Chapter – 6

A Hybrid Model to Enhance the Performance of a Classifier

6.1 Introduction

Data mining, a synonym to knowledge discovery in databases - is a process of analyzing data from different perspectives and summarizing it into useful information. Data mining tasks are generally classified according to the types of data, knowledge, and techniques used. People do not make good or bad decisions; they make decisions on information received. The right information, in the right format at the right time is important. Data can be transformed into good information by efficient methods of processing, filtering, modeling, evaluating, analyzing, and visualizing data. Data mining is applicable in several aspects and can be used to forecast the future.

The complexity and volume of data are increasing day by day; the existing data mining techniques are facing a lot of challenges particularly for classifying large scale multi-class data. Therefore, the combination of clustering and classification becomes an active research area to deal with data in large volumes with high dimensions. One such solution to the problem is the integration of clustering and classification techniques to enhance classification accuracy and scalability. Taking advantage of both methods of clustering and classification techniques is a significant research problem. As, it has been noticed that the integrated model of clustering and classification gives promising results, compared to classification alone. The goal is to find a new classification method that integrates the advantages of clustering and classification.

The integrated model built could reflect the best prediction of class Supervised & unsupervised methods of learning in combination together. Input

features are compared and clustered based on their similarities, prior to the classification task the process of clustering is performed. Therefore, the classifier can be grouped with similar records of input. Each partitioned dataset is considered for performing the classification task. The development of accuracy, the pace of the integrated model & comprehensibility is the result of two combined learned results. The prediction and generalization capacities have increased with the combined methods, as it provided a flexible and strong mapping for a dataset. The integrated unsupervised & supervised learning methods were used to find the significant attributes of target variables and factors.

6.2 Classification

It is a two-stage method where at the first stage we built/trained models from historical training data sets with labeled class attributes. In the second stage, we try to predict the class labels of new test datasets as accurately as possible. Fundamentally, it is a mapping from target function to attribute set of already labeled class. Classification is supervised learning. It tries to assign previously unseen records to a class as accurately as possible. The classification task is to maximize the accuracy rate that is the ratio of the number of correct predictions to the total number of predictions in the test dataset. A confusion matrix is derived to evaluate the performance of the classifiers based on performance evaluating parameters (Accuracy, Precision, Recall, F-measure, AUC, and G-mean).

6.2.1 SVM

The support vector machine is applicable in linear and nonlinear data.SVM is an extremely slow method that takes too much time in learning but its results are highly accurate and are capable to handle complex non-linear data as well. It is less prone to overfitting. A Support Vector Machine uses labeled training data and forms an optimal hyper-plane that can classify testing sets.

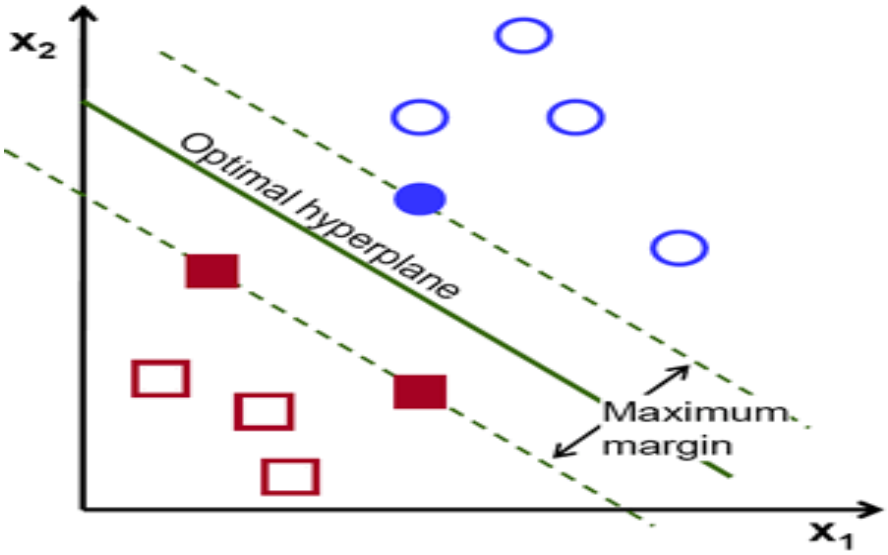

Figure 6.1 Support Vector Machine

Source: (Burges, 1998)

6.3 Clustering

Clustering is the most efficient technique which can be applied to extract useful information from the raw data. The clustering is the technique in which similar and dissimilar types of data can be clustered to analyze useful information from the dataset. Clustering divides the data into groups of similar objects. Each group is called a cluster. Each cluster consists of objects that are similar between themselves and dissimilar to objects of other clusters (Hoppner et al., 2000). Clustering is unsupervised learning since clusters are not predefined & class labels not available. An unsupervised learning technique used to cluster the data into groups based on similarity in two different clusters.

