Scopus

Documents

Poolsawad, N., Moore, L., Kambhampati, C., Cleland, J.G.F.

Issues in the mining of heart failure datasets

(2014) *International Journal of Automation and Computing*, 11 (2), pp. 162-179.

[a] Intelligent Systems Research Group (IS, Department of Computer Science), University of Hull, Hull, United Kingdom
[b] Hull York Medical School, Department of Cardiology, University of Hull, Hull, United Kingdom

Abstract
This paper investigates the characteristics of a clinical dataset using a combination of feature selection and classification methods to handle missing values and understand the underlying statistical characteristics of a typical clinical dataset. Typically, when a large clinical dataset is presented, it consists of challenges such as missing values, high dimensionality, and unbalanced classes. These pose an inherent problem when implementing feature selection and classification algorithms. With most clinical datasets, an initial exploration of the dataset is carried out, and those attributes with more than a certain percentage of missing values are eliminated from the dataset. Later, with the help of missing value imputation, feature selection and classification algorithms, prognostic and diagnostic models are developed. This paper has two main conclusions: 1) Despite the nature of clinical datasets, and their large size, methods for missing value imputation do not affect the final performance. What is crucial is that the dataset is an accurate representation of the clinical problem and those methods of imputing missing values are not critical for developing classifiers and prognostic/diagnostic models. 2) Supervised learning has proven to be more suitable for mining clinical data than unsupervised methods. It is also shown that non-parametric classifiers such as decision trees give better results when compared to parametric classifiers such as radial basis function networks (RBFNs). © 2014 Institute of Automation, Chinese Academy of Sciences and Springer-Verlag Berlin Heidelberg.

Author Keywords
classification; clinical dataset; clustering; feature selection; Heart failure; missing values

References

- Tanwani, A.K., Afridi, J., Shafiq, M.Z., Farooq, M.
  Guidelines to select machine learning scheme for classification of biomedical datasets
  (2009) *Proceedings of the 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pp. 128-139.
  Berlin, Heidelberg, Germany: Springer-Verlag

- Jha, A.K., DesRoches, C.M., Campbell, E.G., Donelan, K., Rao, S.R., Ferris, T.G., Shields, A., Blumenthal, D.
  Use of electronic health records in U. S. hospitals
  (2009) *The New England Journal of Medicine*, 360 (16), pp. 1628-1638.

- Safran, C., Goldberg, H.
  Electronic patient records and the impact of the internet
  (2000) *International Journal of Medical Informatics*, 60 (2), pp. 77-83.

- Cleland, J.G.F., Swedberg, K., Follath, F., Komajda, M., Cohen-Solal, A., Aguilar, J.C., Dietz, R., Mason, J.
  The EuroHeart Failure survey programme - A survey on the quality of care among patients with heart failure in Europe, Part1: Patient characteristics and diagnosis
  (2003) *European Heart Journal*, 24 (5), pp. 442-463.

- Acharya, U.R., Bhat, P.S., Iyengar, S.S., Rao, A., Dua, S.
  Classification of heart rate data using artificial neural network and fuzzy equivalence relation
  (2003) *Pattern Recognition*, 36 (1), pp. 61-68.

- Shi, P., Ray, S., Zhu, Q.F., Kon, M.A.
  Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction
  (2011) *BMC Bioinformatics*, 12, p. 375.

- Mar, T., Zaunseder, S., Martinez, J.P., Llamedo, M., Poll, R.
  Optimization of ECG classification by means of feature selection
  (2011) *IEEE Transactions on Biomedical Engineering*, 58 (8), pp. 2168-2177.

- Sugiyama, M., Kawanabe, M., Chui, P.L.
  Dimensionality reduction for density ratio estimation in high-dimensional spaces
  (2010) *Neural Networks*, 23 (1), pp. 44-59.

