Agricultural Crops Classification Models Based on PCA-GA Implementation in Data Mining

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

***Abstract*—Extraction of knowledge in agricultural data is a challenging task, from discovering patterns and relationships and interpretation. In order to obtain potentially interesting patterns and relationships from this data, it is therefore essential that a methodology be developed and take advantage of the sets of existing methods and tools available for data mining and knowledge discovery in databases. Data mining is relatively a new approach in the field of agriculture. Accurate information in characterizing crops depends on climatic, geographical, biological and other factors. These are very important inputs to generate characterization and prediction models in data mining. In this study, an efficient data mining methodology based on PCA-GA is explored, presented and implemented to characterize agricultural crops. The method draws improvements to classification problems by using Principal Components Analysis (PCA) as a pre processing method and a modified Genetic Algorithm (GA) as the function optimizer. The fitness function in GA is modified accordingly using efficient distance measures. The approach is to asses, the PCA-GA hybrid data mining method, using various agricultural field data sets, generate data mining classification models and establish meaningful relationships. The experimental results show improved classification rates and generated characterization models for agricultural crops. The domain model outcome may have benefits, to agricultural researchers and farmers. These generated classification models can also be utilized and readily incorporated into a decision support system.**

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

I. Introduction

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

Climatic, geographical, biological and other factors affects the historical yield of crops, and these are very important inputs to computer generated crop yield prediction models. Mathematical and statistical modeling are used to discover patterns in the data, thru these observed field and actual experimental data, statistically generated prediction and characterization models are implemented and used by both farmers and researchers. These models are also helpful to government organizations in establishing proper policies for decision making process.

Other than statistics, a very interesting process known as Knowledge Discovery from Databases (KDD) can be used. One of the core tasks involved in the KDD process is Data Mining (DM). The framework of the KDD process shown in Fig. 1, involves several tasks or phases.

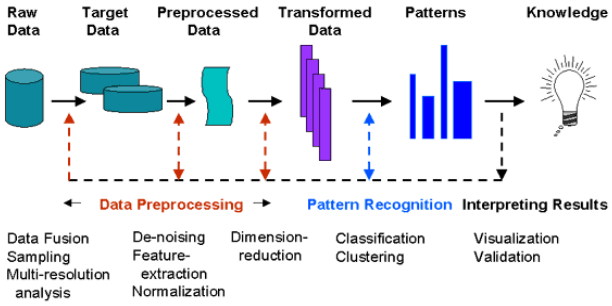


Fig. 1. The KDD process.

The explosion of the information revolution and the proliferation of using computing and information storage made available enormous amounts of data, this led to new methods and techniques such as data mining that can bridge the knowledge discovered in the data. The extraction of knowledge in the data is now a challenging task, from discovering interesting patterns and relationships to interpretation of what the data is. In order to obtain potentially interesting patterns and relationships in the data, it essential that a methodology be developed, taking advantage of the set of existing methods and tools available for data mining and knowledge discovery in databases.

In data mining, classification can be seen as pattern recognition. Each pattern from the data is represented by a set of measurable features or dimensions and viewed as a point in a given dimensional space. The aim is to choose features that allow us to discriminate between patterns belonging to different classes. Often, the optimal set of feature is usually unknown [2], considering every single feature of an input pattern in a large feature set makes classification 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

Manuscript received May 5, 2014; revised July 10, 2014. This work was supported by 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: delacruz.geri@gmail.com).

B. D. Gerardo is with the Administration and Finance at the West 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, Philippine (e-mail: bttanguilig_3@yahoo.com).

number of features as discriminating and small as possible to reduce computational time and complexity [5], [6] in the data mining process.

The focus of this study is to implement an efficient data mining mechanism based on the combination of Principal Component Analysis (PCA) as a preprocessing method and a modified Genetic Algorithm (GA) [7], [8] as the learning algorithm, in order to reduce computational cost and time by keeping a number of features as discriminating and small as possible. 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. The outcome of the data mining modeling can be utilized for decision support in improving agricultural crops productivity.

II. Related Literature

Data mining in agriculture is new, however there are novel ideas and studies conducted to explore its applicability from mining and modeling data to developing applications, using the generated models based on the algorithms used.

