# ENSEMBLED PEARSON CHI-SQUARE FEATURE SELECTION AND KERNEL LEAST SQUARE SUPPORT VECTOR CLASSIFIER FOR AGRICULTURE SEED GROWTH

[1]M.INDIRA, [2]Dr. S.JAYASANKARI
[1]RESEARCH SCHOLAR, [2]ASSISTANT PROFESSOR
Department of Computer Science,
P.K.R. Arts College for Women (Autonomous), Gobichettipalayam, Erode (Dt), India

## ABSTRACT

Data mining is the process of analyzing the data from the large database for extracting the valuable information with the relevant features. Feature selection is a preprocessing step before the classification to improve the mining performance by eradicating the irrelevant features and choosing the relevant features from the database. The several feature selection is done using different data mining techniques but it has high computational complexity. In order to improve the feature selection and classification, a Pearson Chi-Square Based Kernel Least Square Support Vector Classifier (PCS-KLSSVC) is introduced. The PCS-KLSSVC includes two processes namely feature selection and classification for identifying the seed growth in the agriculture field. At first, the feature selection is performed using Pearson chi-squared hypothesis test. Based on the score value, relevant features are selected by removing the irrelevant features for performing classification. This helps to improve feature selection rate and minimizes the classification time. Secondly, the Kernelized Least Square Support Vector Classifier is used for performing the classification process to predict the seed growth. The classifier constructs an optimal and marginal hyperplane for improving the classification accuracy and minimizes the error rate. Experimental evaluation of proposed PCS-KLSSVC and existing methods are carried out with different factors such as feature selection rate, classification accuracy, classification time and space complexity with respect to a number of features and data points. The experimental results reported that the proposed PCS-KLSSVC obtains high classification accuracy and feature selection rate with minimum time as well as space complexity.

***Keywords: Data mining, Pearson chi-squared hypothesis test, feature selection, Kernelized Least Square Support Vector Classifier, hyperplane, seed growth***.

## 1. INTRODUCTION

The database contains a large number of features (i.e. attributes) for performing the certain task. If the numbers of features are high, it degrades the performance of classification. In order to perform efficient classification, feature selection plays a major role to choose the most optimal features by eliminating features that are irrelevant. Moreover, the feature selection reduces the complexity and increases the classification accuracy. The different data mining technique is applied for agriculture field to identify the seed growth. Some of the recent related works regarding the feature selection and classification are reviewed in this section.

The support vector machines discriminant analysis (SVM-DA) was introduced in [1] for classifying the samples with better performances in sensitivity and specificity. However, the performance of the time complexity and space complexity remained unsolved. A new hybrid ensemble approach was developed in [2] that includes a group of two machine learning algorithms namely logistic Regression and naïve Bayes classifier. The gain ratio is employed for choosing the features. The hybrid ensemble approach failed to improve the feature selection rate and minimize the classification time.

Support Vector Machines classifier was presented in [3] for identifying and classifying the plant diseases with high accuracy. The feature selection was not performed to further improve the classification accuracy. The fuzzy rough set theory was presented in [4] for selecting the features from the original dataset. It does not increase the classification accuracy with the selected features. A Distribution Preserving Feature Selection (DPFS) algorithm was presented in [5] to minimize the computational complexity. The DPFS algorithm failed to resolve the space complexity.

Fitting Model Based on Fuzzy Rough Sets was presented in [6] for selecting the features from the dataset. It failed to examine how the proposed model was applied to the field of classification learning. In [7], a Feature Association Map approach was applied to choose the relevant features for effective classification and clustering process. This approach provides high complexity during the classification. An improved-RFC (Random Forest Classifier) approach was introduced in [8] for solving the multi-class classification problem. The approach does not obtain the accurate classification since the relevant features were not selected.

In order to select the relevant feature for classification, an efficient data mining technique called Mutual information criterion was developed in [9]. This approach chooses the dependent feature and eradicates the irrelevant features. This approach failed to handle the categorical feature selection. A neural network classifier using floating centroids method was introduced in [10] to classify the data sample. The classifier failed to resolve the error occurred during the classification.

The various issues are identified from the above-said issues are more time complexity, lack of accuracy, failure of feature selection, more space utilization and so on. In order to resolve such kinds of problems, an efficient machine learning technique called PCS-KLSSVC is introduced.

The main contributions of the PCS-KLSSVC are summarized as follows,

- The contributions PCS-KLSSVC technique is to enhance classification accuracy and minimize the time and space complexity. This contribution is achieved by applying Pearson Chi-Square Based Kernel Least Square Support Vector Classifier. Pearson chi-squared hypothesis test is applied for finding the significant features from the dataset based on the score value. This helps to improve feature selection rate and minimize classification time.
- Kernelized least square support vector classifier constructs the hyperplane to classify the data point as normal or abnormal through the similarities between the training and testing data points. The kernel function is employed to measure the similarity between any pair of data points. Least squares function reduces the squared difference between observed and predicted classification results. This assists to improve the classification accuracy. The classification is done with the selected features to minimize the space complexity.

The rest of the paper is arranged in the following manner. Section 2 discusses the related works. Section 3 provides the description of the PCS-KLSSVC with two algorithms. In section 4, experimental evaluation is described with the dataset. Section 5 provides the results and discussion of various parameters. Section 6 provides the conclusion of the paper.

## 2. RELATED WORKS

In [11], a novel classification rule mining technique was introduced for classifying the rice diseases based on the feature selection. The rough set theory was employed to select the features and minimizes the information loss. Though the method minimizes the computational complexity, the classification accuracy was not improved.

A randomized feature selection and classification (RFSC) approach was developed in [12] to minimize the complexity. The RFSC approach has less computational efficiency in both feature selection and classification. The different feature selection approach and classifiers were presented in [13]. But these methods did not improve the classification accuracy with less error rate.

