Generalized Agreement Statistics over Fixed Group of Experts

Mohak Shah

Accenture Technology Labs
161 N. Clark St, Chicago, IL, 60601, USA
mohak.shah@accenture.com

Abstract. Generalizations of chance corrected statistics to measure inter-expert agreement on class label assignments to the data instances have traditionally relied on the marginalization argument over a variable group of experts. Further, this argument has also resulted in agreement measures to evaluate the class predictions by an isolated classifier against the (multiple) labels assigned by the group of experts. We show that these measures are not necessarily suitable for application in the more typical fixed experts' group scenario. We also propose novel, more meaningful, less variable generalizations for quantifying both the inter-expert agreement over the fixed group and assessing a classifier's output against it in a multi-expert multi-class scenario by taking into account expert-specific biases and correlations.

Keywords: Agreement statistics, classifier evaluation, multiple experts

1 Introduction

Performance evaluation of learning algorithms over data for which deterministic (true) labeling is unknown comes with unique issues. When the ground truth is known, the evaluation consists of calculating a loss function between the true labels and the ones predicted by the classifier. An indicator loss function is used in the case of classification algorithms resulting in an accuracy estimate. However, various relevant application scenarios exist where expert-labels are sought since the true labels cannot be determined due to one or more issues such as inadequate data acquisition, limited knowledge of the application domain and so on. Note that the *expert* can also be an automatic labeling process (e.g., a classifier). Further, labels from *multiple experts* are typically obtained to mitigate the variability in individual estimates.[1] Examples of such applications include tasks such as medical image segmentation or alignment of stock price movement prediction approaches with other market indicators (e.g., market analysts' predictions). With technological advances such as Amazon's Mechanical Turk[2]

[1] In this paper, we do not consider novice or extremely imperfect label generation processes, including experts.

[2] http://www.mturk.com

(AMT), obtaining such labels for various human intelligence tasks is becoming increasingly easier.

It has been widely argued that chance agreements, resulting from experts' natural labeling propensities, should be taken into account while measuring inter-expert agreements. This is also the case when evaluating a classifier's performance, both against true (or even single expert generated) labels and against labels obtained from a group of experts. This argument resulted in various chance corrected agreement statistics such as Scott's π statistic (Scott 1955) and Cohen's kappa (Cohen 1960) measure (see (Kuncheva 2004, Japkowicz and Shah 2011) for discussion).

We consider the general form of this problem. Given a dataset S each of whose instances has been labeled by one of r experts, two quantities of interest need to be quantified. First, the extent of agreement *among* the r experts generating the labels, called the Inter-expert or Inter-rater agreement; and, Second, the extent to which the labeling output by a new classifier $\mathfrak{r}$ agrees with those of r experts as a group.

With regard to measuring the inter-expert agreement, however, the earlier attempts such as Cohen's kappa estimate resulted in statistics that applied only to *binary* classification scenarios over *two* sets of labels. One of the famous generalizations of Cohen's κ statistic was proposed by Fleiss (1971), denoted here by κ_F and has since been projected to be a standard in measuring inter-expert agreement (even hard coded in toolkits such as WEKA (Witten and Frank 2005)). Moreover, attempts motivated by argument along the lines of Fleiss (1971), although few, have also been made to quantify the agreement between a classifier (or an isolated expert) and a group of experts. One of the recent generalizations in this tradition has been that of Vanbelle and Albert (2009) that we will discuss later (we denote the unweighted variant $\hat{\kappa}$ in (Vanbelle and Albert 2009) here by κ_{va}).

In both these cases, the typical approach has been that of marginalizing over the experts comprising the group, under the *variable expert assumption* that each individual expert in the group comes from a (much larger) pool of experts and is interchangeable as long as the size of the group remains fixed. Marginalization over experts refers to obtaining probabilistic estimates of random assignment of a label to an example by a *random* expert. Since the experts need not be the same over instances in the variable expert assumption, this amounts to obtaining such estimates from the pool of all the labels by all the experts taken together. However, we contend that such estimation is not suitable over a *fixed* group of experts. In this case more information on correlated expert behavior is available and needs to be taken into account. By ignoring the expert specific biases and their correlations, the marginalized estimates invariably lead to pessimistic agreement measurements characterized by a higher variance in the *fixed expert group* scenario. In addition, the marginalization approach has more serious implications when there is a high heterogeneity in expert biases. Further, measures such as κ_{va}, motivated by similar arguments, also suffer from similar limitations when applied to the fixed expert scenario.

This paper proposes agreement statistics for measuring agreement *between* and *against* a group of *r fixed* experts. In particular, for inter-expert agreement, we propose a generalization of Cohen's kappa statistic to the case of multiclass classification (nominal scale) by a fixed group of experts. The proposed generalization has the property that it reduces to the classical version of Cohen's kappa in the case of binary classification by two experts. We then use this statistic to obtain a measure of agreement of a new classifier against the fixed group of experts. This argument results not only in tighter agreement statistics but also in more meaningful treatment of chance agreement as we will see below. A point to note here is that we do not assume existence of ground truth and as such neither attempt to learn the raters' behavior nor to obtain an estimate of the ground truth. Attempts along these lines have been made but differ in the inherent assumptions of the framework (see, for instance, (Raykar et al. 2010) which extends the STAPLE approach of (Warfield et al. 2004), or (Whitehill et al. 2009) that, based on different premise, propose estimating ground truth from multiple labels and also model the expertise of each labeler). This work also differs from the recent works in learning from crowds (e.g., (Snow et al. 2008)) settings in that we assume a setting in which the experts are assumed to be fixed and focus on obtaining evaluative estimates, as well as from works in developing probabilistic models (e.g., (Yan et al. 2010)) on annotater expertise to provide an estimate of true label in that we do not assume a determinable ground truth.

