Which Verification Metrics Are Appropriate for Rare-Event Classification Problems?

Azim Ahmadzadeh, Dustin Kempton, Petrus C Martens and *Rafal Angryk*

Strong solar flares are indeed rare events, which make the flare classification task a rare-event problem. Solar energetic particle events are even rarer space weather events as only a few instances of them are recorded each year. With the unprecedented growth in employment of Machine Learning algorithms for rare-event classification/forecast problems, a proper evaluation of rare-event models becomes a necessary skill for domain experts to have. This task remains to be an outstanding challenge as both the learning process and the metrics used for quantitative verification of models can easily obscure or skew the true performance of models and yield misleading and biased results.

To help mitigate this effect we introduce a bounded semimetric space that provides a generic representation for any deterministic performance verification metric. This space, named Contingency Space, can be easily visualized and shed light on models' performance as well as on the metrics' distinct behavior. An arbitrary model's performance can be mapped to a unique point in this space, which allows comparison of multiple models at the same time, for a given metric. Using this geometrical setting we show the difference between a metric's interpretation of performance and the true performance of the model. Using this perspective, models which are seemingly different but practically identical, or only marginally different, can be easily spotted. By tracking down a learner's performance at each epoch, we can also compare different learners' learning paths, which provides a deeper understanding of the utilized algorithms and their challenge in the learning process. Moreover, in the Contingency Space, a given verification metric can be represented by a geometrical surface, which allows a visual comparison between different metrics---a task that without this concept could be done only by the tedious algebraic comparison of metrics' formulae. Moreover, using such a surface, for the first time we can see and quantify the impact of scarcity of data (intrinsic to rare-even problems) on different metrics. This extra knowledge provides us with the information we need to choose an appropriate metric for evaluation of our rare-event models.