A logistic localized feature selection (lLFS) algorithm was designed in [14] to improve the classification. The algorithm failed to choose the less number of features for minimizing the classification time. A Joint feature selection and classification (JFSC) technique was introduced in [15] to choose the discriminative features for an effective classification model. The JFSC technique failed to handle the multiclass problem. Semi-supervised extreme learning machine algorithm was presented in [16] for optimal feature selection. The algorithm was not adapted to obtain the high classification accuracy.

A new immune clonal genetic algorithm based on an immune clonal algorithm (ICGFSA) was introduced in [17] for choosing the features effectively. The algorithm does not apply to more datasets for testing performance. A hybrid feature selection technique depends on Multiple Kernel Learning (MKL) to find the similarity between features for efficient classification. The technique minimizes the error but the classification time was not reduced.

A hybrid genetic algorithm with wrapper Embedded feature approach was introduced in [18] for increasing the performance of feature selection and classification. The approach does not improve the classification with less complexity.

A Dynamic Relevance and Joint Mutual Information Maximization (DRJMIM) approach were introduced in [19] for choosing the less number of features to achieve high classification accuracy. The method failed to improve feature selection and it has high time complexity. A Dynamic Change of Selected Feature with the class (DCSF) was introduced in [20] for finding the variation between the features. The method does not discover the relation between linear features.

The problems identified from the existing methods are overcome by introducing a PCS-KLSSVC. The description of the PCS-KLSSVC is presented in the next section.

## 3. PEARSON CHI-SQUARE FEATURE SELECTION AND KERNEL LEAST SQUARE SUPPORT VECTOR CLASSIFIER FOR AGRICULTURE SEED GROWTH

An efficient technique called Pearson Chi-Square Based Kernel Least Square Support Vector Classifier (PCS-KLSSVC) is introduced with the objective of improving the classification accuracy and minimizes the computational complexity such as space and time. In general, the dataset contains a number of features resulting in the higher computational complexity. The PCS-KLSSVC technique uses the feature selection for extracting optimal features from the dataset. In addition, feature selection is more efficient and minimizes the computation time of classification as well as minimizes the curse of dimensionality. In order to accurately distinguish the data points, the PCS-KLSSVC technique uses the Kernelized Least Square Support Vector Classifier with the selected features. The classification process is employed to evaluate the seed growth in the agriculture field. The architecture diagram of PCS-KLSSVC is shown in below Figure 1.

As shown in figure 1, the architecture of proposed PCS-KLSSVC is described for identifying the seed growth in the agriculture field. The proposed technique includes two major processing steps namely feature selection and classification. The feature selection is carried out using Pearson chi-squared hypothesis test.

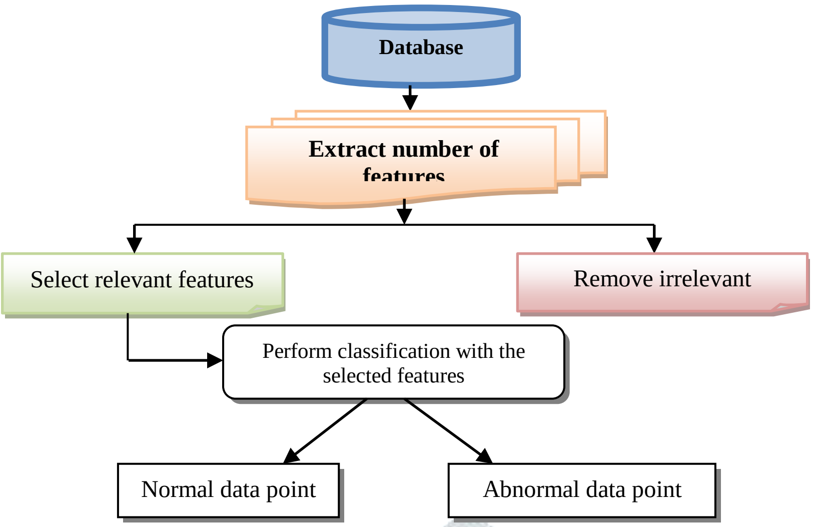

**Figure 1 Architecture of PCS-KLSSVC**

Based on the chi-squared value, the relevant feature and irrelevant features are identified. The relevant features are employed for classification. The classification is done with the help of Kernelized least square support vector classifier to predict seed growth. The brief explanations of these two processes are explained in following sections.

### 3.1 Pearson chi-squared hypothesis test based relevant feature selection

The first process in the proposed technique is a relevant feature selection using Pearson chi-squared hypothesis test (PCHT). The dataset consists of hundreds and thousands of instances and each of which is represented by the number of features. While considering the entire features for processing, it takes more space and time complexity. Therefore, an optimal feature selection is an attributes selection for minimizing these complexities during the classification. The PCHT is the statistical sampling distribution for finding the more relevant features from the dataset. The PCHT is used for testing the two categorical features in some population. Let us consider the number of features in the dataset is expressed in the following equations,

$$f_1, f_{2,} f_3 \dots \dots f_n \in D \ (1)$$

From (1) '$D'$ $d$enotes a dataset contains a number of features $f_1, f_{2,} f_3 \dots \dots f_n$. PCHT is a statistical test which is applied to a sets of categorical data used for finding the dependent and independent features based on the chi squared value. The chi-squared test for categorical features are performed as follows,

$$\chi^2{}_p = \sum_{i=1}^{n} \frac{(a_i - e_i)^2}{e_i} (2)$$

From (2),$\chi^2{}_p$ denotes Pearson chi square scores, $a_i$ denotes an observed value and $e_i$ denotes an expected value. The summation symbol represents a calculation is performed for each single features in the data set. From the equation (2) where '$i$' denotes number of the features varies from 1 to n. In proposed PCS-KLSSVC, there is two hypothesis test are carried out namely null hypothesis and alternative hypothesis to find the more relevant features. These two tests are performed with the obtained Pearson chi square score value. The threshold is assigned for Pearson chi square scores to find the minimum and maximum value of $\chi^2{}_p$. The hypothesis test is carried out as follows,

$$y = \begin{cases} \chi^2{}_p < \delta, & h_o \\ \chi^2{}_p > \delta, & h_1 \end{cases} \quad (3)$$