The rest of the paper is organized as follows: Section 2 proposes a new measure for inter-expert agreement estimation by treating chance agreement in a more coherent manner w.r.t. the observed agreement. Based on this, Section 3 then introduces a novel measure to estimate agreement between a classifier and a fixed group of experts. Both this sections also contrast the proposed measures with their respective marginalization-based counterparts. Section 4 provides an insight into both asymptotic and empirical behavior of the proposed statistics along with the crucial differences from related metrics. These insights are empirically supported by some results on synthetic and real data in Section 5. Section 6 discusses some related approaches along with their limitations with regard to fixed expert group scenario. Finally, we conclude in Section 7.

2 Measuring Inter-expert agreement

Let $S = \{\mathbf{i}_1, \ldots, \mathbf{i}_n\}$ denote a dataset with n instances. Each instance $\mathbf{i} \in S$ is assigned one of the k class labels from $\{l^1, \ldots, l^k\}$ by a group $\mathcal{R}$ of r experts. By c_{ij} we denote the number of experts assigning instance $\mathbf{i}_i$ to class l^j. Also, c_{pj} denotes the number of instances assigned to class l^j by expert p. Note that the measures such as κ_F (as well as κ_{va}) that marginalize over experts assume $\mathcal{R} \subset \mathfrak{R}$ where $\mathfrak{R}$ denotes a pool of experts from which $\mathcal{R}$ is drawn for different instances. However, under our setting $\mathcal{R}$ is considered to be fixed as is the case in many typical applications of the kind mentioned above.

Given any agreement statistic $\mathbb{A}$, a chance corrected agreement estimate can be defined as $\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$ where $\mathbf{E}_S(\mathbb{A})$ denotes the average empirical agreement between the experts on dataset S, $\mathbf{E}(\mathbb{A})$ denotes the true expectation of $\mathbb{A}$ and $\max_S(\mathbb{A})$ denotes the maximum achievable agreement between the experts on dataset S. Various characterizations of the agreement statistic $\mathbb{A}$ lead to different agreement estimates. For instance, Cohen's κ assumes $\mathbb{A}$ to be proportion of instances on which two raters agree in a binary classification scenario. Consequently, the true expectation measures the probability that these raters will agree just by chance (based on their individual labeling probabilities/biases) on a random example. In this sense, different agreement estimates attempt to capture different characteristics of the scenario to obtain assessments of rater agreements obtained above and beyond their coincidental concordance, also referred to as chance agreement (the expectation).

For notational simplicity, let us denote $\mathbb{A}_o = \mathbf{E}_S(\mathbb{A})$, $\mathbb{A}_e = \mathbf{E}(\mathbb{A})$ and $\mathbb{A}_{max} = \max_S(\mathbb{A})$. Hence,

$$\kappa = \frac{\mathbb{A}_o - \mathbb{A}_e}{\mathbb{A}_{max} - \mathbb{A}_e} \tag{1}$$

We adopt a pairwise agreement statistic for $\mathbb{A}$ to model the expert agreements. By taking into account the agreement between all the individual pairs of experts, we can quantify the overall observed agreement among the r experts as:

$$\mathbb{A}_o = \frac{1}{n}\sum_{i=1}^{n} A_o(\mathbf{i}_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} A_o(\mathbf{i}_i, l^j) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \frac{1}{r(r-1)} \sum_{p\in\mathcal{R}} \sum_{p'\in\mathcal{R}, p'\neq p} e_{ij}^{p} \cdot e_{ij}^{p'}$$

where $A_o(\mathbf{i}_i, l^j)$ is nothing but the proportion of pairwise agreement between experts over an instance $\mathbf{i}_i$ assigned to class l^j out of a total of $r(r-1)$ possible expert pairs; $e_{ij}^{p} = 1$ if the expert p assigns instance $\mathbf{i}_i$ to class l^j and zero otherwise. Notice that this includes the duplicate pairs of experts as well. However, we adhere to this more general form since potentially the pairwise costs may be asymmetric. Weights to take into account these asymmetric costs can then be easily integrated in this form.

The above computation yields:

$$\mathbb{A}_o = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \Big[\frac{1}{r(r-1)} c_{ij}(c_{ij}-1)\Big] \tag{2}$$

This agreement criterion has been widely utilized including the case of κ_F. The measures assuming a variable expert case obtain the chance agreement by relying on marginalization over the experts. For instance, Fleiss (1971) used $\mathbb{A}_e = \sum_{j=1}^{k} c_{\cdot j}^2$ where $c_{\cdot j} = \frac{1}{nr}\sum_{i=1}^{n} c_{ij}$. Replacing this in Equation 1 and setting $\mathbb{A}_{max} = 1$ gives the κ_F coefficient. However, this results in an excessively optimistic estimate over chance agreement which in turns gives a very conservative agreement statistic. For unique pairs of experts, the expectation of the pairwise $\mathbb{A}$ statistic has been discussed by (Hubert 1977). We use a similar argument for the more general case of all possible pairs used in $\mathbb{A}_o$ of Equation 2 and obtain, for the fixed experts case,:

$$\mathbb{A}_e = \sum_{j=1}^{k} A_e(l^j) = \sum_{j=1}^{k} \frac{1}{r(r-1)} \sum_{p \in \mathcal{R}} \sum_{p' \in \mathcal{R}, p' \neq p} \left[v_j^p v_j^{p'} \right] \tag{3}$$

where $A_e(l^j)$ quantifies the pairwise chance agreement between the experts on any given class label l^j such that v_j^p denotes the probability with which the expert p assigns a random instance to class l^j. The empirical estimate of $\mathbb{A}_e$ can be obtained from the data as:

$$\mathbb{A}_e = \sum_{j=1}^{k} \frac{1}{r(r-1)} \sum_{p \in \mathcal{R}} \sum_{p' \in \mathcal{R}, p' \neq p} \left[\frac{c_{pj}}{n} \frac{c_{p'j}}{n} \right] \tag{4}$$

