The kappa measure of inter-coder agreement on classification tasks

Andrei Popescu-Belis
IDIAP Research Institute
December 18, 2007

Plan of the talk

• Kappa (κ)
  – origins, definition

• Computing kappa
  – assumptions on annotators’ behavior
  – acceptable values & significance

• Limitations, generalizations

• Applications & conclusion

Kappa (κ)

• Origins of kappa
  – medicine, psychology, behavioral sciences (>1950s): diagnoses
  → Scott’s pi (1955), Cohen’s alpha (1960), Fleiss’ kappa (1971)
  → Landis and Koch (1977), Siegel and Castellan (1988)
  – social sciences: content analysis (> late 1970s)
  – natural language processing: corpus annotation (> 1996)
  → Carletta (1996), discussions > 2000
  – and probably others that I am not aware of…

Motivation

• Measuring agreement using “accuracy” or “raw agreement” (% of instances on which annotators agree) is not sufficient
  – to be corrected by considering agreement by chance
  → e.g., if two annotators classify instances into N classes at random, then they reach… 1/N agreement

• More realistic example with two annotators
  – classify meeting samples as ‘constructive’ / ‘destructive’ / ‘neutral’
  – observed frequencies are around 15% C, 15% D, 70% N
  – the two annotators agree on 70% of samples… are we happy with this annotation?
  → not really; if they answer randomly with above frequencies → 53.5%
Definition

- "Proportion of agreement above chance"
  \[ P(A) = \frac{a + d}{a + b + c + d} \]
  \[ P(E) = \frac{(a + c)(a + b)}{(a + b + c + d)^2} \]
  \[ K = \frac{P(A) - P(E)}{1 - P(E)} \]

\[ K = 1, \text{ perfect agreement} \]
\[ K = -1, \text{ total contradiction} \]
\[ K = 0, \text{ independence / no correlation} \]

How is \( \kappa \) computed?

- Main challenge: estimate \( P(E) \)
  - i.e. the probability of agreeing by chance
  - from a limited number of annotation samples

- Based on the proportions of each category used by each annotator
  - two main options / two versions of \( \kappa \)
    - specific proportions for each annotator (Cohen 1960)
    - same proportion for all annotators (all the others...)

Graphical representation (1):
contingency table / confusion matrix

- The \( a \) priori probability of Coder A to...
  - answer 'Cat1' is \( \frac{a}{a+b+c+d} \)
  - answer 'Cat2' is \( \frac{b}{a+b+c+d} \)
  (and conversely for Coder B)

- Hence, and

Graphical representation (2): agreement matrix

- The \( a \) priori probability is estimated from all coders’ data as

\[ P(E) = \frac{1}{N} \sum_{i=1}^{k} p_i \]

where

\[ p_i = \frac{1}{N} \sum_{j=1}^{n} \frac{n_{ij}}{n} \]

is the probability of each category

Differences between the two versions

- "Generally small, especially when \( \kappa \) is high"
- Despite apparently different formulae, \( P(A) \) is the same, but there is a small difference in \( P(E) \):
Example: annotating meeting samples with ‘constructive’ / ‘destructive’ / ‘neutral’

Two annotators over 500 samples: ~ 16% C, 14% D, 70% N

Results:

<table>
<thead>
<tr>
<th>Item</th>
<th>C</th>
<th>D</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Item 2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Totals</td>
<td>165</td>
<td>140</td>
<td>695</td>
</tr>
</tbody>
</table>

$P(A) = 0.78000$ and $P(A) = 0.78000$

$P(E) = 0.52985$ vs. $P(E) = 0.52950$

$\kappa = 0.53206$ vs. $\kappa = 0.53241$

Statistical significance of kappa

- Determine whether two classification experiments are different or not: “Is $\kappa_1 \neq \kappa_2$ significant?”
- Provide a confidence interval for a value of $\kappa$
- Test an hypothesis, e.g. that $\kappa = 0$ for an experiment
- Formulas for the variance of $\kappa$ are available in (Everitt 1994, p. 29), (Cohen 1960), (Siegel & Castellan 1988)
  - significance can then be computed using Student’s law

What is a good kappa value?