- Wang, P.Y., Chow, T.W.S.
  A new feature selection scheme using data distribution factor for transactional data
  (2007) *Proceedings of the European Symposium on Artificial Neural Networks*, pp. 169-174.
  Bruges, Belgium: ESANN

- Dash, M., Liu, H., Yao, J.
  Dimensionality reduction of unsupervised data
  (1997) *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*, pp. 532-539.
  Newport Beach, CA, USA: IEEE

- Chiang, J.H., Ho, S.H.
  A combination of rough-based feature selection and RBF neural network for classification using gene expression data
  (2008) *IEEE Transactions on Nanotechnology*, 7 (1), pp. 91-99.

- Yan, Z.G., Wang, Z.Z., Xie, H.B.
  The application of mutual information-based feature selection and fuzzy LS-SVMbased classifier in motion classification
  (2008) *Computer Methods and Programs in Biomedicine*, 90 (3), pp. 275-284.

- Muni, D.P., Pal, B.R., Das, J.
  Genetic programming for simultaneous feature selection and classifier design
  (2006) *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36 (1), pp. 106-117.

- Yom-Tov, E., Inbar, G.F.
  Feature selection for the classification of movements from single movement-related potentials
  (2002) *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 10 (3), pp. 170-177.

- Varshavsky, R., Gottlieb, A., Horn, D., Linial, M.
  Unsupervised feature selection under perturbations: Meeting the challenges of biological data
  (2007) *Bioinformatics*, 23 (24), pp. 3343-3349.

- Kelder, J.C., Cramer, M.J., van Wijngaarden, J., van Tooren, R., Mosterd, A., Moons, K.G., Lammers, J.W., Hoes, A.W.
  The diagnostic value of physical examination and additional testing in primary care

patients with suspected heart failure
(2011) *Circulation*, 124 (25), pp. 1865-2873.

- Kelder, J.C., Cowie, M.R., McDonagh, T.A., Hardman, S.M., Grobbee, D.E., Cost, B., Hoes, A.W.
  Quantifying the added value of BNP in suspected heart failure in general practice: An individual patient data meta-analysis
  (2011) *Heart*, 97 (12), pp. 959-963.

- Peterson, P.N., Rumsfeld, J.S., Liang, L., Albert, N.M., Hernandez, A.F., Peterson, E.D., Fonarow, G.C., Masoudi, F.A.
  A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program
  (2010) *Circulation: Cardiovascular Quality and Outcomes*, 3 (1), pp. 25-32.

- Min, K.D., Asakura, M., Liao, Y.L., Nakamaru, K., Okazaki, H., Takahashi, T., Fujimoto, K., Kitakaze, M.
  Identification of genes related to heart failure using global gene expression profiling of human failingmyocardium
  (2010) *Biochemical Biophysical Research Communications*, 393 (1), pp. 55-60.

- Damarell, R.A., Tieman, J., Sladek, R.M., Davidson, P.M.
  Development of a heart failure filter for Medline: An objective approach using evidence-based clinical practice guidelines as an alternative to hand searching
  (2011) *BMC Medical Research Methodology*, 11, p. 12.

- Lee, D.S., Donovan, L., Austin, P.C., Gong, Y.Y., Liu, P.P., Rouleau, J.L., Tu, J.V.
  Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research
  (2005) *Medical Care*, 43 (2), pp. 182-188.

- Lee, D.S., Austin, P.C., Rouleau, J.L., Liu, P.P., Naimark, D., Tu, J.V.
  Predicting mortality among patients hospitalizeed for heart failure, derivation and validation of a clinical model
  (2003) *Journal of the American Medical Association*, 290 (19), pp. 2581-2587.

- Holme, I., Pedersen, T.R., Boman, K., Egstrup, K., Gerdts, E., Kesäniemi, Y.A., Malbecq, W., Gohlke-Bärwolf, C.
  A risk score for predicting mortality in patients with asymptomatic mild to moderate aortic stenosis
  (2011) *Heart*, 98 (5), pp. 377-383.

