Comparative Study of Data Mining Classification Techniques over Soybean Disease by Implementing PCA-GA

Dr. Geraldin B. Dela Cruz

Institute of Engineering, Tarlac College of Agriculture, Philippines, delacruz.geri@gmail.com

Abstract— Data mining is a relatively new approach in the field of agriculture that can be used in the extraction of knowledge and discovery of patterns and relationships in agricultural data. Classification techniques in data mining are used to discover patterns and knowledge agricultural datasets, however, the accuracy of these classification techniques depends on the quality of data that are used as inputs in the data mining process. In this paper, an efficient data mining methodology based on PCA-GA is applied as a data pre processing technique, to reduce the dimensionality of the soybean dataset. The mechanism draws improvements to classification problems by applying Principal Components Analysis (PCA) and subsequently applying Genetic Algorithm (GA) to further reduce the dimensionality of the dataset, and selecting the best representative subsets, thereby improving the performance of classifiers. Different data mining classification techniques are applied to the resulting reduced dataset and classification metrics are compared. This approach is to asses classification rates on the PCA-GA reduced soybean dataset. The learning and validation experiment was performed using WEKA, a workbench containing implementations of the k-NN, Naïve Bayes, J4.8 and MLP classification algorithms, including the PCA and GA. Classification accuracy was validated using 10-folds cross validation..

Keywords—classification, data mining, genetic algorithm, optimization, PCA-GA, soybean,

INTRODUCTION

Databases not only store and provide data but also contain hidden knowledge which can be very important however, human ability to analyze and understand these massive datasets lags far behind his ability to gather and store data in databases. Data in the agricultural domain are robust, comes in different formats, complex, multidimensional and contains noise. Interesting patterns can be mined in this space in discovering knowledge, revealing solutions to specific domain problems [1]. Extraction of the useful set of features is usually unknown from these volumes of data [2], considering every single feature of an input pattern in a large feature set makes classification and knowledge discovery computationally complex. Also, the inclusion of irrelevant or redundant features in the data mining model results in poor predictions and interpretation, high computational cost and high memory usage [3], [4].

In general, it is desired to keep a number of features as discriminating and as small as possible in order to reduce computational time and complexity in the data mining process [5], [6]. This can be addressed by dimensionality reduction [7] method that improves data mining classification and facilitates visualization and data understanding. It is a process that creates a set of features based on transformations or combinations of the original dataset, thereby reducing it into a smaller representative dataset.

The focus of this study is to implement an efficient mechanism based on the combination of Principal Component Analysis (PCA) and Genetic Algorithm (GA) [8], [9] as a data preprocessing method, to reduce the dimensionality of the data by keeping a number of features as small as possible. Apply the k-NN, MLP, NB and J4.8 algorithms in the data mining process and compare the classification results. In so doing, the PCA-GA [10] mechanism is validated as an efficient data reduction method.

THE PCA-GA MECHANISM

The PCA-GA algorithm is a combination of two algorithms, PCA, for data preprocessing and reduction and the GA for feature subset selection method, which makes the whole process a hybrid data mining mechanism. The PCA mechanism is a very useful data dimensionality reduction technique, reducing the number of variables to a few interpretable linear combinations of the data. PCA maps the rows and columns of a given matrix into two or three dimensional points to reveal the structure of dataset. The original data are projected into smaller space. Thus data reduction is performed. The data can be represented by a collection of n points in the z-dimensional space, where each axis corresponds to a measured variable. From this space, a line Y1 can be searched such that the dispersion of n points when projected unto this line is a maximum. The derived variable is denoted by the equation in (1):

$$Y_1 = e_1x_1 + e_2x_2 + \ldots e_px_p \qquad (1)$$

Where, e_i are coefficients satisfying the condition in (2):