Definition: It is used to organize information without any prior knowledge about the distribution of data. A clustering of D dataset is a partition of D into K clusters C_1, C_2, C_3……C_k where i=1,2,3……..k.

6.3.1 K-Means

K-means method of data clustering is one of the oldest and yet very popular among the Data Miners community. One reason for the popularity is it is data-driven so the fewer assumption is required. It uses greedy search strategy so it can divide large datasets into segments based on the number of clusters

supplied as a popular centroid-based algorithm is K-means; the input parameter is taken as K, and then partitions the whole dataset into k clusters resulting into a high intra-cluster similarity and low inter-cluster similarity. The mean value of each cluster is calculated by its all objects. The cluster similarity is determined through the cluster`s centroid. In the k-means algorithm initially, we need to input k objects, each of the objects is considered as the center of cluster or cluster mean. One of the objects is assigned to the cluster according to the similarity distance between the cluster mean and object. For each cluster iterating computation of the new mean occurs till the convergence of the centroid function. Full convergence of clusters uses the square-error criterion; generally, It aims to minimize squared error given by

$$\text{Sum of Squared Error}=\sum_{i=1}^{k}\sum_{j=1}^{ik}\text{dist}(x_{i,}y_{j})$$

The distance can be calculated by following distance metric between xi and yj is given by

$$\text{Euclidean distance}=\sqrt{\sum_{k=1}^{n}(x_{ik},y_{ik})^{2}})$$

6.4 Models for Data mining Tasks

There are various techniques available for accomplishing data mining tasks, but different types of data need different techniques hence, only a single technique cannot be used for proper and thorough data mining studies, as every technique has its specific application. The most widely used data mining techniques are Classification and Clustering.

The three models to accomplish the data-mining task are the Single model where the problem is being solved by applying a single technique that is classification, clustering, or association. Hybrid methods where Heterogeneous Machine Learning approaches like, clustering and classification or clustering are combined in Hybrid learning and the third is Ensemble methods where Homogenous Machine learning approaches are combined in Ensemble learning, using various merging methods.

Creating predictive (classification) models is one of the machine learning applications to uncover novel, interesting, and useful knowledge from large volumes of data in many medical domains such as diagnosis, prognosis, and treatment. They are successfully developed by applying several machine learning techniques. In classification, the goal is the prediction of the target class of every case in the data with predefined classes. A vast amount of data is not possible to process or analyze by one classifier. The integration of the clustering & classification model provides good results than classification alone. Recent researches have shown that if two or more techniques are hybridized or ensembled they can overcome the problems faced by single classifiers. Clustering And Classification in combination is an active research area to deal with huge data with high dimensions.

Ensemble learning generally used to improve the performance of a classifier. It is a process by which predictions of multiple classifiers are combined. The new samples can be classified with better prediction accuracy (Karegowda et al. ,2012). Commonly four types of ensemble techniques are used they are bagging, boosting, stacking, and voting. Recent researches have mentioned some problems in ensemble learning listed below (Kyriakopoulou, 2008).

- The error of each base classifier tends to increase the overall error of the ensemble classifier.
- When many classifiers are used they produce more complex output which is difficult to analyze.
- It is stated that the classification accuracy is not improved in all cases of the ensemble.

Developing ensemble classifiers are a challenging task. Simple base classifiers must be used for ensemble and it should not over-fit. The base learners should be accurate and distinct, Sometimes poor accuracy is observed due to difficulty in the selection of the correct combination of classifiers.

Due to the aforementioned challenges and problems present in selecting classifiers, the technique of Hybridization is a recent development, where two or more heterogeneous techniques are combined. e.g: clustering and classification or clustering and association. A previous research literature study has not focused on hybridization i.e., a merger of more than one technique, and none of the research used hybrid features for selection. Hence, a hybrid model with a combination of clustering and classification is proposed with hybrid features.

The performance of the proposed hybrid model is compared with the literature work and results are displayed and compared.

6.4.1 WEKA

"The Waikato Environment for Knowledge Analysis (WEKA) is a comprehensive suite of Java class libraries that implement many state-of-the-art machine learning and data mining algorithms". The algorithms can be called from your own Java code or directly can be applied from a dataset. The tools in WEKA contain association rules, data preprocessing, regression, classification, visualization & clustering. It is a good tool for the development of new machine languages. Here we have used WEKA 3.8 to implement our algorithm with java class libraries.

Five important features of WEKA are:

1) **Open Source**: It is open-source software under the GNU/GPL. It has dual-licensed; Pentaho Corporation owns the exclusive license to use the platform in their product for business intelligence.
2) **Graphical Interface**: It has a Graphical User Interface. This allows you to work on machine learning projects without programming.
3) **Command Line Interface**: All features of this tool can be used from the command line. This is very useful for scripting large projects.
4) **Java API**: It is written in Java language and provides API, That allows you to integrate WEKA API into your applications.
5) **Documentation**: There are rich documentation is available that can train to use the platform effectively.

