An Improved Data Mining Mechanism Based on PCA-GA for Agricultural Crops Characterization

Geraldin B. Dela Cruz, *Member, IACSIT*, Bobby D. Gerardo, *Member, IACSIT* and Bartolome T. Tanguilig III

***Abstract*—In this study, a data mining method based on PCA-GA is presented to characterize agricultural crops. Specifically it draws improvements to classification problems by using Principal Components Analysis (PCA) as a preprocessing method and a modified Genetic Algorithm (GA) as the function optimizer. The GA performs the optimization process, selecting the most suited set of features that determines the class of a crop it belongs to. The fitness function in GA is studied and modified accordingly using efficient distance measures. The soybean dataset is used in the experiment and results are compared with several classifiers. The experimental results show improved classification rates. This lessens the time consumed of agricultural researchers in characterizing agricultural crops.**

***Index Terms*—Classification, data mining, genetic algorithm, k-NN, principal component analysis.**

I. Introduction

Data comes in different formats, complex, multidimensional, robust and may contain noise. Interesting patterns can be mined from this space in discovering knowledge, revealing solutions to specific domain problems [1].

In data mining, pattern recognition can be seen as a classification process. Each pattern is represented by a set of measurable features or dimensions and viewed as a point in the n dimensional space. The aim of pattern recognition is to choose features that allow us to discriminate between patterns belonging to different classes. Often, optimal set of features is usually unknown [2], considering every single feature of an input pattern in a large feature set makes data mining classification process computationally complex. The inclusion of irrelevant or redundant features in the data mining model results in poor predictions, 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, to reduce computational time and complexity [5], [6] in the data mining process.

In this study, focus is given to improve a data mining mechanism based on the combination of Principal Component Analysis (PCA) as a preprocessing technique and a modified Genetic Algorithm (GA) [7] as the learning algorithm in order to reduce computational cost and time in the data mining process, by keeping a number of features as discriminating and as small as possible. In so doing, it is expected that classification performance is improved.

The data mining mechanism based on PCA-GA will be tested using agricultural crops dataset to identify key attribute combinations and characteristics that determine crop performance. The outcome of the data mining modeling and testing shall be utilized for decision support in improving agricultural crops productivity.

II. Related Literature

Different data mining techniques are available in the literature to improve data mining tasks [1], [2], [4]-[6].

Reference [8] used Genetic Algorithm for feature selection in the context of a neural network classifier. GA was configured to use an approximate evaluation in order to reduce significantly the computation required. The algorithm employed nearest-neighbor (k-NN) classifier to evaluate feature sets and showed that the features selected by this method are effective.

PCA [9] is one of these techniques and performs well in reducing complexity in data by reducing its dimensionality. In [10] they mentioned that, “one of the key steps in data mining is finding ways to reduce dimensionality without sacrificing correctness". They applied PCA 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. In this case, how many components should be considered?

In feature subset selection no new features will be generated but a, subset of the original features are selected and the feature space is reduced. In cases where there are more features than necessary, subset selection helps simplify computational time, enhances and improves predictive power of classifiers [11].

Genetic Algorithm is an evolutionary based stochastic optimization algorithm, proposed by Holland (1973). It is regarded as a function optimizer due to its outstanding performance with optimization. The algorithm comprises of three principal genetic operators: selection, crossover and

Manuscript received November 29, 2013; revised March 3, 2014. This work was supported by a grant from the HRD-Faculty Scholarship Program of the Tarlac College of Agriculture, Camiling,Tarlac, 2306 Philippines.

G. B. Dela Cruz is with the Institute of Engineering, Tarlac College of Agriculture, Camiling, Tarlac, Philippines, he is also with the Technological Institute of Philippines, Cubao, Quezon City, Philippines (e-mail: geridelacruz2002@yahoo.com.hk, delacruz.geri@gmail.com).

B. D. Gerardo is with Administration and Finance at the Western Visayas Stare University, La Paz, Iloilo City, Philippines. He is also with the Department of Information Technology at WVSU. (e-mail: bgerardo@wvsu.edu.ph).

B. C. Tanguilig III is with the Academic Affairs and concurrent Dean of the College of Information end Information Technology Education at the Technological Institute of the Philippines, Quezon City, Philippines (e-mail: bttanguilig_3@yahoo.com).

mutation to form a new generation [12], [13]. It converges to the best chromosome, which hopefully represents the optimum or suboptimum solution to a problem.

