COMPARATIVE ANALYSIS OF DATA CLASSIFICATION TECHNIQUES BY PROCESSING ON MISSING VALUES INSTANCES SUBSTITUTION ON DATA FROM WEB REPOSITORY FOR KNOWLEDGE DISCOVERY

[1] ASST. PROF. MINAXI PRAJAPATI, [2] RICHA MEHTA, [3] DR. B. S. AGRAWAL

[1]Dept of Computer Application (MCA), Gujarat Technological University, Gujarat, India
[2]Dept of Computer Application (BCA), KSV, Gujarat, India
[3]Director of VJKM Institute, Gujarat Technological University, Gujarat, India

bijendra.agrawal@gmail.com,minaxi14287@gmail.com

***ABSTRACT* - *An emerging technology extensively used in every field – Data mining that deals with the innovation and analysis of usage pattern(s) and relationships within trends leading to discover knowledge. This is a way of process to analyze a data from different viewpoints and summarized it into meaningful information. Data can be stored in a range of forms of databases and storage area. This can be superior way to extracting valuable knowledge for decision making. The main objective of this study is to show the classifying data with a reasonable accuracy for improving the performance in data mining by applying missing values instances substitution on a data. . The learning converses data mining techniques to process a medical data set and classify the importance of heart diseases. The massive amounts of data related heart diseases to be gathered, unfortunately that are not "mined" to determine hidden information for valuable decision by healthcare practitioners. A heart disease covers the different diseases that have an effect on the heart. Some categories to be include related heart diseases - Cardiomyopathy and Cardiovascular. The objective will be focused on getting the higher level percentage of accuracy. The implementation of this result carried out with the open source software environment Orange Canvas with three Classification Techniques – NB, KNN and SVM.*

Keywords: Data mining, Data mining Tools – Orange Canvas, Classification Techniques, Performance Evaluation Parameters, Sampling Methods

I. INTRODUCTION

Now a day's numerous data is to be produced every day in various applications. Data mining is collection of techniques of efficient automated discovery of previously unknown in large volume of data [5]. Machine learning refers to a system that has the capability to automatically learn knowledge from experience and other ways [6]. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends [7]. It is seen as an increasingly important tool by modern business to transform data into an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery [2].

The research has been conducted to evaluate the performance of data mining process by comparing the classification techniques – Naïve Bayesian (NB), Support Vector Machine (SVM) and K- Nearest Neighbor (KNN). The problem occurs when the volume of data is large and to estimate that data for human is very difficult to getting valid patterns and intra relationships because of getting human generated errors. To solve this problem the work is to be carried out in the open source environment Orange Canvas. It is depending on the ability for classifying data correctly and accurately. This is used to represent the higher level of accuracy by comparing classification techniques. These tools can include statistical models, mathematical algorithm and machine learning methods. The outcomes of this study are expected to be useful for various areas.

II. CLASSIFICATION TECHNIQUES

Classification is a process of classify the data into systematic manner which is based on the supervised learning techniques. It generally [1] requires that the classes be defined based on the data attributes values. [2] The classification techniques learn from the training set and build a model. Classification is a data mining technique used to map a data item into one of several predefined classes. The objective is to predict the value of a user-specified targeted attribute that is based on the values of other attributes

A. NAÏVE BAYESIAN

Naïve Bayesian uses the concept of Bayesian theorem with strong independence assumptions. A naive Bayesian classifier could be defined as an independent feature model deals with a simple probabilistic classifier. The main goal of this classifier is to build a rule which will allow assigning the viewpoint of the objects to a class objects. The good thing of this technique is to estimate the performance of this technique which needs not a large amount of data to work but it also works well on small amount of data on many composite problems.

B. KNN

KNN algorithm is [3] considered as statistical learning algorithm and easy to implement. It is a technique for classifying objects which is based on closest training data in the feature space. KNN is used for pattern recognition. It is most popular algorithm for text categorization. KNN is a type of instance based learning or lazy learner. It is a simplest algorithm among the entire machine learning algorithm which is used to determine the class label of the object. KNN rules in effect compute the decision boundary in an implicit manner. Also, it is possible to compute decision boundary itself explicitly. The best choice of K depends upon the data; larger values of K reduce the effect of noise on the classification but make boundaries between classes less distance. A good K can be selected by various heuristic techniques. But the obvious drawback of this model is that many test records will not be classified because they do not match exactly with any training instances.

