JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Classification Evaluation Metrics in Machine learning

[1]NagaSubrahmanyeswari.N., [2]G.Satya Suneetha

[1]Lecturer in Computer Science, A.S.D.Govt. Degree College for Women(A), Kakinada.
[2]Lecturer in Computer Applications, A.S.D.Govt. Degree College for Women(A), Kakinada.

Abstract: Machine learning is the field of study where the computers gain the ability to learn without being programmed and this learning happens from the models they build earlier. Classification is a common task in data mining and its applications which is one of the supervised learning techniques. With classification, the data sets are predicted to belong to any one of the set of predefined class labels. Once the classification models are built, they need to be evaluated on the basis of any performance criteria available. Evaluation of the classifiers is an important task for selecting the best model built. This paper presents an overview on the available metrics for evaluating the classifiers.

Index terms: Machine learning, Classification, Evaluation.

I.INTRODUCTION:

There are huge amounts of data available in Information Industry. This data available will not serve us any purpose until it is converted into some useful information. So we need to analyze the raw data available and extract the useful information from it. This process of extracting useful information involves several other processes such as data cleaning, data integration, data transformation, data mining, pattern evaluation etc., [1]. After the completion of these processes we can use this information in many applications wherever it finds its necessity. Data mining is the process of extracting useful information from the available data repositories. Classification is a common task in data mining and its applications. One can rely, based on its accuracy and evaluation. Generally Classification is the task of learning target function *f* that maps every attribute set *x* to one of the predefined class labels *y*. It is a pervasive problem that incorporates many diverse applications. Examples include detecting spam email messages based upon the message header and content, categorizing cells as malignant or benign based upon the outcomes of the CT scans or MRI scans, DNA classification, image classification etc.,

II.MACHINE LEARNING & CLASSIFICATION:

Machine learning is a branch of artificial intelligence that aims at enabling machines to learn from previous experiences and perform their jobs skillfully by using intelligent algorithms. The statistical learning techniques form the basis of the development intelligent software that is used to develop machine intelligence. In the present day scenario, a huge increase in demand for machine learning is seen with the great number of available datasets. Machine learning is the process of knowledge acquisition mechanism with improvement from experience using computational methods. The primary purpose behind its popularity in different applications is its ability to learn once and then can be applied for any type of data or input given to it which is similar to the data learnt.

Supervised machine learning techniques can be broadly categorized into Regression and Classification. Classification is used for predicting the target attributes which are categorical values whereas in Regression the target attributes are continuous values. Linear Regression is the machine learning technique based on Regression whereas Logistic Regression is Supervised classification technique.

A classification technique is a systematic approach to build classification models from a given input dataset. There are several techniques to build classification models which include decision tree classifiers, rule-based classifiers, support vector machines etc. All the techniques employ a learning algorithm to identify the model that suits well for the attribute set and data label of the input data. Hence, the main objective of the learning algorithm is to build models which best fits the input data set as well as accurately predict the class labels of previously unseen records i.e. the central concern of modeling techniques is improvement in accuracy. First, a training set whose class labels are knownis provided and a classification model is built which is applied to the test set with records of unknownclass labels.

Once the models are built, they need to be evaluated to choose the best model that can be usedin our mining application. Evaluation of the classifier is done on the basis of the number of records from the test set correctly and incorrectly predicted. The different criterions on which the classifiers are evaluated are accuracy, speed, robustness and scalability. Accuracy is the ability of the classifier to

correctly predict the class labels of unknown data records. Speed is the computational cost of the class label prediction by the classifier [1]. Robustness is the ability to classify the records with noisydata. Scalability is the ability of the classifier to classify when introduced to huge amounts of data.