6.5 Datasets Description

Thirteen benchmark datasets are used for the study viz. Heart Disease, Wine quality white, Parkinson dataset, Colon Cancer, Breast Cancer, Image Segmentation, Cleveland-0 Dataset, Ionosphere Dataset, Squash Harvest Stored, Bank Dataset, Glass Dataset, Pima and Haberman Dataset. Few datasets belong to Multiclass while others are Binary class. Information about several features present, several instances, and number of classes with their names are given in table-5.1.

Table 6.1 Dataset Description

S. No	**Dataset Name**	**#Features**	**#Instances**	**Class**
1	Heart Disease	14	303	2{Yes, No}
2	Wine quality white	12	1482	2{Negative, Positive}
3	Parkinson dataset	756	757	2{Yes, No}
4	Colon Cancer	2002	62	2{Normal, Abnormal}
5	Breast Cancer	32	569	2{Malignant, Benign}
6	Image Segmentation	19	210	7{Brickface, Sky, Foliage, Cement, Window, Path, Grass}
7	Cleveland-0 Dataset	13	173	2{Negative, Positive}
8	Ionosphere Dataset	35	351	2{Good, Bad}
9	Squash Harvest Stored	25	53	3{Excellent, Ok, Not acceptable}
10	Bank Dataset	17	4522	2{Yes, No}
11	Glass Dataset	10	215	2{Positive, Negative}
12	Pima	9	769	2{Positive, Negative}
13	Haberman Dataset	4	307	2{Positive, Negative}

6.6 Methodology

6.6.1 Data Preparation

In the first phase, the datasets are divided into a training set and a testing set. The output classes are also removed in this phase only from training as well as testing datasets. Removing output class labels is important to make the clusters unbiased of class attributes. after this phase, the dataset is ready for further processing.

6.6.2 Clusters Building

Prior to the classification, the dataset is grouped into several numbers of clusters through the K- means clustering algorithm. Four values for K are considered for the study K (K=2, 3, 4, and 5). When k= 2, two clusters are formed from the training set and two clusters from the testing set. When k=3, three clusters are formed from training and three from testing, and when k=4, four clusters form training, and four from testing and same for k=5 after this whole process fourteen clusters for training and fourteen for testing is formed.

Testing= $\{I_1^{k^2}, I_2^{k^2}, I_1^{k^3}, I_2^{k^3}, I_3^{k^3}, I_1^{k^4}, I_2^{k^4}, I_3^{k^4}, I_4^{k^4}, I_1^{k^5}, I_2^{k^5}, I_3^{k^5}, I_{4,}^{k^5} I_5^{k^5}\}$

Training= $\{I_1^{k^2}, I_2^{k^2}, I_1^{k^3}, I_2^{k^3}, I_3^{k^3}, I_1^{k^4}, I_2^{k^4}, I_3^{k^4}, I_4^{k^4}, I_1^{k^5}, I_2^{k^5}, I_3^{k^5}, I_{4,}^{k^5} I_5^{k^5}\}$

These clusters do not contain output classes. Corresponding output classes and cluster identity are added to each cluster. After this process, the Classification task can be applied. This process is being performed to identify which value of K is most appropriate with which classification algorithm on a given dataset.

6.2.3 Building the classification Models:

The resultant dataset augmented with one more feature originated from the clustering step named cluster-no as an input to the classifier. The expanded data is applied to SVM (Support Vector Machine) classifier. This model will be evaluated on several parameters. The results evidently revealed that this integration process generates a more precise and accurate model. All classifier's accuracy gets affected positively when supervised and unsupervised learning methods are combined.

