International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 15 (2017) pp. 5242-5247
© Research India Publications. http://www.ripublication.com

Data Cleaning and Prototyping Using K-Means to Enhance Classification Accuracy

Lutfi Fanani [1] and Nurizal Dwi Priandani [2]

[1]*Department of Computer Science, Brawijaya University, Malang, Indonesia.*

[2]*Department of Computer Science, Brawijaya University, Malang, Indonesia.*

[1]*Orcid: 0000-0002- 3514-3525 and Orcid: 0000-0002- 0418-7373*

Abstract

This research is to propose a preprocessing method in a data classification process to improve the accuracy of the classification process. The preprocessing method to be used basically aims to clean up the data by forming prototype data generated from the clustering process. It is expected that the proposed preprocessing methods can be used in a variety of data forms and classification algorithms. The difference of this study is the preprocessing stage which consist of 2 processes, which are, the process that eliminate missing value and make data prototyping using clustering approach. After performing the preprocessing steps, the complete and evaluated or clustered data as prototype data, where the cluster will refer to certain classes would be obtained. After the data prototype was obtained, the main steps would be started which consist of the classification and evaluation process. The cycle was generate the classification accuracy results. Based on the comparison of classification accuracy without using and using preprocessing, by using the method proposed in this study could increase approximately 34% accuracy results or reduce the error 75.55%.

Keywords: Data Mining; Clustering; Classification; Data Prototyping; Data Cleaning

INTRODUCTION

Classification is one of the data mining methodologies used to predict and classify predefined data for a particular class. Classification techniques can be applied to classify applied-oriented data [1]. The selection of appropriate classification methods according to the type of application and the dimensions of the data becomes a major challenge for researchers. Appropriate methods can be selected only after analyzing all available classification methods and checking their performance in terms of accuracy.

Noise in a dataset must be dealt with in various situations. Noise may include misclassified data or information such as missing data or information. Simple human errors can also be regarded as misclassified. These errors will cause noise/ outlier data so that it can reduce accuracy of data mining system. With low accuracy, it is not possible for a built system to be used in real cases.

With the background already described, in this study we will propose an effective process/ classification path for dealing with misclassified cases on a dataset [2]. The purpose of this research is to propose a preprocessing method in a data classification process to improve the accuracy of the classification process. The preprocessing method to be used basically aims to clean up the data by forming prototype data generated from the clustering process. It is expected that the proposed preprocessing methods can be used in a variety of data forms and classification algorithms.

CLUSTERING ALGORITHMS

In this section, we give a brief description of the clustering algorithms and its variation. Clustering algorithms is one of data mining techniques that partition data into a certain number of clusters (groups, subsets, or categories). Clustering process carried out so that a pattern within a cluster or group has some similarities in terms and patterns and will be different from the other clusters [3]. The Clustering method can be used for data exploration and to generate prototypes which are then used in supervised classification [4]. Based on the shape of the label data pattern, clustering can be grouped into, labelled and unlabeled.

A. K-means Algorithm

The K-Means algorithm is the most popular and widely used clustering algorithm in the industry. The algorithm is structured on the basis of a simple idea. There was initially determined how many clusters to be formed. Any object or first element in the cluster can be selected to serve as the centroid point of the cluster. The K-Means algorithm will then repeat the following steps until there is stability [5].

1. Determine the coordinates of the midpoint of each cluster,
2. Determine the distance of each object to the coordinates of the midpoint,
3. Grouping the objects according to their minimum distance

CLASSIFICATION LEARNING

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data [6]. Classification learning algorithms are composed of two components; namely, training and prediction (classification). The training phase, using some induction algorithms, forms a model of the domain from the training examples encoding some previous experiences. The classification phase, on the other hand, using this model, tries to predict the class that a new instance (case) belongs to. [7].

