Unsupervised feature selection using feature similarity

Article *in* IEEE Transactions on Pattern Analysis and Machine Intelligence · April 2002

DOI: 10.1109/34.990133 · Source: IEEE Xplore

3 authors, including:

Pabitra Mitra
Indian Institute of Technology Kharagpur
147 PUBLICATIONS **3,073** CITATIONS

SEE PROFILE

Unsupervised feature selection using feature similarity

Mitra P., Murthy C.A., Pal S.K.

IEEE; Machine Intelligence Unit, Indian Statistical Institute, Calcutta 700 035, India

Abstract: In this article, we describe an unsupervised feature selection algorithm suitable for data sets, large in both dimension and size. The method is based on measuring similarity between features whereby redundancy therein is removed. This does not need any search and, therefore, is fast. A new feature similarity measure, called maximum information compression index, is introduced. The algorithm is generic in nature and has the capability of multiscale representation of data sets. The superiority of the algorithm, in terms of speed and performance, is established extensively over various real-life data sets of different sizes and dimensions. It is also demonstrated how redundancy and information loss in feature selection can be quantified with an entropy measure.

Index Keywords: Computational complexity; Correlation methods; Data compression; Data mining; Fuzzy sets; Genetic algorithms; Least squares approximations; Markov processes; Principal component analysis; Regression analysis; Dimensionality reduction; Feature clustering; Feature similarity; Maximum information compression index; Unsupervised feature selection; Feature extraction

Year: 2002

Source title: IEEE Transactions on Pattern Analysis and Machine Intelligence

Volume: 24

Issue: 3

Page : 301-312

Cited by: 228

Link: Scorpus Link

Correspondence Address: Mitra, P.; Machine Intelligence Unit, Indian Statistical Institute, Calcutta 700 035, India; email: pabitra_rsankar@isical.ac.in

Document Type: Article

Source: Scopus

Authors with affiliations:

1. Mitra, P., IEEE, Machine Intelligence Unit, Indian Statistical Institute, Calcutta 700 035, India
2. Murthy, C.A., Machine Intelligence Unit, Indian Statistical Institute, Calcutta 700 035, India
3. Pal, S.K., IEEE, Machine Intelligence Unit, Indian Statistical Institute, Calcutta 700 035, IndiaMachine Intelligence Unit, Indian Statistical Institute, Calcutta 700 035, India,

References:

1. Fayyad, U., Uthurusamy, R., Data mining and knowledge discovery in databases (1996) Comm. ACM, 39 (11), pp. 24-27. , Nov
2. Devijver, P.A., Kittler, J., (1982) Pattern Recognition: A Statistical Approach, , Englewood Cliffs: Prentice Hall
3. Pudil, P., Novovicová, J., Kittler, J., Floating search methods in feature selection (1994) Pattern Recognition Letters, 15, pp.

1119-1125

4. Aha, Dw., Bankert, R.L., A comparative evaluation of sequential feature selection algorithms (1996) Artificial Intelligence and Statistics V, , D. Fisher and J.-H. Lenz, eds., New York: Springer Verlag
5. (1996) Genetic Algorithms for Pattern Recognition, , S.K. Pal and, P.P. Wang, eds. Boca Raton: CRC Press
6. Kudo, M., Sklansky, J., Comparison of algorithms that selects features for pattern classifiers (2000) Pattern Recognition, 33, pp. 25-41
7. Skalak, D., Prototype and feature selection by sampling and random mutation hill climbing algorithms (1994) Proc. 11th Int'l. Machine Learning Conf., pp. 293-301
8. Moore, A.W., Lee, M.S., Efficient algorithms for minimizing cross validation error (1994) Proc. 11th Int'l. Conf. Machine Learning
9. Liu, H., Setiono, R., Some issues in scalable feature selection (1998) Expert Systems with Applications, 15, pp. 333-339
10. Dash, M., Liu, H., Unsupervised feature selection (2000) Proc. Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 110-121
11. Dy, J., Brodley, C., Feature subset selection and order identification for unsupervised learning (2000) Proc. 17th Int'l. Conf. Machine Learning
12. Basu, S., Micchelli, C.A., Olsen, P., Maximum entropy and maximum likelihood criteria for feature selection from multivariate data (2000) Proc. IEEE Int'l. Symp. Circuits and Systems, pp. III-267-III270
13. Pal, S.K., De, R.K., Basak, J., Unsupervised feature evaluation: A neuro-fuzzy approach (2000) IEEE Trans. Neural Network, 11, pp. 366-376
14. Hall, M.A., Correlation based feature selection for discrete and numeric class machine learning (2000) Proc. 17th Int'l. Conf. Machine Learning
15. Heydorn, R.P., Redundancy in feature extraction (1971) IEEE Trans. Computers, pp. 1051-1054
16. Das, S.K., Feature selection with a linear dependence measure (1971) IEEE Trans. Computers, pp. 1106-1109
17. Toussaint, G.T., Vilmansen, T.R., Comments on feature selection with a linear dependence measure (1972) IEEE Trans. Computers, p. 408
18. Kira, K., Rendell, L., A practical approach to feature selection (1992) Proc. Ninth Int'l. Workshop Machine Learning, pp. 249-256
19. Kononenko, I., Estimating attributes: Analysis and extension of relief (1994) Proc. Seventh European Machine Learning Conf., pp. 171-182
20. Koller, D., Sahami, M., Towards optimal feature selection (1996) Proc. 13th Int'l. Conf. Machine Learning, pp. 284-292
21. King, B., Step-wise clustering procedures (1967) J. Am. Statistical Assoc., pp. 86-101
22. Rao, C.R., (1973) Linear Statistical Inference and Its Applications, , John Wiley
23. Blake, C.L., Merz, C.J., (1998) UCI Repository of Machine Learning Databases, , http://www.ics.uci.edu/mlearn/MLRepository.html, Univ. of California, Irvine, Dept. of Information and Computer Sciences
24. Lehmann, E.L., (1976) Testing of Statistical Hypotheses, , New York: John Wiley
25. Aspin, A., Tables for use in comparisons whose accuracy involves two variances (1949) Biometrika, 36, pp. 245-271