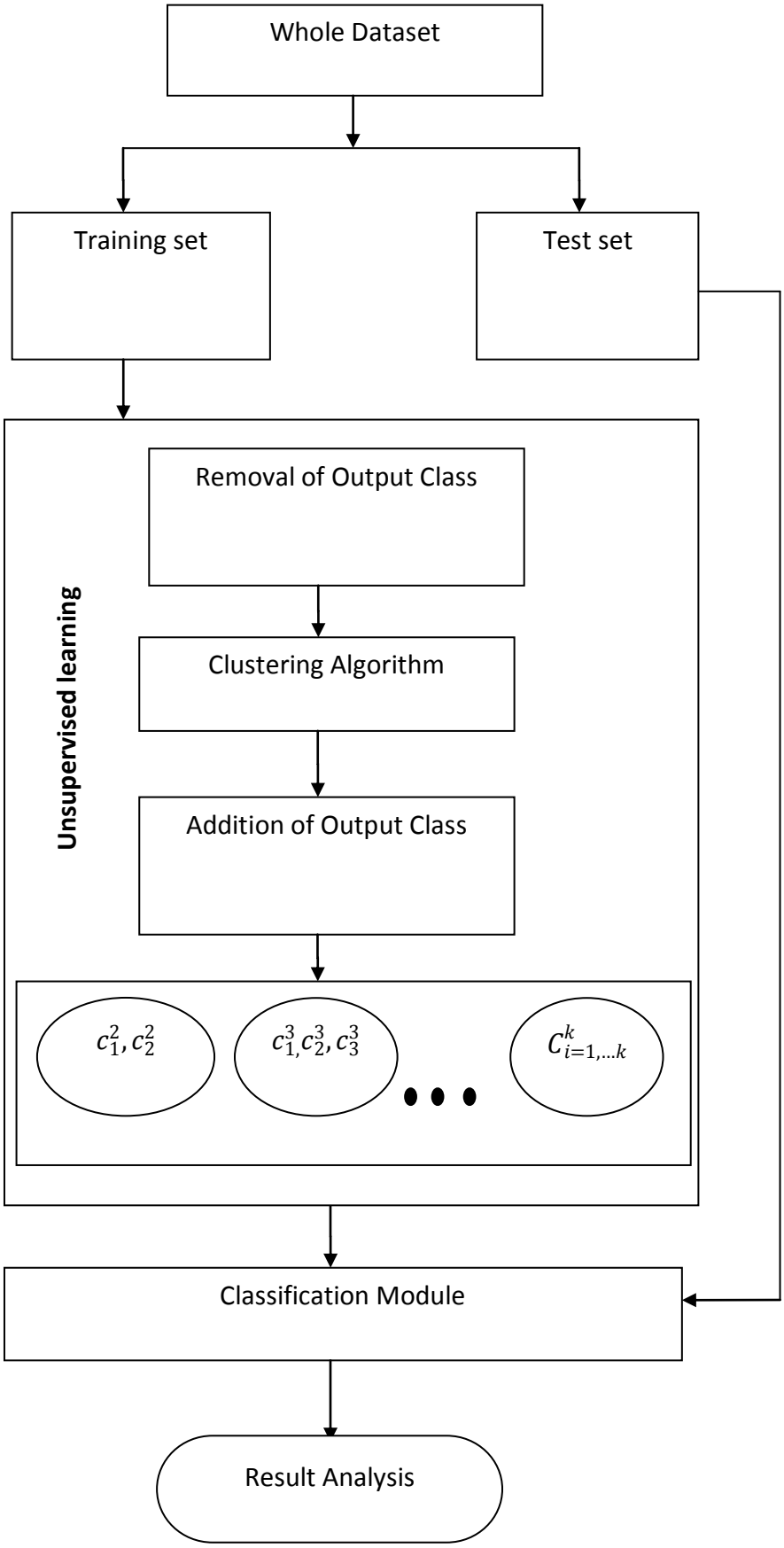

Figure 6.2 Model's Control Flow diagram

6.3 Algorithm

Step 1: Preprocessing the dataset by deleting all missing values instances

D= $\{I^{\text{instance}}, O^{\text{instance}}\}$

Here I stand for Input Instance and O stands for output instances

Step 2: Divide 2/3 of the dataset as training and 1/3 as testing

TrainingSubset =2/3$\{ I^{\text{instance}}, O^{\text{instance}} \}^{\text{training}}$

TestingSubset = 1/3 $\{ I^{\text{instance}}, O^{\text{instance}} \}^{testing}$

Step 3: Remove the Corresponding output class from the dataset

$\mathrm{I}_1^{\mathrm{k}^2}$TrainingSubset = $\{ I^{\text{instance}} \}^{\text{training}}$

Step 4: Apply k-means clustering algorithm for k=2, 3,4 cluster

TrainingSubsets = $\{\mathrm{I}_1^{\mathrm{k}^2}, \mathrm{I}_2^{\mathrm{k}^2}, \mathrm{I}_1^{\mathrm{k}^3}, \mathrm{I}_2^{\mathrm{k}^3}, \mathrm{I}_3^{\mathrm{k}^3}, \mathrm{I}_1^{\mathrm{k}^4}, \mathrm{I}_2^{\mathrm{k}^4}, \mathrm{I}_3^{\mathrm{k}^4}, \mathrm{I}_4^{\mathrm{k}^4}\}$

TestingSubsets = $\{\mathrm{I}_1^{\mathrm{k}^2}, \mathrm{I}_2^{\mathrm{k}^2}, \mathrm{I}_1^{\mathrm{k}^3}, \mathrm{I}_2^{\mathrm{k}^3}, \mathrm{I}_3^{\mathrm{k}^3}, \mathrm{I}_1^{\mathrm{k}^4}, \mathrm{I}_2^{\mathrm{k}^4}, \mathrm{I}_3^{\mathrm{k}^4}, \mathrm{I}_4^{\mathrm{k}^4}\}$

Step 5: Add target classes for each cluster

TrainingSet 1 = $\{ \{\mathrm{C}_1^{\mathrm{k}^2} + \mathrm{O}_{\text{Training}}^{\text{instance}}\}$

TrainingSet 2 = $\{ \{\mathrm{C}_2^{\mathrm{k}^2} + \mathrm{O}_{\text{Training}}^{\text{instance}}\}$

Step 6: Train classifier by the clustered datasets for each combination of K.