Using $\mathbb{A}_o$ and $\mathbb{A}_e$ defined in Equations 2 and 4 respectively, along with $\mathbb{A}_{max} = 1$ in Equation 1, we get the desired inter-expert agreement statistic as:

$$\kappa_S = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} c_{ij} \cdot (c_{ij} - 1) - \frac{1}{n} \sum_{j=1}^{k} \sum_{p \in \mathcal{R}} \sum_{p' \in \mathcal{R}, p' \neq p} [c_{pj} c_{p'j}]}{nr(r-1) \left[1 - \frac{1}{n^2 r(r-1)} \sum_{j=1}^{k} \sum_{p \in \mathcal{R}} \sum_{p' \in \mathcal{R}, p' \neq p} [c_{pj} c_{p'j}] \right]} \tag{5}$$

Note that this statistic reduces to the classical version of Cohen's Kappa for the case of $k = r = 2$.

3 Measuring agreement against a group of experts

Let $\mathfrak{r}$ denote a new classifier (or expert) with $\mathfrak{r}_{ij}$ being unity if $\mathfrak{r}$ assigns a label l^j to instance $\mathbf{i}_i$ and zero otherwise. Since we assume $\mathfrak{r}$ to be a discrete classifier, $\mathfrak{r}_{ij}$ can be interpreted in a probabilistic sense.

In this discrete classification scenario, one way to measure the agreement of $\mathfrak{r}$ against $\mathcal{R}$ would be to measure the agreement of label assigned by $\mathfrak{r}$ against the proportion of experts from $\mathcal{R}$ over certain class of interest l^j while controlling for other classes (that is, mapping it to a binary problem by assigning 1 to l^j and 0 to all other classes), in a fashion similar to the IntraClass kappa Coefficient (ICC) (Kraemer 1979). Similarly, an empirical estimate over the expectation of this statistic (the chance agreement) could be obtained by marginalizing over this proportion (under variable assumption of $\mathcal{R}$ sampled from $\mathfrak{R}$ for each instance) in conjunction with the label assigned by $\mathfrak{r}$. Using these and then adjusting for the maximum achievable agreement can give us the extent to which $\mathfrak{r}$ agrees with $\mathcal{R}$. This one-against-all approach can then be extended to multi-class scenario by iterating over classes with a different l^j being set to 1 in each iteration. This is the approach adopted by Vanbelle and Albert (2009), referred to here as κ_{va}. However, the problems with this approach in the fixed experts' group scenario, are obvious. First is, of course, related to its formulation which ignores interaction biases over classes other than the class of interest since these classes are lumped together while mapping the problem to the binary case (although the parameter estimation takes the general form and is directly computable).

Moreover, due to the implicit assumption over variable $\mathcal{R} \in \mathfrak{R}$, it marginalizes over the expert biases. We propose an alternate measure for the fixed $\mathcal{R}$ case based on the consideration of the κ_S measure derived above.

Extending our notion of pairwise agreement between experts in $\mathcal{R}$ on each example, the overall observed agreement between $\mathfrak{r}$ and $\mathcal{R}$ can be obtained as:

$$\mathbb{A}_o = \mathbf{E}_{\mathbf{i}\sim S}[\mathbb{A}_o(\mathbf{i}_i)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{A}_o(\mathbf{i}_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathfrak{r}_{ij}A_o(\mathbf{i}_i, l^j) \tag{6}$$

where $\mathbb{A}_o(\mathbf{i}_i)$ is the agreement between $\mathfrak{r}$ and the group of raters over all the classes, with $A_o(\mathbf{i}_i, l^j)$ as defined in Equation 2.

Next, $\mathbb{A}_e$ for this case denotes the overall chance agreement between $\mathfrak{r}$ and the group of experts $\mathcal{R}$. Its empirical estimate can be obtained analogous to our previous discussion as:

$$\mathbb{A}_e = \sum_{j=1}^{k}\mathfrak{r}_j \cdot A_e(l^j) = \sum_{j=1}^{k}\left[\mathfrak{r}_j \cdot \frac{1}{r(r-1)}\sum_{p\in\mathcal{R}}\sum_{p'\in\mathcal{R},p'\neq p}\left[\frac{c_{pj}}{n}\frac{c_{p'j}}{n}\right]\right] \tag{7}$$

where $\mathfrak{r}_j = \mathbf{E}_{\mathbf{i}\sim S}\mathfrak{r}_{ij} = \frac{1}{n}\sum_{i=1}^{n}\mathfrak{r}_{ij}$ is the probability of the rater $\mathfrak{r}$ classifying a random example to class l^j.

