IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Comparison of the Performance Evaluations in Classification

So Jung Shin[1], Hyeuk Kim[2], Sang-Tae Han[3]

Master's Student, Department of Applied Statistics, Hoseo University, Asan, Korea [1]

Assistant Professor, Department of Applied Statistics, Hoseo University, Asan, Korea [2]

Professor, Department of Applied Statistics, Hoseo University, Asan, Korea [3]

Abstract: This document gives formatting instructions for authors preparing papers for publication in the Proceedings of an International Journal. The authors must follow the instructions given in the document for the papers to be published. You can use this document as both an instruction set and as a template into which you can type your own text.

Keywords: Include at least 4 keywords or phrases.

I. INTRODUCTION

Data mining is a process for finding relevant information through large amounts of data [1]. Data mining technique is broadly used in every area. Especially in recent years, the range of its application is widened with the era of Big Data. There are four major techniques in data mining: classification, cluster analysis, association analysis, and anomaly detection [2].

Classification is one of the methods in predictive modelling. Predictive modelling is to learn a function that maps each attribute set to a target variable. Both classification and regression belong to predictive modelling. The characteristic of a target variable determines whether predictive modelling is classification or regression. Predictive modelling is classification if a target variable has the predefined class labels which class is categorical. Predictive modelling is classified as regression if a target variable is a quantitative variable.

Many classification methods have been developed as data mining approach becomes popular and new classification methods will be developed continuously in the future since the best classification method cannot exist over all data. Wolpert has derived no free lunch theorems in [3]. A specific classification method can work best on a certain data set, but other method may work best on a different data set. Therefore, it is very important to find which technique works better than others on a specific certain situation [4].

To compare the performances of classification methods, we need to judge the performance exactly. There are several factors which we have to consider in evaluating the performance since the performance of a classification method is a compound characteristic [5]. Accuracy is the most basic measure to evaluate the performance, but there are many other performance evaluation measures.

Comparison of the performance evaluation measures themselves is rarely conducted on research though many classification methods have been developed and their performances have been compared. Therefore, we compare several performance evaluation measures in classification and test them with data sets in the paper. The structure of the article is as follows. We describe several performance evaluation measures in the next section. In the last section, we make a conclusion and describe a future work.

II. PERFORMANCE EVALUATION MEASURES IN CLASSIFICATION

There are many measures for evaluating the performance in classification. Accuracy is the basic measure, but many other measures are also developed. Some measures such as accuracy, recall, and precision are derived from the same tool, a confusion matrix. Other measures are developed from the different concepts. All of them are described one by one in the section.

A. Accuracy and the error rate

Accuracy is the ratio between the number of the instances classified correctly and the number of the while instances. It is the most popular measures for evaluating the performance in classification because of its simplicity and meaning. The error rate is the measure which has the reverse definition against accuracy. It is the ratio between the number of the instances classified incorrectly and the number of the while instances.

$$\text{Accuracy} = \text{C}/\text{N}$$
$$\text{Error Rate} = \text{I}/\text{N}$$

where C is the number of the instances classified correctly, I is the number of the instances classified incorrectly, and N is the number of the whole instances, $\text{C} + \text{I}$.

They can be described from an alternative formula. A confusion matrix which is also known as an error matrix [6] is used for explanation of accuracy and error rate. Suppose that there are two classes in a target variables:

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

plus(+) and minus(-). Each entry in the matrix is the number of the instances which satisfy a specific situation.

Table 1 A confusion matrix

		Predicted Class	
		+	-
Actual Class	+	n_{++}	n_{+-}
	-	n_{-+}	n_{--}

The measures can be described as follows.

$$\text{Accuracy} = \frac{n_{++} + n_{--}}{n_{++} + n_{+-} + n_{-+} + n_{--}}$$

$$\text{Error Rate} = \frac{n_{+-} + n_{-+}}{n_{++} + n_{+-} + n_{-+} + n_{--}}$$

$$\text{Accuracy} + \text{Error Rate} = 1$$

B. Performance evaluation measures derived from a confusion matrix

The numbers of the classes in a target variable are sometimes imbalanced and we focus on the minority class. In the case, accuracy is not an appropriate measure in evaluating performance of the classification method. Suppose that the number of the majority class is 95, and the number of the minority class is 5 in the instances. A certain classification method always classifies an unknown instance as the majority class. Its accuracy is 95 percent and seems to be high. Can we agree that the classification method is good? We introduce other measures to supplement the limitation of the accuracy measure in the situation. They are derived from a confusion matrix which is described in Table 1. The minority class is usually described as the positive class, while the majority class is described as the negative class.