C. SUPPORT VECTOR MACHINE (SVM)

SVM is a set of associated with [8] supervised learning methods used for classification and regression. In today's machine learning applications, support vector machines (SVM) [4] are considered a must try—it offers one of the most robust and accurate methods among all well-known techniques. SVM is a [9] developed techniques for multidimensional function approximation. SVM has the ability to handle the large feature space. It is successfully applied to a real world problem such as face recognition, text categorization, object detection and many other fields.

III. PROPOSED STRUCTURE OF PEP- KD

The proposed PEP – KD Framework is used to implement the planned work. This PEP – KD Framework is designed on the basis of the usual Knowledge Data Discovery process and hence contains five phases initializing from Data Extraction to Data preprocessing which is followed by Processing and Evaluation Phase that finally leads to the proposed concluding Knowledge Discovery Phase. The detail flow of data and information in various phases is shown in FIGURE 3.1

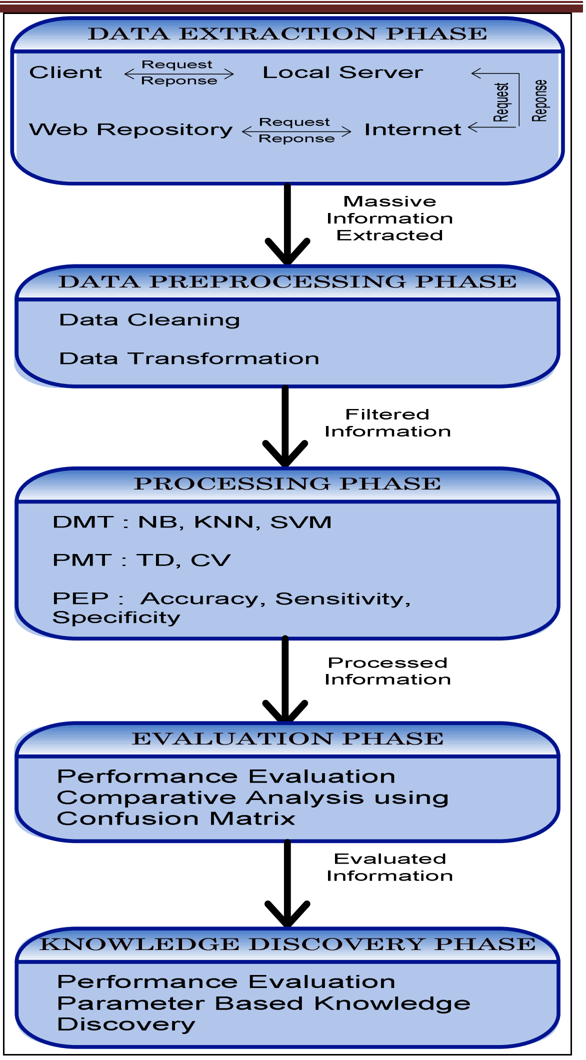

FIGURE 3.1: PEP – KD FRAMEWORK

IV. IMPLEMENTATION AND RESULTS

The proposed work is implemented on a secondary datasets (A Cardiovascular dataset which is the category of heart disease) obtained from the UCI Repositories available on web at URL ("http://orange.biolab.si/datasets.psp"). The dataset is selected from clinical datasets of a status with missing values. The dataset is subjected to the selected three classification techniques – NB, KNN and SVM. The implementation as a part of experimental work is implemented on the platform of Orange Canvas software tool (version 2.7).

A. Comparative Analysis

The outcomes are derived from the comparative analysis using Performance Evaluation Techniques on a Cardiovascular datasets pertaining to heart disease. The dataset is studied and evaluated to get the proposed outcomes.