The next important quantity to evaluate is the maximum achievable agreement $\mathbb{A}_{max}$, *between* $\mathfrak{r}$ and the group of experts $\mathcal{R}$. Note that the earlier measures assumed this to be unity (see, e.g., (Schouten 1982)). This turns out to be a significant limitation since $\mathbb{A}_{max} = 1$ if and only if *all* the raters in $\mathcal{R}$ agree on *all* the instance labels and further when $\mathfrak{r}$ agrees with these labels on all the instances. That is, this assumes κ_S to be unity as a pre-condition for the measure to achieve perfect score. However, this does not reflect the goal of assessing the extent of agreement of the classifier labeling with that of the group $\mathcal{R}$.

This is a crucial point. What we are interested in is the maximum agreement that the classifier $\mathfrak{r}$ can achieve against the *combined* labelings of $\mathcal{R}$ *independent of* the extent of agreement achieved among experts in $\mathcal{R}$ (of course still requiring at least some degree of inter-expert agreement for group qualification). Hence, the maximum agreement that $\mathfrak{r}$ can achieve would be when it assigns a labeling such that each assigned label corresponds to the label on which there is a maximum agreement *actually obtained* among the experts in $\mathcal{R}$.

With this consideration, we define the maximum possible agreement in our case as:

$$\mathbb{A}_{max} = \frac{1}{n}\sum_{i=1}^{n}\max_j A_o(\mathbf{i}_i, l^j) = \frac{1}{n}\sum_{i=1}^{n}\max_j\Big(\frac{1}{r(r-1)}(c_{ij}(c_{ij}-1))\Big) \tag{8}$$

Hence, replacing $\mathbb{A}_o$, $\mathbb{A}_e$ and $\mathbb{A}_{max}$ from Equations 6, 7 and 8 in Equation 1, the new measure, denoted as $\mathcal{S}$, of agreement between a classifier and a fixed group of experts becomes:

$$\mathcal{S} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{k}\mathfrak{r}_{ij}A_o(\mathbf{i}_i, l^j)\right] - \sum_{j=1}^{k}\mathfrak{r}_j \cdot A_e(l^j)}{\frac{1}{n}\sum_{i=1}^{n}\max_j A_o(\mathbf{i}_i, l^j) - \sum_{j=1}^{k}\mathfrak{r}_j \cdot A_e(l^j)}$$

Notice that, unlike the measure of Vanbelle and Albert (2009), $\mathcal{S}$ enables incorporating the expert-specific bias in calculating both $\mathbb{A}_e$ and $\mathbb{A}_{max}$ by considering $\mathcal{R}$ to be fixed. Also, $\mathcal{S}$ can be evaluated directly over all classes unlike the former.

4 Analysis

We present another main result of this work in the form of a theorem upper bounding the variance of the proposed κ_S statistic and showing how this is a more stable measure than κ_F in the fixed-experts case. The arguments for the variance analysis for various agreement measures follow from the large sample estimation of moments in the statistics literature (see, for instance, (Rao 2001)). While we relegate the detailed analysis of empirical variances and associated statistical significance tests for these statistics to the longer version of the paper, we nevertheless deem it important to discuss the following theoretical result.

Theorem 1. *Let κ_F and κ_S denote, respectively, the agreement statistics of Fleiss (1971) and that proposed in Equation 5 computed on a population (dataset) with large sample-size n where each of the sample has been assigned one of k labels by a fixed group of r experts. If $\sigma^2(\kappa)$ denotes the variance of κ then we have that:*

$$\sigma^2(\kappa_S) \leq \sigma^2(\kappa_F)$$

with equality satisfied when the experts emulate the pool.

Proof. The hypothesis of no agreement suggests labeling according to $\mathbf{E}(\mathbb{A})$ (or $\mathbb{A}_e$). Let us analyze this chance agreement $\mathbb{A}_e$ with regard to κ_S as defined in Equation 4. Under our formulation we can model the bias of each expert assigning example $\mathbf{i} \in S$ to a class $l^j, j \in \{1, \ldots, k\}$ as a multinomial b. That is, the multinomial $b_p(l^j)$ models the probability with which the expert p assigns a label l^j to a random example $\mathbf{i}$ chosen from S. The overall bias of expert p can then be modeled by a vector $\mathbf{b}_p = (b_p(l^1), \ldots, b_p(l^k))$. Hence, the chance agreement $\mathbb{A}_e$ essentially models these probabilities of the pairs of experts for each class which for the purposes of our analysis can be considered a constant. Therefore, the variance of κ_S depends basically on the variance of the observed agreement $\mathbb{A}_o$ of Equation 2. Then for large samples, for any agreement statistic $\mathbb{A}$, the variance of the metric $\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$ can be obtained as:

$$\sigma^2(\kappa) = \frac{\sigma^2(\mathbb{A})}{[\max(\mathbb{A}) - \mathbf{E}(\mathbb{A})]^2}$$

For the case of κ_S statistic, disregarding the constants, the expectation of the agreement statistic (denoted with superscript κ_S), is:

$$\mathbf{E}(\mathbb{A}^{\kappa_S}) = \sum_{j=1}^{k} \frac{1}{r(r-1)} \sum_{p \in \mathcal{R}} \sum_{p' \in \mathcal{R}, p' \neq p} \left[v_j^p v_j^{p'} \right] \tag{9}$$

Consequently, the variance becomes:

$$\sigma^2(\mathbb{A}^{\kappa_S}) = \sum_{p \in \mathcal{R}} \sum_{p' \in \mathcal{R}, p' \neq p} \left[\sum_{j=1}^{k} (v_j^p v_j^{p'} (1 - v_j^p - v_j^{p'})) + (\sum_{j=1}^{k} v_j^p v_j^{p'})^2 \right] \quad (10)$$

