UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 8, August 2017

A Comparison of Performance Measures for Classification Methods in Data Mining

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

Master's Student, Division of Big Data and Management Engineering, Hoseo University, Asan, Korea[1]

Assistant Professor, Division of Big Data and Management Engineering, Hoseo University, Asan, Korea[2]

Professor, Division of Big Data and Management Engineering, Hoseo University, Asan, Korea[3]

Abstract: The machine learning area is being developed much as the artificial intelligence develops. It is important to evaluate the performances of the classification methods correctly since many techniques have been developed so far. Many performance measures are also developed for evaluation of the performance of the classification method. The values of the performance measures are changed under the different situations. We review several evaluation measures derived from a confusion matrix in the paper. Then, we investigate the change of the values of the performance measures under the various situations.

Keywords: Classification, Performance, Evaluation, Confusion Matrix.

I. INTRODUCTION

Data mining, which is also called machine learning or artificial intelligence nowadays, is a process to figure out the pattern or useful information through data itself. Classification is one of the major areas in data mining. Many classification methods have been developed and the development of computer accelerates its research. In classification research, no free lunch theorem is applied. It has been introduced by Wolpert and Macready first in 1997. It describes that there is no best classification method which beats others in any situations. Therefore, we need to findwhich a classification technique is better than others in a specific situation. Many performance measures have been developed for the purpose. Machine learning researchers have to know their characteristics thoroughly to use them in the appropriate situation. The paper consists of three chapters excluding the first chapter for introduction. We describe the definitions and the formulas of the performance measures in the chapter two. All of them are used in the next chapter. In the third chapter, we make an experiment and compare the performance measures for the various situations. We make a conclusion in the last chapter.

II. MEASURES

There are many measures to evaluate the performance for the classification method. In the section, we introduce the performance measures which are used in the paper. They are derived from a confusion matrix. First, we define a confusion matrix.

TABLE I A CONFUSION MATRIX

Confusion Matrix		**Predicted Class**	
		Positive	**Negative**
Real Class	**Positive**	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

We define several performance measures for evaluation from the above matrix. A basic performance measure is an accuracy which is the ratio between the number of the correct predicted observations and the number of the whole observations.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

Sensitivity and specificity are the performance measures which focus on a partial class.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 8, August 2017

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

An accuracy is very useful when the dataset is balanced, but it misleads the performance of the classification method when the dataset is unbalanced. We introduce two other evaluation measures which work well when the dataset is unbalanced: precision and recall. Their values are changed in the opposite direction.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$Recall = \frac{TP}{TP + FN}$$

Precision and recall were developed for the special case but it is inconvenient to use two measures to evaluate the performance for the classification method. Therefore, a new measure was developed and was the harmonic mean between precision and recall. It is called F_1 or F_1 score.

$$F_1 = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

$$= \frac{2}{\frac{1}{Precision} + \frac{1}{Sensitivity}}$$

$$= \frac{2 \times TP}{2 \times TP \times FP \times FN}$$

$\kappa$is the last measure which is used in the paper. It compares the accuracy from the classification method with the accuracy by chance. It is calculated from two accuracies which are mentioned in the above sentence.

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

p_o is the accuracy from the classification method and p_e is the accuracy that happens by chance. The exact equations for two accuracies are as follows.

$$p_o = \frac{TP + TN}{N} \; where \; N = TP + FN + FP + TN$$

$$p_e = \left(\frac{TP + FN}{N} + \frac{TP + FP}{N}\right)/2 = \frac{TP + (FN + FP)/2}{N}$$

III. SIMULATION

We investigate how the values of the performance measures are changed in the section. There are two types of the performance measures: One type of the performance measures is derived from a confusion matrix and another type of the performance measures is not related to a confusion matrix. We only handle the performance measures derived from a confusion matrix in the paper. Three parameters are used to represent the various situations.

TABLE II PARAMETERS TO EXPLAIN A SITUATION

Word	Description	Notation
$Accuracy_p$	The accuracy for the observations with a positive class	a_p
$Accuracy_n$	The accuracy for the observations with a negative class	a_n
Balance	The degree of balance about classes.	b

Balance is the number of the observations with a positive class over the number of the whole observations.1-b is the number of the observations with a negative class over the number of the whole observations. All parameters have the values between zero and one. A confusion matrix is described in Table II when the number of the whole observations is N.

