Process of Measuring the Accuracy And Error Measures By Using Classifier In Datamining

Miss. K. Laxminarayanamma[1] , Dr. R. V. Krishnaiah[2], Dr. P. Sammulal[3]

[1]Assoc. [2]Proff, Professor, [3]Asst. Proff, [1,2]Dept. of IT, [3]Dept. of CSE

[1,2]Institute of Aeronautical Engineering.Dundigal,Quthubullapur,Medchal,Telangana,India.

[3]JNTU-H, Kukatpally, Hyderabad, Telngana, India

Abstract - This paper proposes an Accuracy and Error Measures on confusion matrix of a classifier to measure the accuracy of a classifier the percentage of a set of tuples that used to be tested and accurately classified by the classifier is calculated. The accuracy also corresponds to the classifier's total 'recognition rate 'and explains the extent of recognizing tuples belonging to different classes by the classifier.

Keywords: Data mining, Classification, Prediction, accuracy, Error Measures, confusion matrix, Error Rate(ER),

I. INTRODUCTION

Classification and prediction can be viewed as two kinds of data analysis that is used to,

i. Retrieve models that describes impotent data classes
ii. To predict future data trends.

A) Classification

A large database has huge amount of row data, which is analyzed and predicted to retrieve useful information and to make decisions, classification is one of the methods used for data analysis. We analyze the data and classify it, based on our requirement[2].

B) Prediction

Prediction can model the continuous value functions with the help of statistical techniques of regression.

When an input is provided its ordered values can be predicted .This activity is called numeric prediction. A numeric prediction can be performed using regression techniques. Various classifications and prediction techniques has been developed by researches in metalanguage, expect system,statisrics and neurobiology[4].

C) Confusion Matrix

A confusion matrix is a table that is often used to explain the performance of a classification model or a confusion matrix is a technique for summarizing the performance of a classification algorithm[3].

Example confusion matrix for a binary classifier

Table 1: confusion matrix for a binary classifier

n=1652	Predicted No	Predicted No
Actual :No	50	10
Actual:Yes	5	100

In the table 1 two possible predicted classes: "Yes" and "No". If we were predicting the attendance of a bug, for example, "Yes" would mean they have the disease, and "No" would mean they don't have the disease[3].

The classifier made a total of 165 predictions (e.g., 165 patients were being tested for the presence of that disease).

Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times.

In reality, 105 patients in the sample have the disease, and 60 patients do not.

Most basic terms

True Positives (TP): These are cases in which we predicted Yes (they have the disease), and they do have the disease.

True Negatives (TN): We predicted No, and they don't have the disease.

False Positives (FP): We predicted Yes, but they don't actually have the disease.

False Negatives (FN): We predicted No, but they actually do have the disease.

Table 2: Basic terms to the confusion matrix, added the row and column totals:

n=1652	Predicted No	Predicted No	
Actual :No	TN=50	FP=10	60

Actual:Yes	FN=5	TP=100	105
	55	100	

Rates that are often computed from a confusion matrix for a binary classifier

Accuracy: Overall, how often is the classifier correct?

- (TP+TN)/total = (100+50)/165 = 0.91

Misclassification Rate: Overall, how often is it wrong?

- (FP+FN)/total = (10+5)/165 = 0.09
- equivalent to 1 minus Accuracy
- also known as "Error Rate"

True Positive Rate: When it's actually yes, how often does it predict yes?

- TP/actual yes = 100/105 = 0.95
- also known as "Sensitivity" or "Recall"

False Positive Rate: When it's actually no, how often does it predict yes?

- FP/actual no = 10/60 = 0.17

Specificity: When it's actually no, how often does it predict no?

- TN/actual no = 50/60 = 0.83
- equivalent to 1 minus False Positive Rate

Precision: When it predicts yes, how often is it correct?

- TP/predicted yes = 100/110 = 0.91

*Prevalence***:** How often does the yes condition actually occur in our sample?

- actual yes/total = 105/165 = 0.64

II. MEASURES THE ACCURACY OF A CLASSIFIER

The misclassification rate or Error Rate(ER) of a classifier 'R' can also be calculated using the equation

ER=1-A(R)

Where, A(R) =Accuracy of the classifier 'R'

If a training set is used to calculate the error then the resultant error is known as the 'resubstitution error'[4].

To analyze the extent to which a classifier can identify the tuples belonging to various classes in a better way, a 'confusion matrix' is used. A confusion matrix is actually a table with a size n × n where n represents number of classes[5]. when a classifier labels few tuples of class 'i' as class 'j' ,then the composite matrix it can be represented as $CN_{i,j}$ consisting of n rows and n columns, A classifier is supposed to be accurate ,if the composite matrix contains maximum number of tuples only on the diagonals of the matrix(i.e. from $CM_{1,1}$ to $CM_{n,n}$ entries and all the remaining entries are nearly zeroes. Additional rows or columns can also be added to the composite matrix as shown bellow.

Classes	**Student _pass= Yes**	**Student _pass=No**	**Total**	**Recognition %**
Student _pass=Yes Student_pass= No	212 56	88 369	300 425	70.67 86.82
Total	268	457	725	

Table 3: Measures the Accuracy of a Classifier

If 'student_pass=Yes'and 'student_pass=No'are two classes,then ,tuples present in the main class,i.e 'student_pass=yes'are refereed as 'positive tuples'and the tuples present in the ;student_pass=No'class are referred as 'negative tuples'[7].