$$\sum_{j=1}^{z} e_{i^2} = 1 \qquad (2)$$

After obtaining Y1, the (z-1) – dimensional subspace orthogonal to Y1 is considered and line Y2 is found in this subspace such that the dispersion of points when projected onto this line is also maximum, and is perpendicular to Y1 such that the dispersion of points when they are onto this line is the maximum. Having obtained Y2, a line in the (z-2)-dimensional subspace is considered, which is orthogonal to both Y1 and Y2, such that dispersion of points when projected onto this line is as large as possible. The process can be continued until z mutually orthogonal lines are determined. Each of these lines now defines a derived variable shown in equation (3):

$$Y_i = e_{1i}X_1 + e_{2i}X_2 + e_{3i}X_3 + \dots + e_{ni}X_i \qquad (3)$$

where the constants ei_j are determined by the requirement that the variance of Yi is a maximum, subject to the constraint of orthogonality as well as in (4) :

$$\sum_{k=1}^{p} e_{ik^2} = 1 \qquad (4)$$

Thus, the Yi obtained in (3) are called principal components of the system. The process produces a list of linear vectors in (5) called principal components.

$$\begin{aligned} Y_1 &= e_{11}X_1 + e_{12}X_2 + \dots + e_{1p}X_p \\ Y_2 &= e_{21}X_1 + e_{22}X_2 + \dots + e_{2p}X_p \\ &\dots. \qquad\qquad\qquad\qquad (5) \\ Y_p &= e_{p1}X_1 + e_{p2}X_2 + \dots + e_{pp}X_p \end{aligned}$$

Each of the principal components can be thought of as a linear regression predicting Yi in (3) from X_1, X_2, …, X_p. There is no intercept, but e_{i1}, e_{i2}, …, e_{ip} can be viewed as regression coefficients. The first principal component is the linear combination of X variables that has maximum variance among the linear combinations, accounting for as much variation in the data as possible. The remaining principal components, accounts for as much of the remaining variation as possible, thus the principal components are uncorrelated with each other.

GA Process

Genetic algorithms are search algorithms based on natural genetics. It is an iterative process that operates on a population or a set of candidate solutions, in this case, the principal components generated by the PCA mechanism. Each solution is obtained by means of encoding/decoding mechanism, which enables representing the solution. GA is considered as a function optimizer and performs efficient by searching the best representative sets from candidate sets of solutions (principal components). The interest is in the minimization of a set of variables that can represent a dataset with maximum results. Genetic algorithms consist of three essential elements: a coding of the optimization problem, a mutation operator and crossover. The coding of the optimization problem produces the required discretization of the variable values and makes their simple management in a population of search points possible. A binary coding is ideal because in this way the mutation and crossover operators are simple to implement. Thus, the values of the individuals P1…Pn can be encoded with the binary fixed-point coding in (6).

$$\begin{aligned} P_1 &= b_5b_4b_3b_2b_1b_0 \\ &\dots\dots. \qquad\qquad\qquad\qquad (6) \\ P_n &= b_5b_4b_3b_2b_1b_0 \end{aligned}$$

The crossover operator, control the recombination of the individuals in order to generate a new, better population of individuals at each iteration step. Before recombining, the individuals must be evaluated by a fitness function (7) for all data structures in the population. The fitness value is then used as the basis for selection in the crossover or mating.

$$\text{Fitness} = \text{countone(Pn)} \qquad (7)$$

A typical reproduction operator is crossover (8). Before the crossover, two individuals P1 and P2 are selected as “parents”, based on their fitness value. Selection is based on tournament using their fitness. Individuals with higher fitness value, wins the tournament and is chosen to mate. The offspring C1 and C2 are formed so that the left side comes from one parent and the right side from the other. This produces an interchange of the information stored in each parent. The whole process is reminiscent of genetic exchange in living organisms.

$$C1 = Mask1 \,\&\, P1 + Mask2 \,\&\, P2 \qquad (8)$$
$$C2 = Mask2 \,\&\, P1 + Mask1 \,\&\, P2$$

Where :

P1 , P2 – parents chromosomes
C1, C2 - children chromosomes
Mask1, Mask2 – bit masks (Mask2 = NOT(Mask1))

A favorable interchange can produce an offspring with better genes. When the individuals $P1 = b_5b_4b_3b_2b_1b_0$ is recombined with the number $P2 = a_5a_4a_3a_2a_1a_0$. The new individual is then:

$$Cn = b_9b_8 \cdots b_ia_{i-1} \cdots a_0. \qquad (9)$$

Crossover can be interpreted as a variation of optimization. The mutation operator is the simplest. In binary strings, a mutation corresponds to a bit flip. A mutation of the ith bit of the string $Cn = b_5b_4b_3b_2b_1b_0$ produces a change. Thus a new individual is generated and the fitness is again evaluated.

