International Journal of Research Studies in Science, Engineering and Technology
Volume 6, Issue 11, 2019, PP 40-52
ISSN 2349-476X
DOI: https://doi.org/10.22259/2349-476X.0611005

Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement

Stbalinen I Saini[1], Ndgummalla Sirisha[1], Dr. A. Srirama Kanaka Ratnam[2]

[1]*CSE Department, Malineni Lakshmaiah Women"s, Engineering College, Guntur, AP*

[2]*Associate Professor, CSE Department, Malineni Lakshmaiah Women"s, Engineering College, Guntur, AP*

ABSTRACT

In machine learning, cross-validation is a standard method for assessing performance and development. In cross-validation experiments, there are subtle changes in how to compute accuracy, F-measure, and Area Under the ROC Curve (AUC). However, these subtleties are rarely explored in the literature, and different articles and software packages use incompatible methodologies. As a result, there is inconsistency in the research literature. When anomalies in performance estimates for certain folds and conditions are buried in aggregated findings across numerous folds and datasets, they go unnoticed because no one looks at the intermediate performance measurements. This research note explains and demonstrates the differences, as well as offering advice on how to effectively assess classification performance using cross-validation. There are various diverging ways for computing F-measure, which is frequently recommended as a performance measure in situations when there is a class imbalance, such as text classification domains and one-vs.-all reductions of datasets with many classes. We show that, with the exception of one, all of these computing approaches result in biased measurements, especially when there is a large class imbalance. This is a research paper. Those working on machine learning systems will find this of particular interest. Researchers and software libraries focused on high-quality software unbalance.

INTRODUCTION

The field of machine learning has benefited from having more people working on it. A few basic performance measures by which we might assess our performance advances on classification benchmark datasets, such as the Text dataset from Reuters [4]. Many papers in the literature have been published. Literature has cited one another's performance figures in order to prove that a new procedure is better than the old one or at the very least, comparable to previously reported approaches. The significance of being able to cite the work of others Over time, the figures have risen. It can also catch us off guard when, for example, the F-measure was calculated in an incompatible method, or the AUC in one study was calculated in a way that mistakenly necessitated a larger sample size.

As well as a consistently calibrated classifier.

F-measure and AUC are well-defined, widely used performance metrics with definitions available online. Similarly, a large number of publications describe the widely accepted Cross-validation is a technique for evaluating and comparing data. On a given labeled dataset, the quality of classification schemes.

However, there is ambiguity and disagreement over it, which is ironic. How to calculate F-measure and AUC across the board a cross-validation study's folds this was the first time it was mentioned. The quantity of questions we receive from people has brought this to our attention. Other scholars on how to go about measuring things precisely. These are subjected to cross-validation. Following a thorough study, We couldn't locate anything about the subject in the literature. We conducted an informal survey of dozens of papers and discovered that there is much debate on the subject. Not only do different articles employ different computer approaches. Most don't bother to indicate whether they use the F-measure or the AUC. Exactly how they calculated it using cross-validation—possibly oblivious to the fact that there are alternatives. It has done so in the past. not been mentioned in the literature, and especially not in the media shown how certain ordinary decisions

can lead to bias results. one of the article's anonymous reviewers shared in their review that they had to deal with last year two examples of this issue, both of which resulted in experimental failure Positively skewed outcomes are expected. Finally, we've noticed There are a variety of inconsistencies and biases in the strategies provided. Students' research software, as well as software libraries It can be difficult to spot such tiny anomalies. compared to bugs that signal their presence by halting execution.

Not only are the methods of computation different, but \sit also turns out that there might be major disagreement under some test situations, in their outputs This paper lists the various techniques of computation (Section 2), uses examples to show that the differences can be significant (Section 3), and shows that one approach of obtaining F-measure is superior in terms of bias and variance (Section 4).

Interest is rare, which is a regular occurrence in text datasets and a burgeoning area of investigation. When there are a lot of classes in a dataset, there is a lot of class imbalance are accounted for in a slew of one-vs.-all (OVA) subtasks

CROSS-VALIDATION OF PERFORMANCE MEASURES

This section defines and distinguishes the many techniques for calculating performance scores If you're given a the question is what to do with a labelled dataset and a classification method. The task at hand is to determine how effectively the classifier works on the set of data

Preliminaries for Formal Notation

Let's call our instance space X, which is a set that encompasses everything. In our representation, we can express all instances. We believe underlying X is a stable but unknown distribution Dcalculates the likelihood or density of sampling a particular populationx ϵX. Each x corresponds to a label from a book. Y is a finite collection.

A function c: X Y is a hard classifier. After reading a sequence (x1, y1),...,(xt, yt) of t labelled training examples, a learning algorithm outputs a classifier c, where each xi∈ X is an example from the instance space, and yi∈ Y is the associated label of xi.

The training set will be referred to as the sequence of examples, and we'll assume that each labelled example in that set was sampled i.i.d. from D. The overarching goal is to develop learning algorithms that are likely to provide classifiers that behave "well" when compared to the same unknown underlying distribution D. P(x,y)D(c(x) = y).