A. Naïve Bayes

Naïve Bayes Classifier is a classifier method based on probability and Bayesian Theorem with the assumption that each variable X is independence. In other words, Naïve Bayes Classifier assumes that the existence of an attribute (variable) has nothing to do with the presence of another attribute (variable) [3]. Bayes rules as a classification algorithm can be written as follows [3]

$$P(C|f_1,f_2,f_3,\ldots,f_n) = \frac{P(C)P(f_1,f_2,f_3,\ldots,f_n|C)}{P(f_1,f_2,f_3,\ldots,f_n)}$$

B. K-Nearest Neighbour

Nearest Neighbor (NN) is an approach to finding cases by calculating the proximity between a new case and an old case, based on matching the weight of a number of features. One of the variant algorithms of NN is k-NN. The k-NN algorithm is a distance-based algorithm that uses the principle of similarity to solve classification cases [4]. The k-NN algorithm is very simple. This algorithm works based on the minimum distance of new data against the nearest neighboring K that has been set. Once the nearest neighbor K is obtained, the predicted class of the new data, will be determined by the majority of the nearest neighbor's K.

C. Voting Feature Interval

The Voting Feature Interval (VFI5) classification algorithm represents the description of a concept by a set of characteristic or attribute values intervals [8]. The classification of a new instance is based on a voting among the classifications made by the value of each feature separately. From the training examples, the VFI5 algorithm constructs intervals for each feature. An interval is either a range or point interval. A range interval is defined for a set of consecutive values of a given feature whereas a point interval is defined for a single feature value [9].

D. CART

Classification and Regression Trees (CART) is one of the methods based on decision tree techniques. CART is a nonparametric statistic method that can describe the relationship between response variable (dependent variable) with one or more predictor variable (independent variable) [13]. Formation of tree classification consists of 3 stages that require learning sample L [13].

$$p(j_0|t) = \max_j\, p(j|t) = \max_j \frac{N_j(t)}{N(t)}$$

With p(j | t) is the proportion of class j at node t, N_j *(t)* is the number of class j observations at node t, and $N(t)$ is the number of observations at node t. The terminal node label t is j_0 which gives the highest guessed error classifier value of the t knot.

PROPOSED METHOD

The flow of this research can be illustrated in the diagram shown in Figure 1. In the first initiation, each raw data was prepared for data processing. The difference of this study is the preprocessing stage which consist of 2 processes, which are, the process that eliminate missing value and make data prototyping using clustering approach. After performing the preprocessing steps, the complete and evaluated or clustered data (prototype data) where the cluster will refer to certain classes would be obtained. After the data prototype was obtained, the main steps would be started which consist of the classification and evaluation process. The cycle was generate the classification accuracy results.

A. Preprocessing Steps

1) Handling Missing Value

There are several methods to solve incomplete data problem. The easiest way to deal with incomplete data problem is to delete one incomplete line of data. This technique sometimes causes loss of any potential information. Data consists of nominal value and numerical value. One technique for dealing with nominal value lost data is to replace the lost data with mode while for numeric value is to replace lost data with mean [10]. The approach we used was to replace all lost data with the average.

2) *Data Clustering for Prototyping Data*

The process of data prototyping was in the preprocessing step which aims to cleaner data form and outliers-free data form. The data prototyping evaluated each data to the class by using the concept of clustering. The clustering process grouped the data items to made each group had essential simmilarity [4]. On this basis, then the data that became outlier or noise would be revised and changed class into "More Related Class" based on its essential simmilarity without changing the amount of data. In this research, we used K-Means as clustering algorithm. As for some additional rules, we made in this process, which are:

1. The number of groups had been set as the number of classes.
2. The class label on a data was used as an initial cluster assumption.

By using such step, a cluster was able to present a class. At the end of the data prototyping process, the data which has been clustered and cleaner was obtained and ready to be included into the next step, the classification step.

