J. Chosun Natural Sci.
Vol. 11, No. 3 (2018) pp. 161 – 164
https://doi.org/10.13160/ricns.2018.11.3.161

Reconsideration of F1 Score as a Performance Measure in Mass Spectrometry-based Metabolomics

Jaesik Jeong[†], Han Sol Kim, and Shin June Kim

Abstract

Over the past decade, mass spectrometry-based metabolomics, especially two dimensional gas chromatography mass spectrometry (GCxGC/TOF-MS), has become a key analytical tool for metabolomics data because of its sensitivity and ability to analyze complex biological or biochemical sample. However, the need to reduce variations within/between experiments has been reported and methodological developments to overcome such problem has long been a critical issue. Along with methodological developments, developing reasonable performance measure has also been studied. Following four numerical measures have been typically used for comparison: sensitivity, specificity, receiver operating characteristic (ROC) curves, and positive predictive value (PPV). However, more recently, such measures are replaced with F1 score in many fields including metabolomics area without any carefulness of its validity. Thus, we want to investigate the validity of F1 score on two examples, with the goal of raising the awareness in choosing appropriate performance comparison measure. We noticed that F1 score itself, as a performance measure, was not good enough. Accordingly, we suggest that F1 score be supplemented with other performance measure such as specificity to improve its validity.

Keywords: F1 score, Mass spectrometry, Metabolomics, Positive predictive value, Sensitivity, Specificity

1. Introduction

Metabolomics, which aims to understand cellular processes through chemical fingerprints, is the scientific study of the chemical processes. Mass spectrometry-based metabolomics study, especially mass spectrometry coupled with two dimensional gas chromatography (GCxGC/TOF-MS) has been widely used due to its various advantages: increased separation capacity, its sensitivity and ability to analyze complex mixtures. However, the output of GCxGC experiment is still subject to errors such as within/between-experiment variations. Thus, reducing such variations has been a crucial issue. More precisely, various preprocessing steps including metabolite identification and peak alignment has attracted researchers' attention. As a result, large numbers of methodologies have been developed by many researchers[1-12].

Along with the development of various methodologies, comparison studies have gained much attention as well. For comparison, well-known standard measures were considered: sensitivity, specificity, receiver operating characteristic (ROC) curves, and positive predictive value (PPV). However, most recently, these classical performance measures were replaced with F1 score in various fields including metabolomics[2-7,13-15]. But, we think that it is time to ask the following question to ourselves before it is too late: is F1 score good enough to replace such well-known standard measures? In this paper, we try to answer the question. To this end, we investigate the properties of F1 score and relationship among other measures with two illustrating examples. Therefore, the goal of this paper is to raise the awareness when using F1 score as a performance measure instead of standard measures.

2. Standard Performance Measures

Standard performance measures can be easily obtained by using 2 by 2 table (Table 1).

Sensitivity, specificity, positive predictive value (PPV), and F1 are defined as follows:

Department of Statistics, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, Korea

[†]Corresponding author : jjs3098@gmail.com

(Received : June 11, 2018, Revised : September 19, 2018,
Accepted : September 25, 2018

Table 1. Test outcomes based on the results by any method

		Gold standard	
		Positive	Negative
Test outcome	Positive	C11	C12
	Negative	C21	C22

$$Sensitivity = \frac{C11}{C11+C21},$$

$$Specificity = \frac{C22}{C12+C22},$$

$$PPV = \frac{C11}{C11+C12},$$

$$F1 = \frac{2 \cdot Sensitivity \cdot PPV}{Sensitivy + PPV}$$

where *PPV* is 1-*FDR* (false discovery rate) and *F*1 is harmonic mean of sensitivity and *PPV*. In addition, ROC curve is a plot of sensitivity against 1-*specificity*, i.e., a mixture of sensitivity and specificity in a sense.