The study in [9] reviewed the application of data mining techniques and found out that there are several algorithms and techniques being applied in the agriculture domain. Similarly, in [10], data mining techniques was applied to characterize soil data and found that data mining depends on the amount of data used in the process. Their study applied Naïve Bayes in classifying agricultural land soils. That an increase in dataset size improves accuracy, which may improve the verification of valid patterns compared to standard statistical analysis.

On the effectiveness of data mining as a tool, the paper of Raoranne A. A., Kulkarni R. V. [11], discussed how data mining can bridge knowledge of the data to crop yield estimation. 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. Data mining classification techniques applied to soil database can be successful in establishing meaningful relationships from the data [12].

There are different data mining techniques available in the literature to improve data mining tasks. Reference [13] 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 [14] is one of these techniques and performs well in reducing complexity in data by reducing its dimensionality. In [15] 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 [16]-[18].

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 [19] suggested tweaking the GA to a specific problem and correctly set all parameters, conversely, L. Na-Na, G. Jun-Hua, and L. Bo-Ying [20], used the negative selection method and showed promising results.

In the study of A. S. Elden, M. A. Mustafa, H. M. Harb and A. H. Emara [21], 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 [22], [23] PCA was applied, then the k-NN classifier was used as the fitness function for the GA and resulted to reduced classification error rates, they further recommended using different classifiers for similar studies.

III. Concepts and Methods

A. The Data Mining Mechanism

There are two major phases in the data mining mechanism being presented, 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, this process can be considered as an optimization technique. The second phase is to utilize the optimum results and rules in generating models of classification for the characterization of crops. This prediction model is then utilized for decision support.

B. Methods and Procedures

1) Data preprocessing

Data Preprocessing is an important task and technique in the data mining process it transforms data into understandable format. Real world data is incomplete, noisy, inconsistent and lacking certain trends. Data preprocessing resolves these issues which includes cleaning, transformation, normalization, feature extraction and selection.

PCA is a technique that converts 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 task performed to generalize known structure in data mining to apply to new data. It is also the categorization of data for its most effective and efficient use. There are numerous data mining classification algorithms being studied and implemented in different domains. Some of the most popular and common are adapted and presented herein, based on their capabilities simplicity and robustness.

a) K-nearest neighbor (k-NN)

The principle behind this method is to find predefined numbers of training samples closest in the distance to the 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. There are distance measures implemented in the k-NN, Euclidean, Chebysheb, Manhattan and Edit Distance, but the Euclidean distance measure is the most common choice. Despite its simplicity it is successful in large number of classification problems.

b) J4.8

J4.8 decision trees algorithm is an open source Java implementation of the C4.5 [24]. 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. It assumes that one attribute works independently of the other attributes contained by 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) Genetic algorithm (GA)

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

4) Agricultural datasets

Relevant agricultural data was selected from the UCI machine learning repository (https://archive.ics.uci.edu/), specific to agricultural crops, the soybean [25] and mushroom [26] datasets.