B. Results

The outcomes are derived from the comparative analysis using Performance Evaluation Techniques on a dataset pertaining to heart disease.

a. Cardiovascular Dataset with Missing Values

The Cardiovascular is to design a predictive model for presence of heart disease in patients which is obtained from the Cardiovascular dataset. The Cardiovascular dataset contains total 303 instances with 2.0% that is 6 locations of missing values. The dataset is preprocessed in two approaches.

1. Missing values instances substitution

The preprocessing is to be applied on the cardiovascular dataset by filling the missing values with mean value under the attribute. The aggregate value is to be taken from the mean value and put it on the place of missing information. The pre-processed dataset contain 303 instances with 13 attributes. The classifier output of NB, KNN and SVM is carried out with dual sampling methods. In CPESM 1, the testing is to be taken on train data and the classifier is evaluated based on the PEP. In CPESM 2, the Cross Validation is used by selecting the number of folds.

1.1. CPESM 1 - Test on Train Data

In CPESM 1 – Test on Train Data is used to discover the predictive relationships to examine the performance of classification techniques based on PEP- Accuracy, Sensitivity and Specificity.

TABLE 4.1 shows the performance of the classification techniques - NB, KNN and SVM based on the PEP. The outcome of the selected classification techniques based on CPESM 1 – Test on Train Data and PEP is obtained in Test Learner widget.

TABLE 4.1: ACCURACY, SENSITIVITY AND SPECIFICITY FOR NB, KNN, SVM

PEP\Tech.	NB	KNN	SVM
Accuracy	0.85	0.99	0.85
Sensitivity	0.86	1.00	0.87
Specificity	0.84	0.99	0.83

The outcomes of the analyzed result of processing phase is depicted graphically in the below FIGURE 4.1

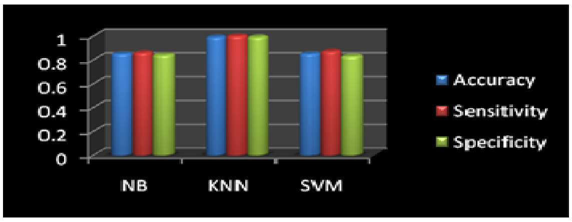

FIGURE 4.1: EVALUATION RESULTS FOR CLASSIFICATION TECHNIQUES BASED ON PEP

The TABLE 4.2 shows the manually computed outcomes of confusion matrix based the PEP for each classification techniques using Kappa Statistics. Here, KNN obtained higher level of Accuracy, Sensitivity and Specificity with 0.99, 1.00, and 0.99 than other techniques.

TABLE 4.2: MANUALLY MANUAL ACHIEVED OUTCOMES FOR NB, KNN, SVM BASED ON PEP FORM CONFUSION MATRIX

	PEP\Tech.	NB	KNN	SVM
Confusion Matrix	**Accuracy**	0.85	0.99	0.85
	Sensitivity	0.86	1.00	0.87
	Specificity	0.84	0.99	0.83

The comparative analysis of CPESM 1 automated computed outcomes shown in TABLE 4.1 and manually computed outcomes depicted in TABLE 4.2 signifies that both the approaches gives the identical results of classification techniques based on PEP justifying the Accuracy of the prediction.

1.2. CPESM 2 - Multi - Fold Cross Validation

In the CPESM 2 - Multi-Fold Cross Validation that is K-Fold Cross Validation is performed on classification techniques - NB, KNN and SVM, by keeping k values ranging from 5 Fold to 10 - Fold Cross Validation used to verify the Accuracy level of the data by step by step increment folds of Cross Validation method. TABLE 4.3, 4.4 and 4.5 illustrate the evaluation outcomes of NB, KNN, and SVM at the growing level from 5 to 10 - Fold Cross Validation.