Step 7: for each value of K get the corresponding constructed model name $\mathrm{M}_{\mathrm{n}}^{\mathrm{k}^{\mathrm{i}}}$where n defines model number and k^i defines cluster number

Step 8: Integrate all cluster values classified to produce the outcome for the whole dataset.

Step 9: Create a confusion matrix for furthest analysis

Table 6.2 Results for the Hybrid model to enhance the performance of the classifier

s. no	**Dataset Name**	**Whole Dataset without Clustering**			**Clustered Data set k=2**			**Clustered Dataset k=3**			**Clustered Dataset k=4**			**Clustered Dataset k=5**		
		F-M	**RO**	**Acc**	**F-M**	**RO**	**Acc**	**F-M**	**RO**	**Acc**	**F-M**	**RO**	**Acc**	**F-M**	RO	Acc.
1	Heart Disease	.85	.82	83.4	.99	.99	99.6	.97	.98	98.6	.97	.99	98.0	.98	.98	98.3
2	Wine-quality	.99	.98	98.3	.99	.50	98.3	.99	.50	98.3	.95	.98	96.6	.99	.50	98.3
3	Parkinson	.90	.78	85.7	.90	.78	85.4	.91	.80	87.5	.91	.80	86.7	.91	.80	87.0
4	Colon	.88	.83	85.4	.88	.83	85.4	.88	.83	85.4	.88	.83	85.4	.85	.78	80.6
5	Breast cancer	.97	.97	97.8	.99	.99	99.8	.99	.99	99.8	.99	.99	99.8	.99	.99	99.8
6	Image	.80	.96	88.5	.96	1.0	95.2	.96	.99	94.7	.91	.97	94.2	.95	.99	96.6
7	Cleland	.97	.80	95.9	.98	.84	97.1	.98	.84	97.1	.98	.84	97.1	.98	.84	97.1
8	Ionosphere	.82	.85	88.6	1	1	100	1	1	100	1	1	100	1	1	100
9	Squash	.76	.83	71.1	.88	.88	80.7	.95	.95	90.3	.92	.93	86.5	.90	.92	88.4
10	Bank data	.94	.57	89.2	.94	.57	89.2	.94	.57	89.2	.94	.57	89.2	.95	.66	90.9
11	Glass	.77	.49	63.5	.76	.48	62.1	1	1	100	1	1	100	.99	.99	99.5
12	pima	.62	.72	77.3	1	1	100	1	1	100	1	1	100	1	1	100
13	Haberman	.84	.50	73.5	.84	.49	73.2	.84	.49	73.2	.84	.50	73.5	1	1	100

6.4 Performance analysis of the generated model

Table 6.2 presents the combined results to show the response of the proposed Hybrid model on the SVM classifier with all 13 datasets. Pictorial representation for the comparative analysis is also presented through graph charts in figure 6.2. Accuracy is taken in the Y-axis and different values of K (K=2, K=3, K=4, and K=5) are taken in the X-axis.

For the heart disease dataset, it is displayed that the SVM classifier's performance is extremely poor (only 83.4) with the original dataset. The proposed model is observing excellent results with all cluster numbers.
and the best results are observed with K=2 on Heart Disease dataset which has two output classes Yes and No.

In the Wine quality white dataset it is depicted in the chart that the performance of SVM is approximately the same for all sets of clusters further it is decreased with K=4. No improvement of the clustering method is observed in the Wine quality white dataset, which has two output classes (Negative and Positive).

The accuracy of the Parkinson dataset with five values of cluster sets are presented in the chart. The first bar in the column chart displays the accuracy of the SVM classifier on the whole dataset without any kind of preprocessing is only 85.6. with two clusters it is further declined but it has improved with bigger cluster numbers. The best accuracy achieved is with K=3 on the Parkinson Disease dataset which has two output classes Yes and No.

For the Colon Cancer dataset, It is shown that the results of the proposed model are quite similar for all the cases except for K=5 for the colon cancer Binary Class dataset.