Genetic Algorithm has been shown in the literature to be an effective tool to use in data mining and pattern recognition. However, GA has problems with premature convergence which inhibit diversity in the population and prevent exploration of the whole search space. To address this problem, the work of A. Hassani, and J. Treijis [14] suggested tweaking the GA to a specific problem and correctly set all parameters, conversely, L. Na-Na, G. Jun-Hua, and L. Bo-Ying [15], used the negative selection method, which showed promising results.

In the study of A. S. Elden, M. A. Mustafa, H. M. Harb and A. H. Emara [16], they designed and evaluated a fast learning algorithm based on GA and proved to have considerable improvements on the accuracy performance, over other classifiers.

And in [17], [18] PCA was applied, then the k-NN classifier was used as the fitness function for the GA and showed promising results in reducing classification error rates, and recommended using different classifiers for similar studies.

III. Design Concepts and Methods

A. The Proposed Data Mining Architecture

The architecture shown in Fig. 1, depicts the data mining process. There are two major phases in the process. The first phase is data preprocessing using PCA and using GA to find the feature subset that is the optimum solution to the problem being addressed. The second phase is to utilize the optimum results and rules generated for the characterization of crops. This prediction model is then utilized for decision support.

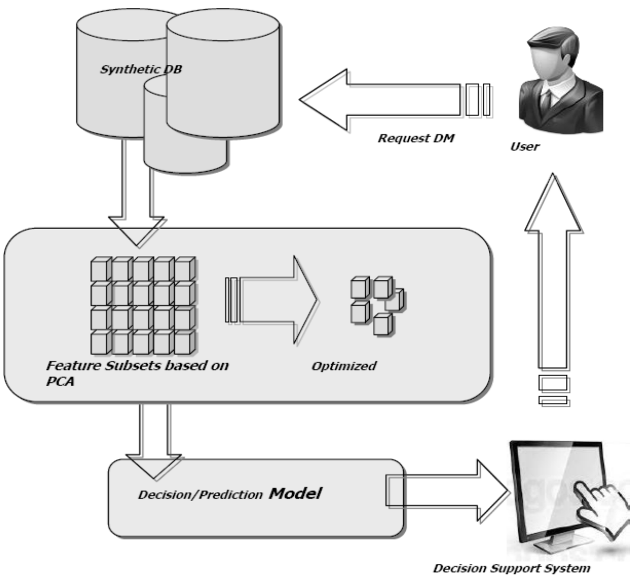

Fig. 1. The general architecture of the data mining process.

B. Methods and Procedures

The idea proposes the application of Principal Component Analysis to reduce the dimensionality of a dataset to a feature set called principal components. The principal components are then the initial population in the search space of the GA in searching for the optimum solution. This mechanism simplifies the data mining process using the representative data of the original dataset, to which reduces computational time and improves classification accuracy of classifiers. (See Fig. 2).

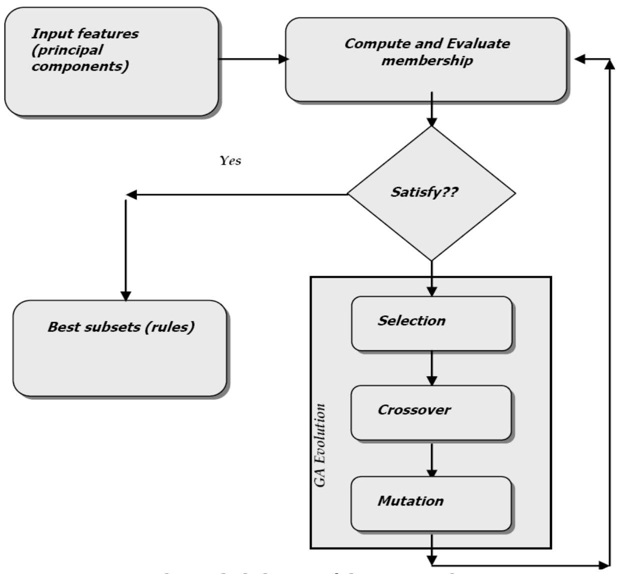

Fig. 2. The exploded view of the DM mechanism.

