Atrial fibrillation detection evaluation - performance measures

Sándor Hargittai

Meditech Ltd., Budapest, Hungary

Abstract

Atrial fibrillation is the most commonly sustained arrhythmia in clinical practice worldwide. It's essential that AF detection algorithms should be powerful. This paper is addressed to study the proper performance metrics for evaluation.

The current ANSI/AAMI standard recommends using two metrics – sensitivity and positive predictive value (PPV). We argue with this combination of indicators of diagnostic performance. We reviewed the possible metrics for evaluation applied in the diverse fields. F-measure, diagnostic odds ratio, accuracy, kappa coefficient, regression and correlation coefficients were closely reviewed.

Sensitivity and specificity are best paired performance measures. We propose to utilize sensitivity, specificity, aggregate phi correlation coefficient and aggregate kappa coefficient as performance measures of atrial fibrillation detection algorithms.

1. Introduction

Atrial fibrillation is the most commonly sustained arrhythmia in clinical practice worldwide. It's essential that AF detection algorithms should be powerful. This paper is addressed to study the proper performance metrics for evaluation.

The current ANSI/AAMI standard recommends using two metrics – sensitivity and PPV [1]. We argue with this combination of indicators of diagnostic performance. Whilst the sensitivity is the intrinsic quality of the algorithms, the PPV is more characteristic of the predictive power at different AF prevalences.

2. Method

The AF segments can be correctly identified (true positive, TP) or incorrectly rejected (false negative, FN) by the algorithms. In the case of absence of AF episode the algorithms can erroneously indicate AF episode (false positive, FP) or correctly reject it (true negative, TN).

The test result can be inserted into a 2x2 contingency table (Table 1). The values in this table are expressed in time duration and percentage.

We systematically reviewed the possible measures for evaluation applied in the diverse fields as medical science, data mining, machine learning, information retrieval and genetic association studies [2-4].

Table 1. 2x2 contingency table

Test annotation, AF present	Reference annotation, AF present		Marginal total and probability
	Yes, x=1	No, x=0	
Yes, y=1	$TP = N * Prev * Se$ $P(xy) = Expected + cov(x,y)$	$FP = N * (1 - Prev) * (1 - Sp)$ $P(\bar{x}y) = Expected - cov(x,y)$	$TP + FP$ $\frac{TP + FP}{N} = Bias$
No, y=0	$FN = N * Prev * (1 - Se)$ $P(x\bar{y}) = Expected - cov(x,y)$	$TN = N * (1 - Prev) * Sp$ $P(\bar{x}\bar{y}) = Expected + cov(x,y)$	$TN + FN$ $\frac{TN + FN}{N} = 1 - Bias$
Marginal total and probability	$TP + FN$ $\frac{TP + FN}{N} = Prev$	$TN + FP$ $\frac{TN + FP}{N} = 1 - Prev$	$TP + TN + FP + FN = N$

2.1. Sensitivity, Specificity, PPV and NPV

The basic measures are the raw and column based indices: sensitivity, specificity, PPV and NPV (negative predictive value). They are defined in Table 6.

Sensitivity and specificity are the properties of the algorithms themselves. Sensitivity and specificity do not depend on prevalence.

Table 2. Same algorithm, very different PPV

	Mitdb	Afdb	Ltafdb
Prevalence	9,16 %	39,94 %	52,96 %
Sensitivity	95,77 %	92,59 %	93,94 %
Specificity	95,26 %	98,27 %	95,61 %
PPV	67,08 %	97,27 %	96,01 %

The predictive values indicate the usefulness of the algorithm on the given databases. They strongly depend on prevalence. The PPV of the algorithm is difficult to

ISSN 2325-8861

compare in case of different databases and it can be deceptively small at low prevalence even for very powerful algorithms (Table 2). Hence the PPV is uninformative without disclosing the prevalence.

We will use for illustration three hypothetical algorithms with different sensitivity and specificity and we will examine their behavior at different prevalences.

On the basis of the values of the basic measures it is hard to decide which algorithm is the best. The highest values are denoted by numbers in bold (Table 3).

Table 3. Hypothetical algorithms

Prev %	Alg	Se %	Sp %	PPV %	NPV %
8.33	Alg1	**90**	90	45.00	**99.00**
	Alg2	80	80	26,67	97,78
	Alg3	61	**99**	**84.71**	96.45
50.00	Alg1	**90**	90	90.00	**90.00**
	Alg2	80	80	80.00	80.00
	Alg3	61	**99**	**98.38**	71.74
91.67	Alg1	**90**	90	99.00	**45.00**
	Alg2	80	80	97,78	26,67
	Alg3	61	**99**	**99.85**	18.76

In case of AF detection obtained by chance PPV and NPV would be equal to prevalence and (1–prevalence). It corresponds to the chance line in the Figure 1. The PPV and (1-NPV) curves correspond to the probability of presence the AF at the positive and negative output of the detector.

Since the pair of sensitivity and specificity characterized the discriminative property and does not depend on prevalence, whereas the predictivity pair indicates the clinical performance and depends on prevalence it is unsuitable to combine them, as it has been done in the standard.

