Scopus

Documents

Khalid, N.H.M., Ismail, A.R., Aziz, N.A., Hussin, A.A.A.

Performance Comparison of Feature Selection Methods for Prediction in Medical Data
(2023) *Communications in Computer and Information Science*, 1771 CCIS, pp. 92-106.

DOI: 10.1007/978-981-99-0405-1_7

Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, PO. Box 10, Kuala Lumpur, 50728, Malaysia

Abstract
Along with technological advancement, the application of machine learning algorithms in industry, notably in the medical field, has grown and progressed quickly. Medical databases commonly contain a lot of information about the medical histories of the patients and patient’s conditions, in addition, it is challenging to identify and extract the information that will be relevant and meaningful for machine learning modelling. Not to mention, the efficacy of the predictive machine learning algorithm can be enhanced by using only useful and pertinent information. Hence, feature selection is proposed to determine the significant features. Thus, feature selection should be fully utilized and applied when building machine learning algorithm. This study analyzes filter, wrapper, and embedded feature selection methods for medical data with the predictive machine learning algorithm, Random Forest and CatBoost. The experiment is carried out by evaluating the performances of the machine learning with and without applying feature selection methods. According to the results, CatBoost with RFE shows the best performance, in comparison to Random Forest with other feature selection methods.

Author Keywords
CatBoost; Feature selection; Lasso; RFE

Index Keywords
Learning algorithms, Medical computing; Catboost, Feature selection methods, Features selection, Lasso, Machine learning algorithms, Medical data, Performance, Performance comparison, Random forests, RFE; Feature Selection

References

- Roy, S.D., Das, S., Kar, D., Schwenker, F., Sarkar, R.
  Computer aided breast cancer detection using ensembling of texture and statistical image features
  (2021) *Sensors*, 21 (11), pp. 1-17.

- Mei, J., Desrosiers, C., Frasnelli, J.
  Machine learning for the diagnosis of Parkinson’s disease: A review of literature. Front
  (2021) *Aging Neurosci*, 13 (May), pp. 1-41.

- Cerri, S., Mus, L., Blandini, F.
  Parkinson’s disease in women and men: What’s the difference?
  (2019) *J. Parkinsons Dis.*, 9 (3), pp. 501-515.

- Knapič, S., Malhi, A., Saluja, R., Främling, K.
  Explainable artificial intelligence for human decision support system in the medical domain
  (2021) *Mach. Learn. Knowl. Extr.*, 3 (3), pp. 740-770.

- Chourib, I., Guillard, G., Farah, I.R., Solaiman, B.
  Stroke treatment prediction using features selection methods and machine learning classifiers
  (2022) *IRBM*, 1 (1-9).

- Zhang, F., Fleyeh, H., Bales, C.
  A hybrid model based on bidirectional long short-term memory neural network and

Catboost for short-term electricity spot price forecasting
(2020) *J. Oper. Res. Soc.*, pp. 1-25.

- Pathan, M.S., Nag, A., Pathan, M.M., Dev, S.
Analyzing the impact of feature selection on the accuracy of heart disease prediction. Healthc. Anal. 2(February)
(2022) *100060*,

- Dissanayake, K., Johar, M.G.M.
Comparative study on heart disease prediction using feature selection techniques on classification algorithms
(2021) *Appl. Comput. Intell. Soft Comput.*, 2021.

- Senan, E.M., Abunadi, I., Jadhav, M.E., Fati, S.M.
Score and correlation coefficient-based feature selection for predicting heart failure diagnosis by using machine learning algorithms
(2021) *Comput. Math. Methods Med.*, 2021.

- Krisnabayu, R.Y., Ridok, A., Budi, A.S.
Hepatitis detection using random forest based on SVM-RFE (recursive feature elimination) feature selection and SMOTE
(2021) *ACM International Conference on Proceeding Series*, pp. 151-156.

- Wolberg, W.H., Mangasarian, O.L.
Multisurface method of pattern separation for medical diagnosis applied to breast cytology
(1990) *Proc. Natl. Acad. Sci. U. S. A.*, 87 (23), pp. 9193-9196.