However, the PCA technique has a tendency to lose data interpretability but has high discriminative power. To overcome the shortcomings of this process, a feature subset selection technique based on a modified GA is used. In this context, the suggestion of [17] using other classifiers is adopted as the fitness function. The fitness function in GA is modified accordingly using efficient variation of distance measures between features, this provides better separation of the pattern classes, which, in turn, reduces complexity and improves the performance of classifiers and reduce computational costs.

1) Data preprocessing

Data preprocessing is an important step and technique in the data mining process, it involves transformation of data into understandable format. Real world data is incomplete, noisy, inconsistent, and lacking certain trends. Data preprocessing is a method of resolving these issues, which includes cleaning, transformation, normalization, feature extraction and selection.

PCA is a procedure to convert a set of observations of possibly correlated variables, into a set of values linearly uncorrelated variables, called principal components. The transformed dataset is defined in such a way that the first principal components account for much of the variance. Principal components are guaranteed to be independent if the data set is jointly normally distributed.

2) Classification

This is a data mining task of generalizing known structure and applying it to new data. It is also the categorization of data for its most effective and efficient use.

a) k-Nearest Neighbor (k-NN)

The principle behind this method is to find predefined numbers of training samples closest in the distance, to a new point and predict label from these. The number of samples can be a user defined constant or varied, based on the local density of points. The distance can be any metric measure.

k-NN uses the Euclidean distance as the most common choice. Despite its simplicity it is successful in large number of classification problems. (Shown in Table I).

TABLE I: DIFFERENT APPROACHES OF DISTANCE MEASURES IMPLEMENTED IN THE K-NN CLASSIFIER.

Euclidean	$D(x,y)=\left(\sum_{i=1}^{m}\lvert x_i-y_i\rvert^2\right)^{1/2}$
Chebysheb	$D(x,y)=\max_{i=1}^{m}\lvert x_i-y_i\rvert$
Manhattan	$D(x,y)=\sum\lvert x_i-y_i\rvert$

b) J4.8

J4.8 decision trees algorithm is an open source Java implementation of the C4.5. 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 us most about the data instances so that it can classify them the best.

c) Naïve Bayes

This classifier is 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.

The Naïve Bayes classifier works on a simple, but comparatively intuitive concept. It makes use of the variables contained in the data sample, by observing them individually, independent of each other. It considers each of the attributes separately when classifying a new instance. The attributes are assumed to work independently from the other attributes contained in the sample.

d) 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.

3) The proposed algorithm

1) [Start] Principal components as population
2) [Fitness] Compute and evaluate the fitness $f(x)$ of each principal component x in the population
3) [New population] Create a new population by repeating following steps until the new population is complete
4) [Selection] Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
5) [Crossover] With a crossover probability, cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents.
6) [Mutation] With a mutation probability mutate new offspring at each locus (position in chromosome).
7) [Accepting] Place new offspring in a new population
8) [Replace] Use new generated population for a further run of algorithm
9) [Test] If the end condition is satisfied, stop, and return the best solution in current population and perform classification
10) [Loop] Go to step b

IV. EXPERIMENTAL RESULTS

The model and recommendation presented in [15] is adopted in the experiment using different classifiers as the fitness function for the GA. The k-NN classification algorithm was also tested and validated using varied distance measures and results are compared accordingly.

Below are the results of the experiment. The experiment used the WEKA version 3.6.10 data mining software in the simulation and testing. A computer with 1 Gigabyte of memory, equipped with an AMD Athlon 2.80 Ghz Processor, and a Windows 32 Bit Operating System was utilized.

The soybean dataset was used in the experiment, which is available that came with the data mining software. It has originally thirty six (36) attributes including the class label. After preprocessing using PCA, the transformed dataset contained forty one (41) principal components. The default settings in WEKA and in the algorithm configurations, was used in the experiment.

TABLE II: PERCENTAGE COMPARISON OF CORRECTLY CLASSIFIED INSTANCES ON SOYBEAN DATASET OF PCA AND MODIFIED GA WITH K-NN AS FITNESS FUNCTION USING DIFFERENT DISTANCE MEASURES

	Original Dataset	PCA Reduced Dataset	PCA-Modified GA (k-NN-Euclidean/Chebysheb/Manhattan)		
k-NN	91.65%	90.77%	99.85%	99.85%	99.85%

Table II shows the performance of the modified GA, using the k-NN as the fitness function and classifier, respectively. It can be seen that classification accuracy has improved, as compared with the original dataset. This can be attributed to the optimization function of the GA. However, using varied distance measures in the k-NN, there is no significant change in the classification performance of the k-NN classifier. This can be attributed to the nature of similarities of the distance measurement functions.