DATA MINING CLASSIFICATION TECHNIQUES

The book in [11], identifies, presents and describes different data mining classification techniques [12]. Classification [13] is a form of data analysis that can be used to construct a model, which can be used in the future to predict the class label of new datasets. It is a two step process, first is the learning step where the classification algorithm builds the classifier by analyzing a training set made up of database tuples and their associated class labels, using a mapping function in the form of classification rules. In the second step, the accuracy of the classifier is predicted.

Some of the most popular and common classification algorithms are adopted and presented herein, based on their capabilities simplicity and robustness. The k-NN and Naïve Bayes were chosen based on the study in [14], which proves to perform excellent using the WEKA [15] data mining tool, likewise the J4.8 and MLP were used for their exemplary performance in classification problems in different datasets [16].

k-Nearest Neighbor (k-NN)

k-NN is a nearest neighbor algorithm that classifies entities taking a class of the closest associated vectors in the training set via distance metrics. The principle behind this method is to find predefined numbers of training samples closest in the distance to the new point and predict the label from these.

Naïve Bayes

Based on the Bayes rule of conditional probability, it uses all of the attributes contained in the data, and analyses them individually as though they are equally important and independent of each other. It considers each of the attributes separately when classifying a new instance. It assumes that one attribute works independently of the other attributes contained by the sample.

J4.8

A popular tree based machine learner, the J4.8 decision trees algorithm is an open source Java implementation of the C4.5 [17]. It grows a tree and uses divide-and-conquer algorithm. It is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. To classify a new item, it creates a decision tree based on the attribute values of the training data. When it encounters a set of items in a training set, it identifies the attribute that discriminates. It uses information gain to tell most about the data instances so that it can classify them the best.

Multi Layer Perceptron (MLP)

MLP is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. It consists of multiple layers of nodes with each layer fully connected to the next one. Each node is a neuron with a nonlinear activation function. It uses a learning technique called back propagation for training the network.