From (3),$y$ represents a hypothesis test outcomes, $\delta$ denotes a threshold value of Pearson chi square scores. If the Pearson chi square scores value less than the threshold value $\delta$, the null hypothesis ( $h_o$) is selected. It means that the observed data fits the expected data extremely well. In other words, two features are dependent. From the equation (3), if the Pearson chi square score value exceeds the threshold value of $\chi^2{}_p$, the alternate hypothesis ( $h_1$) is accepted (i.e. two features are independent). Then these features are accepted as relevant features for classification. As a result, large value of $\chi^2{}_p$ is the two features are independent whereas small value of $\chi^2{}_p$ means two features are dependent. If the two features are dependent, the one feature is selected and removes the other feature. If the two features are independent, two features are selected for classification.The algorithmic description of Pearson chi-squared hypothesis test is described as below

```
Input: Dataset D, Number of features $f_1, f_{2,} f_3 \dots \dots f_n$
Output: Select relevant features for classification
Begin
    For each feature $f_i \in D$
      Calculate $\chi^2{}_p$
       if ( $\chi^2{}_p < \delta$) then
          Two features are dependent
            select one feature and remove another features
        else
```

```
            Two features are independent
               select two features
   end if
 end for
End
```

**Algorithm 1 Pearson chi-squared hypothesis test based feature selection**

Algorithm 1 clearly describes the Pearson chi-squared hypothesis test based feature selection. The less number of features from the dataset is selected for classification. The features selection is done by identifying independent and dependent features. Based on the Pearson chi-square scores, the dependent and independent features are identified. If the score value is higher than the threshold value, then the two features are dependent. From which any one of the features is chosen for classification. If the two features are independent, then the two features are taken for classification. As a result, the classification time is minimized.

### 3.2 Kernelized least square support vector classifier

Once selecting the features from the dataset, the classification is performed to classify the data as normal or abnormal. Data classification is the method of categories the data points into different classes for further processing. Classification is a type of supervised learning approach to predict a seed growth in agriculture field with the selected features. The proposed technique uses Kernelized Least Square Support Vector Classifier (KLSSVC) for classifying the data with minimum error rate. The kernel function is employed for measuring the similarity between any pair of inputs. Least square is used for minimizing the squares of the residuals made in the results i.e. the difference between an observed value and predicted value provided by a model. This is also called as an error. The proposed KLSSVC improves the classification and minimize the error. The flow process of KLSSVC is described as follows,

Figure 2 describes the flow process of KLSSVC based classification. The KLSSVC includes set of training samples $\{(D_1, y_1), (D_2, y_2), \ldots (D_n, y_n)\}$ where $D_n$ denotes a number of data points (i.e. seed) and $y_i$ represents the dependent variable whose value is found by observation. The dependent variable provides the two target class as $y_i \in \{+1, -1\}$.

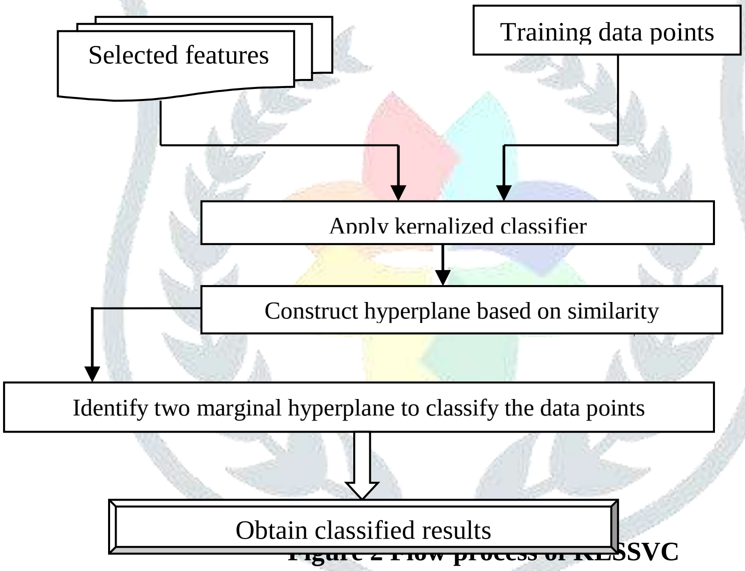
Selected features
Training data points
Apply kernalized classifier
Construct hyperplane based on similarity
Identify two marginal hyperplane to classify the data points
Obtain classified results

**Figure 2 Flow process of KLSSVC**

Therefore, the classification output provides the output $y_i = +1$ is a normal data point, and the output $y_i = -1$ is said to be an abnormal data point. The KLSSVC is a discriminative classifier category the data by using separating hyperplane. The construction of KLSSVC is shown in figure 3.

Figure 3 shows the classification process of KLSSVC where the classification is done by using the optimal separating hyperplane. The hyperplanes are used in KLSSVC to define decision boundaries. The hyperplane (H) partitions the original vector space into two sets. Then the KLSSVC classifies the entire the data points on one side of the decision boundary as belonging to one class and all those on the other side as belonging to the other class. Similarly, the numbers of hyperplanes are employed for classifying the data points into different classes.