a) Soybean dataset

TABLE I: SOYBEAN DATASET DESCRIPTION

	Attribute Name	Contains
1	date	april,may,june,july,august,september, october
2	plant-stand	normal,lt-normal
3	precip	lt-norm,norm,gt-norm
4	temp	lt-norm,norm,gt-norm
5	hail	yes,no
6	crop-hist	diff-lst-year,same-lst-yr,same-lst-two-yrs, same-lst-sev-yrs
7	area-damaged	scattered,low-areas,upper-areas,whole-field
8	severity	minor,pot-severe,severe
9	seed-tmt	none,fungicide,other
10	germination	90-100,80-89,lt-80
11	plant-growth	norm,abnorm
12	leaves	norm,abnorm
13	leafspots-halo	absent,yellow-halos,no-yellow-halos
14	leafspots-marg	w-s-marg,no-w-s-marg,dna
15	leafspot-size	lt-1/8,gt-1/8,dna
16	leaf-shread	absent,present
17	leaf-malf	absent,present
18	leaf-mild	absent,upper-surf,lower-surf
19	stem	norm,abnorm
20	lodging	yes,no
21	stem-cankers	absent,below-soil,above-soil,above-sec-nde
22	canker-lesion	dna,brown,dk-brown-blk,tan
23	fruiting-bodies	absent,present
24	external-decay	absent,firm-and-dry,watery
25	mycelium	absent,present
26	int-discolor	none,brown,black
27	sclerotia	absent,present
28	fruit-pods	norm,diseased,few-present,dna
29	fruit-spots	absent,colored,brown-w/blk-specks, distort,dna
30	seed	norm,abnorm
31	mold-growth	absent,present
32	seed-discolor	absent,present
33	seed-size	norm,lt-norm
34	shriveling	absent,present
35	roots	norm,rotted,galls-cysts
36	class	diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot, phytophthora-rot, brown-stem-rot, powdery-mildew, downy-mildew, brown-spot, bacterial-blight, bacterial-pustule, purple-seed-stain, anthracnose, phyllosticta-leaf-spot, alternarialeaf-spot, frog-eye-leaf-spot, diaporthe-pod-&-stem-blight, cyst-nematode, 2-4-d-injury, herbicide-injury

The cassava dataset was taken from reference [27] and Agrinet [28]. The raw data was selected and cleaned based on relevant fields with the assistance of an expert.

Using a text editor, all the datasets were encoded, formatted and converted into an *attribute relation file format (.arff)* for input compatibility to the DM software and to allow the machine learning algorithms be applied to generate relevant outcomes. The descriptions of the datasets, the

attributes and the data type that it may contain are shown in the following tables.

b) Mushroom dataset

TABLE III: MUSHROOM DATASET DESCRIPTION

	Attributes	Contains
1	cap-shape	bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
2	cap-surface	fibrous=f, grooves=g, scaly=y, smooth=s
3	cap-color:	brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
4	bruises	bruises=t, no=f
5	odor	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
6	gill-attachment	attached=a, descending=d, free=f, notched=n
7	gill-spacing	close=c, crowded=w, distant=d
8	gill-size	broad=b, narrow=n
9	gill-color	black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
10	stalk-shape	enlarging=e, tapering=t
11	stalk-root	bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
12	stalk-surface-above-ring	fibrous=f, scaly=y, silky=k, smooth=s
13	stalk-surface-below-ring	fibrous=f, scaly=y, silky=k, smooth=s
14	stalk-color-above-ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
15	stalk-color-below-ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
16	veil-type:	partial=p, universal=u
17	veil-color:	brown=n, orange=o, white=w, yellow=y
18	ring-number:	none=n, one=o, two=t
19	ring-type:	cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
20	spore-print-color	black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
21	population	abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
22	habitat	grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d
23	class	edible=e, poisonous=p