Similarly, for the case of κ_F, without differentiating between the experts (under the variable expert assumption) and disregarding constants, the expectation of the agreement term becomes,

$$\mathbf{E}(\mathbb{A}^{\kappa_F}) = \sum_{j=1}^{k} c_{\cdot j}^2 \quad (11)$$

Now, the variance $\sigma^2(\mathbb{A}^{\kappa_F})$, using Equation 2 for $\mathbb{A}_o$, can be approximated as:

$$\sigma^2(\sum_j c_{ij}^2) = 2r(r-1)[\sum_j (c_j^2) - (2n-3)(\sum_j (c_j^2))^2 + 2(n-2)\sum_j (c_j^2)] \quad (12)$$

Note, however, the crucial difference between the no agreement hypotheses assumed by κ_S and κ_F. In the case of former, we assume that experts label the instances according to their respective biases while in the case of latter, we assume that the labeling occurs in agreement with the marginals of the pool of experts.

Hence, it can be seen that the expectation $\mathbf{E}(\mathbb{A}^{\kappa_S})$ of Equation 9 is upper bounded by $\mathbf{E}(\mathbb{A}^{\kappa_F})$ of Equation 11. Similarly, Equation 12 upper bounds Equation 10. Now, since both κ_S and κ_F consider all possible expert pairs, the constants would be identical, i.e. $n^2r^2(r-1)^2$ in the denominator for the variance calculation of both measures. This concludes the proof. □

Using similar arguments, the variance of $\mathcal{S}$ can be seen to be upper bounded by the variance of κ_{va}. The sampling variances for $\mathcal{S}$ and κ_{va} can be computed using the Jackknife or leave-one-out method (Efron and Tibshirani 1993, Japkowicz and Shah 2011). For $\mathcal{S}$, let $\mathcal{S}\backslash i$ denote the agreement on the label assignments of all the instances in S except $\mathbf{i}_i$. Calculating $\mathcal{S}\backslash i$ repeatedly n times leaving a different instance each time and subsequently averaging it can then yield the pseudovalues.

4.1 Properties and Behavior

The marginalization argument for estimating $\mathbb{A}_e$, such as that in κ_F, can result in excessively pessimistic agreement estimates. That is, while such measures estimate the observed *agreement*[3] they do not measure the chance probabilities

[3] The pairwise consideration highlight that it would take at least 2 experts to agree on any given instance for the observed agreement to be non zero.

over agreements. As a result the inter-expert correlations, partly as a result of the variable experts assumption, are ignored. This not only results in a loose estimate of $\mathbb{A}_e$ but can also yield less meaningful (even unwarranted negative) values of agreement measure even when the empirical evidence is to the contrary, for smaller values of $\mathbb{A}_o$. We will illustrate this in the next Section. In the fixed expert case, κ_S offers better consistency in the estimation of $\mathbb{A}_o$ and $\mathbb{A}_e$.

Similarly, while κ_{va} depends on the proportion of experts with maximum labels when computing $\mathbb{A}_{max}$, $\mathcal{S}$ depends on the labels on which the pairwise agreement over $\mathcal{R}$ is maximum. Hence, even when all the experts disagree over labels for all the instances, $\mathbb{A}_{max}$ is not zero for κ_{va} while it is zero in the case of $\mathcal{S}$. The latter is indeed desirable in the fixed expert group case since in the event of no agreement among the experts themselves (extreme variability), the agreement of the classifier with any individual expert, being unrepresentative of the *group* agreement, is rendered meaningless (even more so when $k > r$). Similar differences exist in the computation of other quantities. The marginalization argument when applied to the case of calculating $\mathbb{A}_e$ can also result in an overly conservative, and sometimes less meaningful, estimates of κ_{va} in contradiction with the the empirical evidence (as we will see in the next Section). There is an important point to be made here. While, analytically, it can be seen that κ_F is more conservative than κ_S, such a relationship need not exist between κ_{va} and $\mathcal{S}$ since their estimates would depend not only on the expert labels but also on their subsequent agreement with the new classifier. The contribution of individual expert (even when it disagrees with all the others) is not zero for both κ_F and κ_{va}.

5 Empirical Results and Discussion

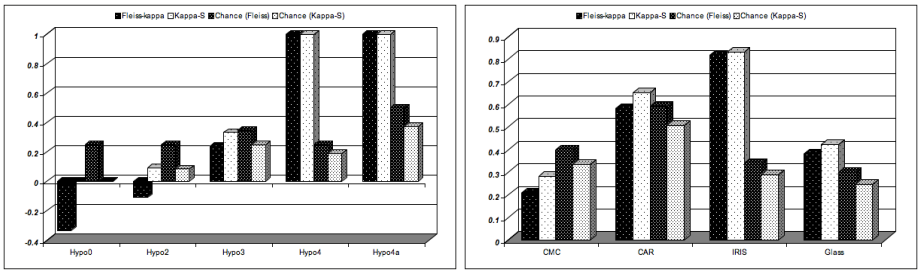
We compare the behavior of the most commonly employed κ_F metric for inter-expert agreement measurement against the proposed measure κ_S. With regard to estimating the agreement of a classifier with group of fixed experts, we compare the generalization proposed by Vanbelle and Albert (2009) denoted as κ_{va} with the proposed $\mathcal{S}$ metric.

