Roy, Asis; **Bhattacharya, Sourangshu**; **Guin, Kalyan**

Prediction of esophageal cancer using demographic, lifestyle, patient history, and basic clinical tests. (English) Zbl 1390.92072

Int. J. Adv. Eng. Sci. Appl. Math. 9, No. 4, 214-223 (2017).

Summary: An early detection of a disease can save many lives but it is impractical to undergo all medical tests for many prevalent diseases. Further these tests are costly, painful, time consuming and may have side effects. We have tried to predict esophageal cancer using demographics, lifestyle, medical history information, and basic clinical tests initially and later removed all clinical tests one after another to study the change of the accuracy without these clinical tests. It is well studied that no single classification technique turns out to be best for all the problems. Here, we test Naive Bayes classification, random forests, support vector machines (SVM) and logistic regression (LR), which perform similarly when all tests are used. However, as we reduce the number of tests, naive versions of these classifiers perform worse than kernelized versions of SVM and LR. We test our methodology with electronic medical record (EMR) of 3500 patients (approx.). The four methods described above, demonstrate a high accuracy with all features, including basic clinical test and show very low accuracy without the basic clinical tests measured by medical practitioner (MP). LR with a polynomial feature transformation of degree three yields an accuracy of 100% (approx), even without the MP features. Further dropping clinical tests one after another we see a decline in the accuracy of detection to 96%. We have also observed high sensitivity to 100% which indicates that no real patients are undetected in this experiment.

MSC:

92C50 Medical applications (general)
62P10 Applications of statistics to biology and medical sciences; meta analysis
62-07 Data analysis (statistics) (MSC2010)

Keywords:

esophageal cancer; electronic medical record; data mining; classification of a disease; demographic data

Software:

LIBLINEAR; LIBSVM; SMOTE; WEKA

Full Text: DOI

References:

[1] Edgren, G., Adami, H.O., Nyren, O., Weiderpass, E.: A global assessment of the oesophageal adenocarcinoma epidemic. Int. J. Gastroenterol. Hepatol. (2012). https://doi.org/10.1136/gutjnl-2012-302412

[2] Scott B., Health W.: Incidence of esophageal cancer linked to gerd, http://www.news-medical.net/news/20150421/Incidence-of-esophageal-cancer-linked-to-GERD-rises-six-fold-in-recent-decades.aspx. Online; Accessed 21 Dec 2015 · Zbl 0994.68128

[3] Cancer-Research-UK, Oesophageal cancer incidence statistics, http://www.cancerresearchuk.org/content/oesophageal-cancer-incidence-statistics#ref-2. Online; Accessed 21 Dec 2015

[4] Blot W., McLaughlin J.: The changing epidemiology of esophageal cancer. Semin. Oncol. 26 (5 Suppl 15). http://europepmc.org/abstract/MED/105

[5] Jensen, PB; Jensen, LJ; Brunak, S, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet., 13, 395-405, (2012)

[6] Alolfe M.A., Youssef A.B.M., Kadah Y.M., Mohamed A.S.: Development of a computer-aided classification system for cancer detection from digital mammograms, In: Radio Science Conference, NRSC 2008. National, 2008, pp. 1-8. https://doi.org/10.1109/NRSC.2008.45423 (2008) · Zbl 1283.62127

[7] Abreu P.H., Hugo Amaro D., C. Silva, Machado P., Abreu M.H., Afonso N., Dourado A.: Overall survival prediction for women breast cancer using ensemble methods and incomplete clinical data, pp. 1366-1369. https://doi.org/10.1007/978-3-319-00846-2_338 (2014) · Zbl 1007.68152

[8] Jacob S.G., Ramani R.G.: Efficient classifier for classification of prognostic breast cancer data through data mining techniques. In: Proceedings of the World Congress on Engineering and Computer Science, vol. 1, pp. 24-26 (2012)

[9] Ramani, RG; Jacob, SG, Improved classification of lung cancer tumors based on structural and physicochemical properties of proteins using data mining models, PLoS ONE, 8, e58772, (2013)