c) Cassava dataset

TABLE II: CASSAVA DATASET DESCRIPTION

	Attributes	Contains
1	f_root_yield	low, moderate, high
2	Root_dry_matter_content	REAL
3	Root_starch_content	REAL
4	Root_skin_color	dark-brown, cream, light-brown
5	Root_HCN_content	low, moderate, high
6	Root_flesh_color	white, yellow, cream
7	unexpanded_apical_leaves	purple, green-purple, dark-green, light-green, green
8	first_fully_expanded_leaves	green-purple, light-green, dark-green, purple, green
9	Petiole_color	purple, green-purple, light-green
10	Stem_color	dark-brown, light-brown, silver-green, reddish-brown
11	Plant_type	slightly-branching, slightly-or-non-branching, erect-and-non-branching, moderately-branching, erect-or-slightly-branching, erect-and-late-branching, erect, slightly-and-late-branching, erect-and-slightly-branching
12	Red_spider_mites	HR, MR, R
13	White_peach_scale_insects	HR, MR, R
14	Cassava_bacterial_blight	HR, MR, R
15	Maturity	low, high
16	Recommended_Uses	1, 2, 3
17	Recommended_for_release	IPB-UPLB, PhilRootcrops-VSU
18	Year_released	1982,1986,1987,1988,1990,1993,1994,1996,1997,1999,2000,2001,2003,2004,2006,2007,2009
19	Class	UPLCa-1(Datu1),UPLCa-2(Lakan1),UPLCa-3(Sultan1)(G50-3),UPLCa-4(Vassourinha),UPLCa-5(Sultan2)(G29r-3),PSBCv-9(VC-4)(CM4014),PSBCv-11(Lakan2)(CMP3419-2A),PSBCv-12(Lakan3)(SM972-20),PSBCv-13(CMP62-15),PSBCv-15(Lakan4)(CM3422-1),PSBCv-17(Sultan3)(CG87-03-01),PSBCv-18(Sultan4)(CG87-02-13),PSBCv-19(SM808-1),PSBCv-20(Sultan5)(CG91-13-01),NSICCv-25(Sultan6)(CG91-08-05),NSICCv-26(Sultan7)(CG91-14-01),NSICCv-27(Datu2)(CM9158-4),NSICCv-28(LSUCv-14)(OMR36-05-09),NSICCv-30(LSUCv-15)(Rayong5),NSICCv-35(LSUCv-18)(CMR37-24-1),NSICCv-36(LSUCv-19)(OMR36-62-03),NSICCv-37(Sultan9)(CG97-17-02),NSICCv-39(Rajah3)(CG97-05Rr-01),NSICCv-40(Sultan10)(CG97-03-04),NSICCv-41(Sultan11)(CG97-09-23),NSICCv-42(Rajah4)(CG97-01r-04),NSICCv-43(LSUCv-21)(OMR40-40-03),NSICCv-44(LSUCv-22)(CMR39-50-18),NSICCv-45(LSUCv-23)(OMR39-48-02),NSICCv-46(VSUCv-24)(CMR40-09-34)

IV. IMPLEMENTATION OF THE PCA-GA MECHANISM

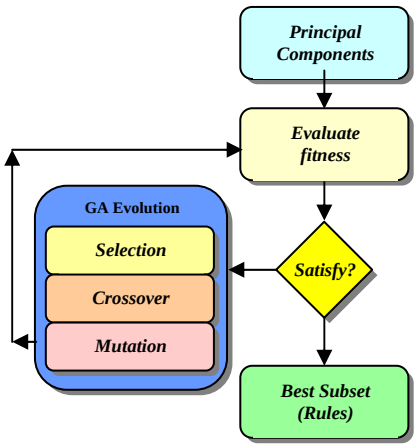


Fig. 2. Exploded view of the PCA-GA method.

The idea is to implement 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 used as input population in the search space of the GA in searching for the optimum solution. This mechanism efficiently simplifies the data mining process using the representative data of the original dataset, to which reduces computational time and improves

classification performance of classifiers

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, using other classifiers is explored and 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.

A. The PCA-GA Algorithm

[Start] Principal components as population

[Fitness] Compute and evaluate the fitness of each principal component in the population

[Test] If the end condition is satisfied, stop and return the best solution in current population, otherwise,

[New population] Create new population by repeating the following steps until the new population is complete

[Selection] Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)

[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.

[Mutate] With a mutation probability mutate new offspring

[Accept] Place new offspring in a new population

[Replace] Use new generated population

[Loop] Go to step b

V. Experimental Results

Using the classifiers presented is adopted in the experiment 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.

The experiment used the WEKA [29] version 3.6.10 data mining software in the implementation and utilization. A computer with 2 Gigabyte of memory, equipped with a 2.80 Ghz Processor, and a proprietary 32 bit Operating System was utilized. The default settings in the data mining software and in the algorithm configurations, was used in the experiment.

Three (3) agricultural crops field data, soybean, mushroom and cassava datasets was used in the experiment. The cassava and mushroom datasets were converted to a compatible file format for input to the DM software.

A. Feature Sets Selected by PCA-GA

The soybean dataset has originally thirty six (36) attributes including the class label, the mushroom and cassava datasets has twenty three (23) and eighteen (18) attributes respectively. After preprocessing using PCA, the transformed dataset contained forty one (41) principal components for the soybean, fifty nine (59) for the mushroom and twenty two (22) for the cassava datasets, including the class labels.

GA was applied to the resulting preprocessed data, further reducing the datasets. In the optimization process, GA selected feature sets that are considered the optimum, thereby further reducing the data into a smaller representative dataset.