In fact, we must rely on test sets to evaluate a classifier's performance with regard to D. It is possible to generate an estimate of several performance measures using a holdout set or test set T sampled i.i.d. from the same D. In this scenario, it is clearly desirable to adopt a method that gives unbiased and low variance estimates of the unknown ground truth performance value over the whole space D. Counts are used to make such estimates. We concentrate on binary (hard) categorization, in which Y only has two labels: "positive" and "negative." Based on both the true label yi and the predicted label c(xi) for each case (xi, yi)∈T, each classifier c divides the test set into four partitions. K-fold cross-validation (typically 10-fold) is the most common method for estimating learning algorithm performance. It separates the training data T into k distinct groups. Subsets $T^{(1)}$ …$T^{(k)}$) equal in size. Every one of the T1 setsis utilized as a test set, and it is compared to a classifier based on all of the other data T \ $T^{(i)}$. As a result, we can get k.the results of various test sets We frequently report the average. of those as the classifier's total estimate on that dataset. The goal of this procedure is to provide accurate estimations. When doing the test, the performance is very near to the real thing.On the entire set T, there is a learning algorithm. However, we will.

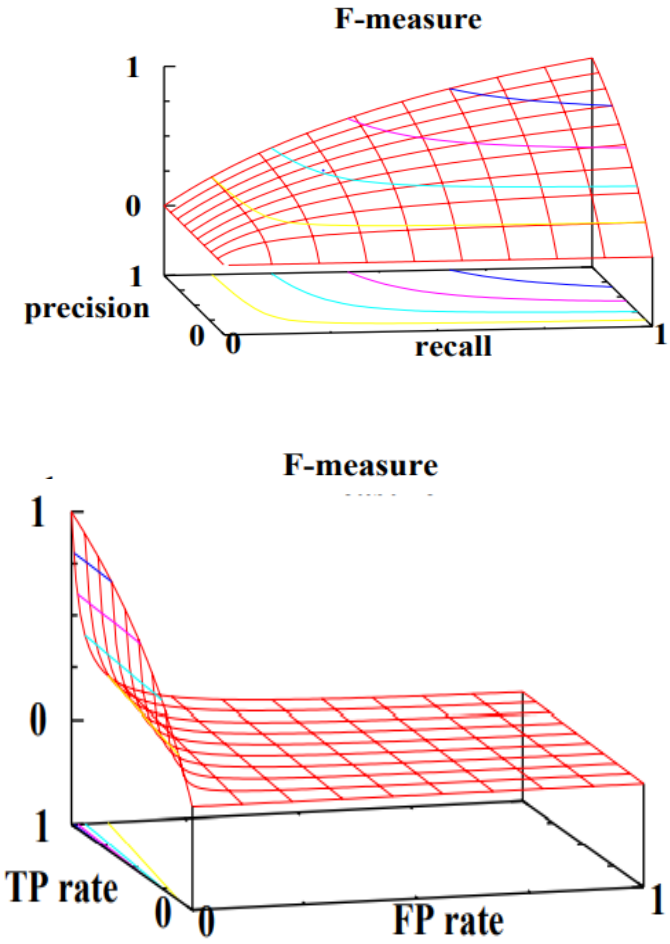

Figure1 shows the F-measure as a function of (a) precision and recall, or (b) true positive rate and false positive rate, with 1% positives assumed.

Demonstrate in the next section that there is a problem with This is how the average F-measure is reported. In this work, we'll refer to values with superscripts. Cross-validation folds that pertain to certain cross-validation folds As an example, Fold I true positives would be referred to as the number of true positives. In the same way that TP (i), the fold j precision as Pr (jAnoption to the cross-validation approach discussed above is stratified cross-validation. The only difference is that it takes care that each subset T (i) contains the same number of examples from each class (±1). This is common practice in the machine learning community, partly as a result of people using integrated learning toolboxes like WEKA [3] or Rapid Miner [6] that provide stratification by default in cross-validation experiments. The main advantage of this procedure is that it reduces the experimental variance, which makes it easier to identify the best of the methods under consideration.

Without Cross-Validation-F-Measure

The F-measure is the most common statistic in the text classification and information retrieval communities, despite the fact that error rate or accuracy dominate much of the classification literature. The reason behind this is that most text mining corpora have a lot of classes and a lot of class imbalance. Precision and recall are balanced in F-measure, whereas accuracy tends to undervalue how well classifiers perform on smaller classes.

Definition 1: The precision Pr and the recall R are two definitions of precision and recall, respectively. The TP true positives, FP false positives, and FN false negatives of a classifier are

$$\begin{aligned} \mathrm{Pr} &:= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \quad \text{and} \\ \mathrm{Re} &:= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \end{aligned}$$

The F-measure combines these two factors into a single number that can be used to rank or compare approaches. It can be thought of as a 'and' function: if precision or recall are both bad, the result is the same.

The resulting F-measure, as represented visually, will be bad.1a in Figure F-measure is the harmonic mean in formal terms. Between recall and precision

Definition 2: The precision of a classifier's F-measure. Pr and recollection The word re is defined as

$$F := 2 \cdot \frac{\mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} \qquad (1)$$

The following is a simplified definition of F-measure seen in many academic papers and software libraries:

$$\begin{aligned} F &= 2 \cdot \frac{\mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} \\ &= 2 \cdot \frac{\left(\frac{TP}{TP+FP}\right) \cdot \left(\frac{TP}{TP+FN}\right)}{\left(\frac{TP}{TP+FP}\right) + \left(\frac{TP}{TP+FN}\right)} \\ &= (2 \cdot TP) \,/\, (2 \cdot TP + FP + FN) \end{aligned}$$

As a result, F-measure is calculated in terms of the number of true and false positives. On the x- and y-axes, Figure 1b depicts this view using the false positive rate and true positive rate. When negatives abound, any significant false positive rate will result in low precision and, as a result, low F-measure. The graph shown assumes 1% positives, resulting in the sharpness of the surface; when negatives abound, any substantial false positive rate will result in low precision and, as a result, low F-measure.