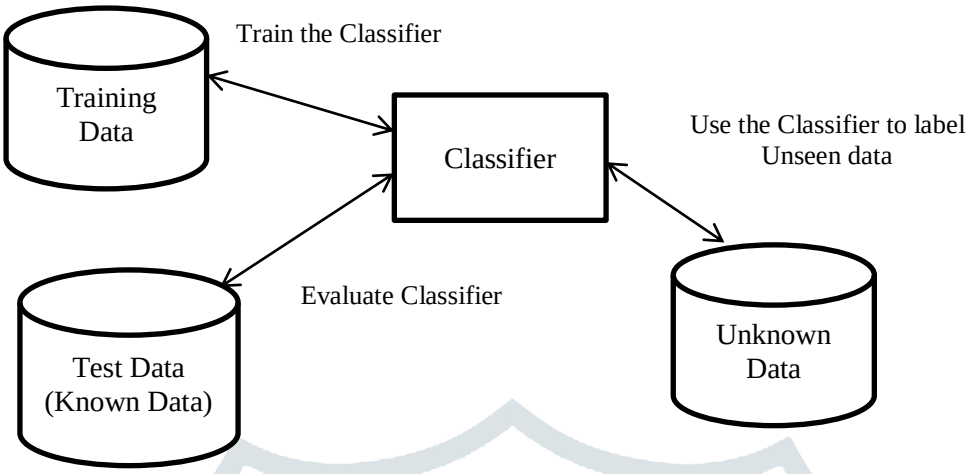

Fig.1: Classification Model

If inappropriate methods are used for evaluating the classifier, well performing models may be measured inappropriately based upon the information available regarding classification error rate and the context of application. An analyst may develop numerous models for a particular problem domain and uses an inaccurate evaluation model for choosing the best model. As a result, the accuracy of the model chosen will be less compared to the other models which can be overcome with the use of an appropriate evaluation method. In the end, poor decisions were made due to theselection of an inappropriate evaluation method.

III.METHODOLOGY OF CLASSIFICATION TECHNIQUES

Decision Tree:
Decision trees are considered as the most popular approaches for representing classifiers. A decision tree is a classifier expressed as a recursive partition of the instance space[2]. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called 'root' that has no incoming edges. All other nodes have exactly one incoming edge. In decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function of the input attribute values. Each leaf is assigned to one class representing the most appropriate target value[2]. The basic algorithm for decision tree induction is a greedy algorithm following recursive divide and conquer strategy [1].

Bayesian Classifiers:
Naive Bayesian classifier is a statistical classifier based on the Bayes' Theorem and the maximum posteriori hypothesis[3]. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. The naive Bayesian classifier estimates the class- conditional probability by assuming that the attributes are conditionally independent, given the class label C_k. Suppose that there are n classes, $C_1, C_2, \ldots, C_n$, the conditional independence assumption can be formally stated as follows:

$P(X|C_k) = \prod_{i=1}^{m} P(Xi|Ck)$ (1)

where every attribute **set** $X = \{X_1, X_2, \ldots, X_m\}$ consists of m attributes. To classify a test record, the naive Bayesian classifier computes the posterior probability for every class C_k.

K-Nearest neighbor Classifiers:
Nearest neighbor classifiers are based on learning by analogy that is by comparing a test record with a set of training records similar to it, with each record having 'n' attributes. Similarity is measured in terms of distance metric such as Euclidean distance etc., It is well suitedfor large scale hierarchical text classification[4]. The major drawback of this approach is it uses all features or attributes in the computation of similarity measures. This can be overcome by learning theweights of all attributes for the computation of similarity measures.

Support Vector Machine:
Support Vector Machine is a new classification method for both linear and non-linear data. It uses a non-linear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyper plane (i.e. "decision boundary"). With an appropriate non-linear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane. SVM finds this hyper plane using support vectors ("essential training tuples") and margins (defined by the support vectors). Features: Training can be but accuracy is high owing to their ability to model complex non-linear decision

boundaries (margins – maximization). SVM used for both classification and prediction [7].