TABLE IV: Number of Feature Sets by Implementing PCA-GA

Datasets	PCA	k-NN-GA	J4.8-GA	Naïve Bayes-GA	MLP-GA
Cassava	22	2	2	3	---
Mushroom	58	2	21	17	---
Soybean	41	2	18	26	---

As can be seen in Table IV, data preprocessing using PCA and feature selection using GA resulted into a smaller number of feature sets, which are considered the best representative feature sets of the data.

B. Classification Performance Results

The following tables are the results of the classification process.

TABLE V: Classification Rates with k-NN Both as Fitness Function in GA and Classifier Using Different Distance Measures

Dataset	k-NN	PCA-Modified GA (Euclidean/Chebysheb/Manhattan)
Soybean	99.85%	99.85%
MAE	0.0029	0.0026
RMSE	0.0152	0.0105
RAE	2.99	2.74
RRSE	6.93	4.77
Tine	0.0sec	0.0sec
Mushroom	100%	99.98%
MAE	0.0001	0.0002
RMSE	0.0001	0.0078
RAE	0.0246	0.0458
RRSE	0.0246	1.572
Time	0.01sec	0.0sec
Cassava	100%	100%
MAE	0.0317	0.0317
RMSE	0.0898	0.0898
RAE	50.8475	50.8475
RRSE	50.8406	50.8406
Time	0.0sec	0.0sec

TABLE VI: Classification Rates Uisng Different Classifiers with J4.8 as Fitness Function in GA

Classifier	Soybean		Cassava		Mushroom	
	Original	PCA-GA	Original	PCA-GA	Original	PCA-GA
k-NN	91.22% 0.0sec	99.85% 0.0sec	100% 0.0sec	100% 0.0sec	100% 0.01sec	99.85% 0.01sec
J4.8	91.51% 0.02sec	98.68% 0.1sec	53.33% 0.0sec	46.67% 0.0sec	100% 0.05sec	100% 0.52sec
Naïve Bayes	92.97% 0.01sec	92.53% 0.0sec	100% 0.0sec	60.00% 0.0sec	95.83% 0.02sec	97.22% 0.12sec
MLP	93.41% 112sec	98.83% 18.2sec	100% 5.32sec	6.67% 0.0sec	100% 1876sec	99.96% 72.4sec

Table V shows the performance of the modified GA, using the k-NN as the classifier and at the same time the fitness function in varied distance measures. It can be seen there is no significant difference in the classification accuracy, compared to the classification rate on the original dataset. This can be attributed to the nature of similarities in the distance measurement functions. However, further analysis,

the observed errors for the soybean dataset was reduced after implementing PCA-GA on the original dataset. For the mushroom dataset, accuracy was negatively affected but processing time improved.

In the case of the cassava dataset, there is no significant difference observed, this maybe attributed to the experts knowledge in selecting relevant attributes in the encoding and creation of the dataset.

Table VI shows otherwise the resulting effect of implementing the J4.8 classifier as a fitness function in GA and using different classifiers in the classification process. The results, implies that classification performance can be improved by using GA as an optimization technique in the classification process. With the exception of Naïve Bayes, further analysis shows that its performance is dependent on the nature of the dataset, and fitness function used.

TABLE VII: CLASSIFICATION RATES SUMMARY BOTH AS CLASSIFIER AND AS FITNESS FUNCTION IN GA

Classifier	Soybean		Cassava		Mushroom	
	Original	PCA-GA	Original	PCA-GA	Original	PCA-GA
k-NN	91.22% 0.0sec	99.85% 0.0sec	100% 0.0sec	100% 0.0sec	100% 0.01sec	99.98% 0.0sec
J4.8	91.51% 0.02sec	98.68% 0.1sec	53.33% 0.0sec	46.67% 0.0sec	100% 0.05sec	100% 0.52sec
Naïve Bayes	92.97% 0.01sec	94.44% 0.01sec	100% 0.0sec	100% 0.0sec	95.83% 0.02sec	97.89% 0.07sec
MLP	93.41% 112sec	---	100% 5.54sec	20% 0.72sec	100% 1876sec	---

The MLP performed exceptional on the datasets, specific to the speed of processing on the mushroom and soybean datasets, which shows a very significant difference between the original and the PCA-GA reduced dataset. Interesting also to note, the classification rates on the mushroom dataset, though the classifier performs outstanding with the original dataset, the indicated classification process took longer to perform as compared to the other classifiers but is exceptional in speed on the PCA-GA reduced dataset.