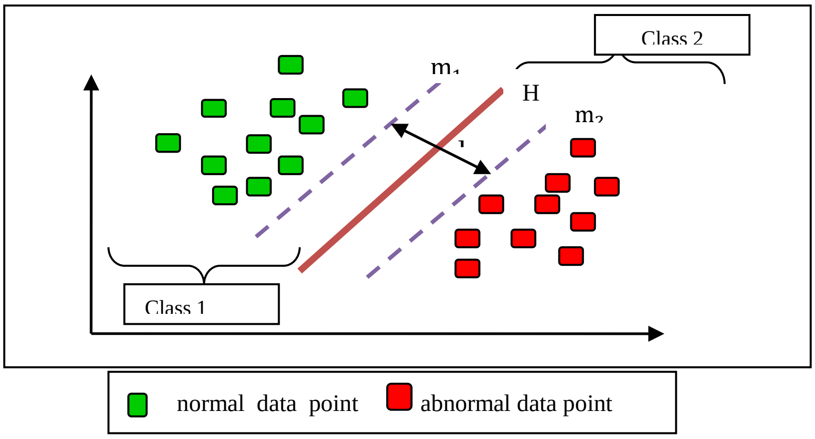

**Figure 3 Kernelized Least Square Support Vector Classifier**

The hyperplane maximizes the margin of training samples. The region surrounded by the hyperplane is called as marginal hyperplane (i.e. $\mathrm{m_1, m_2}$). The support vectors belonging to the margin that affects the direction of the discriminate hyperplane. At each occurrence, the seeds are classified on either side of the hyperplane. In other words, the seeds are separated as both upper and lower side of the margins. The upper sides of the samples are called as normal data points whereas lower labeled samples are abnormal data points. These two classes are obtained by using following mathematical equations.

A separating hyperplane in KLSSVC is mathematically represented as,

$$y_i = \vec{w}.\vec{f} + \vec{b_l} \quad (4)$$

$$H \rightarrow \vec{w}.\vec{f} + \vec{b_l} = 0 \quad (5)$$

From (5),$H$ denotes a hyperplane (i.e. boundary), $f$ denotes training samples (i.e. data points), $\vec{b_l}$ represents a bias and $\vec{w}$ is the normal weight vector to the hyperplane (H). If the training samples are linearly separable, two parallel hyperplanes (i.e. marginal hyperlane) are selected that separate the two classes (+1, -1) of data, so that the distance between them is as large as possible. Therefore, the data points assigned to a positive class and negative class is expressed as follows,

$m_1 \rightarrow \vec{w}.\vec{f} + \vec{b_l} > 0$ for $\mathrm{D_i}$having the class $+1$ (6)

$m_2 \rightarrow \vec{w}.\vec{f} + \vec{b_l} < 0$ for $\mathrm{D_i}$having the class $-1$ (7)

From (6) (7),$m_1, m_2$ denotes a lower and upper marginal hyperplanes to classify the data points as above and below the boundary. The KLSSVC comparing training data points using boundary set (i.e. test data) and classifies it as either normal or abnormal. From the figure 3, '$d$' denotes a distance between the two marginal hyperplane and it is calculated as follows,

$$d = 2 * \left(\frac{1}{\|\vec{w}\|}\right) \quad (8)$$

From (8),$d$ denotes a distance between two marginal hyperplane$m_1$ *and* $m_2$ . The predicted output ($y'$) of the KLSSVC is obtained with the kernel function is expressed as follows,

$$y' = sign \sum_{i=1}^{n} w\, y_i K(f_i, f') \qquad (9)$$

From (9), $y'$ represents a kernalized predicted classification results, "$sign$" determines whether the predicted classification output as positive or negative. The positive output results provide the higher similarity whereas negative similarity denotes less similarity. $K(f_i, f')$ denotes a kernel function that measures the similarity between any pair of data points. $w_i$ designates the weights of the training data points. From (9), the sum ranges over the '$n$' data points in the training classifier set and $y_i$ denotes a dependent variable (i.e. output) whose value is determined by observation. The kernalized predicted classification results is described as follows,

$$y' = \begin{cases} +1 & normal\ data\ point \\ -1 & abnormal\ data\ point \end{cases} \quad (10)$$

After performing the classification, the Least squares minimize the sum of the difference between an actual value, and the predicted value provided by a model which is called error. The Error rate is measured as follows,

$e_r = \sum_{i=1}^{n}(a_i - p_i)^2$ (11)

Ls → arg min $e_r$ (12)

From (11) (12),$a_i$ denotes an observed value and $p_i$ denotes a predicted value (i.e. $\boldsymbol{y'}$) and $e_r$ denotes an error rate, Ls denotes a least square. The least square in the KLSSVC minimizes the error function using $arg\ min$ (i.e argument minimum function) and improve the classification results with minimum time. Similarly, the number of hyperlane is used to categorize the data points into different classes. The algorithm of KLSSVC is described as follows,

```
Input: Data points $\mathbf{D_1, D_2, D_2, \dots . D_n}$
Output: Improve classification accuracy
Begin
    for each $D_i$
        Construct optimal hyperplane $H$
        Find marginal hyperplane $m_1, m_2$
        Find distance between $m_1$ and $m_2$
        The output of classifier is $y' = sign \sum w\, y_i K(f_i, f')$
            If ($y' = +1$) then
                Data point is classified as 'normal'
                else
                Data point is classified as 'abnormal'
            end if
        $Ls$ minimizes $e_r$
    End for
End
```

**Algorithm 2 Kernelized Least Square Support Vector Classifier**

The above algorithm clearly defines the classification of data points with the selected features. For each data points, the hyperplane is constructed to classify the data point as normal or abnormal. Here, the hyperplane act as a boundary and it verifies the training data points with the testing results. Based on the results, the KLSSVC classifies the data points based on the similarity and it uses the least square function to minimize the error. As a result, data points are classified effectively with minimum time.

