A New Approach for Automated Feature Selection

Presentation of work originally published in the Proc. of the 2018 IEEE International Conference on Big Data [GLS18]

Andreas Gocht[1] Christoph Lehmann[1] Robert Schöne[1]

Keywords: Data mining; Feature selection; Mutual information

While more and more data is collected for machine learning, validation and exploration of these data become increasingly challenging. Moreover, if the collection of feature data is expensive, or the amount of usable features is limited, *feature selection* is often employed.

In [GLS18] we present an algorithm, which is able to select the most relevant features and stops automatically once the information added does not improve the quality of the result anymore. It is possible to specify an upper limit of features, and our algorithm is independent of any following machine learning task. The selection algorithm is based on the so-called Historical JMI (HJMI) score, as it uses the information from already selected features:

$$J_{HJMI}(X_k, S) = J_H + I(X_k; Y) - \frac{\sum_{X_j \in S}[I(X_k; X_j) - I(X_k; X_j|Y)]}{|S|}.$$

The set S contains already selected features. X_k refers to the currently investigated feature and X_j to an already selected feature out of S. X_k and X_j are features out of $X = \{X_1, X_2, ..., X_l\}$, where l denotes the number of all features. Y defines the target variable, which we like to predict. J_H is the historical information about the selected features, $I(X_k; Y)$ and $I(X_k; X_j)$ refers to the mutual information. $I(X_k; X_j|Y)$ specifies the conditional mutual information.

The algorithm to calculate the HJMI is given by:

1. Set $J_H = 0$
2. Calculate $J_{HJMI}(X_k, S)$ for all $X_k \in X \backslash S$
3. Save the largest result for $J_{HJMI}(X_k, S)$ as J_H and add the associated X_k to S
4. Repeat 2 and 3 until the stopping criterion is met or the maximum amount of features is reached

[1] Center for Information Services and High Performance Computing (ZIH) Technische Universität Dresden, 01062 Dresden, Germany, {andreas.gocht|christoph.lehmann|robert.schoene}@tu-dresden.de

doi:10.18420/inf2019_42

As for stopping criterion, we propose $\delta > \frac{J_{HMI}-J_H}{J_H}$, i.e., if the information added by a new feature X_k does not increase the information of the already selected features in S by more than a given δ, the algorithm will stop. However, as in a typical exploratory setting, the choice of the stopping criterion is to be reflected case-dependent as well as its interpretation. Compared to PCA, which only considers the pairwise dependence of features, the JMI is more general as it approximates the mutual information of the conditional distribution $Y|(S)$.

To evaluate our approach, we used the NIPS Feature Selection Challenge [Gu04], similar to [Br12]. The results are shown in Table 1. A detailed analysis can be found in [GLS18].

Benchmark	JMI Validation Error [%]	JMI Amount of Features (l)	HJMI Validation Error [%]	HJMI Amount of Features (l)
ARCENE	21.19	20	19.64	32
DEXTER	15.0	60	13.0	21
DOROTHEA	32.99	200	25.63	24
GISETTE	4.1	200	8.0	26
MADELON	10.67	20	10.67	20

Tab. 1: Results for NIPS Feature Selection Challenge. The first column shows the smallest error for the validation set, with features selected using JMI. The second column shows the amount of features used to achieve this result. The third and fourth column present the same information for the newly introduced HJMI-based algorithm. In most cases, HJMI is as good or better than the JMI. For GISETTE, JMI outperforms HJMI. However, it depends on the application if a reduction of only 3.9% in prediction error is worth selecting 174 additional features.

Acknowledgments This work is supported by the European Union's Horizon 2020 program in the READEX project (grant agreement number 671657) and the German Federal Ministry of Education and Research (BMBF, 01IS14014A-D) by funding the competence center for Big Data "ScaDS Dresden/Leipzig".

References

[Br12] Brown, G.; Pocock, A.; Zhao, M.-J.; Luján, M.: Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. The Journal of Machine Learning Research/, ACMID: 2188387, 2012.

[GLS18] Gocht, A.; Lehmann, C.; Schöne, R.: A New Approach for Automated Feature Selection. In: 2018 IEEE International Conference on Big Data (Big Data). DOI: 10.1109/BigData.2018.8622548, 2018.

[Gu04] Guyon, I.; Gunn, S. R.; Ben-Hur, A.; Dror, G.: Result Analysis of the NIPS 2003 Feature Selection Challenge. In: Advances in Neural Information Processing Systems. ACMID: 2976109, 2004.