Exceptions: In some cases, Equation (1) is undefined. If the classifier makes no positive predictions, precision is undefined (TP+FP = 0). This can happen with tiny or unbalanced test sets, as well as with particularly conservative classifiers, such as those that learn to always vote for the majority class during training. When there are no positives in the test set, recall is undefined. If random sampling or unstratified cross-validation are performed on extremely imbalanced datasets, this can happen in rare cases.

In almost all instances, Equation (2) smoothly extends the definition of F-measure to be well-defined (specifically, zero). EvenSo, if a test set includes no positives (TP + FN = 0) and the classifier agrees—that is, it produces no positive predictions (TP + FP = 0), it still results in division-by-zero. When test harness software finds one of these uncertain scenarios, it has two options.

It can substitute a zero for an otherwise undefined number, or it can omit the occasional, difficult test fold from the final computations, which is less usual. These options result in a negative or positive bias in the measurement of F-measure, as we'll see later. If unstratified cross-validation or random samples are utilized.

F-measure Cross-Validation

We examined cross validation and F-measure individually in the preceding two sections. The majority of researchers do not think about the concept of cross-validation is a combination of these two. To be unclear, F-measure. We'll go over this in this part.a description of three different strategy combinations All of these terms are often used in the literature. Two of them make it possible to multiple approaches to dealing with the undefined corner cases, so in the end, you'll have five different aggregation strategies. altogether. If we double the number of strategies to ten,Cross-validation, both unstratified and stratified, should be considered.

All of the following scenarios have one thing in common: we train k classifiers and evaluate the classifier c. (i)(which we obtained during iteration I when training on T\ T)I)

T is only available on the hold-out setI. The terms in italics are superscripted.TP(i)by way of Tennessee (i)F.,I, and Pr (i)Re, or (i)refer to the examination set c's performance Ion the TI, as specified in Sections 2.1, 2.2, and 2.3as well as 2.We'll use the precise notation and framework we've created. We are now in a position to define the three categories. The most common methods for aggregating F-measure findings across the boardk cross-validation folds.

1. Let's start with the simple example of averaging the F-measure. We keep track of the F-measure in each fold. F\s(i)andas the mean of all folds, compute the final estimate:

$$\overline{F_{\mathbf{avg}}} := \frac{1}{k} \cdot \sum_{i=1}^{k} \overline{F^{(i)}}$$

2. Another option is to average precision and recall over the folds and use the final values to construct F measure using Equation 1:

$$\begin{aligned} \mathrm{Pr} &:= \frac{1}{k} \cdot \sum_{i=1}^{k} \mathrm{Pr}^{(i)} \\ \mathrm{Re} &:= \frac{1}{k} \cdot \sum_{i=1}^{k} \mathrm{Re}^{(i)} \\ F_{\mathbf{pr,re}} &:= 2 \cdot \frac{\mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} \end{aligned}$$

3. Alternatively, one can add up the number of true and false positives across the folds and compute F-measure using either Equations 1 or 2:

$$\begin{aligned} \mathrm{TP} &:= \sum_{i=1}^{k} \mathrm{TP}^{(i)} \\ \mathrm{FP} &:= \sum_{i=1}^{k} \mathrm{FP}^{(i)} \\ \mathrm{FN} &:= \sum_{i=1}^{k} \mathrm{FN}^{(i)} \\ F_{\mathbf{tp,fp}} &:= (2 \cdot \mathrm{TP}) / (2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{F} \end{aligned}$$

$$\tilde{F}_{\mathbf{avg}} := \frac{1}{\sum_{i=1}^{k} V^{(i)}} \cdot \sum_{i=1}^{k} F^{(i)}$$

Exceptions: As previously stated, we may confront the problem of indeterminate accuracy or recall in some folds. LetsVI:= 1 if Pr is a positive integer (i)as well as (i)both have been defined, and VI:= zero otherwise. When a classifier is used, precision is unknown. C I does not anticipate any of the fold T test examplesI as optimistic Only if a fold does not exist may recall be undefined. include any positives This isn't possible with stratified data. Unless the number of folds is greater than the maximum, cross-validation is used. It is uncommon for unstratified to have a large number of positives, and it is regarded rare for unstratified to have a large number of positives cross-validation. Substituting is one approach for dealing with this issue. Zero, based on an F-measure reformulation; see statement (2). This will be the default meaning for the rest of the paper, so F I = 0 when V I = 0.As an alternative, declare all folds with indeterminate precision and recall as incorrect measurements and skip them entirely. In a later part, the absurdity of such a decision will be revealed. This could happen as a result of the software throwing an exception inadvertently. When we use the terms Favg or Fpr,re, we'll use a tilde to indicate that we're talking about the latter calculation.

AUC, Accuracy, and Error Rate

Under cross-validation, accuracy and error rate do not have the same issue: You receive the same outcome. whether you calculate accuracy on each fold separately and then combine them If you tally the error count and then compute, you'll get an average. Just once, at the conclusion, calculate the accuracy rate. As a result, the issue exists. For many learning articles that have been written, this has not been an issue. Previously, performance was evaluated solely on the basis of error rate. or precision AUC under cross-validation, on the other hand, can be calculated. in two contradictory ways The first step is to classify each individual. All folds' scores are combined onto a single ROC curve. and then find the area of this curve, which we refer to as AUC merge. The other option is to calculate the AUC for each fold separately. independently, then averaging across all folds:

$$AUC_{\mathbf{avg}} := \frac{1}{k} \cdot \sum_{i=1}^{k} AUC^{(i)}$$

The issue with AUC merge is that it assumes that by grouping various folds together, the classifier would give well-calibrated probability estimates. A researcher is a person who studies something.

