

Class Imbalance

A data set said to be highly skewed if sample from one class is in higher number than other. In an ***imbalanced data set*** the class having more number of instances is called as ***major class*** while the one having relatively less number of instances are called as ***minor class***.

Applications such as medical diagnosis prediction of rare but important disease is very important than regular treatment. Similar situations are observed in other areas, such as detecting fraud in banking operations, detecting network intrusions, managing risk and predicting failures of technical equipment.

In such situation most of the binary classifier are **biased** towards the major classes and hence show very poor classification rates on minor classes. It is also possible that classifier predicts everything as major class and ignores the minor class completely.

The Accuracy measure is an example of an metric that is affected by this bias. As the F-measure is not computed using the ***True Negatives***, it is less Biased.