Available online at www.sciencedirect.com

ScienceDirect

Engineering Applications of Artificial Intelligence 20 (2007) 875–885

Engineering Applications of
ARTIFICIAL INTELLIGENCE

www.elsevier.com/locate/engappai

A lot of randomness is hiding in accuracy

Arie Ben-David*

Management Information Systems, Department of Technology Management, Holon Institute of Technology,
52 Golomb St. P.O. Box 305, Holon 58102, Israel

Received 3 November 2006; received in revised form 5 January 2007; accepted 10 January 2007
Available online 23 February 2007

Abstract

The proportion of successful hits, usually referred to as "accuracy", is by far the most dominant meter for measuring classifiers' accuracy. This is despite of the fact that accuracy does not compensate for hits that can be attributed to mere chance. Is it a meaningful flaw in the context of machine learning? Are we using the wrong meter for decades? The results of this study do suggest that the answers to these questions are positive.

Cohen's kappa, a meter that does compensate for random hits, was compared with accuracy, using a benchmark of fifteen datasets and five well-known classifiers. It turned out that the average probability of a hit being the result of mere chance exceeded one third (!). It was also found that the proportion of random hits varied with different classifiers that were applied even to a single dataset. Consequently, the rankings of classifiers' accuracy, with and without compensation for random hits, differed from each other in eight out of the fifteen datasets. Therefore, accuracy may well fail in its main task, namely to properly measure the accuracy-wise merits of the classifiers themselves.

Keywords: Accuracy; Kappa; "Chance driven" hits; Machine learning; AUC

1. Introduction

Accuracy measures the number of successful hits relative to the total number of classifications. It is by far the most commonly used metric for assessing the accuracy of classifiers for years (Lim et al., 2000; Alpaydin, 2004; Witten and Frank, 2005; Demsar, 2006).

This research deals with a very serious anomaly of the accuracy. Here is a simple example: Table 1 shows a binary confusion matrix with 1000 classifications.

The accuracy in the confusion matrix of Table 1 is 0.5; Fifty percent of the classifications were correct. But what can be said about the classifier that produced these predictions? One can hardly think of a worse classifier. This is due to the fact that a randomly tossed fair coin will produce approximately similar results. In other words, all the classifier's predictions of Table 1 may be due to mere chance. A good accuracy meter should explicitly measure the added value, if any, of a classifier relative to a random, or a majority-based, outcome. In this respect, the classifier that produced the confusion matrix of Table 1 has no added value at all. Saying that the accuracy is 50%, though arithmetically correct, does not explicitly convey this meaning. Similar examples can be given for any multi-class case.

The machine-learning community has long been aware of the fact that accuracy is far from being a perfect meter. Usually, several classifiers are competing against each other. Baseline classifiers (typically, majority based) are often used too. There would have been nothing wrong with this method provided that the effect of random hits was similar across all classifiers for any given dataset. However, this hidden assumption was never put to a real test. Consider the following hypothetical example: Classifiers A and B are applied to a single dataset. Classifier A scores on the average 80% success rate, and classifier B (which can be a baseline) only 70%. Assume further that a proper statistical test on accuracy has concluded that A is more accurate than B. This conclusion, however, would not

*Tel.: +972 3 7317977; fax: +972 3 5716481.
E-mail address: hol_abendav@bezeqint.net.

0952-1976/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2007.01.001

Table 1
A simple confusion matrix

Correct class	Predicted class		
	Good	Bad	Total
Good	250	250	500
Bad	250	250	500
Total	500	500	1000

make much intuitive sense, should one also knew that 50% of A's successes may be due to mere chance, and only 10% of B's. This research clearly shows that such scenarios are possible, because chance differently affects various classifiers, even when they are applied to a similar dataset. Classifiers' accuracy should be compared after compensating for random hits, and this compensation may vary with each classifier, even when a single dataset is used. By ignoring the effects of random hits, one unavoidably risks arriving at the wrong conclusions.

An alternative to accuracy, a meter that does compensate for random hits, is known for decades. It is called Cohen's kappa (Cohen, 1960). Cohen's kappa is routinely used in disciplines such as Statistics, Psychology, Biology and Medicine for a long period. However, for one reason or another, it has received only very little attention in machine-learning circles.

This research was focused at answering the following two questions:

A. Is the problem of counting random hits meaningful in the context of machine leaning?
B. Are rankings according to accuracy always identical to those that are arrived at using Cohen's kappa? In other words, can we arrive at different conclusions about classifiers accuracy when chance considerations are taken into account?

To answer these questions, an empirical study was conducted. Fifteen datasets were tested using five well-known classifiers. The results are quite interesting:

A. On the average, more than one third of the hits in the benchmark could be attributed to chance alone. Accuracy ignores this high proportion altogether.
B. The rankings by accuracy and via Cohen's kappa differed from each other in eight out of the fifteen datasets. Different rankings may lead to different conclusions.

The findings of this research strongly suggest that we, the machine-learning community, are traditionally using the wrong meter, namely accuracy. We do that without being fully aware of the fact that a significant portion of the so-called "accuracy" is merely the product of chance. In this respect, Cohen's kappa is a more accurate meter for measuring classifiers' own merits than accuracy.

2. Cohen's kappa and its very rare use in machine learning

Cohen's kappa (Cohen, 1960) was first introduced as a measure of agreement between observers of psychological behavior. The original intent of Cohen's kappa was to measure the degree of agreement, or disagreement, between two people observing the same phenomenon. Cohen's kappa can be adapted to machine learning, as shown in the example of Table 2.

The accuracy shown in Table 2 is 97% ((70 + 900)/1000). Can all these 97% be attributed to the sophistication of the classifier alone? Does chance have anything to do with it?

Cohen's kappa is defined as

$$K = \frac{p_0 - p_c}{1 - P_c}, \qquad (1)$$

where P_0 is the total agreement probability, or accuracy, and P_c is the "agreement" probability which is due to chance.

For the data of Table 2 kappa is computed as follows:

$$P_0 = \frac{70}{1000} + \frac{900}{1000} = 0.97 \quad \text{(i.e., accuracy)},$$

$$P_c = \frac{80}{1000} \times \frac{90}{1000} + \frac{920}{1000} \times \frac{910}{1000} = 0.84$$

and the value of kappa is thus

$$K = \frac{0.97 - 0.84}{1 - 0.84} = 0.81.$$

According to the kappa statistic, the classifier that produced the confusion matrix of Table 2 has a less impressive "accuracy": 0.81 and not 0.97.

What the kappa statistic expresses can be explained in a nutshell as follows: kappa evaluates the portion of hits that can be attributed to the classifier itself (i.e., not to mere chance), relative to all the classifications that cannot be attributed to chance alone.

What about a case of a perfect agreement?

In this case, shown in Table 3, α, β are integers and C_1 and C_2 are class values.

$$p_0 = \frac{\alpha}{\alpha + \beta} + \frac{\beta}{\alpha + \beta} = 1,$$

$$p_c = 2\left(\frac{\alpha}{\alpha + \beta}\right)^2 \geqslant 0$$

Table 2
Another confusion matrix

Correct class	Predicted class		
	Good	Bad	Total
Good	70	10	80
Bad	20	900	920
Total	90	910	1000