TABLE 4.3: ACCURACY, SENSITIVITY, SPECIFICITY AT DIFFERENT LEVEL FOR NB

NB			
Value\PEP	**Accuracy**	**Sensitivity**	**Specificity**
K=5	0.84	0.86	0.81
K=6	0.82	0.85	0.78
K=7	0.83	0.87	0.78
K=8	0.83	0.84	0.81
K=9	0.82	0.84	0.80
K=10	0.83	0.87	0.79

TABLE 4.4: ACCURACY, SENSITIVITY, SPECIFICITY AT DIFFERENT LEVEL FOR KNN

KNN			
Value\PEP	**Accuracy**	**Sensitivity**	**Specificity**
K=5	0.77	0.80	0.73
K=6	0.77	0.79	0.74
K=7	0.78	0.81	0.75
K=8	0.77	0.79	0.75
K=9	0.78	0.80	0.75
K=10	0.77	0.79	0.73

TABLE 4.5: ACCURACY, SENSITIVITY, SPECIFICITY AT DIFFERENT LEVEL FOR SVM

SVM			
Value\PEP	**Accuracy**	**Sensitivity**	**Specificity**
K=5	0.84	0.85	0.82
K=6	0.82	0.85	0.78
K=7	0.83	0.86	0.80
K=8	0.82	0.85	0.78
K=9	0.83	0.85	0.79
K=10	0.83	0.85	0.81

That overall result of each NB, KNN, and SVM techniques are obtained by taking an average of PEP to get an appropriate result is shown in the TABLE 4.6. The average Accuracy PEP outcomes of SVM and NB are 0.83 which is higher than KNN with 0.77 respectively. The average Sensitivity of PEP of NB is 0.86 which is higher than SVM and KNN with 0.85 & 0.80 outcomes and even the average Specificity of PEP of SVM and NB are 0.80 which is also higher in comparison to KNN with the value of 0.74.

TABLE 4.6: AVERAGE RESULT OF EACH CLASSIFICATION TECHNIQUES

PEP\Tech.	NB	KNN	SVM
Accuracy	0.83	0.77	0.83
Sensitivity	0.86	0.80	0.85
Specificity	0.80	0.74	0.80

The outcomes of the analyzed result of processing phase is depicted graphically in the below FIGURE 4.2

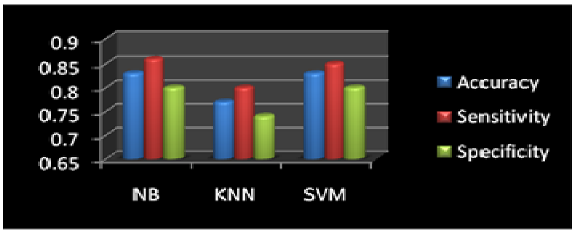

FIGURE 4.2: AVERAGE OUT THE EVALUATION RESULTS FOR CLASSIFICATION TECHNIQUES BASED ON PEP

The comparative outcomes of CPESM 2 when analyzed for the Accuracy (depicting the aptness of the dataset), Sensitivity(defining the heart problem of the patients) and Specificity(focusing the normal status of a patients) summarizes that the SVM achieved higher Accuracy, Sensitivity and Specificity in implementing CPESM 2.

TABLE 4.7: MANUAL AVERAGE EVALUATION RESULT OF CONFUSION MATRIX

PEP\Tech.	NB	KNN	SVM
Accuracy	0.83	0.77	0.83
Sensitivity	0.86	0.80	0.85
Specificity	0.80	0.74	0.80

The average Accuracy PEP outcomes of SVM and NB are 0.83 which is higher than KNN with 0.77. The average Sensitivity of PEP of NB is 0.86 which is higher than SVM and KNN with 0.85 & 0.80 outcomes and even the average Specificity of PEP of SVM and NB are 0.80 which is also higher in comparison to KNN with the value of 0.74.

The outcomes of the analyzed result of processing phase is depicted graphically in the below FIGURE 4.3

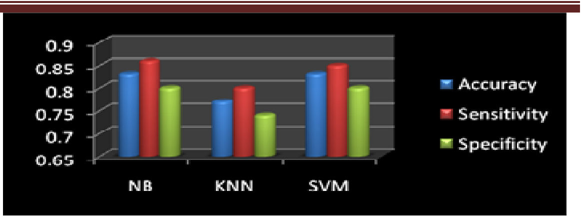

FIGURE 4.3: AVERAGE RESULT OF CLASSIFICATION TECHNIQUES

The comparative analysis the used techniques manually computed outcomes in CPESM 2 is similar for Accuracy, Sensitivity and Specificity to the automated computed outcomes when analyzed till 2 decimal level. Therefore, both the computation methods conclude that SVM contain higher aptness in Accuracy and Specificity.

