Abstract

Attribute subset selection refers to the method of choosing the set of attributes that best describes the dataset. The attributes obtained from the attribute subset selection method when applied to machine learning operations such as clustering, classification etc., should provide the same result as that of the original dataset. The method employed for attribute subset selection must be efficient in terms of selecting the relevant attributes and must also be accurate in terms of eliminating the redundant attributes.

With the aim of satisfying the above two goals we have designed a feature subset selection method using the precise relevance measures. We first efficiently select the relevant attributes using the relevance measure “symmetric uncertainty (SU)”. The selected relevant attributes are then divided into clusters based on “graph-theoretic” clustering method using the relevance measure “conditional mutual information (CMI)”. Then the relevance measure “symmetric uncertainty” is used to select the attributes that are strongly related to the target class and also which best represents each cluster, thus giving us an accurate and independent subset of
features. The above developed method not only produces smaller more accurate subset of features but also improves the performance of the machine learning operations such as naive base classifier.

References

13. Krier C., Francois D., Rossi F. and Verleysen M., Feature clustering and mutual information for the selection of variables in spectral data, In Proc European Symposium on Artificial Neural Networks Advances in Computational Intelligence and Learning, pp 157-162 (2007).
17. Pereira F., Tishby N. and Lee L., Distributional clustering of English words, In Proceedings of the 31st Annual Meeting on Association For Computational Linguistics,
An Ameliorated Methodology for Feature Subset Selection on High Dimensional Data using Precise Relevance Measures

183-190 (1993).


**Index Terms**

Computer Science

Artificial Intelligence

**Keywords**
Relevant feature, Redundant features, relevance measures, symmetric uncertainty (SU), conditional mutual information (CMI).