When a classifier labels the positive tuples without any error then they are called false-negative and when classifier lables the negative tuples without any errors then they are called False _positives[7]. the corresponding confusion matrix for the positive and negative tuples is shown bellow.

C1	C2
C1 true positives C2 false Positive	False negative True negative

The extents to which the tuples can be recognized I \s measures using the following:

1. Sensitivity or the true positive recognition rate

2. Specificity or the true negative recognition rae.

Thus, Sensitivity $=t_p/p$ and Specified $=t_n/n$

Moreover, precision is used to retrieve the percentage of tuples which are labeled incorrectly

i.e., Precision $=t_p / t_p+f_p$

where

t_p=Number of true positives

t_n = Number of true negatives

n = Number of negative tuples

p = Number of positives tuples

Thus, the accuracy which is a function of both sensitivity and specifiecity measures can be given as ,

Accuracy=Sensitivity p/(p+n) + Specificity n/(p+n)

When the cost and benefits are considered, then the accuracy of a classifiers can be calculated by finding the average of cost or benefit for each tuple. Then accuracy is given as,

Accuracy=sensitivity

P/avg(cost or benefit)+Specificity n/avg(cost or benefit)

III. PROCEDURE FOR MEASURING PREDICTION ACCURACY

Consider a test set T^s such that,

$T^s=\{(P_1,Q_1)(P_2,Q_2)\ldots(P_t, Q_t)\}$

Where P_i represents n-dimensional test tuples related to q_i known values (where q represented response variable) and t denotes number of test tuples in the test set $T^{s.}$

Here, it is difficult to guess whether the predicted value q_i is correct for its associated value p_i . This is because the predictors yields continuous value instead of categorical labled [4].Hence ,rather than concentrating on exact value of procedure aims at knowing how far the predicted value is from the actual known value[3]. For this a loss function is used in order to calculated the error between q_i and the predicted value $q_{i.}$ The two most common loss function used in practice are as follows,

Absolute error : $|q_i-q_i|$

Squared error : (q_i-q_i)

Depending on the above loss functions the test error rate or generazlition error gives the average loss that occurs over the test set. Therefore, the error rates obtained are as follows,

Mean absolute error: $\sum_{i=0}^{t} |qi - qi|$

Mean Squared error: $\sum_{i=0}^{t} |qi - qi|^2$

In the above error rates, the mean squared error is capable in greatly notifying the occurrence of outliers, whereas the mean absolute error cannot notify them, also calculating square root of mean squared error would yield error measure known as “root mean squared error”. This “root mean squared error “helps in making sure that the magnitude of measures error is equivalent to the magnitude of the predicted value [3].

Furthermore, it is possible that the error relative to the predicted value q^- should d be measured for the mean value q i.e., the total loss can be normalized by dividing it with the total loss incurred from predicting the value of mean[4].

The two relative error rates available are as follows,

$$\text{Relative absolute error: } \frac{\sum_{i=0}^{t} |qi - qi|}{\sum_{i=0}^{t} |qi - qi|^2}$$

$$\text{Relative squared error: } \frac{\sum_{i=0}^{t} |qi - qi|^2}{\sum_{i=0}^{t} |qi - qi|^2}$$

Here q- represents the mean value for the q from the trained data T such that,
Also, calculating the root of relative squared error provides root relative squared error, which helps in obtaining the same magnitude for both, resulting error and predicted value.

IV. CONCLUSION

Accuracy and Error Measures on confusion matrix of a classifier and procedure for measuring prediction accuracy of a classifier the percentage of a set of tuples that used to be tested and accurately classified by the classifier is calculated for different domains which is used in the data mining . The accuracy also corresponds to the classifier‘s total ‘recognition rate ‘and explains the extent of recognizing records belonging to different classes by the classifier.

REFERENCES

[1] J. A. Benediktsson, M. Pesaresi, and K. Arnason, “Classification and feature extraction for remote sensing images from urban areas based on morphological transformations,” IEEE Trans. Geosci. Remote Sens., vol. 41, no. 9, pp. 1940–1949, Sep. 2003.

[2] L. Bruzzone and L. Carlin, “A multilevel context-based system for classification of very high spatial resolution images,” IEEE Trans. Geosci. Remote Sens., vol. 44, no. 9, pp. 2587–2600, Sep. 2006.

[3] Model predictive control of a Doubly Fed Induction Generator with an Indirect Matrix Converter M. Rivera; J. L. Elizondo; M. E. Macias; O. M. Probst; O. M. Micheloud; J. Rodriguez; C. Rojas; A. Wilson.

[4] Hybrid discrete event simulation with model predictive control for semiconductor supply-chain manufacturing .H.S. Sarjoughian; Dongping Huang; G.W. Godding; Wenlin Wang; D.E. Rivera; K.G. Kempf; H.D. Mittelmann

[5] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed. New York: Academic, 1990.
[6] D. A. Landgrebe, Signal Theory Methods in Multispectral Remote Sensing. Hoboken, NJ: Wiley, 2003.
[7] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," IEEE Trans. Pattern Anal. Machine Intell., vol. 15, pp.388–400, Apr . 1993.
[8] "Decision boundary feature extraction for neural networks," IEEE Trans. Neural Networks, vol. 8, pp. 75–83, Jan. 1997.