- Little, M.A., McSharry, P.E., Roberts, S.J., Costello, D.A.E., Moroz, I.M.
Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection
(2007) *Biomed. Eng. Online*, 6, pp. 1-19.

- Chen, C.W., Tsai, Y.H., Chang, F.R., Lin, W.C.
Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results
(2020) *Expert Syst*, 37 (5), pp. 1-10.

- Siemieniec, R., Mente, R.
600 V Power Device Technologies for Highly Efficient Power supplies,º in 2021 23Rd European Conference on Power Electronics and Applications,

- V, P., V, J.
Hybrid feature selection technique for prediction of cardiovascular diseases
(2021) *Mater. Today Proc.*,

- Verploegh, I.S.C., Lazar, N.A., Bartels, R.H.M.A., Volovici, V.
Evaluation of the use of P values in neurosurgical literature: From statistical significance to clinical irrelevance
(2022) *World Neurosurg*, 161, pp. 280-283.

- Muñoz Montoya, J.E., Carreño Rodríguez, J.N., Ardila Duarte, G., Maldonado Moran, M.Á., Luque Suarez, J.C.
Correlation of carbon dioxide and systolic velocity of the middle cerebral artery in patients with spontaneous subarachnoid hemorrhage of aneurysmal origin
(2022) *Interdiscip. Neurosurg. Adv. Tech. Case Manag.*, 27.

- Chen, J., Song, L., Wainwright, M.J., Jordan, M.I.
Learning to explain: An information-theoretic perspective on model interpretation
(2018) *35Th International Conference on Machine Learning, ICML 2018, Vol. 2, Pp. 1386–1418*,

- Tsalatsanis, A., Hozo, I., Djulbegovic, B.
(2020) *Meta-Analysis of Mutual Information Applied in EBM Diagnostics*, pp. 1-14.

- Zaidan, M.A.
Exploring non-linear associations between atmospheric new-particle formation and ambient variables: A mutual information approach
(2018) *Atmos. Chem. Phys.*, 18 (17), pp. 12699-12714.

- Benish, W.A.
A review of the application of information theory to clinical diagnostic testing
(2020) *Entropy*, 22 (1), p. 97.

- Arun Kumar, C., Sooraj, M.P., Ramakrishnan, S.
A comparative performance evaluation of supervised feature selection algorithms on microarray datasets
(2017) *Proc. Comput. Sci.*, 115, pp. 209-217.

- Nair, R., Bhagat, A.
Feature selection method to improve the accuracy of classification algorithm
(2019) *Int. J. Innov. Technol. Explor. Eng.*, 8 (6), pp. 124-127.

- Pires, A.C., Mendes, G.R., Santos, G.F.M., Dias, A.P.C., Santos, A.A.
Indirect identification of wheel rail contact forces of an instrumented heavy haul railway vehicle using machine learning
(2021) *Mech. Syst. Sig. Process.*, 160.

- Sharan, R.V., Moir, T.J.
Pseudo-color cochleagram image feature and sequential feature selection for robust acoustic event recognition
(2018) *Appl. Acoust*, 140 (May), pp. 198-204.

- Gu, N., Fan, M., Du, L., Ren, D.
Efficient sequential feature selection based on adaptive eigenspace model
(2015) *Neurocomputing*, 161, pp. 199-209.

- Mostafiz, R., Uddin, M.S., Alam, N.A., Mahfuz Reza, M., Rahman, M.M.
Covid-19 detection in chest X-ray through random forest classifier using a hybridization of deep CNN and DWT optimized features
(2021) *J. King Saud Univ.-Comput. Inf. Sci.*, 34 (6), pp. 3226-3235.

- Ahmad, G.N., Ullah, S., Algethami, A., Fatima, H., Akhter, S.M.H.
Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection
(2022) *IEEE Access*, 10, pp. 23808-23828.

