Feature Selection

Martin Sewell

2007

1 Definition

Feature selection (also known as *subset selection*) is a process commonly used in machine learning, wherein a subset of the features available from the data are selected for application of a learning algorithm. The best subset contains the least number of dimensions that most contribute to accuracy; we discard the remaining, unimportant dimensions. This is an important stage of preprocessing and is one of two ways of avoiding the curse of dimensionality (the other is feature extraction).

There are two approaches:

forward selection Start with no variables and add them one by one, at each step adding the one that decreases the error the most, until any further addition does not significantly decrease the error.

backward selection Start with all the variables and remove them one by one, at each step removing the one that decreases the error the most (or increases it only slightly), until any further removal increases the error significantly.

To reduce overfitting, the error referred to above is the error on a validation set that is distinct from the training set.

2 Chronological Literature Review

Kira and Rendell (1992) described a statistical feature selection algorithm called RELIEF that uses instance based learning to assign a relevance weight to each feature.

John, Kohavi and Pfleger (1994) addressed the problem of irrelevant features and the subset selection problem. They presented definitions for irrelevance and for two degrees of relevance (weak and strong). They also state that features selected should depend not only on the features and the target concept, but also on the induction algorithm. Further, they claim that the filter model approach to subset selection should be replaced with the wrapper model.

Pudil, Novovičová and Kittler (1994) presented "floating" search methods in feature selection. These are sequential search methods characterized by a dynamically changing number of features included or eliminated at each step. They were shown to give very good results and to be computationally more effective than the branch and bound method.

Koller and Sahami (1996) examined a method for feature subset selection based on Information Theory: they presented a theoretically justified model for optimal feature selection based on using cross-entropy to minimize the amount of predictive information lost during feature elimination.

Jain and Zongker (1997) considered various feature subset selection algorithms and found that the sequential forward floating selection algorithm, proposed by Pudil, Novovičová and Kittler (1994), dominated the other algorithms tested.

Dash and Liu (1997) gave a survey of feature selection methods for classification.

In a comparative study of feature selection methods in statistical learning of text categorization (with a focus is on aggressive dimensionality reduction), Yang and Pedersen (1997) evaluated document frequency (DF), information gain (IG), mutual information (MI), a χ^2-test (CHI) and term strength (TS); and found IG and CHI to be the most effective.

Blum and Langley (1997) focussed on two key issues: the problem of selecting relevant features and the problem of selecting relevant examples.

Kohavi and John (1997) introduced wrappers for feature subset selection. Their approach searches for an optimal feature subset tailored to a particular learning algorithm and a particular training set.

Yang and Honavar (1998) used a genetic algorithm for feature subset selection.

Liu and Motoda (1998) wrote their book on feature selection which offers an overview of the methods developed since the 1970s and provides a general framework in order to examine these methods and categorize them.

Weston, *et al.* (2001) introduced a method of feature selection for SVMs which is based upon finding those features which minimize bounds on the leave-one-out error. The method was shown to be superior to some standard feature selection algorithms on the data sets tested.

Xing, Jordan and Karp (2001) successfully applied feature selection methods (using a hybrid of filter and wrapper approaches) to a classification problem in molecular biology involving only 72 data points in a 7130 dimensional space. They also investigated regularization methods as an alternative to feature selection, and showed that feature selection methods were preferable in the problem they tackled.

See Miller (2002) for a book on subset selection in regression.

Forman (2003) presented an empirical comparison of twelve feature selection methods. Results revealed the surprising performance of a new feature selection metric, 'Bi-Normal Separation' (BNS).

Guyon and Elisseeff (2003) gave an introduction to variable and feature selection. They recommend using a linear predictor of your choice (e.g. a

linear SVM) and select variables in two alternate ways: (1) with a variable ranking method using a correlation coefficient or mutual information; (2) with a nested subset selection method performing forward or backward selection or with multiplicative updates.

For a summary of feature selection methods see Figure 1, and for a taxonomy of algorithms see Figure 2.