- Ho, K.K.L., Moody, G.B., Peng, C.K., Mietus, J.E., Larson, M.G., Levy, D., Goldberger, A.L.
  Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics
  (1997) *Circulation*, 96 (3), pp. 842-848.

- Fonarow, G.C., Abraham, W.T., Albert, N.M., Stough, W.G., Gheorghiade, M., Greenberg, B.H., O'Connor, C.M., Young, J.B.
  Association between performance measures and clinical outcomes for patients hospitalized with heart failure
  (2007) *Journal of the American Medical Association*, 297 (1), pp. 61-70.

- Bohacik, J., Davis, D.N.
  Data mining applied to cardiovascular data
  (2010) *Journal of Information Technologies*, 3 (2), pp. 14-21.

- Bohacik, J., Davis, D.N.
  Alert rules for remote monitoring of cardiovascular patients
  (2012) *Journal of Information Technologies*, 5 (1), pp. 16-23.

- Bohacik, J., Davis, D.N.
  Estimation of cardiovascular patient risk with a Bayesian network
  (2011) *Proceedings of the 9th European Conference of Young Research and Scientific Workers*, pp. 37-40.
  Žilina, Slovakia: University of Žilina

- Jain, A., Zongker, D.
  Feature selection: Evaluation, application, and small sample performance
  (1997) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (2), pp. 153-158.

- Saeys, Y., Abeel, T., van de Peer, Y.
  Robust feature selection using ensemble feature selection techniques
  (2008) *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 313-325.
  Berlin, Heidelberg, Germany: Springer-Verlag

- Yu, L., Liu, H.
  Feature selection for high-dimensional data: A fast correlation-based filter solution
  (2003) *Proceedings of the 20th International Conference on Machine Learning*, pp. 856-863.
  Washington DC, USA: AAAI

- Zhou, N., Wang, L.
  A modified T-test feature selection method and its application on the HapMap genotype data
  (2007) *Genomics, Proteomics & Bioinformatics*, 5 (3-4), pp. 242-249.

- Fayyad, U.M., Irani, K.
  Multi-interval discretization of continuous-valued attributes for classification learning
  (1993) *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 1022-1029.
  San Francisco, CA, USA: Morgan Kaufmann Publishers Inc

- Liu, H., Li, J., Wong, L.
  A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns
  (2002) *Genome Informatics*, 13, pp. 51-60.

- Hsu, C.N., Huang, H.J., Dietrich, S.
  The ANNIGMA-wrapper approach to fast feature selection for neural nets
  (2002) *IEEE Transactions Systems, Man, and Cybernetics, Part B*, 32 (2), pp. 207-212.

- Bohácik, J., Davis, D.N., Benediković, M.
  Risk estimation of cardiovascular patients using Weka
  (2012) *Proceedings of the International Conference OSSConf 2012*, pp. 15-20.

(The Society for Open Information Technologies - SOIT in Bratislava, Slovakia, Žilina, Slovakia)

- Acuña, E., Rodriguez, C.
  The treatment of missing values and its effect in the classifier accuracy
  (2004) *Classification, Clustering, and Data Mining Applications*, pp. 639-648.
  D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul (Eds.), Berlin, Heidelberg: Springer

- Lin, J.H., Haug, P.J.
  Data preparation framework for preprocessing clinical data in data mining
  (2006) *Proceedings of AMIA Annual Symposium*, pp. 489-493.
  American: AMIA

- Poolsawad, N., Kambhampati, C., Cleland, J.G.F.
  Feature selection approaches with missing values handling for data mining - A case study of heart failure dataset
  (2011) *World Academy of Science, Engineering and Technology*, 60, pp. 828-837.

- Poolsawad, N., Moore, L., Kambhampati, C., Cleland, J.G.F.
  Handling missing values in data mining - A case study of heart failure dataset
  (2012) *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 1934-2938.
  Chongqing, China: IEEE

- Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J.
  Knowledge discovery in databases: An overview
  (2011) *Artificial Intelligence Magazine*, 13 (3), pp. 57-70.