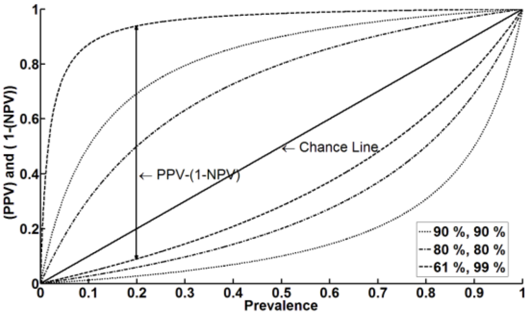

Figure 1. Probability of presence the AF

Sensitivity and specificity are best paired performance measures which are the inherent properties of algorithms. However, using paired indicators can be a drawback in comparing the performance of detection algorithms, particularly if one of them does not outperform the other on both indicators. If sensitivity is not equal to specificity then it is very important, how the algorithms behave at different prevalences namely in different clinical conditions.

This problem can be eliminated using one aggregate measure.

2.2. Aggregate measures

If the sensitivity, specificity and prevalence are known we can calculate any other performance measures (Table 6). The sensitivity and specificity fully characterize the algorithm. At the same time, prevalence defines the clinical situation where the algorithm is being used.

2.2.1. F-measure

F-measure is an integrated measure of the sensitivity and PPV. It is a balanced harmonic mean of them. However, it does not take into account the TN cases.

It behaves like the PPV, but at high prevalence its value is limited by the sensitivity.

2.2.2. Accuracy

Accuracy is a weighted average of the sensitivity and the specificity by prevalence. At zero prevalence it is equal to specificity, at the value one it is equal to sensitivity. The main problem with it is that a useless algorithm could have a high accuracy value in case of unbalanced database simply by always predicting the majority class.

2.2.3. Diagnostic odds ratio (DOR)

The DOR rises steeply when either the sensitivity or the specificity becomes nearly perfect. In this situation the other measure can be unacceptably low, even at a very high DOR value.

Table 4. Diagnostic odds ratio

	Se %	Sp %	DOR
Alg1	**90**	90	81
Alg2	80	80	16
Alg3	61	**99**	**155**

According to Table 4 Alg3 is much better than the other two, which is very suspicious.

2.3. Covariance based aggregate measures

The previous aggregate performance measures are inappropriate for AF detection evaluation.

Cell values in the confusion table can be divided into two parts. The first part is the product of marginal probabilities which provides the part due to chance. The second part is the covariance between the two binary variables, which reveals the strength of the association between them. The covariance shows how strongly the two variables are linearly related. The value of covariance ranges between -1/4 and +1/4. If the algorithm works as a random choice then the covariance will be zero.

The next measures are the normalized version of covariance.

2.3.1. Measures related to linear regression

Linear regression reveals the degree of relationship between the independent variable x and dependent variable y.

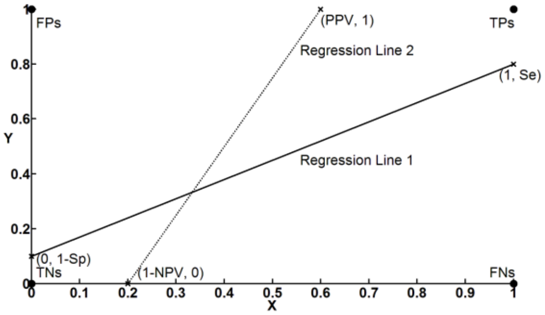

Figure 2. Linear regression

The result of linear regression is two regression coefficients and one correlation coefficient. They can be used as a performance measure of AF detection algorithms.

The first regression coefficient is equal to (Se+Sp-1), the second is equal to (PPV+NPV-1). They correspond to the slopes of the regression lines (Figure 2). They are sometimes called Youden index and Predictive Summary index (PSI) correspondingly. The correlation coefficient is the geometric mean of the regression coefficients. It is known as the Phi correlation coefficient.

2.3.2. Cohen’s kappa

Cohen’s kappa can be considered as the degree of association between two dichotomous variables. Its zero value denotes the absence of the relationship. The value of one implies a perfect association between reference annotations and test annotations.

2.3.3. Comparison of covariance based aggregate measures

The covariance based measures as the function of prevalence is presented in Figure 3, except Youden index, since it does not depend on prevalence.

All these measures can be considered chance corrected values as they are the scaled version of covariance. They indicate how much better the classification than would be expected by a random assignment of classes. All four measures are equal to zero when there is no association between AF present and the detector output.

The Youden index is not sensitive to differences in the sensitivity and specificity. The second and third algorithm examples have the same value, although they behave very dissimilarly at different prevalences.