Based on the above said two processes, experimental evaluation is carried out to show the performance of proposed PCS-KLSSVC.

## 4. EXPERIMENTAL EVALUATION

Experimental evaluations of proposed PCS-KLSSVC and existing methods namely support vector machines discriminant analysis (SVM-DA) [1] and hybrid ensemble approach [2] are implemented using Java language. The main objective of the data set is to identify the seed disease through feature selection and classification in the agriculture field. The dataset includes 35 categorical attributes such as date, plant-stand, precip, temp,hail, crop-hist, area-damaged, severity, seed-tmt, germination, plant-growth, leaves, leafspots-halo, 14. leafspots-marg, leafspot-size, leaf-shread, leaf-malf, leaf-mild, stem, lodging, stem-cankers, canker-lesion, fruiting-bodies, external decay, mycelium, int-discolor, sclerotia, fruit-pods, fruit spots, seed, mold-growth, seed-discolor, seed-size, shriveling and roots. Based on these features, 19 different classes are obtained such as 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 and herbicide-injury. This dataset have 307 instances. Among the 35 attributes, less number of attributes is selected for diagnosing the seed disease at an earlier stage through the classification process. The experimental is carried out with three different methods namely, proposed PCS-KLSSVC and existing methods namely support vector machines discriminant analysis (SVM-DA) [1] and hybrid ensemble approach [2].

## 5. RESULTS AND DISCUSSION

The results and discussion of the proposed PCS-KLSSVC and existing SVM-DA [1] and hybrid ensemble approach [2] are described in this section with the different parameters such as feature selection rate, classification accuracy, classification time and space complexity. Performance is evaluated based on these metrics with the help of tables and graph values.

### 5.1 Performance results of feature selection rate

Feature selection rate is defined as the ratio of number of features selected to the total number of features in the dataset. The formula for calculating the feature selection rate is measured as follows,

$$FSR = \frac{n-Number\ of\ features\ selected}{n} * 100 \quad (13)$$

From (13), where $FSR$ denotes a feature selection rate and 'n' denotes a number of features. Feature selection rate is measured in terms of percentage (%).

**Sample mathematical calculation for feature selection rate**

**PCS-KLSSVC**: Total number of feature is 5 and the number of features is correctly selected is 2. Then the FSR $= \frac{5-2}{5} * 100 = 60\%$

**SVM-DA**: Total number of feature is 5 and the number of features is selected are 3. FSR $= \frac{5-3}{5} * 100 = 40\%$

**Hybrid ensemble approach:** Total number of feature is 5 and the number of features is selected are 3. FSR $= \frac{5-3}{5} * 100 = 40\%$

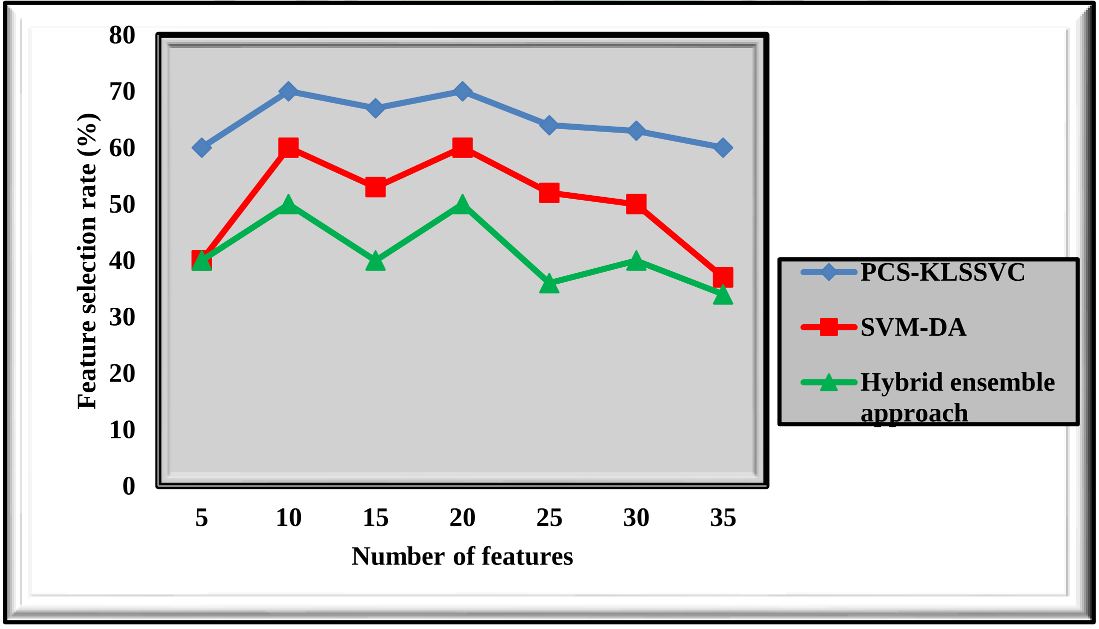

**Figure 4 performance results of feature selection rate**

Figure 4 depicts the performance results of feature selection rate with respect to a number of features taken from the dataset. The number of features is taken for the experimental evaluation is varied from 5 to 35. Totally 7 different runs are carried out to show the performance of the proposed and existing methods. For each runs different number of features is taken as input to measure the feature selection rate. The feature selection is a process of selecting the less number of features from the dataset. The Soybean (Large) Dataset includes the 35 features for diagnosing the soybean disease. From the selected features, less number of features is selected for disease classification. In figure 3, three different colors indicate the feature selection rate of three methods namely proposed PCS-KLSSVC and two existing methods. The above figure clearly shows that the feature selection rate is significantly improved using proposed PCS-KLSSVC when compared to existing SVM-DA [1] and hybrid ensemble approach [2]. This is because, the PCS-KLSSVC uses the Pearson chi-squared hypothesis test. It is used for testing the dependency and independence between the features based on score value. If the two soybeans features are dependent, the one feature is selected. Otherwise two features are selected for classification. The Pearson chi-squared hypothesis test score value selects the features and removes the redundant features from the dataset. After performing the 7 runs, the comparison of the proposed and existing classifier is performed. The average comparison results show that the feature selection rate of the PCS-KLSSVC is improved by 32% and 58% when compared to existing SVM-DA [1] and hybrid ensemble approach [2] respectively.