Brier score or something similar will be used to assess the quality of probability estimates. Researchers, on the other hand, AUC is commonly used to assess performance. oblivious to the need for calibration or precise threshold levels focusing solely on the classifier's ability to rank Positives take precedence over negatives. As a result, AUC merge adds a typically The study will be downgraded due to an unforeseen requirement. Rank-high classifiers with inadequate calibration throughout the board As seen in Section 3.2, folds are possible. WEKA [3] employs the AUCmerge approach as of version 3.6.1.in its Explorer user interface, as well as in its Evaluation core class It employs AUCavg in its Experimenter for cross-validation, but not for cross-validation interface.

Exceptions: Although not generally a problem, it would be impossible to compute AUC for any fold that had no positives. Under stratified conditions This will never be a problem with cross-validation. However, without it, stratification (for example, in a multi-label environment) and For some classes, there is a significant imbalance, and this problem could be exacerbated. emerge. Some software libraries may fail under this situation. Others may discreetly swap a zero or skip such information folds.

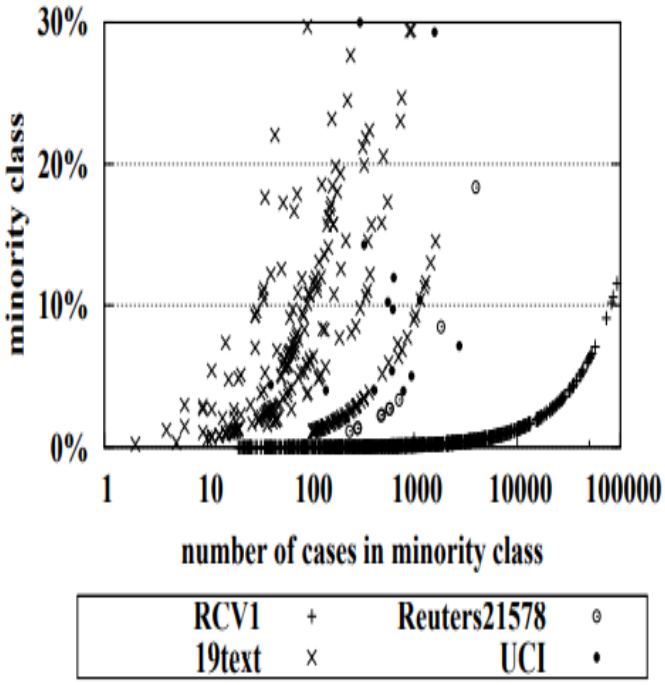

Figure2. *In the literature, class imbalance and minority class size are shown for a range of binary classification problems [1,2,5,6].*

ILLUSTRATION

In this section, we offer specific instances of cross-validation findings that illustrate a wide range of performance depending on the condition. on the computation method We'll start with F-measure and work our way up.AUC should be followed. To keep things simple, we only utilise four folds. nonetheless, the exposition and eliminate visual clutter. The gap between the methodologies can be substantially greater. With a standard 10-fold cross-validation or a larger number the number of folds Although more complex, we use stratified cross-validation. Unstratified results could be shown to be severe. Situations in which recall isn't always clear. We To be more specific, choose examples that avoid all

possible scenarios. Potentially persuasive (we'll return to this later).matter). The statistics on performance represent the actual outcomes of the tests. a linear SVM (with options) (WEKA[3] SMO implementation)Platt scaling (-M -N 2) on binary text classification jobs reprinted with permission from Reuters (dataset re0 in [1]).

In order to accentuate the difference, the examples above choose tasks that are significantly imbalanced. In text studies or research that focuses on women, the degree of imbalance we evaluate (1 percent positives and 2.5 percent negatives) is not uncommon. unbalance. Figure 2 depicts the disparity as well as the number of people affected. for a set of binary examples of the minority (positive) class

Activities based on the previous Reuters benchmark [4], and the new Reuters benchmark 19 multiclass text datasets [1], Reuters RCV1 benchmark [5], in addition to a variety of UCI and other datasets that were used in the imbalance [7] research

F-measurement

The detailed numbers for each fold of a paper are shown in Table 1.On a task with 1% positives, stratified cross-validation was used. There are 1504 data rows in total. This level of social inequityis deemed difficult, especially given the small number of participants.a lot of positives Despite this, tiny classrooms do appear. When it comes to text and UCI benchmarks, and the objective of this example is just to show how the procedures differ significantly.

In the table, we can see that the classifier made a significant contribution. On the latter two folds, there were a lot of false positives, which led todue to a lack of accuracy in those folds Whenever precision or accuracy is required, If recall is low, F-measure will be low as well. folds. We get the following result by averaging the four per-fold F-measures:

Table 1: Example 4-fold stratified cross-validation shows F-measure can differ widely depending on how it is computed.

Fold	Negatives	Positives	TP	FP		Precision	Recall	F-measure
1	373	3	3	0		100%	100%	100%
2	372	4	4	1		80%	100%	89%
3	372	4	4	13		24%	100%	38%
4	372	4	3	5		38%	75%	50%
Totals:	1489	15	14	19	Averages:	60%	94%	69% F_{avg}
			58% $F_{tp.fp}$			73% $F_{pr.re}$		

Table 2: A second example where the F-measure calculation methods disagree because the classifier predicted no positives in the second fold. Precision here(†) is set to zero to avoid division by zero; the metrics with a tilde instead skip this fold.