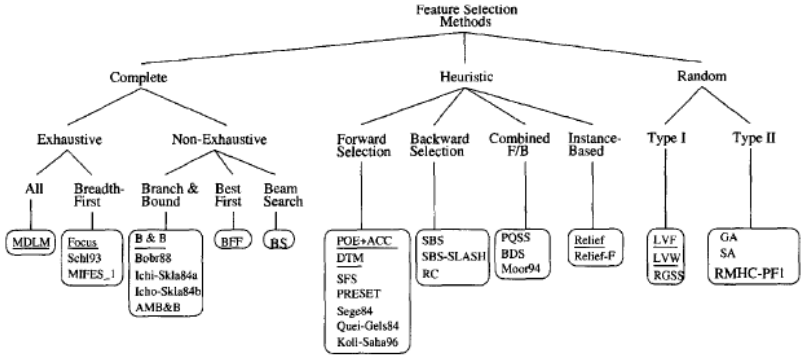

Figure 1: Summary of feature selection methods. Dash and Liu (1997)

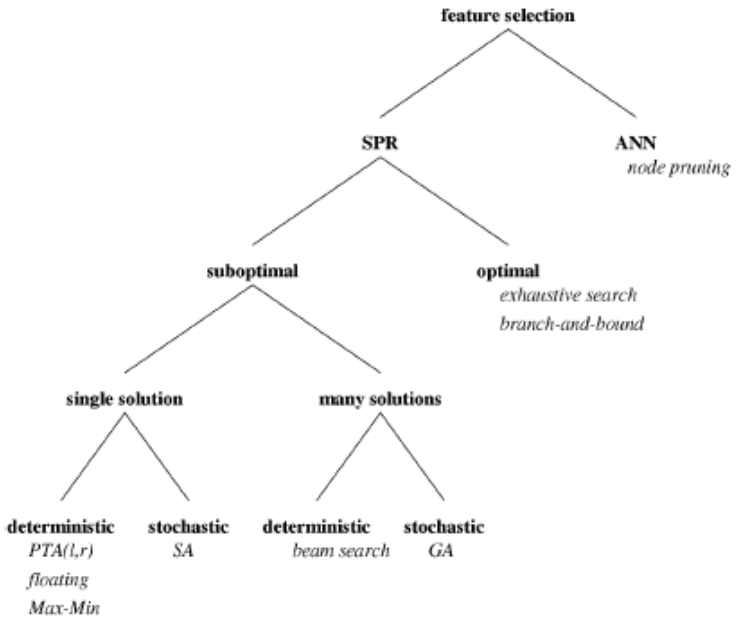

Figure 2: A taxonomy of feature selection algorithms. Jain and Zongker (1997)

References

BLUM, Avrim L., and Pat LANGLEY, 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, **97**(1–2), 245–271.

DASH, M., and H. LIU, 1997. Feature selection for classification. *Intelligent Data Analysis*, **1**(1–4), 131–156.

FORMAN, George, 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, **3**, 1289–1305.

GUYON, Isabelle, and André ELISSEEFF, 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.

JAIN, Anil, and Douglas ZONGKER, 1997. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(2), 153–158.

JOHN, George H., Ron KOHAVI, and Karl PFLEGER, 1994. Irrelevant features and the subset selection problem. *In*: William W. COHEN and Haym HIRSH, eds. *Machine Learning: Proceedings of the Eleventh International Conference*. San Francisco, CA: Morgan Kaufmann Publishers, pp. 121–129.

KIRA, Kenji, and Larry A. RENDELL, 1992. A practical approach to feature selection. *In*: Derek H. SLEEMAN and Peter EDWARDS, eds. *ML92: Proceedings of the Ninth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 249–256.

KOHAVI, Ron, and George H. JOHN, 1997. Wrappers for feature subset selection. *Artificial Intelligence*, **97**(1–2), 273–324.

KOLLER, Daphne, and Mehran SAHAMI, 1996. Toward optimal feature selection. *In*: *Proceedings of the Thirteenth International Conference on Machine Learning*. Morgan Kaufmann, pp. 284–292.

LIU, Huan, and Hiroshi MOTODA, 1998. *Feature Selection for Knowledge Discovery and Data Mining*. The Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers.

MILLER, Alan, 2002. *Subset Selection in Regression*. Second ed. Chapman & Hall/CRC.

PUDIL, P., J. NOVOVIČOVÁ, and J. KITTLER, 1994. Floating search methods in feature selection. *Pattern Recognition Letters*, **15**(11), 1119–1125.

WESTON, Jason, *et al.*, 2001. Feature selection for SVMs. *In*: Todd K. LEEN, Thomas G. DIETTERICH, and Volker TRESP, eds. *Advances in Neural Information Processing Systems 13*. Cambride, MA: The MIT Press, pp. 668–674.

XING, Eric P., Michael I. JORDAN, and Richard M. KARP, 2001. Feature selection for high-dimensional genomic microarray data. *In*: *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann, pp. 601–608.

YANG, Jihoon, and Vasant HONAVAR, 1998. Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems*, **13**(2), 44–49.

YANG, Yiming, and Jan O. PEDERSEN, 1997. A comparative study of feature selection in text categorization. *In*: *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 412–420.