Properties of the sample estimators used for statistical normalization of feature vectors.

Summary: Normalization of feature vectors is often used as a step of data preprocessing for clustering. A unified statistical approach to feature vector normalization has been proposed recently by the authors. After the proposed normalization, the contributions of both numerical and categorical attributes to a specified objective function are statistically the same. In spite of the importance for estimators to be consistent, the consistency of the sample estimators used for normalization, has never been considered. A mathematical justification of the statistical normalization procedure is given here. The sample estimators proposed for normalization of attributes of feature vectors are proven to have desirable properties, namely they are consistent and unbiased. Some other mathematical questions related to clustering have got here a rigorous treatment. In particular, the statistical normalization procedure is discussed in detail in the cases of the objective functions being based on the Chebyshev, attribute mismatch categorical and Minkowski mixed p-metrics. As an application of the normalization procedure, clustering of several benchmark datasets is performed with non-normalized and introduced normalized mixed metrics using either the k-prototypes (for \( p = 2 \)) or another algorithm (for \( p \neq 2 \)).

MSC:

62H30 Classification and discrimination; cluster analysis (statistical aspects)

62G05 Nonparametric estimation

Keywords:
estimators; normalization; Minkowski metrics; mixed databases

Software:

UCI-ml; ElemStatLearn

Full Text: DOI

References:


Edited by FIZ Karlsruhe, the European Mathematical Society and the Heidelberg Academy of Sciences and Humanities © 2021 FIZ Karlsruhe GmbH


This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.