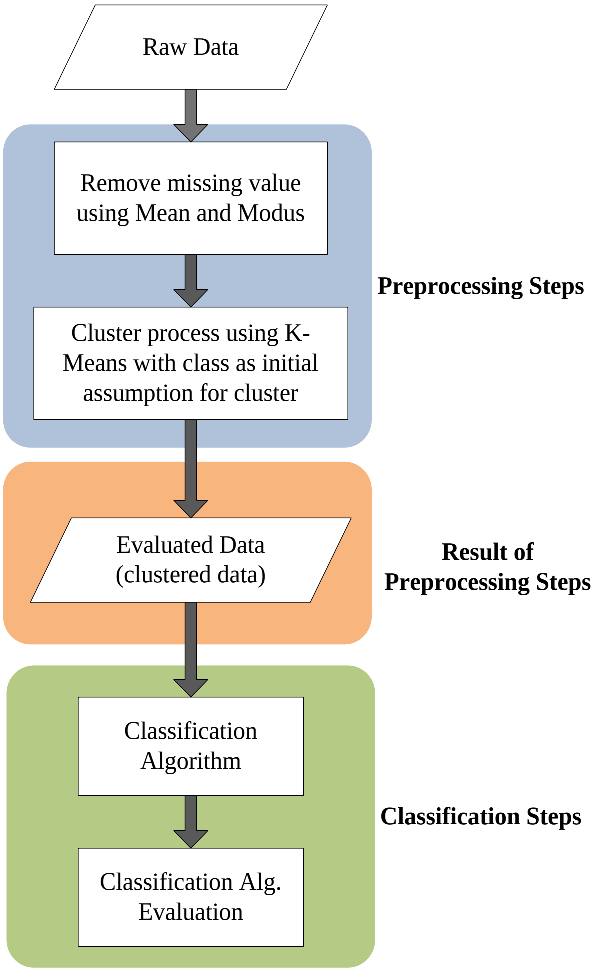
Raw Data
Remove missing value using Mean and Modus
Cluster process using K-Means with class as initial assumption for cluster
Preprocessing Steps
Evaluated Data (clustered data)
Result of Preprocessing Steps
Classification Algorithm
Classification Alg. Evaluation
Classification Steps

Figure 1: Research Flow Diagram

B. *Classificatisn Step*

At the classification stage, the four algorithms described in the previous chapter, which are, Naïve Bayes, KNN, VFI, and CART was used. The utilization of multi-algorithms was intended to provide a variety of classification approaches. Naïve Bayes represented the classification by statistical approach, KNN represented the classification with Nearest-Neighbour approach, VFI represented the classification with Feature Space data approach, and CART represented the classification with tree-based approach.

C. *Evaluation Step*

Each process/classification cycle was included into the evaluation step to know the accuracy of the classification of the data used. The common measure used to measure the quality of the classification data is accuracy. Accuracy states how close the value of the measurement results to the true value or the value that is considered true (accepted value) [11].

In this research, evaluation step used k-Fold Cross Validation method. The K which used was 10, it can be said that the method which used was 10-Fold Cross-Validation. The selection of k in the amount of 10, due to many experimental results indicates that this is the best option for obtaining accurate estimation and 5-fold or 20-fold often also yields almost identical results.

In K-fold cross-validation, the intact datasets were randomly divided into 'k' subset of almost equal size and mutually exclusive to each other. The model in 'classification' was trained and tested as much as 'k' times. All folds was trained in each training except one fold which was left for testing [12]. The cross-validation assessment of the overall model accuracy was calculated by taking the average of all individual k 'accuracy results, as shown by the following expression [12]:

$$CVA = \frac{1}{k}\sum_{i=1}^{k} A_i$$

CVA is cross-validation accuracy, k is the number of folds used, and A is a measure of the accuracy of each fold.

D. *Data*

In this research, seven types of datasets obtained from UCI was used. The dataset which used had the number of attributes, attribute types, and the number of data which vary. The use of this dataset was intended to found how capable the proposed method to resolve the differences in the number of attributes, attribute types, and the number of data each data has.