- (2011) *EM Imputation and Missing Data: Is Mean Imputation Really So Terrible?*,
  Analysis Factor

- Silva-Ramírez, E.L., Pino-Mejías, R., López-Coello, M., Cubiles-De-la-Vega, M.D.
  Missing value imputation on missing completely at random data using multilayer perceptrons
  (2011) *Neural Networks*, 24 (1), pp. 121-129.

- Han, J., Kamber, M.
  (2006) *Data Mining: Concepts and Techniques*,
  2nd ed.th edn., San Francisco: Morgan Kaufman Publishers

- Aha, D.W., Bankert, R.L.
  A comparative evaluation of sequential feature selection algorithms
  (1995) *Proceedings of the 5th International Workshop on Artificial Intelligence and Statistics*, pp. 1-7.

- Yu, L., Liu, H.
  Efficient feature selection via analysis of relevance and redundancy
  (2004) *Journal of Machine Learning Research*, 5, pp. 1205-1224.

- Jirapech-Umpai, T., Aitken, S.
  Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes
  (2005) *BMC Bioinformatics*, 6, p. 148.

- Coetzee, F.M.
  Correcting the Kullback-Leibler distance for feature selection
  (2005) *Pattern Recognition Letters*, 26 (11), pp. 1675-1683.

- Wu, B.L., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Zhao, H.Y.
  Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data
  (2003) *Bioinformatics*, 19 (13), pp. 1636-1643.

- Levner, I.
  Feature selection and nearest centroid classification for protein mass spectrometry
  (2005) *BMC Bioinformatics*, 6, p. 68.

- Jäeger, J., Sengupta, R., Ruzzo, W.L.
  Improved gene selection for classifcation of Microarrays
  (2003) *Pacific Symposium on Biocomputing*, 8, pp. 53-64.

- Su, Y., Murali, T.M., Pavlovic, V., Schaffer, M., Kasif, S.
  RankGene: Identification of diagnostic genes based on expression data
  (2003) *Bioinformatics*, 19 (12), pp. 1578-1579.

- (2011),
  The University of Waikato. WEKA: The Waikato Environment for Knowledge Acquisition. [Online], Available, 30 August

- Gardner, M.W., Dorling, S.R.
  Artificial neural networks (the multilayer perceptron) - A review of applications in the atmospheric sciences
  (1998) *Atmospheric Environment*, 32 (14-15), pp. 2627-2636.

- Autio, L., Juhola, M., Laurikkala, J.
  On the neural network classification of medical data and an endeavour to balance non-uniform data sets with artificial data extension
  (2007) *Computers in Biology and Medicine*, 37 (3), pp. 388-397.

- Khemphila, A., Boonjing, V.
  Parkinsons disease classification using neural network and feature selection
  (2012) *World Academy of Science, Engineering and Technology*, 64, pp. 15-18.

- Cortes, C., Vapnik, V.
  Support-vector networks
  (1995) *Machine Learning*, 20 (3), pp. 273-297.

- Platt, J.C.
  Fast training of support vector machines using sequential minimal optimization
  (1998) *Advances in Kernel Methods - Support Vector Learning*, pp. 185-208.
  B. Schoelkopf, C. Burges, and A. Smola (Eds.), Cambridge, MA, USA: MIT Press

- Hastie, T., Tibshirani, R.
  Classification by pairwise coupling
  (1998) *Advances in Neural Information Processing Systems*, pp. 507-513.
  Cambridge, MA, USA: MIT Press

- Breiman, L.
  Random forests
  (2001) *Machine Learning*, 45 (1), pp. 5-32.

- Kim, W.D., Lee, H.K., Lee, D.
  Fuzzy clustering of categorical data using fuzzy centroids
  (2004) *Pattern Recognition Letters*, 25 (11), pp. 1263-1271.