### 5.2 Performance results of classification accuracy

Classification accuracy is measured as the ratio of a number of data points are correctly classified to the total number of data points. The mathematical formula for classification accuracy is expressed as,

$$CA = \frac{Number\ of\ data\ points\ correctly\ classified}{N} * 100 \quad (14)$$

From (14) where 'CA' denotes a classification accuracy and 'N' denotes a number of data points. Classification accuracy is measured in terms of percentage (%).

**Sample mathematical calculation for classification accuracy**

**PCS-KLSSVC**: Number of data points is correctly classified is 25 and the total number of data points (N) is 30. Then the CA = $\frac{25}{30} * 100 = 83\%$

**SVM-DA:** Number of data points is correctly classified is 19 and the total number of features (n) is 30. Then the CA $= \frac{19}{30} * 100 = 63\%$

**Hybrid ensemble approach**: Number of data points is correctly classified are 23 and the total number of data points (N) is 30. Then the CA $= \frac{23}{30} * 100 = 77\%$

**Table 1 Tabulation for classification accuracy**

| Number of data points | Classification accuracy (%) | | |
|---|---|---|---|
| | PCS-KLSSVC | SVM-DA | Hybrid ensemble approach |
| 30 | 83 | 63 | 77 |
| 60 | 87 | 67 | 72 |
| 90 | 86 | 50 | 69 |
| 120 | 92 | 61 | 77 |
| 150 | 91 | 55 | 67 |
| 180 | 88 | 68 | 73 |
| 210 | 86 | 63 | 75 |
| 240 | 89 | 65 | 70 |
| 270 | 96 | 78 | 86 |
| 300 | 97 | 85 | 94 |

Table 1 describes a classification accuracy using Soybean (Large) Dataset with different data points. By using Soybean (Large) Dataset, soybean seed disease is classified into nineteen different classes. The seeds are vulnerable to different diseases caused by bacteria, viruses and so on. These diseases are diagnosed using kernelized least square support vector classifier (KLSSVC). For each soybean, the hyperplane is constructed to classify the data point in different classes. In KLSSVC, the hyperplane is a boundary and it finds the similarities of the data points. Based on the similarities, the hyperplane is employed to identify the different disease by comparing the training data points with the testing results. Based on the results, the KLSSVC effectively diagnosis the soybean disease. But in case of existing support vector classifier failed to find the similarities between the data points resulting classification error may occur. On the contrary, the PCS-KLSSVC uses the least square to minimize the sum of the square difference between the actual and predicted value using argument minimum function. This helps to further improve the classification accuracy.

Experimental evaluation is carried out with 10 different runs to calculate the classification accuracy. After performing the ten runs, the classification accuracy of the proposed classifier is compared to the existing classifier. Finally, the comparison results show that the classification accuracy of PCS-KLSSVC is increased by 39%, and 19% when compared to existing SVM-DA [1] and hybrid ensemble approach [2] respectively.

### 5.3 Performance results of classification time

Classification time is measured as an amount of time taken for classifying the disease with different data points. The classification time is measured as follows,

$$CT = N * time\ (classifying\ the\ data)\ (15)$$

From (15),$CT$ denotes a classification time which is measured in terms of milliseconds (ms) and $N$ denotes a data points.

**Sample mathematical calculation for classification time**

**PCS-KLSSVC**: Number of data points is 30 and the time for calculating one data point is 0.8ms, then the $CT = 30 * 0.8ms = 24ms$

**SVM-DA:** Number of data points is 30 and the time for calculating one data point is 1.1ms, then the $CT = 30 * 1.1ms = 33ms$

**Hybrid ensemble approach:** Number of data points is 30 and the time for calculating one data point is 1.3ms, then the $CT = 30 * 1.3ms = 39ms$

Figure 5 shows the classification time based on the number of data points. The numbers of data points 30 to 300 are taken as an input for measuring the classification time. The Soybean (Large) Dataset is used for experimental evaluation to classify the data points as abnormal (i.e. disease). The above graphical results clearly illustrate that the proposed PCS-KLSSVC minimizes the classification time when compared to existing classifier. This significant improvement is achieved by applying a feature selection process. The proposed PCS-KLSSVC uses Pearson chi-squared hypothesis test to select the features to form the dataset to minimize the classification time.