TABLE III: CORRECTLY CLASSIFIED INSTANCES OF CLASSIFIERS ON SOYBEAN DATASET WITH J4.8 AS FITNESS FUNCTION IN GA

Classifier	Original Dataset	PCA Reduced	PCA-J4.8-GA
J4.8	91.65%	90.77%	99.85%
Naïve Bayes	93.70%	93.26%	92.53%
MLP	99.85%	99.71%	98.83%

Table III shows otherwise the resulting effect of implementing a different classifier as a fitness function in the GA. The results, implies that classification performance can be improved by using GA as an optimizer in the classification process. However, it can also be analyzed from the table that using a specific classifier, as a fitness function implies that the same fitness function should be used in the classification process in order to have considerable improvements in the results of the classification process. This can also be attributed to the characteristics of the GA.

Interesting to note is the performance of the MLP classifier, though the classifier performs outstanding with the original dataset, it was observed that the classification process took longer to perform the indicated operation on the original dataset as compared to the other classifiers. The classification accuracy rate also degrades as the operations were done, this can be attributed to the characteristics of the GA, as the fitness function was not both the same as the classifier.

TABLE IV: SUMMARY OF PERFORMANCE OF CLASSIFIERS AS FITNESS FUNCTIONS IN GA

	k-NN	J4.8	Naïve Bayes	MLP
Original Dataset	91.65%	91.65%	93.70%	99.85%
PCA Reduced Dataset	90.77%	90.47%	93.26%	99.71%
PCA-mGA	99.85%	99.85%	93.99%	---

It can be shown from the summary in Table IV, that a combination of PCA and a modified GA improves classification accuracy, using classifiers as fitness functions, even with varied distance measures in the k-NN. However, it can be seen from the table that after preprocessing using PCA and applying classification directly without optimization, accuracy is affected.

The MLP classifier as it has been observed in the experiment, however poorly performed as a fitness function in terms of processing time with the GA in the optimization process.

V. SUMMARY AND FUTURE WORK

We have shown in this paper that the proposed hybrid data mining method based on PCA-GA is considerable and shows improvement on classification performance of classifiers as fitness function in GA, thereby improving the data mining process. Likewise, the work in [17] is further validated and shows significant results with other distance measures in the k-NN.

Based on the results of the experiment, it was shown that the proposed algorithm can be used to optimize the results of classification process for agricultural crops characterization. The results are preliminary, using actual field data of agricultural crops to further validate and evaluate the proposed method is being considered. Future work involves further study, to use other efficient distance measures in the k-NN data mining classification algorithm not presented in this study and using only efficient distance measures as the fitness function.

It is suggested that similar studies can also be undertaken, using other preprocessing techniques and further study on the PCA.

REFERENCES

[1] S. Beniwal and J. Arora, "Classification and feature selection techniques in data mining," *International Journal of Engineering Research and Technology*, vol. 1, no. 6, pp. 1-6, August 2012.

[2] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 2, pp. 164–171, July 2000.

[3] G. Qu, S. Hariri, and M. Yousif, "A new dependency and correlation analysis for features," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1199-1207, September 2005.

[4] A. Janecek, W. N. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," *Journal of Machine Learning Research-Proceedings Track* 4, pp. 90-105, 2008.

[5] B. D. Gerardo and J. Lee, "Principal component analysis mechanism for association rule mining," School of Electronic and Information Engineering, Kunsan National University.

[6] D. Diepeveen and L. Armstrong, "Identifying key crop performance traits using data mining," presented at the IAALD – AFITA - WCCA2008, World Conference on Agricultural Information and IT, Tokyo, Japan, August 2008.

[7] 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.

[8] F. Z. Brill, D. E. Brown, and W. N. Martin, "Fast generic selection of features for neural network classifiers," *IEEE Transactions on Neural Networks*, vol. 3, no. 2, pp. 324-328, March 1992.

[9] C. J. Burges, "Dimension reduction: a guided tour," *Machine Learning*, vol. 2 no. 4, pp. 275-365, 2009.

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

[11] H. Liu and H. Motoda, "Feature transformation and subset selection," *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 2, pp. 26-28, 2008.