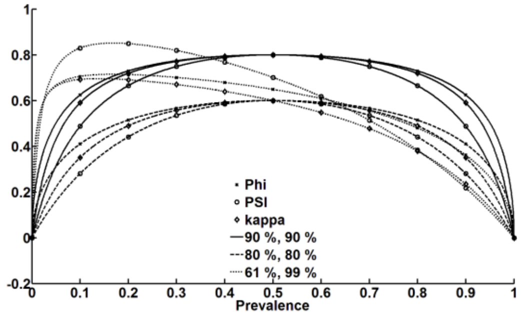

Figure 3. Covariance based aggregate measures

In the case of different sensitivity and specificity the PSI could be improperly high at low or high prevalences as shown in Figure 3.

The Phi correlation coefficient and the kappa coefficient behave better than others.

2.4. Averaging

The problem is their dependence on prevalence. The solution can be to average them over the prevalence. The average values are presented in Table 5.

Table 5. Final result

	Alg1	Alg2	Alg3
Se	**0.90**	0.80	0.61
Sp	0.90	0.80	**0.99**
Aver_Youden	**0.80**	0.60	0.60
Aver_Phi	**0.70**	0.50	0.58
Aver_Kappa	**0.68**	0.47	0.52
Aver_PSI	**0.63**	0.43	0.60

According to Table 5 the best algorithm is the Alg1 as it would be expected.

3. Result

Sensitivity and specificity are best pair measures for performance evaluation. However, using paired metrics can be a drawback in comparing the performance of detection algorithms.

Phi correlation coefficient and kappa coefficient are the best aggregate measures, but they depend on prevalence.

4. Discussion and conclusions

We propose to utilize sensitivity and specificity in the standard as the main performance measures. Additionally, the average value of phi correlation coefficient and the kappa coefficient over the prevalence can be used as an aggregate measure. Another possibility is providing their values at different prevalences.

Table 6.

Sensitivity	$\frac{TP}{TP + FN}$	Se
Specificity	$\frac{TN}{TN + FP}$	Sp
PPV	$\frac{TP}{TP + FP}$	$\frac{Se * Prev}{Se * Prev + (1 - Sp) * (1 - Prev)}$
NPV	$\frac{TN}{TN + FN}$	$\frac{Sp * (1 - Prev)}{Sp * (1 - Prev) + (1 - Se) * Prev}$
Expected joint probability	$(TP + FN) * (TP + FP)/N^2$ $(TP + FN) * (TN + FN)/N^2$ $(TN + FP) * (TP + FP)/N^2$ $(TN + FP) * (TN + FN)/N^2$	$Prev * Bias$ $Prev * (1 - Bias)$ $(1 - Prev) * Bias$ $(1 - Prev) * (1 - Bias)$
Accuracy	$\frac{(TP + TN)}{N}$	$Se * Prev + Sp * (1 - Prev)$
Error rate	$\frac{(FP + FN)}{N}$	$1 - Accuracy$
F-measure	$\frac{TP}{TP + \frac{FP + FN}{2}}$	$\frac{2 * Se * PPV}{Se + PPV} = \frac{Se * Prev}{Se * Prev + errorrate/2}$
DOR	$\frac{TP * TN}{FP * FN}$	$\frac{Se * Sp}{(1 - Se) * (1 - Sp)}$
cov(x,y)	$\frac{(TP * TN - FP * FN)}{N^2}$	$Youden * Prev * (1 - Prev)$
var(x)	$\frac{(TP + FN) * (TN + FP)}{N^2}$	$Prev * (1 – Prev)$
var(y)	$\frac{(TP + FP) * (TN + FN)}{N^2}$	$(Youden * Prev + 1 - Sp) * (Youden * (1 - Prev) + 1 - Se)$
Youden index	$\frac{TP * TN - FP * FN}{(TP + FN) * (TN + FP)}$	$Se + Sp - 1 = \frac{cov(x, y)}{var(x)} = Youden$
Predictive Summary index	$\frac{TP * TN - FP * FN}{(TP + FP) * (TN + FN)}$	$PPV + NPV - 1 = \frac{cov(x, y)}{var(y)} =$ $\frac{cov(x, y)}{(Youden * Prev + 1 - Sp) * (Youden * (1 - Prev) + 1 - Se)}$
Phi coefficient	$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$	$\frac{cov(x, y)}{\sqrt{(cov(x, y) + (1 - Prev) * (1 - Sp)) * (cov(x, y) + Prev * (1 - Se)}}$
Cohen's kappa	$\frac{TP * TN - FP * FN}{TP * TN - FP * FN + N * (FN + FP)/2}$	$\frac{cov(x, y)}{cov(x, y) + errorrate/2}$

References

[1] ANSI/AAMI EC57:2012. Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms.

[2] Powers DWM. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. School of Informatics and Engineering, Flinders University, Adelaide, Australia. 2007.

[3] Rubanovich AV. Theoretical Analysis of the Predictability Indices of the Binary Genetic Tests. Russian Journal of Genetics: Applied Research, 2014;4:146–158.

[4] Kraemer HC. Periyakoil VS. Noda A. Kappa coefficients in medical research. Stat Med. 2002;21:2109-29.

Address for correspondence.

Sándor Hargittai
Meditech Ltd.
Mikszáth Kálmán utca 24.
H-1184. Budapest
Hungary
sandor.hargittai@gmail.com