Table 1. Dataset Specification

Names	Number of Attributes	Attribute Types	Instances
Abalone	8	Categorical & Numerical	4177
Breast Tissue	10	Numerical	106
Glass	10	Numerical	214
ILPD	10	Categorical & Numerical	583
Red Wine Quality	12	Numerical	1599
White Wine Quality	12	Numerical	4898
WPDBC	34	Numerical	198

EXPERIMENTATION

A. Experiment Setup

At the experiment, we used WEKA Data Mining tools. To know the difference of experiment result, there were 2 data treatment used which were not using preprocessing (directly conducted the classification and evaluation process) and using preprocessing (data prototyping) first. The "Cluster Evaluation derived from Class" mode facility which already presented in WEKA was used to keep the track of the connectedness between clusters and classes. There were fixed variables set to keep the results fair, which are:

- Euclidean distance was used to calculate the similarity
- The number amount of K which used on the k-NN was 5

B. Experment Results

After preparing the data and tools that used in the testing phase, the results of classification accuracy was obtained, as shown in Table 2 and 3. Table 2 shows the result of classification accuracy without using data prototyping, while Table 3 shows the results of classification accuracy after using preprocessing. To visualize the details of Classification accuracy improvement, the charts of each data accuracy results on all classification algorithms is shown in Figure 2 to 8.

Table 2. The result of classification accuracy without using data prototyping

Data	Classification Algorithm Accuracy (%)			
	NB	**KNN**	**VFI**	**CART**
Abalone	26.86	29.67	24.38	31.33
Breast Tissue	70.75	69.81	66.98	68.86
Glass	49.53	67.28	57.00	70.56
ILPD	55.74	64.32	47.51	71.01
Wine Quality (white)	44.26	55.41	28.07	59.30
Wine Quality (Red)	55.03	58.16	34.89	60.16
WPDBC	67.17	74.24	58.08	76.26

Table 3. The results of classification accuracy after using preprocessing

Data	Classification Algorithm Accuracy (%)			
	NB	**KNN**	**VFI**	**CART**
Abalone	90.89	97.70	65.01	95.80
Breast Tissue	89.62	91.50	84.90	88.67
Glass	88.78	90.65	82.24	90.18
ILPD	98.11	100	100	100
Wine Quality (white)	85.87	88.95	71.15	85.34
Wine Quality (Red)	85.74	86.55	73.92	82.23
WPDBC	97.97	96.96	88.88	94.44

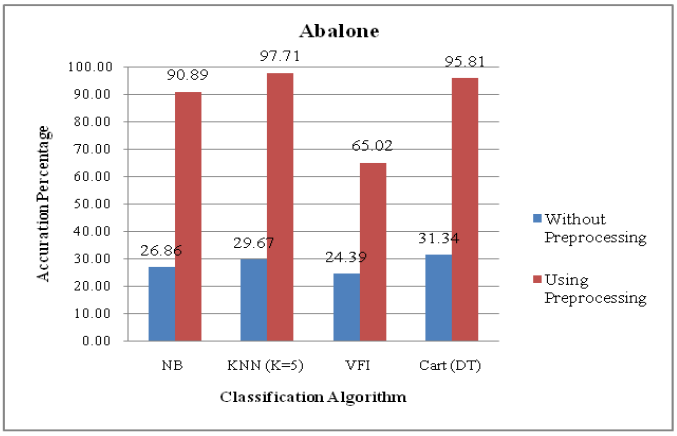

Figure 2: Comparison of our proposed method classification accuracy result on Abalone data