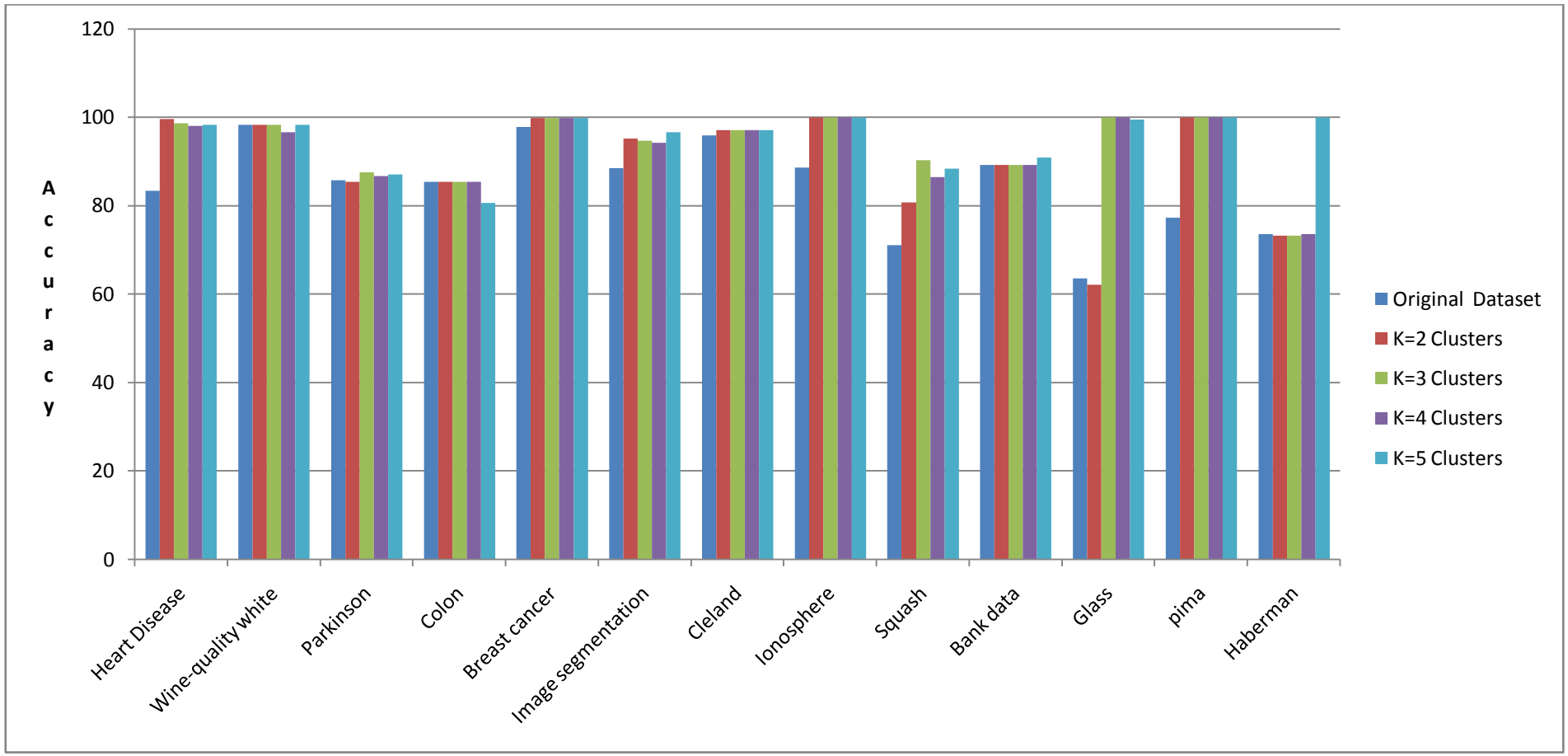

Figure 6.3 Chart for Accuracy on different cluster no for 13 Datasets

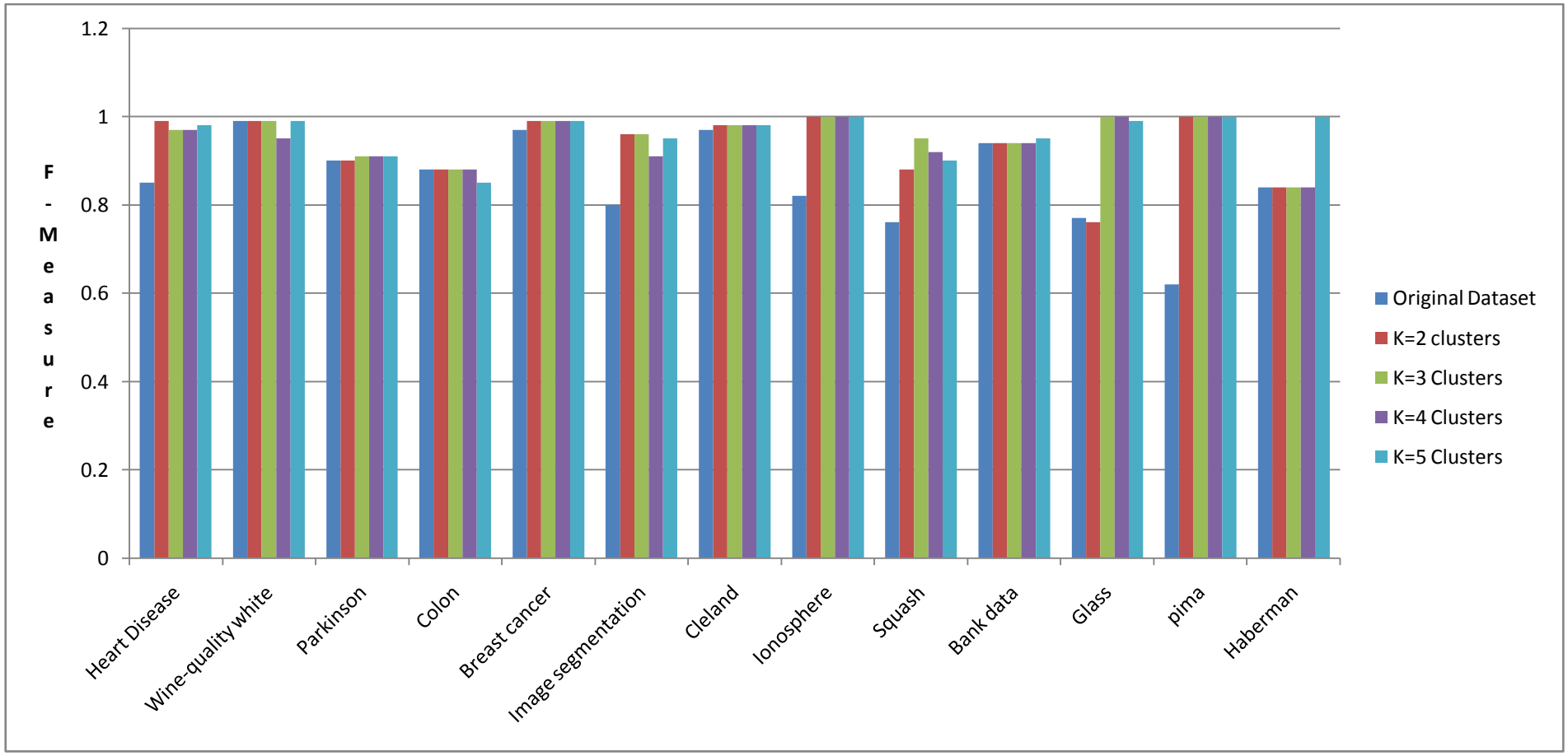

Figure 6.4 Chart for F-Measure on different cluster no for 13 Datasets

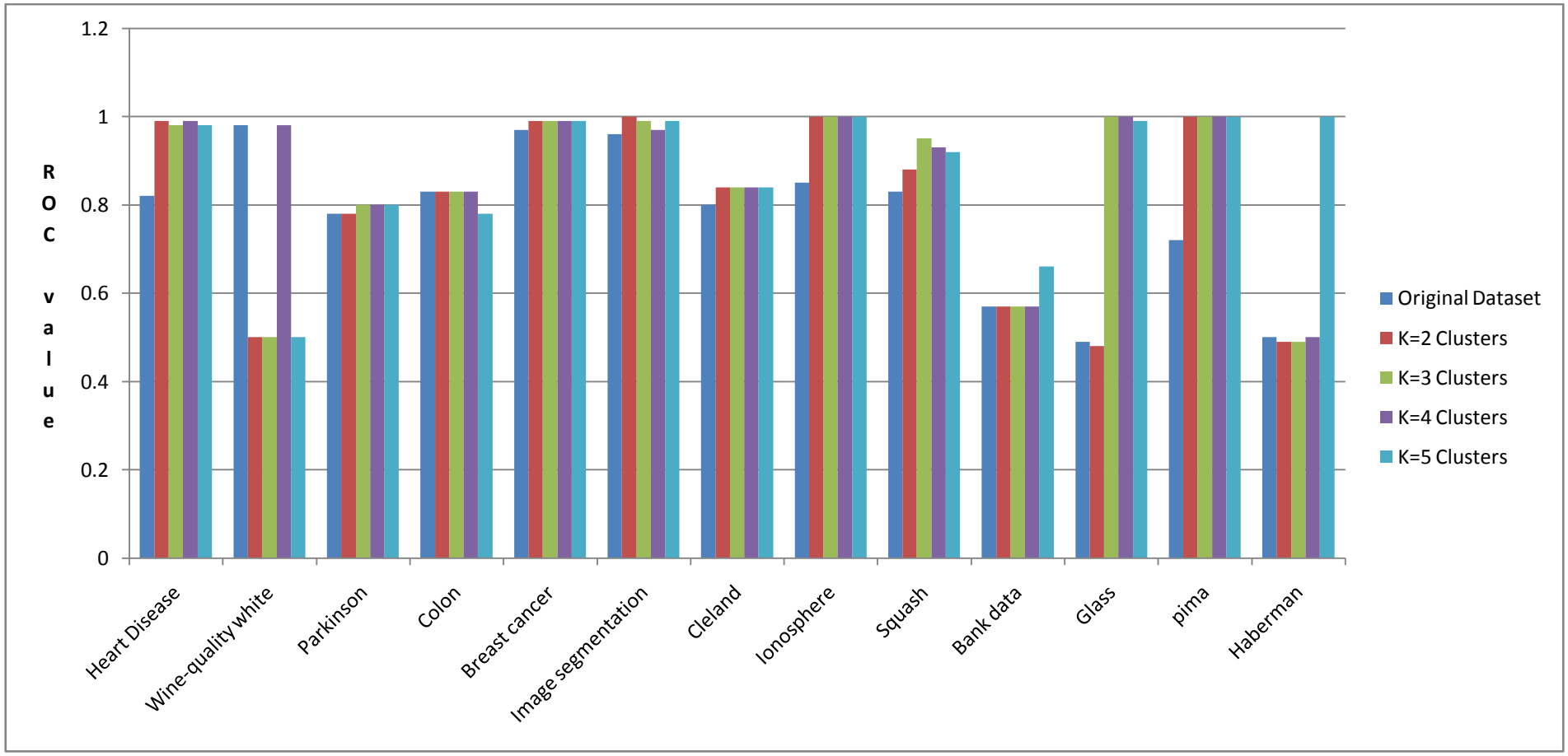

Figure 6.5 Chart for ROC value on different cluster no for 13 Datasets

Figure 6.2 is displaying slightly improved results of proposed clustering prior to the classification approach on the Breast Cancer dataset .similar kinds of performance with all five-cluster numbers are displayed on the Breast Cancer dataset which is a binary classification problem having two output classes Malignant and Benign.

Image Segmentation results are displayed in figure 6.2. It can be seen that the proposed model with cluster K=5 is giving the best accuracy with the improvement of 15% over SVM accuracy on the original dataset. Image segmentation is a multi-class classification problem. It has 7 output classes named Brickface, Sky, Foliage, Cement, Window, Path, and Grass.

The chart is also showing the accuracy of SVM classifiers on the Cleveland dataset. Which is a binary class classification problem having two output classes named Cleveland positive and Negative. The chart is showing a huge effect on the accuracy of the proposed model having the same lift of accuracy on all five-cluster numbers.