For both sets of comparisons, we use 4 different sized multi-class datasets from UCI repository (Asuncion and Newman 2007) over WEKA implementations of 6 different classifiers in addition to their true labels. Further, we also illustrate the limitations of κ_F in the fixed experts case with the help of synthetic data. The main aims of the simulations presented here are two-fold: i) illustrating the differences between the compared measures; and ii) highlighting the discrepancies in the estimation of variable expert assumption based measures when applied to the fixed experts cases making them unsuitable for the purpose. The datasets used include CMC (1473 instances, 3 classes), Car (1728 instances, 4 classes), Iris (150 instances, 3 classes) and Glass (214 instances, 7 classes) while the learning algorithms used are Support Vector Machine (with linear kernel), Naive Bayes, C4.5 Decision Trees, 3-Nearest Neighbor, Ripper and a Conjunction Rule learning algorithm. The reported results are over 10-

fold Cross Validation with default parameter values over learning algorithms.[4] Finally, we illustrate these differences in a real world example of Syphilis Serogen data (Williams 1976). Novel venues such as AMT can also yield relevant data for such simulations. While data from learning from crowds scenario are sometime publicly available (Snow et al. 2008), we are not aware of any relevant AMT datasets available yet for the fixed experts case.

5.1 Evaluating Inter-expert Agreement

Let us first consider a simple synthetic dataset of 200 instances labeled by 4 experts (E1, E2, E3 and E4) into one of the 4 classes (L1, L2, L3 and L4) and consider 5 different illustrative scenarios. The first label configuration "Hypo0" denotes the case when on each instance all four experts disagree. We do this by simply making E1 assign L1, E2 assign L2 and so on to all instances. Next, we flip the first 100 labels of E1 from L1 to L2 and the last 100 labels of E2 from L2 to L1 so that these two experts agree on all the labels while still disagreeing with E3 and E4 who themselves are in disagreement. This case is denoted "Hypo2". Next, from "Hypo2", we let E3 assign L1 to the first 100 instances and L2 to last 100 instances so that E3 agrees with both E1 and E2 yielding dataset "Hypo3". We then obtain a dataset "Hypo4" where all experts assign L1, L2, L3 and L4 to the respective subsets of 50 instances and are in complete agreement. Finally, We obtain a dataset "Hypo4a" that too simulates all experts in agreement but this time all the experts assign L1 to the first 100 instances and L2 to last 100 instances. (Note that the suffix in the name of each synthetic variation denote the number of experts in complete agreement). The results are presented in Figure 1(a).



(a) Synthetic Cases (b) UCI Data

Fig. 1. Comparison of the proposed κ_S (kappa-S in figures) statistic against the Fleiss' Kappa coefficient κ_F. Also shown are the corresponding estimates of chance agreements in both cases. An absence of a bar indicates a zero value.

In the case of "Hypo0", $\mathbb{A}_e > 0$ for κ_F so that $\kappa_F < 0$. Note here that there is a strict heterogeneity among experts' biases in this case (e.g. E1 never

[4] Note that model selection is not our main concern here since we aim to show the difference in the agreement estimates.

assigns any label other than L1, and so on) and the estimation of $\mathbb{A}_e$ using the marginalization argument is not meaningful since it does not reflect the probabilities of random label assignments. Unlike κ_F, $\mathbb{A}_e = 0$ for κ_S. Similarly, the marginalization argument over $\mathbb{A}_e$ results in negative value for κ_F in the event of partial agreement in the case of "Hypo2" data where E1 and E2 agree on all the instances. Again, this is not desirable since, here, both the $\mathbb{A}_o$ and $\mathbb{A}_e$ are solely based on E1 and E2 for classes L1 and L2. Keeping in view the label assignment in these cases, κ_S gives a more realistic estimate. Note that $\kappa_F \to \kappa_S$ as $\mathbb{A}_o \to 1$. This can be seen in the case of "Hypo4" and "Hypo4a" datasets. However, even when $\mathbb{A}_o$ for both the measures is 1, the chance agreement is not treated in the similar manner in these two cases.

Next, we compare the inter-expert agreement between the set of 7 label sets obtained on the 4 UCI datasets, one from each of the 6 classifiers, and the true labels of the instances, using both the κ_F and κ_S measures in Figure 1(b). As a result of optimistically estimating the chance agreement by marginalization over experts, κ_F is consistently more conservative than κ_S. However, the measures seem to converge with increasing levels of agreement with $\kappa_F \to \kappa_S$ as the agreement approaches unity. An example can be seen in the case of Iris datasets where classifiers typically obtained a very high accuracy rate and are in high agreement. However, the gap between the two measures is higher for moderate to low agreement values.

5.2 Evaluating agreement against a group of fixed experts

We consider the UCI datasets to compare $\mathcal{S}$ against κ_{va}. For each case, the experts' group is simulated by taking into account the true labels along with the two classifiers achieving highest 10-fold accuracy on each dataset. The (unweighted) κ_{va} and $\mathcal{S}$ are then estimated for each of the remaining classifiers against the group. The results are presented in Figure 2. As can be seen, κ_{va} consistently results in a conservative agreement estimate as compared to $\mathcal{S}$ in these cases (at least partly due to marginalization). Note, in particular, the case of Car and CMC datasets over the comparison of conjunction rule learner against expert labels. While in both cases Conjunction rule learner obtains a trivial classifier assigning class 1 to all the instances, it should be noted that this class is highly overrepresented in these datasets (about 70% in CAR and 42% in CMC). While κ_{va} gives a less meaningful null estimate in both these cases, this scaling is better captured by the $\mathcal{S}$ measure as can be seen in the last column of Figure 2. Also, the two measures seem to converge as $\mathbb{A}_o$ approaches unity (see e.g., the Iris dataset over C1, C2 and C3 in Figure 2).

