Cohen's kappa [1] has originally been introduced in applied statistics as a measure of so-called inter-rater agreement. Given two raters who classify $N$ items into $C$ categories, the measure is defined as

$$\kappa = \frac{p - p_e}{1 - p_e},$$

where $p$ is the probability that both raters assign an item to the same category, and $p_e$ is the probability of an agreement by chance, i.e., under the assumption of independent assignments (both raters assign items according to the respective marginal distributions on the set of categories).

In recent years, the kappa measure has also been advocated as a useful performance measure in machine learning [2, 3]; for classification problems, it is now even routinely computed by software tools such as the WEKA framework. Yet, despite some appealing properties, especially for class-imbalanced problems, the kappa measure might be criticized for a somewhat non-intuitive behavior in specific situations. More generally, its reasonableness in the context of machine learning might be questioned on the grounds of an important difference between a true inter-rater situation and the setting of supervised machine learning, in which a learner (the first rater) deliberately attempts to mimic a data-generating process (the second rater).

The goal of this thesis is to elaborate on the usefulness of the kappa measure in machine learning. Based on a thorough review of the literature and the problems for which it has been used, the properties of the measure ought to be analyzed in a critical manner. Moreover, in order to gain a deeper understanding of the measure, suitable experimental studies should be conducted.

**Requirements:** Survey of literature on the kappa measure; analysis of formal properties of kappa; design and realization of experimental studies, in which kappa is used as a performance measure.

**Prerequisites:** Basic knowledge in machine learning; programming skills.

For more information, please contact Prof. Eyke Hüllermeier (eyke@upb.de).

**References**

