling and Optimization, Vol. 4, No. 5, October 2014.

[8.] A. Raoranne and R. V. Kulkarni, "Data Mining: An effective tool for estimation in the agricultural sector," International Journal of Emerging Trends and Technology in Computer Science, vol. 1, no. 2, pp. 75-79, July-August 2012.

[9.] D. Gerardo, J. Lee, I. Ra, and S. Byun, “Association rule discovery in data mining by implementing principal component analysis,” Artificial Intelligence and Simulation, Springer, Berlin Heidelberg, 2005, pp. 50-60.

[10.] Farina, R., Seddaiu, G., Orsini, R., Steglich, E., Roggero, P.P., Francaviglia, R., (2011). Soil carbon dynamics and crop productivity as influenced by climate change in a rainfed cereal system under contrasting tillage using EPIC. Soil Till. Res. 112, 36–46.

[11.] Dr. D. Ashok Kumar, N. Kannathasan, “A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining” IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011.

[12.] S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh, “Data mining Techniques for Predicting Crop Productivity –A review article”, International Journal of Computer Science and technology, march 2011.