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 classification process. This can also be attributed to the characteristics of the GA.

It can be seen from Table VII, that a combination of PCA and a modified GA improves classification accuracy, specific to Naïve Bayes for all the datasets, using it both as fitness function and classifier. The k-NN and J4.8 also performed well for the mushroom and soybean dataset. The J4.8 and MLP classifiers performed negatively after the PCA-GA was applied to the cassava dataset.

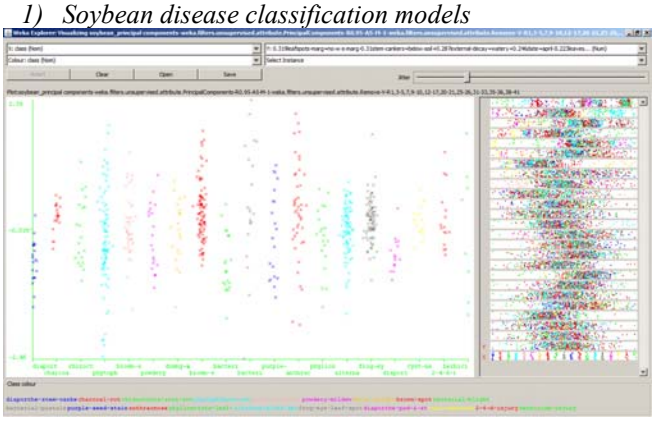
The MLP classifier as it has been observed in the experiment, poorly performed in processing time both as classifier and as fitness function with the GA in the optimization process, although exceptionally accurate in classifying the original datasets, this maybe attributed to the MLP characteristics.

The speed of processing as can be seen does not have significant difference for all the datasets, with the exception of the MLP. The performance rates observed may be attributed to the dependency and characteristics of the classifiers on the nature of the datasets: large, small, clean or noisy.

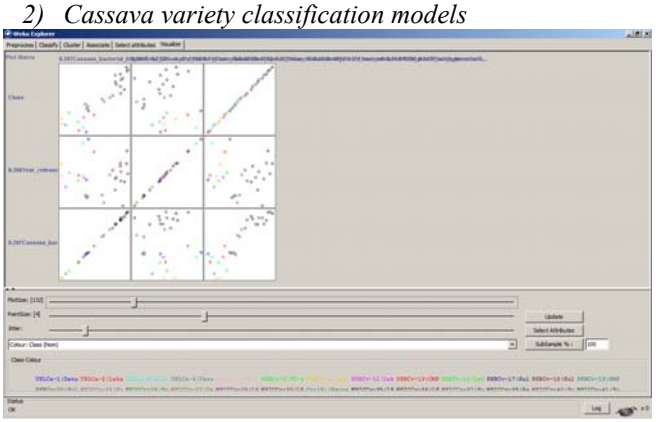
C. Classification Models Visualization Using Naïve Bayes as Fitness Function in GA and as Classifier after applying PCA-GA

Now, that we have shown the performance results of the data mining process, for its exceptional performance we select the Naïve Bayes classifier in the presentation of the models generated in implementing the PCA-GA data mining mechanism for the characterization of the agricultural crops. Shown in the following sections, are models of nineteen (19) soybean disease classification, cassava varietal productivity and mushroom edibility classification, using the Naïve Bayes classifier which performed exceptional in the data mining process with near perfect accuracy for all of the datasets.