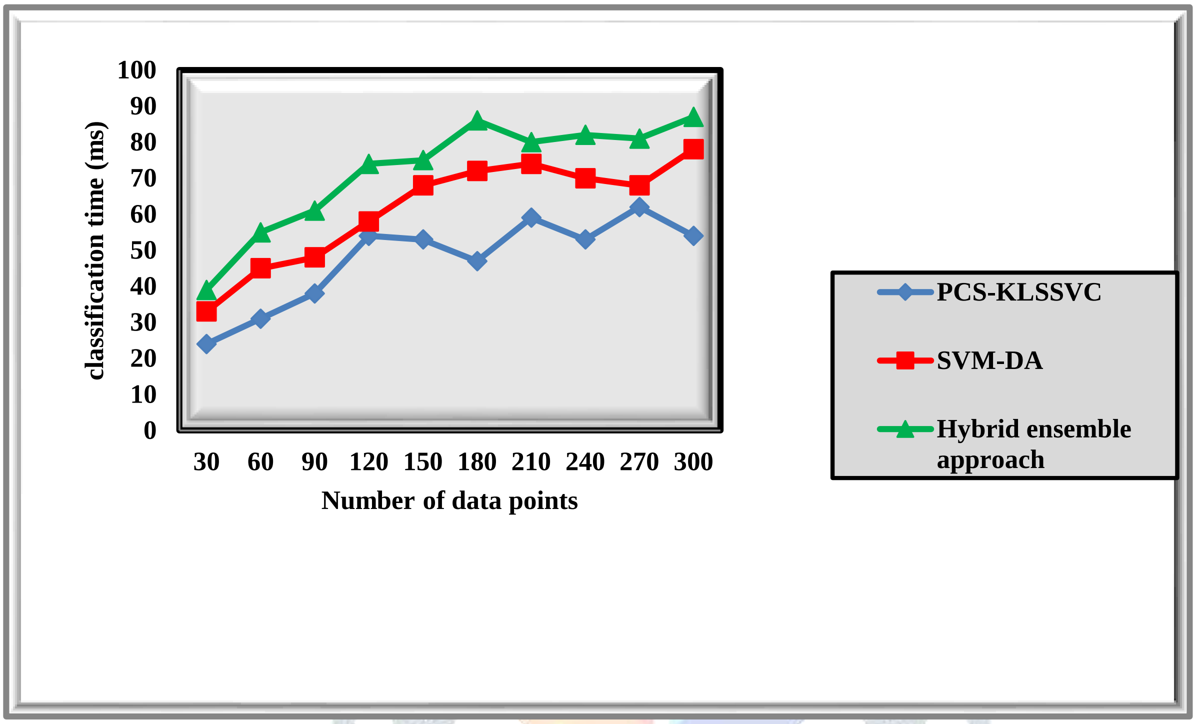

**Figure 5 Performance results of classification time**

The dataset includes 35 features for soybean disease diagnosis. Among the several features, the less number of features are selected for disease diagnosis with minimum time. After that, the classification is performed using a kernelized least square support vector to identify disease affected data points through the similarity between them. An optimal hyperplane identifies the disease affected seeds and it classifies into a particular class. This helps to minimize the classification time as well as minimizes the disease diagnosis time. As a result, this similarity measure improves the classification accuracy and minimizes the time. The comparisons of proposed and existing methods show that the PCS-KLSSVC technique minimizes the classification time by 23% and 34% than the existing SVM-DA [1] and hybrid ensemble approach [2] respectively.

### 5.4 Performance results of space complexity

The space complexity is defined as an amount of storage space required for an algorithm stores the data points. The mathematical formula for space complexity is expressed as follows,

$SC = N * space\ (Storing\ one\ data\ point)$ (16)

From (16), Where $SC$ denotes a space complexity and 'N' denotes a number of data points. The space complexity is measured in the unit of Kilobytes (KB).

**Sample mathematical calculation for space complexity**

**Proposed technique**: Number of data points is 30 and the space for calculating one data point 0.4KB is, then the $SC = 30 * 0.4KB = 12KB$

**Existing technique 1**: Number of data points is 30 and the space for calculating one data point is 0.5KB, then the $SC = 30 * 0.5KB = 15KB$

**Existing technique 2:** Number of data points is 30 and the space for calculating one data point is 0.8, then the $SC = 30 * 0.8 = 24KB$

**Table 2 Tabulation for Space complexity**

| Number of data points | Space complexity (KB) | | |
|---|---|---|---|
| | **PCS-KLSSVC** | **SVM-DA** | **Hybrid ensemble approach** |
| **30** | 12 | 15 | 24 |
| **60** | 18 | 24 | 42 |
| **90** | 23 | 41 | 48 |
| **120** | 36 | 46 | 49 |
| **150** | 35 | 50 | 54 |
| **180** | 38 | 45 | 59 |
| **210** | 42 | 46 | 61 |
| **240** | 36 | 48 | 55 |
| **270** | 38 | 49 | 57 |
| **300** | 39 | 45 | 60 |

The performance results of the space complexity versus a number of data points are clearly described in table 2. The experimental results of three different methods namely PCS-KLSSVC, SVM-DA [1] and hybrid ensemble approach [2]. The above table clearly shows that the space complexity using PCS-KLSSVC is reduced when compared to existing classifier. The PCS-KLSSVC performs two processes namely feature selection and classification for identifying the soybean disease. The seed features such as seed-size, seed discolor, shriveling and so on. This kind of similar features is selected to classify the seed disease. The less number of features are selected to minimize the space complexity for storing the different data points. In addition, the improved support vector classifier categories the seed disease into different classes with the selected features. In order to handle a large number of data, the existing classification algorithm does not minimize the space complexity at a required level. This problem is minimized by introducing a novel classifier to minimize the space complexity. The space complexity of the proposed PCS-KLSSVC and existing classifier are compared. After the comparison, the space complexity of the PCS-KLSSVC is considerably minimized by 23% and 39% than the existing SVM-DA [1] and hybrid ensemble approach [2] respectively.

The above results and discussions clearly describe the proposed PCS-KLSSVC efficiently improve the seed disease classification accuracy with minimum time and space complexity in the agriculture field.

## 6. CONCLUSION

An efficient machine learning technique namely Pearson Chi-Square Based Kernel Least Square Support Vector Classifier (PCS-KLSSVC) is introduced to improve the classification accuracy with minimum time. The PCS-KLSSVC performs Pearson Chi-squared hypothesis test to find the dependent and independent features based on chi-square score. The feature which is more relevant is selected for classification and removes the other features from the dataset. Secondly, the classifications of different data points are done with the selected features. The kernelized least square support vector classifies the data points into various classes with minimum time. Experimental evaluation of proposed PCS-KLSSVC and existing methods are carried out using Soybean (Large) Dataset. The proposed PCS-KLSSVC is applied for diagnosing the soybean diseases. The experimental results of proposed PCS-KLSSVC improve feature selection rate, the classification accuracy and minimize the classification time as well as space complexity when compared to existing SVM-DA and hybrid ensemble approach. Hence it is concluded that PCS-KLSSVC is an efficient classifier for accurate multiclass classification problems in the agriculture field.