- Bean, C.L., Kambhampati, C.
  Knowledge-oriented clustering for decision support
  (2003) *Proceedings of the International Joint Conference on Neural Networks*, pp. 3244-3249.
  Portland, OR, USA: IEEE

- Steinbach, M., Karypis, G., Kumar, V.
  A comparison of document clustering techniques
  (2000) *Proceedings of KDD Workshop on Text Mining*, pp. 1-2.

- Huang, Z.X.
  Extensions to the k-means algorithm for clustering large data sets with categorical values
  (1998) *Data Mining and Knowledge Discovery*, 2 (3), pp. 283-304.

- Kanungo, T., Mount, M.D., Netanyahu, S.N., Piatko, D.C., Silverman, R., Wu, Y.A.
  An efficient k-means clustering algorithm: Analysis and implementation
  (2002) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 (7), pp. 881-892.

- Alsabti, K., Ranka, S., Singh, V.
  An efficient k-means clustering algorithm
  (1998) *Proceedings of IPPS/SPDP Workshop on High Performance Data Mining*, pp. 1-7.

- Mirkin, B.
  (2005) *Clustering for Data Mining: A Data Recovery Approach*,
  Florida: Chapman and Hull/CRC

- Sridhar, A., Sowndarya, S.
  Efficiency of k-means clustering algorithm in mining outliers from large data sets
  (2010) *International Journal on Computer Science and Engineering*, 2 (9), pp. 3043-3045.

- Napoleon, D., Lakshmi, G.P.
  An efficient k-means clustering algorithm for reducing time complexity using uniform distribution data points
  (2010) *Proceedings of the Trendz in Information Sciences & Computing*, pp. 42-45.
  Chennai, India: IEEE

- Zhao, Y., Karypis, G., Fayyad, U.
  Hierarchical clustering algorithms for document datasets
  (2005) *Data Mining and Knowledge Discovery*, 10 (2), pp. 141-168.

- Lee, J.S.J., Hwang, J.N., Davis, D.T., Nelson, A.C.
  Integration of neural networks and decision tree classifiers for automated cytology screening
  (1991) *Proceedings of the IJCNN-91-Seattle International Joint Conference on Neural Networks*, pp. 257-262.

Seattle, WA, USA: IEEE

- Zhang, Y., Kambhampati, C., Davis, D.N., Goode, K., Cleland, J.G.F.
A comparative study of missing value imputation with multiclass classification for clinical heart failure data
(2012) *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 2840-2844.
Sichuan, China: IEEE

- Al-Najiar, Y., Goode, K.M., Zhang, J., Cleland, J.G., Clark, A.L.
Andrew. Red cell distribution width: An inexpensive and powerful prognostic marker in heart failure
(2009) *European Journal Heart Failure*, 11 (12), pp. 1155-1162.

- (2011),
Atherotech Diagnotics Lab. Atherotech Panels. [Online], Available, 13 June

- (1998) *International Journal of the Computer, the Internet and Management*, 6 (2).

- Herrero, J., Valencia, A., Dopazo, J.
A hierarchical unsupervised growing neural network for clustering gene expression patterns
(2000) *Bioinformatics*, 17 (2), pp. 126-136.

- Myers, W.R.
Handling missing data in clinical trials: An overview
(2000) *Drug Information Journal*, 34 (2), pp. 525-533.

- Grinstead, C.M., Snell, J.L.
(1998) *Introduction to Probability*,
Rhode Island: American Mathematical Society

- Rahman, M.M., Davis, D.N.
Machine learning-based missing value imputation method for clinical datasets
(2013) *IAENG Transactions on Engineering Technologies*, pp. 245-257.
Netherlands: Springer

Correspondence Address
Poolsawad N.; Intelligent Systems Research Group (IS, Department of Computer Science), University of Hull, Hull, United Kingdom; email: N.Poolsawad@2008.hull.ac.uk

Publisher: Chinese Academy of Sciences

ISSN: 17518520
DOI: 10.1007/s11633-014-0778-5
Language of Original Document: English
Abbreviated Source Title: Int. J. Autom. Comput.
Document Type: Article
Source: Scopus