[12] P. Chadha and G. N. Singh, "Classification rules and genetic algorithm in data mining," *Global Journal of Computer Science and Technology Software and Engineering*, vol. 12, no. 15, pp. 50-54, 2012 .

[13] R. Malhorta, N. Singh, and Y. Singh, "Genetic algorithms: concepts, design for optimization of process controllers," *Computer and Information Science* , vol. 4, no. 2, pp. 39-54, March 2011.

[14] A. Hassani, and J. Treijis, "Overview of standard and parallel genetic algorithms," paper presented at the IDT Workshop on Interesting Results in Computer Science and Engineering, Mälardalen University, Sweden, October 30, 2009.

[15] L. N. Na, G. J. Hua, and L. B. Ying, "A new genetic algorithm based on negative selection," in *Proc.* 2006 *International Conference on Machine Learning and Cybernetics*, 2006, pp. 4297-4299.

[16] A. S. Elden, M. A. Mustafa, H. M. Harb, and A. H. Emara, "AdaBoost ensemble with simple genetic algorithm for student prediction model," *International Journal of Computer Science & Information Technology*, vol. 5, no. 2, pp. 73-85, April 2013.

[17] M. Pei, W. F. Punch, and E. D. Goodman, "Feature extraction using genetic algorithms," in *Proc. International Symposium on Intelligent Data Engineering and Learning'98*, October 14-16, 1998, pp. 371-384.

[18] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognition Letters*, vol. 10, no. 5, pp. 335-347, November 1989.

Geraldin B. Dela Cruz is currently pursuing the doctor degree in information technology at the Technological Institute of the Philippines, Quezon City. He finished his bachelor of science in computer science at the Colegio de Dagupan, Dagupan City, Pangasinan, Philippines in 1997. He finished his masters in information technology degree at Hannam University, Daejon, South Korea in 2003. This author became a Member of IACSIT in 2012.

He is currently the chief of academic programs and associate professor of information technology at the Tarlac College of Agriculture-Institute of Engineering.

Mr. Dela Cruz is also a member of IAENG and the Data Mining and Computer Societies.

Bobby D. Gerardo is currently the vice president of Administration and Finance of West Visayas State University, Iloilo City, Philippines. His dissertation is "Discovering driving patterns using rule-based intelligent data mining agent (RiDAMA) in distributed insurance telematic systems." He has published 54 research papers in national and international journals and conferences. He is a referee of international conferences and journal publications such as *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Knowledge and Data Engineering*. He is interested in the following research fields: distributed systems, telematics systems, CORBA, data mining, web services, ubiquitos computing and mobile communications.

Dr. Gerardo is a recipient of CHED Republica Award in Natural Science Category (ICT field) in 2010. His paper entitled SMS-based Automatic Billing System of Household Power Consumption based on Active Experts Messaging was awarded Best Paper on December 2011in Jeju, Korea. Another Best Paper award for his paper Intelligent Decision Support using Rule-based Agent for Distributed Telematics Systems" Asia Pacific International Conference on Information Science and Technology on December 18, 2008. An Excellent Paper award was given for his paper "Principal Component Analysis Mechanism for Association Rule Mining," Korean Society of Internet Information's (KSII) 2004 Autumn Conference on November 5, 2004. He was given a University Researcher Award by West Visayas State University in 2005.

Bartolome T. Tanguilig III was born on February 24, 1970 in Baguio City, Philippines. He took his bachelor of science in computer engineering in Pamantasan ng Lungsod ng Maynila, Philippines in 1991. He finished his master degree in computer science from De la Salle University, Manila, Philippines in 1999. His doctor of philosophy in technology management was awarded by the Technological University of the Philippines, Manila in 2003.

He is currently the assistant vice president of Academic Affairs and concurrent dean of the College of Information Technology Education and Graduate Programs of the Technological Institute of the Philippines, Quezon City. His research entitled "J-master: an interactive game-based tool for teaching and learning basic java programming" was awarded the best research in the 10th National Convention for IT Education held in Ilocos Norte, Philippines in 2012. He published a research entitled "Predicting faculty development trainings and performance using rule-based classification algorithm" in Asian Journal for Computer Science and Information Technology.

Dr. Tanguilig is a member of Commission on Higher Education Technical Panel for IT Education, Board Chairman of Junior Philippine IT Researchers, member of Computing Society of the Philippines and Philippine Society of IT Educators-NCR.