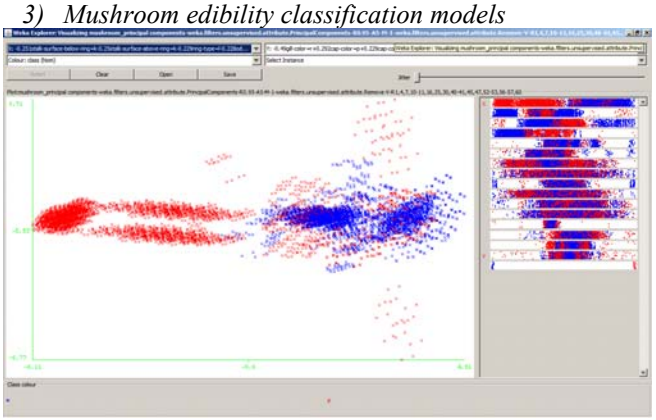
1) Soybean disease classification models



2) Cassava variety classification models



3) Mushroom edibility classification models



The reader is presented a view of the results of the experiment, which classifier is best suited in characterizing crops based on the PCA-GA mechanism. It illustrates that these are possible techniques and the choice is to have the most efficient and accurate model in characterizing crops.

The visuals presented, proves the capability of the classifiers based on the PCA-GA reduced dataset to discriminate and categorize the data on different classes presented in the original datasets. This implies that, data mining classification based on the PCA-GA mechanism is efficient and advantageous as compared to raw, large and complex datasets. Further analysis of the models, validates the efficiency and simplicity of implementing the mechanism in mining on representative set resulting to improved classifier performance and establishing significant relationships among the variables that influence higher accuracy rates, thus simplifying the task of characterizing crops.

D. Extracted Classification Rules Using JRIP and PART

To further demonstrate, shown in Table VIII, are the number of extracted rules using JRIP and PART from the PCA-GA mechanism using the Naïve Bayes as the fitness function in GA with the corresponding accuracy rates. It can be seen that the extracted classification rules from the mushroom dataset is highly accurate, hence establishing valid relationships among the variables from the reduced representative dataset based on PCA-GA.

TABLE VIII: DISCOVERED CLASSIFICATION RULES FROM THE PCA-GA REDUCED DATASET BASED ON THE NAÏVE BAYES AS FITNESS FUNCTION

Classifier	Soybean		Cassava		Mushroom	
	Rules	% Accuracy	Rules	% Accuracy	Rules	% Accuracy
JRIP	25	91.25	1	96.67	9	99.99
PART	32	98.68	11	63.33	9	99.99

To illustrate the established relationships that influence higher accuracy rates in the characterization process, let us consider the mushroom dataset. Careful analysis of the rules generated by JRIP and PART, the summary in Table IX shows the attributes that characterize the mushroom crop, with near perfect accuracy. The results imply that the rules extracted using the attributes presented can be used efficiently in the implementation of an intelligent system for agricultural crops characterization.

TABLE IX: VARIABLES THAT HIGHLY INFLUENCE THE CLASSIFICATION OF MUSHROOM BASED ON PCA-GA MECHANISM

Classifier	Rules	Attributes Involved
JRIP/PART	9	stalk-surface-below-ring, stalk surface above ring, ring-type, stalk-root, cap-color, odor, habitat, population, gill-size, stalk-color-above ring, stalk-color-below-ring, gill-spacing, bruises, cap-shape, cap-surface, gill-color

The simplest rules were of the JRIP involving sixteen (16) out of twenty two (22) attributes. The rules for edible mushrooms are obtained as negation of the extracted rules. In the case of PART, the same attributes were found to have high influence in the edibility of mushroom, though slight variations on the rules exist for non-edible and edible mushrooms.

VI. SUMMARY AND CONCLUSION

Presented in this paper, is a proposed hybrid data mining method based on PCA-GA. The mechanism was shown to have considerable influence in improving classification performance rates of classifiers. Using both the classifiers as fitness function in GA and in the data mining classification process, improves the performance of the data mining process.

The Naive Bayes and k-NN classifiers both performed exceptional as fitness functions and classifiers in the PCA-GA data mining mechanism. Likewise, the findings available in the literature is further validated which showed significant results with other distance measures in the k-NN.

Based on the results of the experiment, the implementation of the algorithm based on PCA-GA is efficient 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. Further observation, the classifiers may have dependencies on the nature of the datasets, specific to the attribute data types and in the pre-processing technique used in using actual field data of agricultural crops.