Fold	Negatives	Positives	TP	FP		Precision	Recall	F-measure
1	372	4	2	0		100%	50%	67%
2	372	4	0	0		0%†	0%	0%
3	372	4	4	0		100%	100%	100%
4	372	4	4	0		100%	100%	100%
Totals:	1488	16	10	0	Averages:	75%	63%	67% F_{avg}
			77% $F_{tp.fp}$			68% $F_{pr.re}$		89% $\tilde{F}_{avg}$
						91% $\tilde{F}_{pr.re}$		

69 percent of people like it. When we average the precision and recall columns, however, any particularly low precision or recall value is smoothed out rather than emphasised. Thus, Despite the low precision of 24 percent on one fold, the The average precision and recall are moderate, resulting 73 percent in

Fpr, re = (2) x (0.60×0.94)/0.60+0.94,

Finally, if we add all the true positives and false positives, we get Favg.positives across the folds (at the bottom left), and then compute58 percent Ftp, fp= (2×14)/(2*14+19+1),This is significantly less than Favg. This demonstrates that the difference between Fpr, re and Ftp, fp can be significant: Fpr,re = 1.26 Ftp, fp. In the fourth section, We characterise each's bias and variance, demonstrating which is which. is, in fact, the superior estimator.

We discovered that the classifier made no positive predictions for one of the four folds for a separate class (not shown) with exactly four positives in each of the four folds (1 percent positive).folds. This resulted in a lack of precision and was penalised as a result. despite the fact that the classifier has zero F-measure for that fold On the other folds, the classifier worked admirably. Finally, you can choose to skip any folds that lead to. Accuracy that is not defined A tilde is used to indicate these versions. Naturally, they give higher marks to those who are able to communicate well. A tough fold was eliminated from the

test set. This is a natural result. As a result, the scoring function has a substantial positive bias:F˜pr,re = 1.34 XFpr,re

AUC

Next, we'll look at the Area Below the ROC Curve. The main problem here is that the soft score outputs from Each of the fold classifiers is not need to be calibrated. to each other For instance, we ran a four-fold stratified study. cross-validation for a different class on the same dataset38 positives in a dichotomy (2.5 percent). The AUC scores are a measure of how well a person knows The percentages for each fold were 96 percent, 91 percent, 94 percent, and 87 percent, respectively, yielding AUCavg is 92 percent on average. These four classifiers, on the other hand, were not calibrated with one another, as shown in Graph 3. The graph on the left depicts the false positive rate vs. The right graph depicts the classifier score threshold and the classifier score threshold. The genuine positive rate is the same. It's worth noting that just two of the Two other curves are considerably shifted as a result of the alignment of the folds. Horizontally. As a result, when the soft scores of all four folds are equal, are combined to generate a single ROC curve, It has an overall AUC merge score of only 80%. Unless the classifier is set up to output probabilities rather than merely scores, When scores fall below a certain threshold, comparing them is pointless. folds of several kinds This is true for ranking as well. Precision at 20 and Mean Average Precision are two examples of measurements. For each fold, such metrics must be computed independently. after which it was averaged In the event that the classifiers, on the other hand, are meant to be calibrated, and one wants to punishIf you use approaches that result in poor calibration, you might want to reconsider. After that, sort all of the soft classifier outputs together and compute the unit of measurement Again, our goal here is to just illustrate. a significant difference

BIAS AND VARIANCE IN F-MEASURES

The following questions are addressed in this section:

• Why should we expect F-measure results that have been cross-validated to be prejudiced?

•What are the various approaches for estimating F-measure introduce several kind of biases?

• Which method introduces the least absolute bias and has the smallest variance?

• What effects do class imbalance and altering target F-measures have on bias and variance?

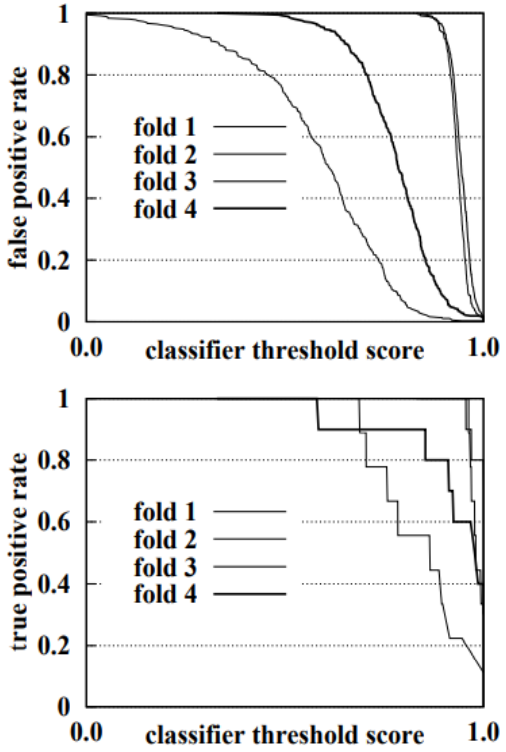

Figure3. ***(a)*** *Classifier false positive rate vs. output score (b) output score vs. genuine positive rate*

Why Do We Expect Biased Results?