- Aggrawal, R., Pal, S.
Sequential feature selection and machine learning algorithm-based patient’s death events prediction and diagnosis in heart disease
(2020) *SN Comput. Sci.*, 1 (6), pp. 1-16.

- Aziz, R., Verma, C.K., Srivastava, N.
Dimension reduction methods for microarray data: A review
(2017) *AIMS Bioeng*, 4 (2), pp. 179-197.

- Chen, Q., Meng, Z., Su, R.
WERFE: A gene selection algorithm based on recursive feature elimination and ensemble strategy. Front. Bioeng
(2020) *Biotechnol*, 8 (May), pp. 1-9.

- Siemieniec, R., Mente, R.
600 V Power Device Technologies for Highly Efficient Power supplies,º in 2021 23Rd

European Conference on Power Electronics and Applications,

- Jović, A., Brkić, K., Bogunović, N.
  A review of feature selection methods with applications
  (2015) *Proceedings of the 2015 38Th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015, Pp. 1200–1205*,

- Schonlau, M., Zou, R.Y.
  The random forest algorithm for statistical learning
  (2020) *Stata J*, 20 (1), pp. 3-29.

- la Cava, W., Bauer, C., Moore, J.H., Pendergrass, S.A.
  Interpretation of machine learning predictions for patient outcomes in electronic health records
  (2019) *Arxiv*, pp. 572-581.

- Menze, B.H.
  A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data
  (2009) *BMC Bioinform*, 10, pp. 1-16.

- Khalid, N.H.M., Ismail, A.R., Aziz, N.A.
  Interpretation of machine learning model using medical record visual analytics
  (2022) *Proceedings of the 8Th International Conference on Computational Science and Technology. LNEE*, 835, pp. 633-645.
  Alfred, R., Lim, Y. (eds.) , Springer, Singapore

- Gao, W., Zhou, L., Liu, S., Guan, Y., Gao, H., Hui, B.
  Machine learning prediction of lignin content in poplar with Raman spectroscopy. Bioresour. Technol. 348(February)
  (2022) *126812*,

- Kang, Y., Jang, E., Im, J., Kwon, C., Kim, S.
  Developing a new hourly forest fire risk index based on Catboost in South Korea
  (2020) *Appl. Sci.*, pp. 4-6.

- Ambe, K., Suzuki, M., Ashikaga, T., Tohkin, M.
  Development of quantitative model of a local lymph node assay for evaluating skin sensitization potency applying machine learning CatBoost
  (2021) *Regul. Toxicol. Pharmacol.*, 125.

- Khan, P.W., Byun, Y.C., Lee, S.J., Park, N.
  Machine learning based hybrid system for imputation and efficient energy demand forecasting
  (2020) *Energies*, 13 (11).

- Jani, D.
  (2022) *An Efficient Gait Abnormality Detection Method Based on Classification*, pp. 1-22.

- Tibshirani, R.
  Regression shrinkage and selection via the lasso
  (1996) *J. R. Stat. Soc. Ser. B*, 58 (1), pp. 267-288.

- Heiskanen, M.A.
  Different predictors of right and left ventricular metabolism in healthy middle-aged men
  (2015) *Front. Physiol*, 6 (DEC).

- Chintalapudi, N.
  LASSO regression modeling on prediction of medical terms among seafarers’

health documents using tidy text mining
(2022) *Bioengineering*, 9 (3), pp. 1-14.

Correspondence Address
Ismail A.R.; Department of Computer Science, PO. Box 10, Malaysia; email: amelia@iium.edu.my

Editors: Yusoff M., Kassim M., Mohamed A., Hai T., Kita E.
Publisher: Springer Science and Business Media Deutschland GmbH

Conference name: 7th International Conference on Soft Computing in Data Science, SCDS 2023
Conference date: 24 January 2023 through 25 January 2023
Conference code: 292079

ISSN: 18650929
ISBN: 9789819904044
Language of Original Document: English
Abbreviated Source Title: Commun. Comput. Info. Sci.
2-s2.0-85151166299
Document Type: Conference Paper
Publication Stage: Final
Source: Scopus