5.3 Illustration on real data

We illustrate the proposed indices of agreement on the Syphilis Serogen data of Williams (1976) who presented result obtained by three reference laboratories (denoted by Ref-1, Ref-2 and Ref-3) and an additional participant laboratory (denoted by T) on 28 specimens (data is shown in Table 1). Each specimen

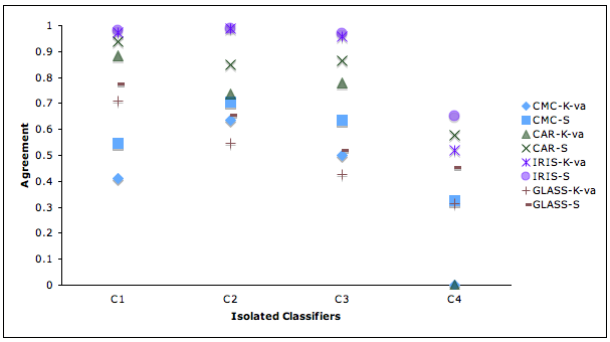


Fig. 2. Comparison of the proposed $\mathcal{S}$ measure against the κ_{va} statistic for measuring the agreement of isolated classifier (on horizontal axis) against expert labels composed of the actual labels and outputs of SVM and Decision Trees for the case of Car, CMC and Iris datasets; and of actual labels along with outputs of Ripper and 3-NN in the case of Glass dataset. C1 and C4 represent 3-NN and Conjunction Rule respectively in all datasets. C2 and C3 represents, respectively, SVM and NB in the case of Glass data while NB and Ripper in the case of the other three datasets.

was classified into one of the three classes viz. Non Reactive (NR), Borderline (BL) and Reactive (RE). The additional participant laboratory also classified the 28 specimen into one of the three classes. The same dataset was also used by Vanbelle and Albert (2009, Table 3). For both κ_F and κ_S, we get $\mathbb{A}_o = 0.81$ between the three reference laboratories. The difference, analogous to the synthetic cases, appears in terms of optimistic estimate of $\mathbb{A}_e = 0.412$ in the case of κ_F. In the case of κ_S, we obtain $\mathbb{A}_e = 0.272$. Hence, we obtain $\kappa_F = 0.676$ and $\kappa_S = 0.738$. We can see how κ_F results in a pessimistic agreement estimate due to overestimating chance agreement. Also, note the difference in the results obtained for κ_S as compared to the agreement statistic such as ICC which was found to be 0.68 as reported by Vanbelle and Albert (2009).

Similarly, when comparing the three reference laboratories to laboratory T, we obtain, for κ_{va}: $\mathbb{A}_o = 0.655$, $\mathbb{A}_e = 0.362$ and $\mathbb{A}_{max} = 0.893$. On the other hand, for $\mathcal{S}$, we get: $\mathbb{A}_o = 0.571$, $\mathbb{A}_e = 0.105$ and $\mathbb{A}_{max} = 0.81$. This gives $\kappa_{va} = 0.551$ and $\mathcal{S} = 0.662$. These results too demonstrate the manner in which the two measures differ in the estimation of various quantities.

6 Related Work

In addition to Fleiss' coefficient, various other general inter-expert agreement measures such as the well known ICC (Kraemer 1979) or context-specific measure of Schouten (1982), have also appeared (e.g., see (Kuncheva 2004) for a discussion on some such measures in the context of classifier fusion). However, these measures too typically marginalize over the experts.

In this respect the proposed κ_S statistic is more in line with the arguments of Berry and Mielke Jr (1988) who propose a generalization over interval measurements and multiple experts by way of measuring the extent of disagreements

#	**T**	**Ref-1**	**Ref-2**	**Ref-3**
1	RE	RE	RE	RE
2	RE	RE	RE	RE
3	BL	NR	NR	NR
4	BL	NR	NR	NR
5	BL	NR	NR	NR
6	RE	RE	RE	RE
7	BL	NR	NR	NR
8	RE	RE	RE	RE
9	NR	NR	NR	NR
10	NR	NR	NR	NR
11	RE	RE	RE	RE
12	RE	RE	BL	BL
13	RE	RE	RE	RE
14	RE	RE	BL	BL
15	RE	RE	RE	RE
16	RE	RE	NR	BL
17	RE	RE	NR	BL
18	RE	RE	RE	RE
19	RE	RE	RE	RE
20	BL	BL	NR	NR
21	RE	RE	RE	RE
22	BL	NR	NR	NR
23	BL	BL	NR	NR
24	BL	BL	NR	NR
25	RE	RE	RE	RE
26	NR	NR	NR	NR
27	RE	RE	RE	RE
28	NR	NR	NR	NR

Table 1. Syphilis Serogen data of (Williams 1976) used in Section 5.

between the experts in the l_2-norm setting. However, the disagreement measured by the Δ function there need not reflect the corresponding agreement under the l_2-norm. Furthermore, it requires rescaling the label assignments.

With regard to measuring the agreement against the group of experts, another commonly applied approach is the consensus approach where, for each instance, the label assigned by a majority (defined by a consensus threshold) of the experts is considered as the true label (see, for instance, (Soeken and Prescott 1986, Smith et al. 2003)). This simplifies the subsequent evaluation against a classifier by mapping the problem to a deterministic label matching problem amenable to more conventional techniques such as Cohen's kappa. However, such consensus labeling makes the output dependent on the consensus threshold and has serious limitations. In addition, issues such as not accounting for experts' dispersion as well as difficulty in dealing with instances with no consensus makes this approach highly contentious (Eckstein et al. 1998, Salerno et al. 2003, Miller et al. 2004).