True positive rate is the number of positive instances classified correctly divided by the number of actual positive instances. It is also called sensitivity.

$$True\ Positive\ Rate = \frac{n_{++}}{n_{++} + n_{+-}}$$

True negative rate or specificity is the number of negative instances classified correctly divided by the number of actual negative instances.

$$True\ Negative\ Rate = \frac{n_{--}}{n_{-+} + n_{--}}$$

There are the measures which focus on the number of the instances which are classified incorrectly.

False positive rate is the number of negative instances classified incorrectly divided by the number of actual negative instances.

$$False\ Positive\ Rate = \frac{n_{-+}}{n_{-+} + n_{--}}$$

Finally, false negative rate is the number of positive instances classified incorrectly divided by the number of actual positive instances.

$$False\ Negative\ Rate = \frac{n_{+-}}{n_{++} + n_{+-}}$$

C. Precision and recall

Precision and recall [7] are very useful for an imbalanced data set and also derived from a confusion matrix.

$$Precision = \frac{n_{++}}{n_{++} + n_{-+}}$$

$$Recall = \frac{n_{++}}{n_{++} + n_{+-}}$$

Precision is the number of positive instances classified correctly divided by the number of the instances predicted positively. Recall is the number of positive instances classified correctly divided by the number of actual positive instances. Therefore, recall is equivalent to sensitivity.

Both measures are usually calculated together and move reversely. Suppose that a classification method always predicts any instance into positive class. Its recall is 1, the highest value, but the precision is low. Conversely, there is a classification method which predicts only one instance which looks definitely positive into positive class. The precision is 1, but the recall is low.

D. F_1 score

We use precision and recall together when we consider them as the performance evaluation measures. However, their values move reversely and it is convenient to use one measure instead of multiple measures for evaluation. Therefore, F_1 score has been developed which is the harmonic mean of precision and recall.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \times n_{++}}{2 \times n_{++} + n_{-+} + n_{+-}}$$

High F_1 score means that both precision and recall are somewhat high because the measure is the harmonic mean of them. The harmonic mean is one of the Pythagorean means and has a tendency to be the value which is closer to the smaller value between two numbers. For example, there are 2 and 4. The ordinary mean, an arithmetic mean, is $\frac{2+4}{2} = 3$. The harmonic mean is $\frac{1}{1/2+1/4} = \frac{4}{3} \approx 1.33$ which is closer to 2.

There is the general form of F_1 score which is called F_β score.

$$F_\beta = \frac{(\beta^2 + 1) \times precision \times recall}{\beta^2 \times precision + recall}$$

Then, precision and recall are special cases in F_β. We figure out precision when β is zero and recall when β goes to infinity. Two specific F_β except F_1 are additionally used among many possible F_β scores. $F_{0.5}$ score emphasizes precision than recall, and F_2 score emphasizes recall than precision. F_β can be derived from an effectiveness

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

measure [8]. The relationship between F_β and an effectiveness measure is described as below.

$$F_\beta = 1 - E$$

where $E = 1 - \left(\frac{\alpha}{P} + \frac{1-\alpha}{R}\right)^{-1}$ and $\alpha = \frac{1}{1+\beta^2}$

E. Youden's J statistic

It is a single measure to evaluate the performance in classification [9]. The formula is as follows.

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

Its range is between -1 and +1. J statistic must be greater zero if the classification method works. Zero of J statistic means that the classification method has a 50 percent chance to predict an instance with positive class correctly and a 50 percent chance to predict an instance with positive class incorrectly. The classification method works perfectly if J statistic is 1 since its sensitivity is 1 and the specificity is also 1.