VII. RECOMMENDATION AND FUTURE WORK

It is suggested that similar studies may also be undertaken in using other preprocessing techniques and further study on the PCA. Further validation and evaluation of the proposed method is also suggested using other agricultural datasets with varying attribute data types, and using the results herein as benchmark data. The classification models and rules generated and presented can be used as baseline framework in the development of intelligent systems for precision agriculture.

Future work involves further study to use other efficient distance measures in the k-NN data mining classification algorithm not presented herein and using only efficient distance measures as the fitness function is being considered.

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*, (Antwerp, Belgium, September, 2008, pp. 90-105.

[5] B. D. Gerardo and J. Lee. Principal Component Analysis Mechanism for Association Rule Mining. [Online]. Available: www.dbpia.co.kr/Journal/ArticleDetail/669692

[6] D. Diepeveen and L. Armstrong, "Identifying key crop performance traits using data mining," in *Proc. the IAALD AFITA WCCA2008, World Conference on Agricultural Information and IT*, pp.1-21, 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] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Reading Menlo Park: Addison-Wesley, vol. 412, 1989.
[9] N. G. Yethiraj, "Applying Data Mining Techniques in the field of agriculture and allied sciences," *International Journal of Business Intelligents*, vol. 1, no. 2, December 2012.
[10] P. Barghavi and S. Jyothi, "Applying naïve bayes data mining technique for classification of agricultural land soils," *International Journal of Computer Science and Network Security*, vol. 9, no. 8, pp 117-122, August 2009.
[11] A. 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.
[12] R. Vamanan and K. Ramar, "Classification of agricultural land soils a data mining approach," *International Journal of Computer Science and Engineering*, vol. 3, no. 1, pp. 379-384, 2011.
[13] 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.
[14] C. J. Burges, "Dimension reduction: A guided tour, Machine Learning," *Foundations and Trends in Machine Learning*,vol. 2 no. 4, pp. 275-365, 2009.
[15] B. 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.
[16] H. Liu and H. Motoda, "Feature transformation and subset selection," *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 2, pp. 26-28, 2008.
[17] 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 .
[18] 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.
[19] A. Hassani and J. Treijis, "Overview of standard and parallel genetic algorithms," in *Proc. IDT Workshop on Interesting Results in Computer Science and Engineering (IRCSE '09),* Mälardalen University, Sweden, October 30, 2009.
[20] 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*, pp. 4297-4299, 2006.
[21] 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 (IJCSIT)*, vol. 5, no. 2, pp. 73-85, April 2013.
[22] 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*, Hong Kong, October 14-16, 1998, pp. 371-384.
[23] 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.
[24] J. R. Quinlan, *C4. 5: Programs for Machine Learning*, vol.1. Morgan Kaufmann, 1993.
[25] Soybean (Large) Data Set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Soybean+%28Large%29
[26] Mushroom Data Set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Mushroom
[27] W. M. G Fukuda, C. L. Guevarra, R. Kawuki, and M. E. Fuerguson, "Selected morphological and agronomic descriptors for the characterization of cassava," *International Institute of Agriculture*, (IITA), Ibadan, Nigeria, pp. 1-19, 2010.
[28] Agripinoy. [Online]. Available: http://www.agripinoy.net/
[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Rtemann, and I. H. Witten, *The WEKA Data Mining Software: An Update; SIGKDD Explorations*, vol. 11, no. 1, 2009.

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 master's of science in information technology degree at Hannam University, Daejon, South Korea in 2003. He became a member of IACSIT in 2012. He is currently the assistant dean of the institute of engineering and an associate professor of information technology at the Tarlac College of Agriculture in Camiling, Tarlac, Philippines.

Mr. Dela Cruz is also a member of IAENG and its data mining and computer societies.

Bobby D. Gerardo is currently with the 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 to international conferences and journal publications such as in 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 masters degree in computer science from De La Salle University, Manila, Philippines in 1999, and his doctor of philosophy in technology management was earned at the Technological University of the Philippines, Manila in 2003. He is currently the assistant vice president for academic affairs and concurrent the dean of the College of Information Technology Education and Graduate Programs of the Technological Institute of the Philippines, Quezon City. 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.