IV.METRICS FOR CLASSIFIER EVALUATION

Evaluating the performance of any supervised or unsupervised technique is the most fundamental aspect of machine learning. It is very important for selecting the most appropriate model from a given set of models. When inappropriate evaluation methods are applied, well-performing models may be measured inappropriately based upon the information available regarding classification error rate and the context of application. An analyst may develop multiple models to address a problem domain, but use an incorrect evaluation method to select a "best" model. The result may be the selection of a sub- optimal model, which leads to less accurate classification than may have been possible with a more appropriate evaluation method. In the end, poor decisions are made because an incorrect model was selected, using an inappropriate evaluation method.[5]
Based on the metrics used, we can distinguish between various evaluation methods used for evaluating the classifier. It is very common that a classifier is evaluated based on the ***Error rate or Accuracy.*** The ***Error rate*** is nothing but the rate of data items misclassified to the total number of data items to be classified where as ***Accuracy*** is the rate of total no. of correct classifications made.
Other performance measures used are:

Gini Index: In decision tree induction, Gini Index is used as a measure of purity. It is used to select the feature at each internal node of the decision tree. This measure is used to identify the partition purity of the data set. The Gini index can be defined as:

$$Gini(t) = \sum_{i=0}^{c-1}[p(i|t)]^2 \quad (2)$$

where *'c'* is the no. of classes and $[p(i|t)]^2$ is the fraction of records belonging to class '*i*' at node '*t*' [1].

Minimum Description Length Principle (MDL): Another way to find the complexity of a model built using decision tree using information-theoretic approach is Minimum Description Length Principle or MDL principle. As per this principle, we seek to build a model that minimizes the overall cost function.

Confusion matrix: In supervised classification with two classes, performance can be measured with the help of four values called true-positives (tp), false-positives (fp), true-negatives (tn) and false- negatives (fn) which correspond to the correct and incorrect classifications made [6]. These values are entered as entries into a 2 × 2 matrix called confusion matrix. The total of all values i.e., tp+tn+fp+fn=n, the total of test records.

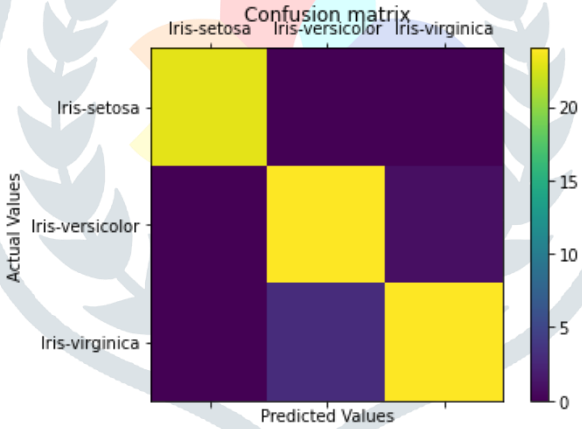

Fig 2. Confusion matrix for Iris Dataset with 3 labels

Accuracy: It is the ratio of total no. of correct predictions to the total no. of predictions made.

$$Accuracy = \frac{|tp|+|tn|}{n} \quad (4)$$

Precision: The number of class members classified correctly over total number of instancesclassified as class members.

$$Precision = \frac{|tp|}{|tp|+|fp|} \quad (5)$$

Recall: The number of class members classified correctly over total number of class members.

$$Recall = \frac{|tp|}{|tp|+|fn|} \quad (6)$$

ROC Curve: An Receiver Operating Characteristics (ROC) graph is a way for visualizing and selecting the classifier based on its performance. It is a two-dimensional graph with tp(true-positives)rate on y-axis and fp(false-positives) rate plotted on x-axis[17]. To compare two classifiers, we can use AUC (Area under ROC Curve), Since the AUC is a portion of the area of the unit square, and its value will always be between 0 and 1.0. A sample ROC curve for Diabetes Dataset using Logistic Regression is shown below with Test set = 0.56 and seed = 8