Youden's J statistic is interpreted in view of ROC (Receiver Operating Characteristic) curve which is described later [10]. We can modify the formula as below.

$$\begin{aligned} J &= \text{Sensitivity} + \text{Specificity} - \\ &= \text{Sensitivity} - (1 - \text{Specificity}) \end{aligned}$$

The left term is the value at y-axis and the right term is the value at x-axis on ROC chart. Therefore, it is the difference between two values which are located on x-axis and y-axis respectively on ROC chart. Graphically, it is the difference between ROC curve and the diagonal line on ROC chart.

F. Kappa measure

The kappa measure is a tool which compares an observed accuracy with an expected accuracy that happens by chance. It is very useful for comparing the performances of the multiple classification methods. An observed accuracy is derived as follow. We describe the kappa measure in considering positive class since positive class (minority class) is usually more important than negative class (majority class).

$$p_o = \frac{n_{++} + n_{--}}{N} \text{ where } N = n_{++} + n_{+-} + n_{-+} + n_{-}$$

An expected accuracy is the mean between the ratio of positive class in actual class and the ratio of positive class in predicted class.

$$\begin{aligned} p_e &= \left(\frac{n_{++} + n_{+-}}{N} + \frac{n_{++} + n_{-+}}{N}\right)/2 \\ &= \frac{n_{++} + (n_{+-} + n_{-+})/2}{N} \end{aligned}$$

The kappa measure is defined by p_o and p_e.

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

The kappa measure which is described is called Cohen's kappa [11]. It is a statistic for a binary class. There is also the kappa measure for multi class. It is called Fleiss' kappa [12] whose concept is same with Cohen's kappa. For interpretation, Landis and Koch [13] suggested the guideline which is described in Table 2 for kappa.

y a specific situation.

Table 2 An interpretation of kappa measure

Kappa value	Degree of agreement
Less than 0	Poor
0.01~0.20	Slight
0.21~0.40	Fair
0.41~0.60	Moderate
0.61~0.80	Substantial
0.81~1.00	Almost perfect

We have to be cautious to apply a rule in Table 2 since the guideline is not determined by logical reason. Each range and the corresponding description are determined by experience.

G. Matthews correlation coefficient

Matthews developed Matthews correlation coefficient to evaluate the performance of a classification method [14]. MCC (Matthews correlation coefficient) is derived directly from a confusion matrix.

$$MCC = \frac{TP/N - SP}{\sqrt{PS(1-S)(1-P)}}$$

where

$$N = n_{++} + n_{+-} + n_{-+} + n_{-}$$
$$S = \frac{n_{++} + n_{+-}}{N}$$
$$P = \frac{n_{++} + n_{-+}}{N}$$

MCC is also expressed by χ^2 statistic from a confusion matrix.

$$|MCC| = \sqrt{\frac{\chi^2}{N}} \quad \text{where } N = n_{++} + n_{+-} + n_{-+} + n_{-}$$

The possible range of MCC is between -1 and +1. +1 means that a classification method predicts all observations correctly, 0 means that its performance is same with the performance of a random choice. It predicts all observations wrong if MCC is -1.

H. ROC curve and AUC

ROC (Receiver Operation Characteristic) curve is the graph with 1-specificitiy as x-axis and sensitivity as y-axis. ROC curve is especially useful when we compare the performances of the multiple classification methods visually. It is a trade-off between true positive rate (sensitivity) and false positive rate. True positive rate increases if the false positive rate decreases and vice versa.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

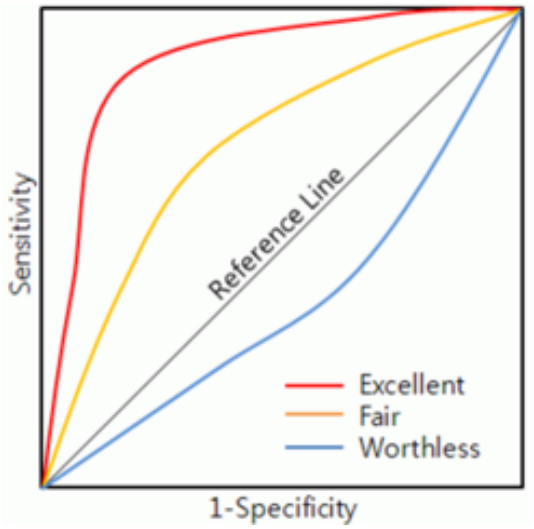

Fig. 2 ROC chart

A reference line in Figure 1 means the performance at random. Therefore, a classification model is valid if ROC curve of the specific classification method is located above a reference line like red and yellow lines in Figure 1. A classification model is worthless if its ROC curve is drawn below a reference line like blue line in Figure 1.