MATERIALS AND METHODS

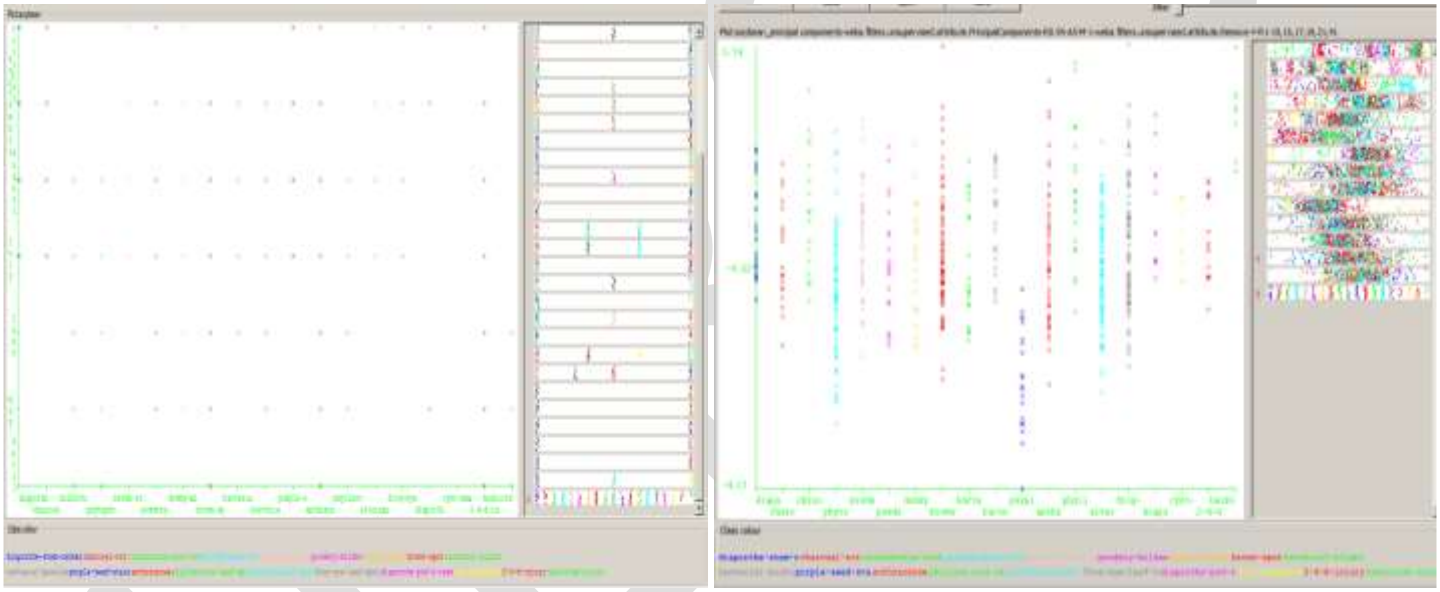
The soybean dataset in [18], was used in the experiment. The data mining software used in the experiment is the WEKA version 3.6.10, which contained the implementations of the classification algorithms presented. A computer with two (2) Gigabytes of memory, with a 32 bit 2.80 Ghz processor, and a proprietary 32 bit Operating System was utilized. The default settings in the data mining software and in the configurations of the algorithms were used.

The dataset was cleaned, encoded and saved as attribute relation file format (arff) file using a text editor, the dataset were loaded in WEKA, the PCA was applied as a data preprocessing method to transform and simplify the dataset into smaller representative sets called principal components. The GA was then applied to the PCA transformed dataset that resulted to an optimized dataset. Subsequently, each of the machine learning algorithms in WEKA, the J4.8, Naïve Bayes, MLP and k-NN were then applied to the resulting PCA-GA reduced datasets and results were recorded and compared accordingly.

Classification accuracy was validated using 10-fold cross validation. This validation method is a standard method to estimate classification accuracy over unseen data.

RESULTS AND DISCUSSION

This section presents and discusses the results of the experiment after applying the PCA-GA mechanism used in the study. There are two parts; first, the outcome of applying the PCA-GA mechanism that was used in generating the visualization model of the reduced dataset. Second the comparison of classification results of classifiers between the original and PCA-GA reduced dataset.



(a) *(b)*

Figure 1. Visualization of the Soybean Disease (a) original dataset, (b) PCA-GA reduced dataset

The soybean dataset originally has thirty six (36) attributes. As can be seen in Figure 1b, after applying the PCA-GA mechanism to the soybean dataset, resulted to fifteen (15) feature sets. This reduced dataset is now considered the smaller representative soybean dataset.

As can be observed, the resulting feature sets in Figure 1b are simplified in structure compared to the original dataset in Figure 1a. The reduced dataset is the optimized smaller representative dataset of the original, thru the optimization technique of GA. The presented visualization in Figure 1b, confirms the PCA-GA efficiency as a dimensionality reduction method, this implies that extracting knowledge from this smaller and optimized dataset is more accurate, efficient and faster. Based on analysis of Figure 1b, the features sets are combinations of the original attributes, that resulted in process of pre-processing through the PCA-GA mechanism. Further the figure shows a distinct variation of the feature sets, classifying the different soybean disease.