Before we go into the technicalities, let's talk about why the F-measure is prone to erroneous estimates. Let us begin by looking at the behaviour of accuracy. Because it is "naturally" neutral, accuracy tends to be "naturally" unbiased. In terms of a binomial distribution, this can be represented as follows: A "win" in the underlying Bernoulli trial would be considered a "success" .a sampling of an example for which a classifier is used. The correct prediction is made after careful consideration. As a rule of thumb, The probability of success is proportional to the accuracy of the prediction classifier. The i.i.d. assumption states that each example must be unique. The test set is sampled individually, resulting in the expected results. The percentage of correctly categorized samples is the same. Above is the likelihood of seeing a success. With an average of over Increasing the number of folds is the same as increasing the number of folds. the number of times the Binomial trial is repeated. If the test sets are of identical size, or if we weight each estimate by the size of each test set, this has no effect on the posterior distribution of accuracy. F-measure, on the other hand, has the limitation of not being able to be split down into F-measures of arbitrary example subsets.

Equation (2) shows that the impact of each individually sampled case on the total estimate is highly dependent on which other examples are already included in the test set. This prevents a precise calculation of the global F-measure in terms of the F-measures of each cross-validation fold. The presence of random variables in the denominator adds complexity, akin to "context dependencies." Even if we assume we acquire the same classifier for all the test sets of folds, the averaged result will likely change whenever we switch examples across the test sets of folds. F-measure is concave in the number of true positives T P, and steepest near T P = 0. As shown in Equation (2).Missing even a single true friend may be devastating, especially when there is a socioeconomic divide. positive (in comparison to the ground truth expectation) A contingency table) could significantly reduce the F-measure of a crossvalidation fold. Including a bonus, on the other hand, is a good idea. Because true positives have a significantly lower influence, the overall bias is substantially lower. is unfavorable Clearly, this is an unfavorable property. Cross-validation. Calculating the bias for the methods used in this study Analyzing a document is a difficult task. Simulations are being run.is similarly straightforward and provides similar insights into the matter.

Simulation Specifications

Over a 1000-case dataset, we regularly simulated 10-fold cross-validation: 900 hours of instruction and 100 hours of testing for every fold The binary classifier's performance was simulated in such a way that it has ground-truth control The F-measure has a precision that is precisely equal to its recall. As a result, we may propose a classifier with an F-measure of 80%. In ground-truth, it has an accuracy of 80% and a recall of 80%. For We begin by allocating our simulated test set results. the folds' positives and negatives, stratified or not For unstratified, choose at random. Then we sample within each fold. To determine the number of people, use the binomial distribution The amount of positives that become true positives, as well as the number of positives that become true positives False positives from false negatives There is no such thing as a costly item. It is necessary to take a learning step. If you run the simulation a million times, you'll get a million different results. We were able to determine the distribution of scores several times generated for each of the five Fmeasure computation techniques For two reasons, this experiment approach makes things easier. First, it provides a sense of ground truth because we already know the correct result (the ground truth F-measure). Clearly, we need a validation method that works. reports the truth as it is with no or very little bias. as well as low variability Second, assuming that the i.i.d. assumption is correct, Given our classifiers' "ground truth" contingency table, We can evaluate each method's bias and variance. In our simulations, we looked at possibilities ranging from 1% to 1%.Positive instances account for 25% of all cases. Because there are only 1000 of them, In 1% of situations, there are just 10 positives in the dataset. This extreme scenario was chosen to emphasize the point. When no positives are expected in a situation, exceptional behaviour is expected. Occasionally, there are some folds. With so few positives in their dataset, most researchers would obviously avoid drawing any inferences. However, there are two significant exceptions. First and foremost, in Conclusions about classifiers are frequently

reached in the medical arena. based on datasets with a small number of cases; for example, The Golub et al. [2] Leukemia dataset has been extensively explored. There are only 74 cases in all, grouped into four groups. Second, some learning-focused machine learning research 'Underclass Inequality' derives conclusions from research ona large number of diverse datasets or classification jobs with a small number of variable search with a certain number of positives When the data is compiled, it is intended that The superior classifiers will win over a large number of unbalanced jobs. become well-known These findings must be correct in order for them to be accurate. It's crucial to be consistent and similar across the literature. to accurately estimate F-measure, despite what some may think Extreme situations are referred to as. And, of course, when it comes to writing, We can't control all test circumstances with software. It could be used later.

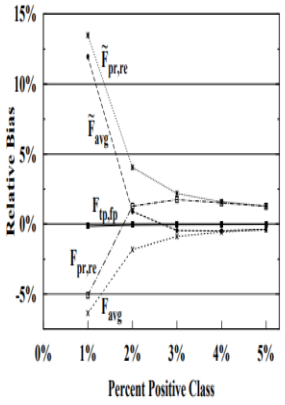

Figure 4: Bias under stratified 10-fold cross-validation.

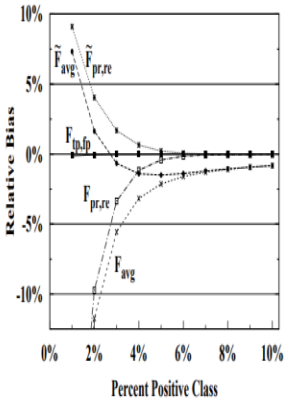

Figure 5: Bias under *unstratified* 10-fold cross-validation.