1.3. Results

The comparative evaluation analysis of selected Classification Techniques – NB, KNN and SVM using the dual CPESMs based on three PEP – Accuracy, Sensitivity & Specificity concludes that when automated and manually computed in the CPESM 1 : Test on Train Data - the KNN achieved higher level of Accuracy and in CPESM 2 : Multi-Fold Cross Validation - the SVM achieved higher level of Accuracy.

The overall consolidate result of the Case 1 is evaluated by filling mean values instead of missing values is summarized in TABLE 4.8

TABLE 4.8: SUMMARY TABLE FOR CASE 1 (CARDIOVASCULAR) BY FILLING MEAN VALUES INSTEAD OF MISSING VALUES

CPESM	(CARDIOVASCULAR) by Filling Mean Values					
PEP\Tech.	Test on Train Data			Multi- fold cross validation		
	NB	KNN	SVM	NB	KNN	SVM
Accuracy	0.85	0.99	0.85	0.83	0.77	0.83
Sensitivity	0.86	1.00	0.87	0.86	0.80	0.85
Specificity	0.84	0.99	0.83	0.80	0.74	0.80

The comparative analysis based on TABLE 4.8 justifies that dual CPESMs, the CPESM 1 and CPESM 2 when evaluated concludes that CPESM 1 achieved optimum result for Accuracy in KNN with 0.98 than other technique on the other hand the CPESM 2 achieved optimum result for Accuracy in SVM with 0.84 than other techniques. Therefore, it can be stated that selecting the different CPESMs may affect the outcomes of the results, so an optimum choice of the Classification Technique based on a particular CPESM can be used to obtain the targeted outcome.

V. CONCLUSION

There are two results achieved for Cardiovascular by pre-processing, processing, comparative analysis evaluation of the datasets, both the approaches are used two Classification Performance Evaluation Sampling Methods CPESM 1 - Test on Train Data and CPESM 2 - Multi-Fold Cross Validation.

And the result for (Cardiovascular by filling mean values under attributes) indicates that the KNN achieved more robust result on dataset with train data and not only at accuracy level but also at higher level for sensitivity and specificity. With the multi-fold cross validation SVM gives higher level accuracy on dataset and also gives higher level for specificity but in sensitivity NB gives higher level for sensitivity.

VI. FUTURE EXTENSION

For the extension of this work may consider different combinations of classification techniques with a purpose to increase the level of accuracy using Multi - Fold Cross Validation. To develop dipper understanding of analysis the classification techniques may be used with suitable visualization tools with different datasets.

REFERENCES

[1] "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", IJEIT, Volume 2, Issue 3, September 2012

[2] "COMPARISON OF CLASSIFICATION TECHNIQUES", IJRIM Volume 1, Issue 2 (June, 2011)

[3] "Comparison of Different Classification Techniques Using WEKA for Breast Cancer", Mohd Fauzi bin Othman, Thomas Moh Shan Yau, Springer, 2007

[4] Vapnik V (1995) The nature of statistical learning theory. Springer, New York.

[5] D.K. Roy, L.K. Sharma, (2010) "Genetic k-Means clustering algorithm for mixed numeric and categorical data sets", International Journal of Artificial Intelligence & Applications, Vol 1, No. 2, pp 23-28.

[6] H. Jiawei and K. Micheline, (2008) Data Mining-Concepts and Techniques, Second Edition, Morgan Kaufmann - Elsevier Publishers, ISBN: 978-1-55860-901-3

[7] "DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE", Durgesh K. Shrivatava and Lekha Bhambhu, JATIT, 2005-2009

[8] "Classification Methods", Aijun Ai, 2005