3. Validity of *F*1 Score as an Accuracy Measure

We investigate the validity of F1 score as an accuracy measure. For comparison, we define a simple and naive overall accuracy measure (OA), proportion of correct test outcome (Table 1):

$$OA = \frac{C11+C22}{N}$$

where $N = C11 + C12 + C21 + C22$. Here, we focus on the variability of F1 score when OA does not change at all. To make OA fixed, we fix diagonal sum ($C11 + C22 = 80$) and $N = 100$. Table 2 provides F1 score and OA for three different cases, implying that F1 does change while OA does not. Graphical presentation of F1 scores for each combination of $C11(1 \leq C11 < 80)$ and $C12(1 \leq C12 < 20)$ are provided in Fig. 1.

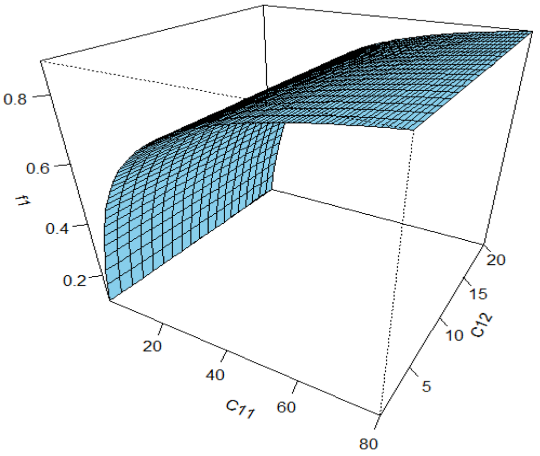

Fig. 1. Plot of f1 score for the combination of C11 and C12.

Unlike fixed OA (=0.8), F1 score varies too much from 0.09 to 0.89, implying that there is no relationship between F1 score and overall accuracy. Most importantly, it should be kept in mind that the value of *C*12 would not be involved in the calculation of F1.

4. Relationship between F1 Score and Specificity

We investigate the relationship between F1 score and specificity through a simple example. In Table 3, we fixed three cells.

As a result, sensitivity and F1 score are fixed (say 0.9 here). However, specificity and OA change along with the value of $1 \leq x \leq 100$. Fig. 2 shows how variable they are as *x* changes. From this figure, we notice that specificity varies from 0.167 to 0.995 while F1 score is fixed. Clearly, F1 score is not dependent on specificity.

Table 2. Variability of F1 score: each scenario has the same OA, but different F1 score.

	Case 1		Case 2		Case 3	
	Positive	Negative	Positive	Negative	Positive	Negative
Positive	10	10	6	6	2	2
Negative	10	70	14	74	18	78
F1	0.500		0.375		0.166	
OA	0.800		0.800		0.800	

Table 3. Relationship between specificity and F1: column presents the status of Gold standard and row presents the claimed status.

	Positive	Negative
Positive	C11 (=45)	C12 (=5)
Negative	C21 (=5)	C22 (=*x*)

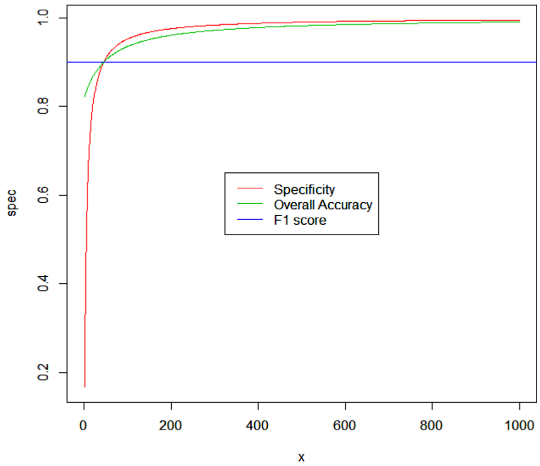

Fig. 2. Plot of specificity, overall accuracy, and F1 score

In other words, any information about specificity is not reflected in F1 score.

5. Conclusion

As a result of the increased popularity of spectrometry-based metabolomics study over the past decade, various methodologies dealing with preprocessing have been developed. To find better methodology, comparative studies have also been done by using standard numerical performance measures: sensitivity, specificity, and PPV. Recently, instead of such classical measures, F1 score gets popularity in various fields.

Just like controlling type I and type II errors is very important in classical testing problem, using proper performance measures combined with sensitivity and specificity is very crucial in comparative study. Thus, it is necessary to summarize good properties extracted from various performance measures in evaluating the performance of methodology. In this regard, we investigate the appropriateness of F1 score as a performance comparison measure on two examples and notice that F1 score is not directly related with both overall accuracy and specificity. In this respect, F1 does not represent other performance measures very well.