Approaches proposed to bypass such consensus requirement such as those of Schouten (1982), Williams (1976) and Light (1971) either do not address the problem of interest directly or pose issues such as introduction of bias or ignoring

interdependence of experts (Vanbelle and Albert 2009). Note, in particular, that the approach of Schouten (1982), even though applied in fixed expert settings, disregards the interdependence of experts when measuring agreement of one expert against others in the *same* group.

Another important caveat in above approaches lies in the assumption over the maximum achievable agreement between the classifier and the group of experts being unity. This caveat has profound implications since it makes the assessment of classifier performance *dependent on the inter-expert agreement*. Such measures, hence, can achieve a perfect score for the classifier *only when* the inter-expert agreement is unity which essentially obviates the need for (and utility of) multiple experts altogether. Vanbelle and Albert (2009) also noted these limitations and proposed an alternative general measure (κ_{va} discussed earlier). As mentioned above, κ_{va} too followed the marginalization argument in a binary classification case. It was then extended to the multi-class case in an indirect manner using an iterative one-against-all strategy.[5] We compared κ_{va} against $\mathcal{S}$ on various criteria above.

7 Conclusion and Future Work

In this paper, we noted the main limitations of measures based on marginalization over experts rendering them unsuitable for application in the typical fixed experts' group scenario. Among the crucial issues lie the excessively conservative agreement estimate obtained by the inter-expert agreement measures such as κ_F. Moreover, these measures, as seen in both theoretical arguments and empirical results, can yield less meaningful values when the heterogeneity in the expert biases is high. We also proposed two novel statistics, respectively, to measure inter-expert agreement (κ_S) *between*, and agreement of a classifier *against*, a fixed group of experts ($\mathcal{S}$) in the general case of multiple classes and multiple experts. The main advantage of the proposed measures can be seen in terms of their accounting for expert specific biases and correlations yielding tighter agreement assessments. The proposed measure $\mathcal{S}$ also scales the maximum achievable agreement in accordance thereby allowing more meaningful characterization of classifier's performance that is *independent* of the agreement achieved *within* the expert group. Finally, in contrast to the marginalization based measures, κ_S reduces to the classical Cohen's κ in the binary classification case over two label sets. The future work includes investigating the behavior and dependence of proposed statistics, as well as extending them, to testing scenarios such as asymmetric loss, bias, prevalance and class imbalance. Another area worth investigating is the sample size requirement for the data over classes since the expert specific biases are obtained from the data empirically. A sparse class can in principle affect such estimates adversely (of course, even in this case, the assessed biases are best that can be obtained in accordance with both the maximum likelihood as well as information-theoretic arguments). Finally, the proposed measures can also be generalized for probabilistic classifiers.

[5] This effectively generalizes Fleiss' kappa, or alternatively Scott's π statistic and not Cohen's kappa.

Bibliography

A. Asuncion and D. J. Newman. UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science, 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

K. J. Berry and P. W. Mielke Jr. A generalization of cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurements*, 48:921–933, 1988.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurements*, 20:37–46, 1960.

M. P. Eckstein, T. D. Wickens, G. Aharonov, G. Ruan, C. A. Morioka, and J. S. Whiting. Quantifying the limitation of the use of consensus expert commitees in roc studies. In *Proceedings SPIE: Medical Imaging 1998*, volume 3340, pages 128–134, 1998.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, New York, 1993.

J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

L. Hubert. Kappa revisited. *Psychological Bulletin*, 84(2):289–297, 1977.

N. Japkowicz and M. Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, New York, 2011.

H. C. Kraemer. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika*, 44:461–472, 1979.

L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004. ISBN 0471210781.

R. L. Light. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin*, 76:365–377, 1971.

D. P. Miller, K. F. O'Shaughnessy, S. A. Wood, and R. A. Castellino. Gold standard and expert panels: a pulmonary nodule case study with challenges and solutions. In *Proceedings SPIE: Medical Imaging 2004: Image Perception, Observer Performance and Technology Assessment*, volume 5372, pages 173–184, 2004.

C. R. Rao. *Linear Statistical Inference and its Applications, 2nd Ed.* Wiley, New York, 2001.

V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11: 1297–1322, 2010.

S. M. Salerno, P. C. Alguire, and S. W. Waxman. Competency in interpretation of 12-lead electrocardiograms: a summary and appraisal of published evidence. *Annals of Internal Medicine*, 138:751–760, 2003.

H. J. A. Schouten. Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica*, 36:45–61, 1982.

W. A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Q*, 19:321–325, 1955.

R. Smith, A. J. Copas, M. Prince, B. George, A. S. Walker, and S. T. Sadiq. Poor sensitivity and consistency of microscopy in the diagnosis of low grade non-gonococcal urethrisis. *Sexually Transmitted Infections*, 79:487–490, 2003.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. ACL, 2008.

K. L. Soeken and P. A. Prescott. Issues in the use of kappa to estimate reliability. *Medical Care*, 24:733–741, 1986.

S. Vanbelle and A. Albert. Agreement between an isolated rater and a group of raters. *Statistica Neerlandica*, 63(1):82–100, 2009.

S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*, 23(7):903–921, 2004.

J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pages 2035–2043, 2009.

G. W. Williams. Comparing the joint agreement of several raters with another rater. *Biometrics*, 32:619–627, 1976.

I. H. Witten and E. Frank. *Weka 3: Data Mining Software in Java*. http://www.cs.waikato.ac.nz/ml/weka/, 2005.

Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, and J. Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS), Vol. 9 of JMLR*, pages 932–939, 2010.