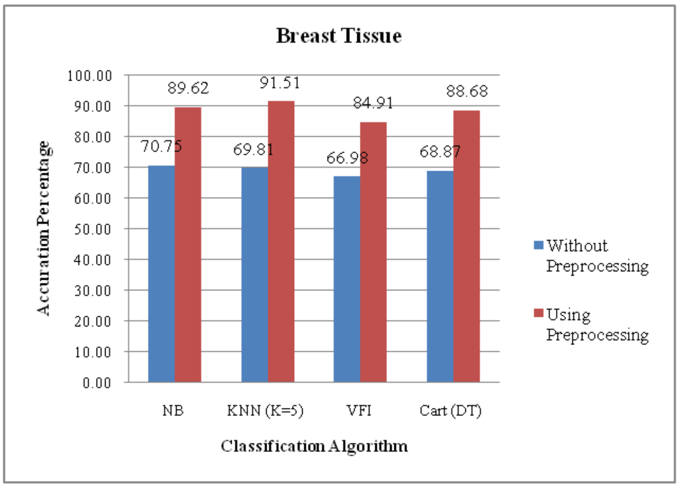

Figure 3: Comparison of our proposed method classification accuracy result on Breast Tissue data

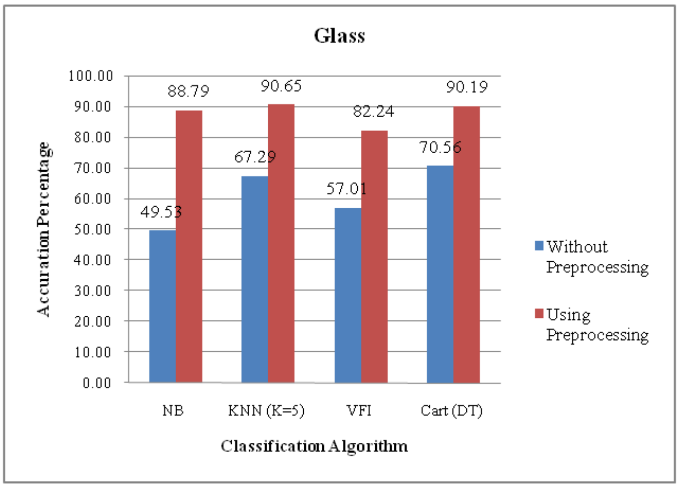

Figure 4: Comparison of our proposed method classification accuracy result on Glass data

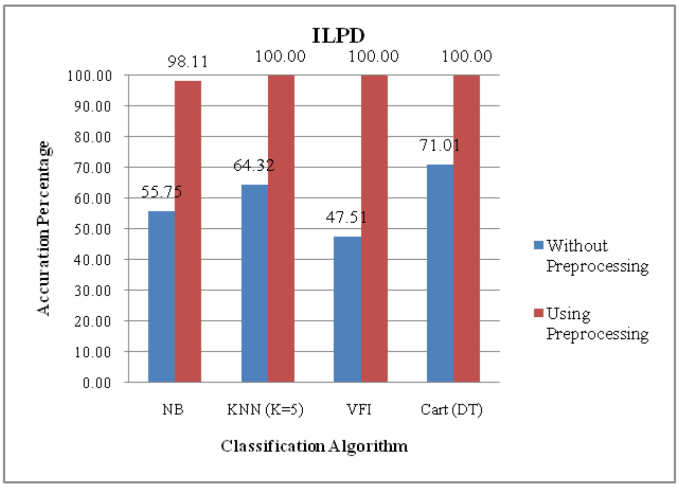

Figure 5: Comparison of our proposed method classification accuracy result on ILPD data