Classifier	Original Dataset		PCA-GA Reduced Dataset	
	Accuracy	Time	Accuracy	Time
k-NN	91.22%	0.00 sec	99.85%	0.00 sec
J4.8	91.51%	0.02 sec	98.68%	0.01 sec
Naïve Bayes	92.97%	0.01 sec	94.44%	0.00 sec
MLP	93.41%	112.0 sec	98.83%	18.20 sec

Table 1. Classification Rates of Soybean Dataset When Applied with Different Classifiers

Table 1 shows the comparison of classification rates of various classifiers over the original and PCA-GA reduced soybean dataset. It can be seen that the classification rates between the original and reduced dataset, have noticeable improvements in the J4.8, Naïve Bayes, k-NN and MLP classifiers. Among the classification algorithms tested, the fastest was the k-NN classifier and a significant improvement on accuracy can be observed. Interesting to note also is the speed of the MLP, which is significantly faster on the PCA-GA reduced dataset, the accuracy also significantly increased. Although the MLP improved on its processing time on the reduced dataset, it can be seen that its processing time took longer compared to the other classifiers

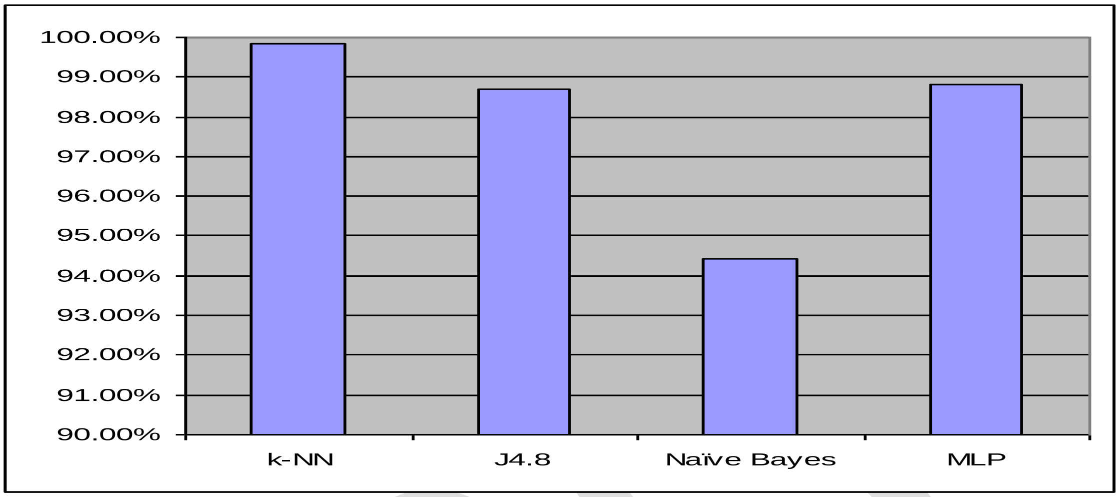


Figure 2. Comparison of Accuracy Rates of the PCA-GA Reduced Dataset When Applied with Different Classifiers

In comparing the classifiers accuracy rates on the PCA-GA reduced dataset, it can be seen in Figure 2, that the k-NN classifier is the most accurate. Naïve Bayes is the least accurate among the classifiers, with MLP and J4.8 are as nearly accurate as the k-NN. Generally, the results presented imply that the PCA-GA method significantly improves classification rates, over the soybean disease dataset with the PCA-GA dimensionality reduction method.

ACKNOWLEDGMENT

The researcher would like to extend his gratitude to the Tarlac College of Agriculture for the support of this study.

SUMMARY AND CONCLUSION

Presented in this study, is the PCA-GA hybrid data reduction method and comparison of classification rates of various data mining classification techniques over the soybean dataset. The k-NN, J4.8. Naïve Bayes ad MLP classifiers were applied to the resulting PCA-GA reduced datasets. Based on the results, the PCA-GA mechanism reduced the original dataset into smaller representative datasets. Visualization model was generated based on the result of the data reduction mechanism based on PCA-GA to present a clearer view of its potential as a hybrid method. Results also show that classification accuracy and processing speed improved for all of the classifiers.