## REFERENCES

[1] Bruna Tassi Borille, Marcelo Caetano Alexandre Marcelo, Rafael Scorsatto Ortiz, Kristiane de Cássia Mariotti, Marco Flores Ferrao, Renata Pereira Limberger, “Near-infrared spectroscopy combined with chemometrics for growth stage classification of cannabis cultivated in a greenhouse from seized seeds”, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, Elsevier, Volume 173, 2017, Pages 318–323

[2] Archana Chaudhary, Savita Kolhe and Raj Kamal, “A hybrid ensemble for classification in multiclass datasets: An application to oilseed disease dataset”, Computers and Electronics in Agriculture, Elsevier, Volume 124, 2016, Pages 65–72

[3] T. Rumpfa, A.-K. Mahlein, U. Steiner, E.-C. Oerke, H.-W. Dehne, L. Plümer, “Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance”, Computers and Electronics in Agriculture, Elsevier, Volume 74, 2010, Pages 91–99

[4] Sheeja T.K and A.Sunny Kuriakose, “A novel feature selection method using fuzzy rough sets”, Computers in Industry, Elsevier, Volume 97, 2018, Pages 111–116

[5] Ting Xie, Pengfei Ren, Taiping Zhang, Yuan Yan Tang, “Distribution preserving learning for unsupervised feature selection”, Neurocomputing, Elsevier, Volume 289, 2018, Pages 231-240

[6] Changzhong Wang , Yali Qi , Mingwen Shao , Qinghua Hu , Degang Chen , Yuhua Qian,Yaojin Lin, “A Fitting Model for Feature Selection With Fuzzy Rough Sets”, IEEE Transactions on Fuzzy Systems, Volume 25, Issue 4, 2017, Pages 741 – 753

[7] Amit Kumar Das, Saptarsi Goswami, Amlan Chakrabarti, Basabi Chakraborty, “A new hybrid feature selection approach using feature association map for supervised and unsupervised classification”, Expert Systems with Applications, Elsevier, Volume 88, 2017, Pages 81-94

[8] Archana Chaudhary, Savita Kolhe and Raj Kamal, “An improved random forest classifier for multi-class classification”, Information Processing in Agriculture, Elsevier, Volume 3, 2016, Pages 215–222

[9] Wenbin Qiana and Wenhao Shuc, “Mutual information criterion for feature selection from incomplete data”, Neurocomputing, Elsevier, Volume 168, 2015, Pages 210-220

[10] Lin Wang, Bo Yang, Yuehui Chen, Ajith Abraham, Hongwei Sun, Zhenxiang Chen, Haiyang Wang, “Improvement of neural network classifier using floating centroids”, Knowledge and Information Systems, Volume 31, Issue 3, 2012, Pages 433–454

[11] Santanu Phadikar, Jaya Sil , Asit Kumar Das, “Rice diseases classification using feature selection and rule generation techniques”, Computers and Electronics in Agriculture, Elsevier, Volume 90, 2013, Pages 76–85

[12] Aida Brankovic , Alessandro Falsone , Maria Prandini , Luigi Piroddi, “A Feature Selection and Classification Algorithm Based on Randomized Extraction of Model Populations”, IEEE Transactions on Cybernetics, Volume 48, Issue 4, 2018, Pages 1151 - 1162

[13] L.O.L.A. Silva, M.L. Koga, C.E. Cugnasca, A.H.R. Costa, “Comparative assessment of feature selection and classification techniques for visual inspection of pot plant seedlings”, Computers and Electronics in Agriculture, Elsevier, Volume 97, 2013, Pages 47–55

[14] Narges Armanfard , James P. Reilly , Majid Komeili, “Logistic Localized Modeling of the Sample Space for Feature Selection and Classification”, IEEE Transactions on Neural Networks and Learning Systems , Volume 29, Issue 5, 2018, Pages 1396 – 1413

[15] Jun Huang, Guorong Li , Qingming Huang , Xindong Wu, “Joint Feature Selection and Classification for Multilabel Learning”, IEEE Transactions on Cybernetics ,Volume 48, Issue 3, 2018 , Pages 876 – 889

[16] Jun Pang , Yu Gu , Jia Xu , Ge Yu, “Semi-supervised multi-graph classification using optimal feature selection and extreme learning machine”, Neurocomputing, Elsevier, Volume 277, 2018, Pages 89-100

[17] ShaohuaWu, Yong Hu, WeiWang, Xinyong Feng, and Wanneng Shu, “Application of Global Optimization Methods for Feature Selection and Machine Learning”, Mathematical Problems in Engineering, Hindawi Publishing Corporation, Volume 2013, October 2013, Pages 1-8

[18] Xiao-Ying Liu , Yong Liang, Sai Wang , Zi-Yi Yang, Han-Shuo Ye, “A Hybrid Genetic Algorithm With Wrapper-Embedded Approaches for Feature Selection”, IEEE Access ,Volume 6, Pages 22863 – 22874

[19] Liang Hu, Wanfu Gao, Kuo Zhao, Ping Zhang, FengWang, “Feature selection considering two types of feature relevancy and feature interdependency”, Expert Systems with Applications, Elsevier, Volume 93, 2018, Pages 423-434

[20] Wanfu Gao, Liang Hu, Ping Zhang, “Class-specific mutual information variation for feature selection”, Pattern Recognition, Elsevier, Volume 79, 2018, Pages 328-339