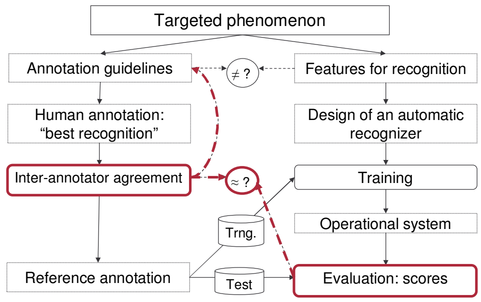
The kappa measure of inter-coder agreement on classification tasks

Andrei Popescu-Belis
IDIAP Research Institute

December 18, 2007

Reference data and evaluation in automatic categorization tasks



2

Plan of the talk

- Kappa (κ)
  - origins, definition
- Computing kappa
  - assumptions on annotators' behavior
  - acceptable values & significance
- Limitations, generalizations
- Applications & conclusion

3

Kappa (κ)

- Goal
  - measure agreement between annotators (or raters) on classification tasks
  - often when the classes/data have nominal values
    - e.g. psychologists or students rating subjects as 'bipolar', 'depressed' or 'normal'
  - often when the ground truth is difficult to determine
    - hence the importance of human observers

4

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)
    - → Krippendorff (1980, 2004)
  - natural language processing: corpus annotation (> 1996)
    - → Carletta (1996), discussions > 2000
  - and probably others that I am not aware of…

5

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%

6

Definition

- "Proportion of agreement above chance"
  $P(A)$ = observed agreement (as a percentage of the total number of classified instances)
  $P(E)$ = agreement due to chance
- Corrected measure: $\kappa = \dfrac{P(A) - P(E)}{1 - P(E)}$

  maximum: κ = 1, perfect agreement
  minimum: κ = -1, total contradiction
  κ = 0, independence / no correlation

How is κ 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 κ
    - 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 (*a*+*c*) / total
  - answer 'Cat2' is (*b*+*d*) / total
    (and conversely for Coder B)
- Hence, $P(A) = \dfrac{a+d}{a+b+c+d}$
  and
  $P(E) = \dfrac{(a+c)(a+b)}{(a+b+c+d)^2} + \dfrac{(b+d)(c+d)}{(a+b+c+d)^2}$

		Coder A		
		Cat1	Cat2	
Coder B	Cat1	*a*	*b*	*a+b*
	Cat2	*c*	*d*	*c+d*
		a+c	*b+d*	total

Graphical representation (2): agreement matrix

- The *a priori* probability is estimated from all coders' data as $P(E) = \sum_{j=1}^{k} p_j^2$

where $p_j = \dfrac{1}{N} \sum_{i=1}^{N} \dfrac{n_{ij}}{n}$

is the probability of each category

	Number of assignments (total = *n* per item)			
	Cat.1	Cat.2	.. *j* ..	Cat.k
Item 1	0	0	…	*n*
Item 2	0	3	…	1
Item *i*	0	0	n_{ij}	1
Item *N*	…	…	…	…
"Totals"	p_1	p_2	p_j	p_k

Graphical representation (2): agreement matrix *(continued)*

- Then, the proportion of observed agreement $P(A)$ is computed using π_i , the average proportion of agreement for each item (computed over all *k* categories, for *n* annotators)

$$\pi_i = \sum_{j=1}^{k} \frac{n_{ij}}{n} \cdot \frac{n_{ij} - 1}{n - 1}$$

- So, $P(A) = \dfrac{1}{N} \sum_{i=1}^{N} \pi_i$, and again $\kappa = \dfrac{P(A) - P(E)}{1 - P(E)}$

Differences between the two versions

- "Generally small, especially when κ is high"
- Despite apparently different formulae, $P(A)$ is the same, but there is a small difference in $P(E)$:
- First case: $P(E) = \sum_{j=1}^{k} \left(\dfrac{1}{N} \sum_{i=1}^{N} \dfrac{n_{ij}}{n} \right) \left(\dfrac{\left(\sum_{i=1}^{N} n_{ij}\right) - 1}{Nn - 1} \right)$

- Second case: $P(E) = \sum_{j=1}^{k} \left(\dfrac{1}{N} \sum_{i=1}^{N} \dfrac{n_{ij}}{n} \right)^2$

Example: annotating meeting samples with ‘constructive’ / ‘destructive’ / ‘neutral’

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

	Nb. of codings (n=2)		
	C	D	N
Item 1	0	2	0
.. *i* ..	1	0	1
Item *N*	0	0	2
Totals	165	140	695

		Coder A			
		C	D	N	
Coder B	C	50	5	25	80
	D	10	40	15	65
	N	25	30	300	355
		85	75	340	500

➔ Results:

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

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

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

What is a good kappa value?

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

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	$0.67 < \kappa < 0.8$	Allows tentative conclusions
$0.8 < \kappa < 1$	Almost perfect	$0.8 < \kappa < 1$	Good reliability

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 κ
- Test an hypothesis, e.g. that $\kappa = 0$ for an experiment
- Formulas for the variance of κ are available in (Everitt 1994, p 29), (Cohen 1960), (Siegel & Castellan 1988)
  - significance can then be computed using Student’s law

Two problems for kappa

- 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 ↔ lower $P(E)$ ↔ larger κ
    - ➔ κ increases as distributions become less similar!!!
  - possible test for bias: Cochran’s Q-test
    - (no bias ➔ Q is a chi-square)
- Prevalence
  - for a given $P(A)$, the variations of $P(E)$ (from $1/N$ to 1) change κ
    - larger $P(E)$ (= prevalence of a class) ↔ lower κ

➔ Desired behaviors of κ or parasitic effects?
What do they really show about inter-coder agreement?

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
    - NB: might even include segmentation problems
  - good for “uncertain” annotations: no obvious ground truth, coding schemes under development, taxonomies in progress, etc.
    - ➔ higher-level abstractive annotations

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

http://kappa.chez-alice.fr
- very clear explanations… but in French

http://en.wikibooks.org/wiki/Algorithm_implementation/Statistics/Fleiss%27_kappa
- java implementation of Fleiss’ kappa

http://www.dmi.columbia.edu/homepages/chuangj/kappa/
- other ‘Kappa calculators’ are available via Google search

References (1)

Scott, William A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, 19:127-141.

Cohen, Jacob (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1): 37-46.

Fleiss, Joseph L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378-381.

Fleiss, Joseph L. and Cohen, Jacob (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33: 613-619.

Fleiss, Joseph L. (1981). *Statistical Methods for Rates and Proportions*. 2nd ed. New York: John Wiley. (See p.38-46).

Siegel, Sandy and Castellan, N. John Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill International Editions, 2nd ed.

Landis, J.R. and Koch G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159-174.

Everitt, Brian (1994). *Statistical Methods in Medical Investigations*, Wiley. [0 EVE]

References (2)

Krippendorff, Klaus (1980, 2nd ed. 2004). *Content Analysis: An Introduction to Its Methodology*. Beverly Hills, CA: Sage Publications.

Krippendorff, Klaus and Hayes, Andrew F. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and* Measures, 1(1):77-89.

Carletta, Jean (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249-254.

Di Eugenio, Barbara (2000). On the usage of kappa to evaluate agreement on coding tasks. *Proc of LREC 2000*, Athens, 441-444.

Di Eugenio, Barbara, and Glass, Michael (2004). The kappa statistic: a second look. *Computational Linguistics*, 30(1):95-101.

Craggs, Richard and McGee Wood, Mary (2005). Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 31(3):289-295.

Dietterich, Thomas G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms, *Neural Comp.*, 10:1895-1923.