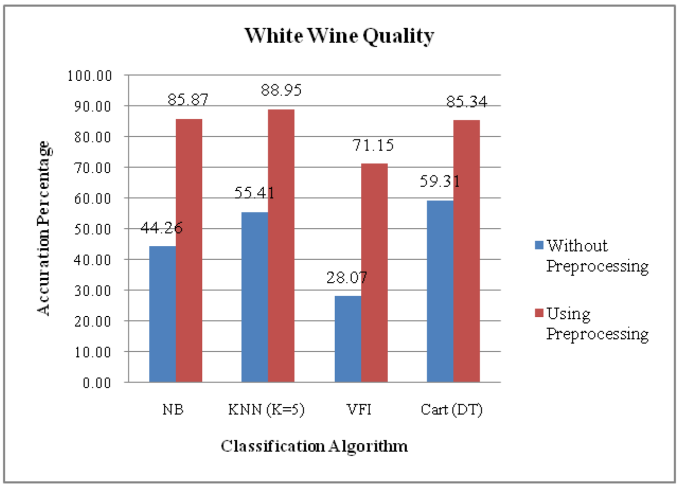

Figure 6: Comparison of our proposed method classification accuracy result on White Wine Quality data

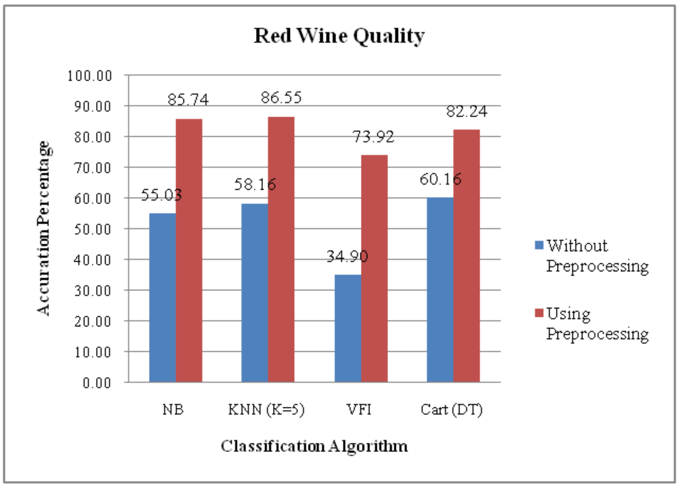

Figure 7: Comparison of our proposed method classification accuracy result on Red Wne Quality data

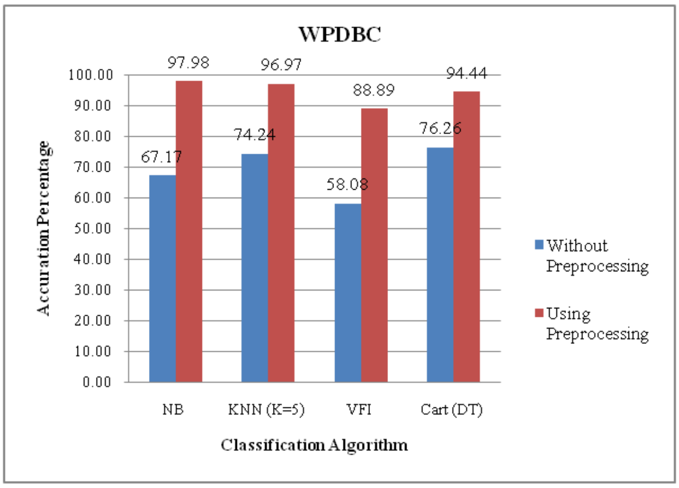

Figure 8: Comparison of our proposed method classification accuracy result on WPDBC data

C. Discussion

Based on the comparison of Table 2 and 3, the results on all data and the classification algorithm showed there was an increase of the magnitude accuracy as shown in Table 4. The increase of classification accuracy on each data was different. To determine the effect of the use of data cleaning and data prototyping, the average increase in classification accuracy based on the data used was included in Table 4. The highest average increase of accuracy results was in the Abalone data. This was possible because the Abalone data consisted a lot of noise or misclassification so that after the data cleaning and prototyping, the data became more normal.