TABLE IIII A CONFUSION MATRIX

Confusion Matrix		Predicted Class	
		Positive	Negative
Real Class	Positive	Nba_p	$Nb(1 - a_p)$
	Negative	$N(1 - b)(1 - a_n)$	$N(1 - b)a_n$

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 8, August 2017

The major performance measures are from a confusion matrix in Table II as follows.

$$Accuracy = \frac{Nba_p + N(1-b)\mathrm{a}_n}{Nba_p + Nb(1-a_p) + N(1-b)(1-a_n) + N(1-b)a_n}$$
$$= \frac{Nba_p + N(1-b)a_n}{N}$$
$$= ba_p + (1-b)a_n$$

Therefore, the accuracy means the weighted average between the accuracy for the observations with a positive class and the accuracy for the observations with a negative class.

$$Sensitivity = \frac{Nba_p}{Nba_p + Nb(1-a_p)} = a_p$$

$$Specificity = \frac{N(1-b)a_n}{N(1-b)(1-a_n) + N(1-b)a_n} = a_n$$

a_p in parameters is a sensitivity and a_n in parameters is a specificity.

$$Precision = \frac{Nba_p}{Nba_p + N(1-b)(1-a_n)} = \frac{ba_p}{ba_p + (1-b)(1-a_n)}$$

$$Recall = Precision = a_p$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$
$$= \frac{2Nba_p}{2Nba_p + Nb(1-a_p) + N(1-b)(1-a_n)}$$
$$= \frac{2ba_p}{b + ba_p + (1-b)(1-a_n)}$$

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

$$p_o = Accuracy = ba_p + (1-b)a_n$$

$$p_e = \frac{Sensitivity + Precision}{2} = \frac{Nba_p + \{Nb(1-a_p) + N(1-b)(1-a_n)\}/2}{N}$$

$$= \frac{1}{2}\{2ba_p + b(1-a_p) + (1-b)(1-a_n)\} = \frac{1}{2}\{b + ba_p + (1-b)(1-a_n)\} = \frac{ba_p}{F_1}$$

3.1. Balanced Dataset ($b = 0.5$)

We investigate the values of the performance measures when the dataset is balanced. The balanced dataset means that the number of the observations with positive class and the number of the observations with negative class are same in the dataset. We assume four situations based on the different accuracies. Each situation is as follows. The classification model is appropriate when the accuracy is 0.7, good when the accuracy is 0.9, same with a random choice when the accuracy is 0.5, and bad when the accuracy is 0.3.

3.1.1. Accuracy=0.7

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.5	0.900	0.500	0.700	0.900	0.500	0.643	0.900	0.750	0.250
0.5	0.700	0.700	0.700	0.700	0.700	0.700	0.700	0.700	0.400
0.5	0.500	0.900	0.700	0.500	0.900	0.833	0.500	0.625	0.500

Acc, Sen, Spe, Pre, and Rec are the abbreviations for an accuracy, a sensitivity, a specificity, a precision, and a recall, respectively.

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 8, August 2017

3.1.2. Accuracy=0.9

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.5	0.990	0.810	0.900	0.990	0.810	0.839	0.990	0.908	0.780
0.5	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.900	0.800
0.5	0.810	0.990	0.900	0.810	0.990	0.988	0.810	0.890	0.817

3.1.3. Accuracy=0.5

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.5	0.700	0.300	0.500	0.700	0.300	0.500	0.700	0.583	-0.250
0.5	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.000
0.5	0.300	0.700	0.500	0.300	0.700	0.500	0.300	0.375	0.167

3.1.4. Accuracy=0.3

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.5	0.500	0.100	0.300	0.500	0.100	0.357	0.500	0.417	-0.750
0.5	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300	-0.400
0.5	0.100	0.500	0.300	0.100	0.500	0.167	0.100	0.125	-0.167

The values of precision and recall are changed in the opposite direction when the dataset is balanced. The values of an accuracy, F_1, and κ are changed in the same direction. The situation with a high sensitivity is better even though accuracies are same in several situations. The direction of F_1 follows the direction of a sensitivity but κ has the opposite direction.

3.2. Unbalanced Dataset ($b = 0.2$)

Many classification methods and evaluation measures work well when the distribution of classes in the dataset are is balanced. However, many datasets are unbalanced in the social and natural environment. Especially the number of the observations with positive class is usually less than the number of the observations with negative class. We investigate the values of several performance measures in the case. There is no clear definition which distinguishes balanced data from unbalanced data. Researchers usually define that the dataset is unbalanced when the degree of imbalance, b, is between 0.1 and 0.2. In the paper, we use $b = 0.2$ as the status of unbalanced dataset. The proportion between the number of the observations with positive class and the number of the observations with negative class is 2:8. We set four situations like the section 3.1.

3.2.1. Accuracy=0.7

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.2	0.990	0.628	0.700	0.900	0.628	0.399	0.990	0.569	0.540
0.2	0.700	0.700	0.700	0.700	0.700	0.368	0.700	0.483	0.577
0.2	0.100	0.850	0.700	0.100	0.850	0.143	0.100	0.118	0.639

3.2.2. Accuracy=0.9

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.2	0.990	0.878	0.900	0.990	0.878	0.669	0.990	0.798	0.867
0.2	0.900	0.900	0.900	0.900	0.900	0.692	0.900	0.783	0.870
0.2	0.540	0.990	0.900	0.540	0.990	0.931	0.540	0.684	0.881

3.2.3. Accuracy=0.5

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.2	0.990	0.378	0.500	0.990	0.378	0.284	0.990	0.442	0.094
0.2	0.500	0.500	0.500	0.500	0.500	0.200	0.500	0.286	0.231
0.2	0.100	0.600	0.500	0.100	0.600	0.059	0.100	0.074	0.315

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 8, August 2017

3.2.4. Accuracy=0.3

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.2	0.990	0.128	0.300	0.990	0.128	0.221	0.990	0.361	-0.549
0.2	0.300	0.300	0.300	0.300	0.300	0.097	0.300	0.146	-0.186
0.2	0.100	0.350	0.300	0.100	0.350	0.037	0.100	0.054	-0.111

The characteristic of the case is as follows. The values of precision and recall are changed in the opposite direction like the balanced dataset. The values of an accuracy, F_1, and κ are changed in the same direction. F_1 has a high value and κ has a low valuewhen its sensitivity is high.