- $\kappa = 1 \rightarrow$ identical annotations / $\kappa = 0 \rightarrow$ independence
- Strictly below 1, only subjective considerations relate $\kappa$ values and annotation acceptability: no general scale!
- Landis and Koch 1977 Krippendorff 1980

- $\kappa < 0$ Poor
- $0 < \kappa < 0.2$ Slight
- $0.2 < \kappa < 0.4$ Fair
- $0.4 < \kappa < 0.6$ Moderate
- $0.6 < \kappa < 0.8$ Substantial
- $0.8 < \kappa < 1$ Almost perfect

Desired behaviors of $\kappa$ or parasitic effects?

- Inter-observer bias
  - annotators could be internally biased towards a given class
  - when $P(E)$ is computed from individual distributions, if $P(A)$ constant
    - less similar distributions $\rightarrow$ lower $P(E) \rightarrow$ larger $\kappa$
    - $\kappa$ increases as distributions become less similar!!!
  - possible test for bias: Cochran’s Q-test
    - (no bias $\Rightarrow$ Q is a chi-square)
- Prevalence
  - for a given $P(A)$, the variations of $P(E)$ (from $1/N$ to 1) change $\kappa$
    - larger $P(E)$ ($= prevalence of a class$) $\rightarrow$ lower $\kappa$

Two problems for kappa

Generalizations of kappa

- More than two annotators
  - use agreement tables (not contingency) and formulas above
  - or compute kappa per class (one class against all others) and do the average (same result)
    - do not average pairwise values of kappa
- Ordered-category data
  - i.e., some classes are closer than others
  - weighted version of kappa
    - difficulty: assign proper weights in a contingency table
- Scalar data
  - beyond kappa: Pearson chi-square, t-test, ANOVA

Other potential measures

- Raw agreement, or accuracy
  - but kappa was designed to overcome some of their limits...
- Intraclass correlation
  - how one annotator differs from the average
  - several versions, extending towards ANOVA
- McNemar test
  - compare two classification algorithms
    - null hypothesis: the algorithms are similar
    - see (Dietterich 1998) for an application to machine learning
- They do not apply to the same situations!
Using kappa to measure classification performance

- Mutatis mutandis
  - compare ground truth with system output
  - compute kappa as if these were two annotators

- Advantages
  - relatively ‘cheap’ if already computed for inter-coder agreement
  - compare system’s kappa with inter-coder one
    - if they are close, work on improving the data, not the system!

- Limitations
  - kappa is symmetric, its use for evaluation is not
  - rather ‘strict’ – high values are difficult to reach
  - could also use per-class accuracy, or recall/precision

Bottom line

- Kappa is frequently used in some domains
  - so, even if it has limitations, this provides good references to which one can compare their score
  - its behavior is quite well understood
  - comparison & scales still a problem
  - example: annotation → kappa → adjudication → better kappa?

- Applications (often in Cohen’s 1960 version)
  - mono- and multimodal annotations that can be defined as classification (labeling) problems
  - good for ‘uncertain’ annotations: no obvious ground truth, coding schemes under development, taxonomies in progress, etc.
    - Higher-level abstractive annotations

References (1)


References (2)


Websites

http://faculty.vassar.edu/lowry/VassarStats.html
- explanations and online calculator / for kappa, check “Frequency data”
http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm
- Statistical Methods for Rater Agreement
- guidelines for selecting metrics, excellent list of references
- very clear explanations… but in French
http://en.wikibooks.org/wiki/Algorithm_implementation/Statistics/Fleiss%27s+kappa
- java implementation of Fleiss’ kappa
http://www.dmi.columbia.edu/homepages/chuang/kappa/
- other ‘Kappa calculators’ are available via Google search