The proposed model observed 100% accuracy with all five-cluster numbers on the Ionosphere dataset. Accuracy observed with the original dataset is only 88.6 %. The Ionosphere dataset is a binary class classification problem having only two classes named Good and Bad.

With the Squash dataset, the best accuracy achieved is by K=3 and an elevated performance is reported with each case of the proposed hybrid model. The Squash dataset has three output classes named Excellent, Ok, Not acceptable.

In the chart, it can be seen the accuracy of bank data of the SVM is very poor with the original dataset but it has increased using the proposed model with k=5. The Bank data is also a binary classification problem having two Yes and No classes.

A chart for the Glass dataset which has Positive and negative classes. The performance of the proposed model is reported the same with three values of K they are 3,4 and 5.100% accuracy is achieved with the proposed model.

Figure 6.2 also contains the accuracy of the Pima India diabetic patients dataset which is classified into two classes diabetic positives and Negatives patients. The chart represents the performance of the proposed model is very good over the original dataset. The accuracy is 100% with all values of K.

A comparative sight of the accuracy of the SVM classifier on unprocessed data and the data processed using the proposed model on the Haberman dataset. The highest accuracy reported with K= 5. Haberman dataset is also a binary dataset having two classes positive and negative.

All classifier's accuracy gets affected positively when supervised and unsupervised learning methods are combined.

Figure 6.3 is the graph chart for the representation of F-measure on the original dataset and clustered dataset using the proposed model. F-measure is taken on the Y-axis and different values of K (K=2, K=3, K=4, and K=5) are taken on the X-axis. The performance of the proposed model is excellent with Heart disease dataset, Parkinson dataset, Breast cancer dataset, Image segmentation dataset, Cleland dataset, Ionosphere dataset, Squash dataset, Bank dataset, Glass dataset, Pima dataset, and Haberman datasets. No improvement in the performance of the proposed model is found with the Wine quality white dataset and Colon cancer dataset.

Figure 6.4 is the graph chart for the representation of ROC value on the original dataset and clustered dataset using the proposed model. ROC values are taken on the Y-axis and different values of K (K=2, K=3, K=4,

and K=5) are taken on the X-axis. An improved ROC value is reported with all datasets except with the Wine quality white dataset.

6.5 Summary

The research experiments observed that the hybrid model is more refined and accurate than a single individual classifier. The objective is to utilize the strength of one method to complement the weaknesses of another. If we are interested in the best possible classification accuracy, it might be difficult to find a single classifier that performs as good as a hybrid model.