Table 4. The detail of Classifcaton Accuracy Increase

Data	Classification Algorithm Accuracy (%)				Average (%)
	NB	KNN	VFI	CART	
Abalone	64.02	68.03	40.62	64.47	**59.29**
Breast Tissue	18.86	21.69	17.92	19.81	**19.57**
Glass	39.25	23.36	25.23	19.62	**26.86**
ILPD	42.36	35.67	52.48	28.98	**39.87**
Wine Quality (white)	41.60	33.54	43.07	26.03	**36.06**
Wine Quality (Red)	30.70	28.39	39.02	22.07	**30.05**
WPDBC	30.80	22.72	30.80	18.18	**25.63**

The increase of accuracy results can reach 100% in the ILPD data using KNN, VFI, and CART classification algorithms. The lowest accuracy result after the data cleaning and prototyping was in Abalone data using VFI classification algorithm that was 65%. This was caused by the incompatibility of the use of classification algorithm to the data, so it was inadequate in the process of classification. Despite being the lowest accuracy result, after the data cleaning and prototyping, the result of accuracy Abalone data classification using VFI increased by 40.6299%. The highest increase of accuracy result was on Abalone data using KNN classification algorithm that was from 29.672% increase to 97.7095% or increased by 68.0375.

Based on Table 2, the average classification result without the data cleaning and prototyping on all data and the classification algorithm was about 55%. Based on Table 3, the average accuracy result on all data and classification algorithm after the data prototyping was 89%. Then, it can be concluded that by using proposed method in this research can increase about 34% accuracy result or can reduce error equal to 75,55%.

CONCLUSION

After doing the experiment process, the conclusion that can be obtained from this research was by adding the preprocessing step that we proposed, the increase of accuracy of classification result on all data and algorithm was obtained. The increase of accuracy of classification results reached 100% on the data ILPD using KNN, VFI, and CART classification algorithm. Based on the comparison of classification accuracy without using and using preprocessing, by using the method proposed in this study could increase approximately 34% accuracy results or reduce the error 75.55%.

REFERENCES

[1] S. Nagaparameshwara, B. Rama, "Analysis of Classification Technique Algorithms in Data mining- A Review" IJCSE, vol-4,issue-6, pp. 180-185, 2016.

[2] *Gulsen, etc., "Non-Incremental Classification Learning Algorithms Based On Voting Feature Intervals", 1997.*

[3] M.N. Murty, V.S. Devi, "Pattern Recognition: An Algoritmic Approach". Springer, pp.275. 2011.

[4] A.R. Webb, K.D. Copsey, "Statistical Patern Recofnition 3rd Edition". A John Wiley & Sons,Ltd. 2011.

[5] G. Dougherty, "Pattern Recognition and Classification: An Introduction". Springer. 2012.

[6] J. Han, M. Kamber, "Data Mining Concepts and Techniques". 2001.

[7] H.A. Guvenir, G. Demiroz, N. Ilter, "Learning Differential Diagnosis of Erythemato-Squamous Diseases using Voting Feature Intervals". Artificial Intelligence in Medicine 13 (3): 147-165. 1998.

[8] G. Demiros, etc., "Non-Incremental Classification Learning Algorithms Based On Voting Feature Intervals". 1997.

[9] H.A. Guvenir, N. Emeksiz, "An Expert System for the Differential Diagnosis of Erythemato-Squamous Diseases". Expert Systems with Applications 18(1): 43-49. 2000.

[10] Mei-Ling, Shyu. 2005. Handling Missing Values Via Decomposition of the Conditioned Set.

[11] Labatut, Vincent dan Cherifi, Hocine. 2011. Accuracy Measures for the Comparison of Classifiers.

[12] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," in Encyclopedia of Database Systems, L. Liu and M. T. ÖZSU, Eds. New York: Springer US, 2009, pp. 532–538.

[13] Soman, K.P., S. Diwakar and V. Ajay, "Insight into Data Mining-Theory and Practice". Prentice Hall of India, New Delhi, ISBN: 81-203-2897-3.2006.