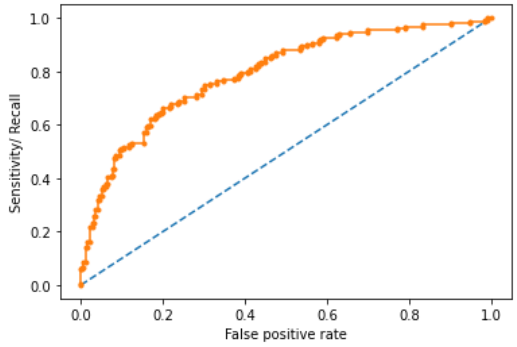

Fig. 3(a): ROC curve for sample Diabetes Dataset using Logistic Regression with AUC - Test Set: 79.27%

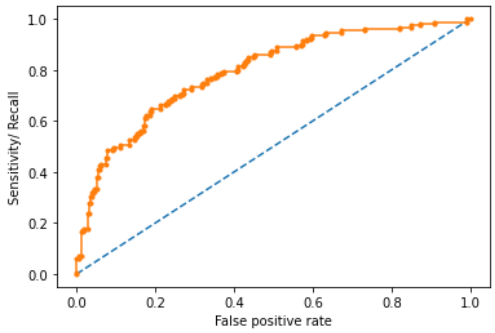

Fig. 3(b): ROC curve for sample Diabetes Dataset using Linear Regression with AUC Test Set: 79.41%

Various Classification Evaluation metrics for classification of Diabetes dataset using Logistic Regression are shown in the below table

Classification Model	Precision	Recall	F1-Score	Log-Loss
Logistic Regression	0.666667	0.529801	0.590406	8.92

Table 1: Classification Evaluation Metrics - Logistic Regression

Various Evaluation metrics for classification of Diabetes dataset using Linear Regression are shown in the below table

Classification Model	Mean Absolute Error	Mean Square Error	Root Mean Square Error	R2 Score
Linear Regression	0.346620	0.178182	0.422116	0.217978

Table 2: Evaluation Metrics - Linear Regression

F1 Score: It is a machine learning metric used in classification models and is defined as the harmonic mean of Precision and Recall.

$$\text{F1 score} = 2 * \frac{\text{precision} * \text{Recall}}{\text{precision} + \text{Recall}}$$

Log-Loss: It is a classification metric used in machine learning based on probabilities. It is the negative average of the log of corrected predicted probabilities for each instance.

Mean Absolute Error: It is the difference between the actual value and the predicted value.

Mean Square Error: It is the average of the square of the difference between the original and predicted values of the data.

Root Mean Square Error: It is the standard deviation of the errors that occur when a prediction is made.

R2 Score: It is also called as Co-efficient of determination and is an indication of how good a model which is built, fits the data.

V.CONCLUSION

In this, we have presented various classification techniques, the various performance measures used for evaluating the classifiers and analyzed some of the metrics on various datasets like Iris, diabetes etc., and compared various Classification metrics available using Logistic and Linear Regression. Depending on the perspective and the context being used for evaluating the classifier; we employ the best technique that suits our need.

VI.REFERENCES

[1] Survey of Classification Techniques in Data Mining Thair Nu Phyu, International MultiConference of Engineers and Computer Scientists 2009

[2] Classification trees. Lior Rokach, Oded Maimon, Springer, 2009.

[3] Enhanced Classification Accuracy on Naive Bayes Data Mining Models, Md. Faisal Kabir ,Chowdhury ,Mofizur Rahman Alamgir Hossain, Keshav Dahal, International Journal of Computer Applications, August 2011.

[4] An Optimized K-Nearest Neighbor Algorithm for Large Scale Hierarchical Text Classification,Xiaogang Han, Junfa Liu, Zhiqi Shen, and Chunyan Miao, 2011

[5] Namit Katariya, Arun Iyer, Sunita Sarawagi "Active Evaluation of Classifiers on Large Datasets",in IEEE 12th International Conference on Data Mining, 2012.

[6] Evaluating Classifiers, Charles Elkan ,elkan@cs.ucsd.edu January 20, 2012.

[7] A Survey of Classification Techniques in Data Mining, M. Sujatha, S. Prabhakar, Dr. G. LavanyaDevi, IJIET, 2013.