The implementation of the algorithm based on PCA as a preprocessing technique and GA as a feature subset selector is efficient in reducing the dimensionality of the soybean dataset. Results imply that all classifiers can be implemented and are efficient in classifying soybean disease using the PCA-GA pre processing mechanism. Classification speed is further improved for all the classifiers used. On the other hand, results of comparison between the classifiers accuracy rates show that the k-NN is the most accurate and the least accurate is the Naïve Bayes. Generally, using the PCA-GA data reduction technique proves to have significant results in characterizing soybean disease using the presented data mining classification techniques. Thus, simplifies the process of extracting knowledge, discovering patterns and relationships and the interpretation of soybean disease.

REFERENCES:

[1] Arora, Rohit and Suman Suman. "Comparative analysis of classification algorithms on different datasets using WEKA." International Journal of Computer Applications Vol 54, No 13, pp. 21-25, 2012

[2] Raymer, Michael L., William F. Punch, Erik D. Goodman, Leslie Kuhn, and Anil K. Jain. "Dimensionality reduction using genetic algorithms." Evolutionary Computation, IEEE Transactions Vol 4, No 2, pp. 164-171, 2000.

[3] Qu, Guangzhi, Salim Hariri, and Mazin Yousif. "A new dependency and correlation analysis for features." Knowledge and Data Engineering, IEEE Transactions Vol 17, No. 9, pp. 1199-1207, 2005.

[4] Janecek, Andreas, Wilfried N. Gansterer, Michael Demel, and Gerhard Ecker. "On the relationship between feature selection and classification accuracy". Journal of Machine Learning Research-Proceedings Track 4, Antwerp, Belgium, pp. 90-105, 2008.

[5] Gerardo, Bobby D., Jaewan Lee, Inho Ra, and Sangyong Byun. "Association rule discovery in data mining by implementing principal component analysis." In Artificial Intelligence and Simulation, pp. 50-60. Springer Berlin Heidelberg, 2005.

[6] Diepeveen, D. & Armstrong, L. "Identifying key crop performance traits using data mining". IAALD AFITA WCCA2008, World Conference on Agricultural Information and IT, 1-21, 2008

[7] Burges, Christopher JC. "Dimension reduction: A guided tour, Machine Learning". Foundations and Trends in Machine Learning, Vol. 2, No. 4, pp. 275-365, 2009.

[8] Yang, Jihoon, and Vasant Honavar. "Feature subset selection using a genetic algorithm." In Feature extraction, construction and selection, pp. 117-136. Springer US, 1998.

[9] Goldberg, David E., and John H. Holland. "Genetic algorithms and machine learning." Machine learning Vol 3, No. 2 pp. 95-99, 1988.

[10] Cruz, Geraldin B. Dela, Bobby D. Gerardo, and Bartolome T. Tanguilig III. "An Improved Data Mining Mechanism Based on PCA-GA for Agricultural Crops Characterization." International Journal of Computer and Communication Engineering Vol 3, No. 3, pp. 221-225, 2014

[11] Han, Jiawei, Micheline Kamber. Data Mining: Concepts and Techniques, 2nd Edition, Morgan Kaufmann, pp 285-289, 2006

[12] Phyu, Thair Nu. "Survey of classification techniques in data mining." In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, pp. 18-20. 2009.

[13] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas.. "Supervised Machine Learning : A review of Classification Techniques". Informatica 31, pp 246-268, 2007

[14] Wahbeh, A. H., Q. A. Al-Radaideh, M. N. Al-Kabi, and E. M. Al-Shawakfa. "A comparison study between Data Mining Tools over some classification methods. IJACSA, Special Issue on Artificial Intelligence." *SAI Publisher* Vol 2, No 8 (): 18-26, 2010.

[15] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update." *ACM SIGKDD explorations newsletter* Vol 11, No. 1, pp. 10-18, 2009.

[16] Beniwal, Sunita, and Jitender Arora. "Classification and feature selection techniques in data mining." *International Journal of Engineering Research & Technology (IJERT)* Vol 1, No. 6, 2012.

[17] Quinlan, J. Ross. C4. 5: Programs for Machine Learning, Vol.1. Morgan Kaufmann. 1993

[18] Bache, Kevin, and Moshe Lichman. "UCI machine learning repository, 2013." *URL http://archive. ics. uci. edu/ml* (1990): 92.