Results of the Simulation

Figure 5 depicts the relative bias of each technique using a classifier with a 10-fold stratified cross-validation. In ground-truth, the F-measure is exactly 80%. There is only one approach. Ftp,fp is almost perfectly neutral, and as a result it is the method for calculating F-measure that is approved. This is the situation. The most important outcome of this investigation. We then proceed to make an offer. For the biases of the other methods, use intuition. The x-axis is the horizontal axis. varies the class prior to from 1% to 5% positives in order to to demonstrate various impacts As we go to the left, we come across a Undefined accuracy is found in a higher percentage of test folds: the two ways that are used to replace zero in these instances(the smallest F-measure feasible) have a pessimistic attitude, The two ways that skip instead are Favg and Fpr,re.Favg and F pr,re, for example, exhibit a positive bias.that zero substitution is not a haphazard decision: The As we go closer to zero, the function converges to zero.

any point with an ambiguous precision or recall As a result, 0 is the right answer, and the negative bias may come as a surprise. first. The reason for this is due to the concave contour of the surface. See Section 4.1 for the F-measure function.As we move to the right, the folds get more precise.as a result, the differentiation between them vanishes.as though they were lines. The Fpr,re technique has a right-hand side.relative bias >1%, and the Favg technique has a lesser relative bias.there is a negative bias Why? Because F-measure works in the same way as a calculator, Any fold with a and-function between accuracy and recall, by chance, especially if the precision or recall are lowwill be given a failing gradeI score. There are tenfolds there are tenfolds there are tenfolds there are10 chances to receive a particularly poor gradeI score by accident, lowering the average Favg; on the other hand, averaging the

Over ten folds, precision and recall are usually achieved. Their harmonic mean Fpr,re is calculated from less extreme values. As a result, Fpr,re is significantly less likely to have a particularly low precision or recall score, and it demonstrates a significant amount of There is a positive bias. Then we look at how the bias varies depending on the groundtruth F-measure, which we alter from 60% to 95%. The trio The results of 10-fold stratified

crossvalidation for datasets with 1%, 5%, and 25% positives are shown in panels in Figure 6.As the ground-truth F-measure decreases for each dataset, Each method's bias tends to become more extreme over time.The same is seen in Figure 7 for unstratified 10-fold cross-validation. Except for the leftmost dataset, where the range of bias is substantially enlarged, the y-axis is kept the same (note its y-axis). Undefined precision without stratification Undefined recall might occasionally impact measures, as well. as previously described With a 5 percent positive dataset already, The zero-substitution algorithms Favg and Fpr,re are demonstrated. There is a significant negative bias. (In the graph on the right.) Fpr,re and Fpr,re are not shown because they are 25% positives. Ftp,fp is superimposed.) Ftp,fp is designed to handle all of these scenarios. Obviously, this is the recommended method.

Finally, we'll talk on Ftp, prejudice. fp's The same argument about the concavity of the F-measure holds true here, explaining a (very minor) negative bias. We sample on a regular basis. form a contingency table based on ground truth (our simulation) and The biases are then averaged. Underestimating the percentage of the population Over estimating it has a greater impact than actual positives. Especially around the vicinity of 0. The primary distinction between Ftp, fp, and The average cross-validation folds is one of the ways that average cross-validation folds. The former avoids the F-very measure's non-linear regions. By taking aggregates into account, functions near 0 can be found. As a result, then our tests, there was a two-order-of-magnitude skew. After we've looked at bias, we'll look at variance. Figure 8 depicts the standard deviation in terms of the mean. F-measure with a ground truth. We're at 5% positives and higher. Notice how Ftp, fp has the least amount of variance. It does not, however, At 1%, always exhibit the least variance, the other techniques Here, the bias is intolerable.

CONCLUSIONS AND DISCUSSION

The empirical study concludes that (a) Ftp,fp is the most unbiased method by far and should be utilised.

(a) This distinction becomes important when computing F-measure, and

(b)necessary for higher levels of class disparity as well as

for classifiers that are less accurate The Favg technique is a method for calculating averages.in general usage, penalises approaches that may be used on a sporadic basis. For some test folds, forecast zero positives. This results in a In some research publications, there is accidental and unwanted bias.to favourpractises that lean on the side of increased productionF alse positives are common. Naturally, this is of greater significance for Researchers who are interested in class disparities. However, software programmers should be concerned as well. whose software may one day be employed in unequally distributed classes conditions, as well as to researchers who are investigating large groups of people.

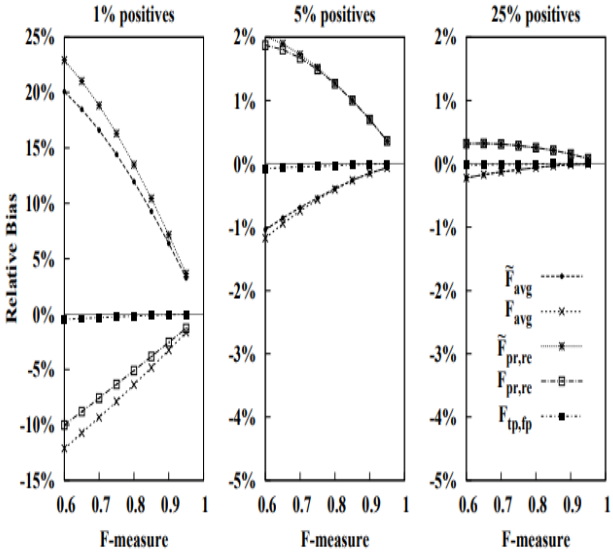

Figure 6: Relative bias under stratified 10-fold cross-validation.