To conclude, we want to answer the question raised earlier: Is F1 score a good enough to replace such well-known standard measures? As can be seen by two examples, it is clear that F1 score is not good enough because (1) it does not reflect specificity at all and (2) it does not show any relationship with overall accuracy measure. As a suggestion, a performance measure, which is a mixture of F1 score and specificity, would be a good alternative, but not F1 score alone.

References

[1] C.G. Frage, B. Prazen, and R. Synovec, “Objective data alignment and chemometric analysis of comprehensive two dimensional separations with run-to-run peak shifting on both dimensions”, Anal. Chem., Vol. 73, p. 5833, 2011.

[2] J. Jeong, S. Xue, X. Zhang, S. Kim, and C. Shen, “An empirical Bayes model using a competition score for metabolite identification in gas chromatography mass spectrometry”, BMC Bioinformatics, Vol. 12, p. 392, 2011.

[3] J. Jeong, X. Xue, X. Xiang, S. Kim, and C. Shen “Model-based peak alignment of metabolic profiling from comprehensive two dimensional gas chromatography mass spectrometry”, BMC Bioinformatics, Vol. 13, p. 27, 2012.

[4] J. Jeong, X. Zhang, X. Shi, S. Kim, and C. Shen “An efficient post-hoc integration method improving peak alignment of metabolomics data from GCxGC/TOF-MS”, BMC Bioinformatics, Vol. 14, p. 123, 2013.

[5] S. Kim, A. Fang, B. Wang, J. Jeong, and X. Zhang “An optimal peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry using mixture similarity measure”, Bioinformatics, Vol. 27, p. 1660, 2011.

[6] S. Kim, I. Koo, A. Fang, X. Zhang “Smith-Waterman peak alignment for comprehensive two-dimensional gas chromatography-mass spectrometry”, BMC Bioinformatics, Vol. 12, p. 235, 2011.

[7] S. Kim, M. Ouyang, C. Shen, and X. Zhang “A new method of peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data”, Ann. Appl. Stat., Vol. 8, p. 1209, 2014.

[8] V. G. Mispelaar, A. C. Tas, A. K. Smilde, P. J. Schoenmakers, and A. C. Asten “Quantitative anal-

ysis of target components by comprehensive two-dimensional gas chromatography", J. Chromatogr. A, Vol. 1019, p. 15, 2003.

[9] C. Oh, X. Huang, F. Regnier, C. Buck, and X. Zhang "Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry peak sorting algorithm", J. Chromatogra., Vol. 1179, p. 205, 2008.

[10] K. Pierce, L. Wood, B. Wright, and R. Synovec "A comprehensive two-dimensional retention time alignment algorithm to enhance chemometric analysis of comprehensive two-dimensional separation data", Anal. Chem., Vol. 77, p. 7735, 2005.

[11] B. Wang, A. Fang, J. Heim, B. Bogdanov, S. Pugh, M. Libardoni, and X. Zhang "Disco: distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics", Anal. Chem., Vol. 83, p. 5069, 2010.

[12] X. Zhang, C. Oh, C. Riley, and C. Buck "Current status of computational approaches for protein identification using tandem mass spectra", Curr. Proteomics, Vol. 4, p. 121, 2007.

[13] M. Mohiyuddin, J. Mu, J. Li, N. Asadi, M. Gerstein, A. Abyzov, and W. Wong, H. Lam, "Metasv: an accurate and integrative structural-variant caller for next generation sequencing", Bioinformatics, Vol. 31, p. 2741, 2015.

[14] A. Pesaranhader, S. Matwin, M. Sokolova, and R. Beiko, "simdef: Definition-based semantic similarity measure of genontology terms for functional similarity analysis of genes", Bioinformatics, Vol. 32, p. 1380, 2016.

[15] D. Xu, M. Zhang, Y. Xie, F. Wang, M. Chen, K. Zhu, and J. Wei, "Dtminer: identification of potential disease targets through biomedical literature mining", Bioinformatics, Vol. 32, p. 1, 2016.