ROC has been first developed during World War II for signal detection. Then, this tool is used broadly in such areas as psychophysics, epidemiology, medical research, and machine learning.

AUC (Area Under the Curve) is the area under curve on ROC chart. AUC is the measure how a classification method performs well. A classification method is perfect if AUC is 1.0 which means that specificity is 1 and sensitivity is 1. AUC of a reference line is a half since the area is a triangle of which a hypotenuse is a reference line in the rectangle. The exact abbreviation is AUROC (Area under the Receiver Operating Characteristic Curve) because AUC is ambiguous.

I. Gain and Lift charts

The measures which we describe so far are derived from a confusion matrix. The concepts of the next two measures are developed without a confusion matrix.

Gain or Lift is a measure of the performance evaluation of a classification model which is the ratio between the number of the instances predicted by a model and the number of the instances predicted randomly. A cumulative lift chart is defined as the cumulative response rate divided by the overall response rate. They consider a part of data instead of whole data. The measure is common on in Marketing where we would like to know the performance for parts of customers. The budget or other standard determine how many customers are considered. We usually focus on the cumulative response rate for the top decile or the cumulative response rate for the top two deciles if we use the specific value instead of the chart.

III. CONCLUSION AND A FUTURE WORK

Classification is one of the most popular areas in data mining; so many classification methods are developed. There is no classification method which is superior over other methods in any cases because of no free lunch theorem. Therefore, it is important to evaluate how a classification method performs in a specific situation to find the better or best tool in that case. We introduce many measures for evaluating the performances of the classification methods in the paper. A confusion matrix is a fundamental tool to denote the performance of the classification method and many of performance measures are derived from a confusion matrix.

We explain the characteristics of each measure briefly, but we will compare several measures introduced in the paper directly for a future work. Through comparison of measures in a theoretical view and an experimental view, we will recognize the pros and cons of the evaluation measures in classification in detail.

ACKNOWLEDGMENT

The paper is based on the draft of So Jung Shin's dissertation. She is 2nd year graduate student and will graduate on Summer 2017.

REFERENCES

[1] D. J. Hand, "Statistics and data mining: Intersecting disciplines," ACM SIGKDD Explorations Newsletter, vol. 1(1), pp. 16–19, Jun. 1999.
[2] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, 1st ed., Pearson, May 2005.
[3] D. Wolpert, "The lack of a priori distinctions between learning algorithms," Neural Computation, vol. 8(7), pp. 1341–1390, Oct. 1996.
[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, 1st ed., Springer, Aug. 2013.
[5] L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, 1st ed., Wiley-Interscience, Jul. 2004.
[6] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," Remote Sensing of Environment, vol. 62 (1), pp. 77–89, 1997.
[7] J. W. Perry, A. Kent, and M. M. Berry, "Machine literature searching X. Machine language; factors underlying its design and development," Journal of the Association for Information Science and Technology, vol. 6 (4), pp. 242–254, 1955.
[8] C. J. Van Rijsbergen, Information Retrieval, 2nd ed., Butterworth-Heinemann, Mar. 1979.
[9] W. J. Youden, "Index for rating diagnostic tests," Cancer, vol. 3, pp. 32–35, 1950.
[10] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies, vol. 2(1), pp. 37–63, 2011.
[11] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20(1), pp. 37–46, 1960.
[12] J. L. Fleiss, "Measuring nominal scale agreement among many raters," Psychological Bulletin, vol. 76(5), pp. 378–382, 1971.
[13] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33(1), pp. 159–174, 1977.
[14] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochimica et Biophysica Acta – Protein Structure, vol. 405 (2), pp. 442–451, 1975.