When we compare the performance measures for the unbalanced dataset with the performance measures for the balanced dataset, the value of $F_1$for the unbalanced dataset is much less than the value of F_1forthe balanced dataset and the value of $\kappa$for the unbalanced dataset is much higherthan the value of κforthe balanced dataset. Therefore

3.3. Unbalanced Dataset ($b = 0.8$)

As we mention in the section 3.2, many datasets are unbalanced in the social and natural environment. The number of the observations with positive class is usually less than the number of the observations with negative class. In the section, we investigate the values for the performance measures when the number of the observations with positive class is higher than the number of the observations with negative class even though the situation is abnormal.

3.3.1. Accuracy=0.7

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.8	0.850	0.100	0.700	0.850	0.100	0.791	0.850	0.819	-0.765
0.8	0.700	0.700	0.700	0.700	0.700	0.903	0.700	0.789	-0.034
0.8	0.628	0.990	0.700	0.628	0.990	0.996	0.628	0.770	0.138

3.3.2. Accuracy=0.9

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.8	0.990	0.540	0.900	0.990	0.540	0.896	0.990	0.941	0.367
0.8	0.900	0.900	0.900	0.900	0.900	0.973	0.900	0.935	0.565
0.8	0.878	0.990	0.900	0.878	0.990	0.997	0.878	0.934	0.597

3.3.3. Accuracy=0.5

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.8	0.600	0.100	0.500	0.600	0.100	0.727	0.600	0.658	-0.852
0.8	0.500	0.500	0.500	0.500	0.500	0.800	0.500	0.615	-0.429
0.8	0.378	0.990	0.500	0.378	0.990	0.993	0.378	0.547	-0.116

3.3.4. Accuracy=0.3

b	a_p	a_n	Acc	Sen	Spe	Pre	Rec	F_1	κ
0.8	0.350	0.100	0.300	0.350	0.100	0.609	0.350	0.444	-0.892
0.8	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.407	-0.707
0.8	0.128	0.900	0.300	0.128	0.990	0.981	0.128	0.226	-0.277

The pattern of the performance measures in the section 3.3 are same with the pattern of the performance measures when b = 0.5 and b = 0.2. However, the value of F_1 from the unbalanced dataset with b = 0.8is greater than the value of F_1 from the balanced dataset. The value of κ from the unbalanced dataset with b = 0.8is less than the value of κ from the balanced dataset.

The value of F_1 from the unbalanced dataset with b = 0.2is less than the value of F_1 from the balanced dataset. The value of κ from the unbalanced dataset with b = 0.2is greater than the value of κ from the balanced dataset. Therefore, κ is the good measure for evaluating the performance of the classification method when it classifies the dataset with positive minor class.

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 8, August 2017

IV. CONCLUSION

There are several measures for evaluating the performance of the classification method. For example, an accuracy is widely used since it is easy to calculate and understand. However, it is useless when the dataset is unbalanced. An accuracy, sensitivity, specificity, recall, precision, F_1, and κ are derived from a confusion matrix. They are classified into two categories. Sensitivity, specificity, precision, and recall are the measures which evaluate the performance for only one class. Accuracy, F_1, and κ are the comprehensive measures which consider all classes for evaluation of the classification method. We investigate the change of the performance measures under the various simulation cases. We find that the direction of the change of each measure can be different based on the situation. Therefore it is recommended to use several measures for evaluating the performance of the classification method correctly.

ACKNOWLEDGMENT

The paper is based on the So Jung Shin's dissertation. She is 2nd year graduate student and will graduate on August 18 2017.

REFERENCES

[1] J. L. Fleiss, "Measuring nominal scale agreement among many raters," Psychological Bulletin, vol. 76(5), pp. 378-382, 1971.
[2] D. J. Hand, "Statistics and data mining: Intersecting disciplines," ACM SIGKDD Explorations Newletter, vol. 1(1), pp. 16-19, Jun. 1999.
[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, 1st ed., Springer, Aug. 2013.
[4] L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithm, 1sted., Wiley-Interscience, Jul. 2004.
[5] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33(1), pp. 159-174, 1977.
[6] 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.
[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] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," Remote Sensing of Environment, vol. 62 (1), pp. 77–89, 1997.
[9] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, 1st ed., Pearson, May 2005.
[10] C. J. Van Rijsbergen, Information Retrieval, 2nd ed., Butterworth-Heinemann, Mar. 1979.
[11] D. Wolpert, "The lack of a priori distinctions between learning algorithms," Neural Computation, vol. 8(7), pp. 1341–1390, Oct. 1996.