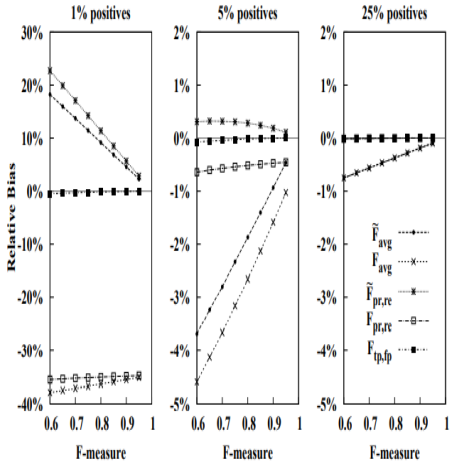

Figure 7: Relative bias under *unstratified* 10-fold cross-validation.

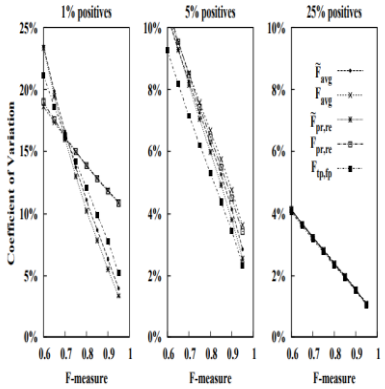

Figure 8: Relative Standard Deviation under stratified 10-fold cross-validation.

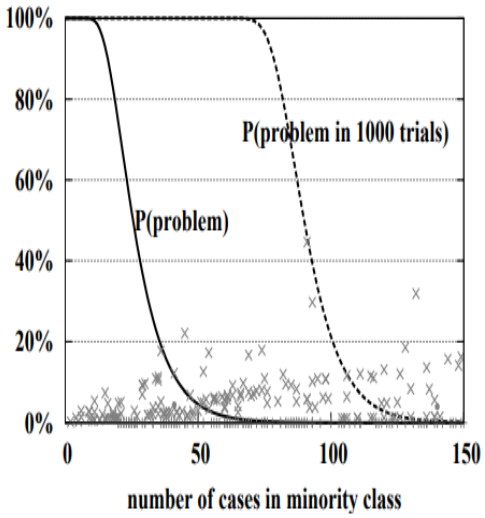

Figure 9: In 10-fold unstratified cross-validation, the likelihood of having at least one fold with no positives, resulting in undefined recall. The second curve depicts this probability increasing over time as a result of numerous separate trials, such as testing several different classes, studying many distinct datasets, or randomly splitting the same dataset.

Datasets in aggregate, especially datasets with several classes or multi-label configurations, without rigorous examination. Normally, the stratification option is used to reduce the number of people in a group. However, in certain research, the experimental variance is omitted. We face the risk of having no stratification if we don't stratify.positives in one or more of the folds, resulting in an ambiguous result AUC isn't defined and recall

isn't defined. If there is a significant increase in this risk, There are only a few positives in the dataset. Graph 9 reveals a tenfold increase in the likelihood of this condition developing. changing the number of positives in unstratified cross-validation available. The actual number of people is shown by the grey data points. Some binary classification jobs have positives available.as previously demonstrated in Figure 2. Given that each study is unique, endeavour is concerned with a large number of trials and/or multiple trials. within each dataset, and/or numerous classes being researched The right-hand curve in the datasets depicts the likelihood that In 1000 separate trials, the problem appears.

The idea is that when investigating datasets with fewer than 100 samples for a given class, it's likely that some of them may be missing. Experiments that aren't stratified will come across some folds that aren't stratified. positives to investigate AUC and maybe F-measure are the only options. undefined. Now, the simple solution is to always use caution. To circumvent this potential issue, employ stratification. But Only single-label datasets can benefit from stratification. In settings with several labels It is impossible to assure that each and every one of these requirements is met. Every fold (equally) represents every class. Thus, the possibility of meeting recall and AUC values that aren't defined is primarily a concern for multi-label setups, which is a growing field Interest in research is increasing.

To put things in perspective, there are a number of well-known hazards that are far more common than the minor computing issues mentioned in this paper: use simply a single, frequently ill-chosen baseline approach; failure to make certain the baselines have a suitable number of alternatives and tuning; as well as mistakenly releasing data from the test set, Occasionally, due to twinning in datasets containing

In training and testing, there are nearly identical circumstances. Altogether, Our multidisciplinary scientific community must continue to innovate. progress and, in general, implement best practices for high-quality production research into machine learning.

REFERENCES

[1] T. Raeder, G. Forman, and N. V. Chawla. Data Mining: Foundations and Intelligent Paradigms, chapter Learning with Imbalanced Data: Evaluation Matters. Intelligent Systems Reference Library. Springer Verlag, 2010.

[2] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. SIGKDD Explorations, 11(1), 2009.

[3] D. D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. volume 5, pages 361–397, 2004. http://www.jmlr.org/papers/volume5/lewis04a/ lewis04a.pdf.

[4] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Caasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531–537, 1999.

[5] D. D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In Symposium on Document Analysis and Information Retrieval, pages 81–93, Las Vegas, NV, Apr. 1994. ISRI; Univ. of Nevada, Las Vegas.

[6] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. Yale: Rapid prototyping for complex data mining tasks. In L. Ungar, M. Craven, D. Gunopulos, and T. Eliassi-Rad, editors, KDD ’06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 935–940, New York, NY, USA, August 2006. ACM.

[7] G. Forman. BNS feature scaling: an improved representation over TF-IDF for SVM text classification. In Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM), pages 263–270, New York, NY, 2008. ACM.

Citation: *Stbalinen I Saini et al, “Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement”, International Journal of Research Studies in Science, Engineering and Technology. 2019; 6(11): 40-52. DOI: DOI: https://doi.org/10.22259/2349-476X.0611005*