[10] Alizadehsani, R; Habibi, J; Hosseini, MJ; Mashayekhi, H; Boghrati, R; Ghandeharioun, A; Bahadorian, B; Sani, ZA, A data mining approach for diagnosis of coronary artery disease, Comput. Methods Progr. Biomed., 111, 52-61, (2013)

[11] Peter T.J., Somasundaram K.: An empirical study on prediction of heart disease using classification data mining techniques. In: 2012 International Conference on Advances in Engineering, Science and Management (ICAESM), pp. 514-518 (2012) · Zbl 0825.62593

[12] Nahar, J; Imam, T; Tickle, SK; Chen, PYP, Association rule mining to detect factors which contribute to heart disease in males and females, Expert Syst. Appl., 40, 1086-1093, (2013)

[13] Austin, P.C., Tu, J.V., Ho, J.E., Levy, D., Lee, D.S.: Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J. Clin. Epidemiol. (2013). https://doi.org/10.1016/j.jclinepi.2012.11.008

[14] Shouman M., Turner T., Stocker R.: Using data mining techniques in heart disease diagnosis and treatment. In: 2012 Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC), 2012, pp. 173-177. https://doi.org/10.1109/JEC-ECC.2012.6186978

[15] Wu, J; Roy, J; Stewart, WF, Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches, Med. Care, 48, s106-s113, (2010)

[16] Penny K.I., Smith G.D.: The use of data-mining to identify indicators of health related quality of life in patients with irritable bowel syndrome. In: Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces, 2009. ITI '09, pp. 87-92 (2009). https://doi.org/10.1109/ITI.2009.5196059

[17] Leke-Betechuoh B., Marwala T., Tim T., Lagazio M.: Prediction of HIV status from demographic data using neural networks. In: 2006 IEEE International Conference on Systems, Man and Cybernetics, vol. 3, pp. 2339-2344 (2006). https://doi.org/10.1109/ICSMC.2006.38521

[18] Altikardes, Z.A., Erdal, H., Baba, A.F., Tezcan, H., Fak, A.S., Korkmaz, H.: A study to classify non-dipper/dipper blood pressure pattern of type 2 diabetes mellitus patients without holter device, In: 2014 World Congress on Computer Applications and Information Systems (WCCAIS), pp. 1-5 (2014) https://doi.org/10.1109/WCCAIS.2014.6916555

[19] Raju, D; Su, X; Patrician, PA; Loan, LA; McCarthy, MS, Exploring factors associated with pressure ulcers: a data mining approach, Int. J. Nurs. Stud., 52, 102-111, (2015)

[20] Dai, W; Brisimi, TS; Adams, WG; Mela, T; Saligrama, V; Paschalidis, IC, Prediction of hospitalization due to heart diseases by supervised learning methods, Int. J. Med. Inf., 84, 189-197, (2015)

[21] Fawcett, T, An introduction to ROC analysis, Pattern Recogn. Lett., 27, 861-874, (2006)

[22] Swets, J.A.: Signal Detection Theory and Roc Analysis in Psychology and Diagnostics: Collected Papers. Lawrence Erlbaum Associates, Mahwah (1996) · Zbl 0913.92041

[23] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, vol. 2. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7 · Zbl 1273.62005

[24] Davis J., Goadrich M.: The relationship between precision-recall and roc curves, In: Proceedings of the 23rd International Conference on Machine Learning, ACM, pp. 233-240 (2006)

[25] Zhu J., Hastie T.: Kernel logistic regression and the import vector machine. In: Advances in neural information processing systems, pp. 1081-1088 (2001)

[26] Cessie, S; Houwelingen, J, Ridge estimators in logistic regression, Appl. Stat., 41, 191-201, (1992) · Zbl 0825.62593

[27] Cortes, C; Vapnik, V, Support-vector networks, Mach. Learn., 20, 273-297, (1995) · Zbl 0831.68098

[28] Scholkopf, B; Sung, K-K; Burges, CJC; Girosi, F; Niyogi, P; Poggio, T; Vapnik, V, Comparing support vector machines with Gaussian kernels to radial basis function classifiers, IEEE Trans. Signal Process., 45, 2758-2765, (1997)

[29] Zhang H.: The optimality of Naive Bayes. In: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach (2004)

[30] Breiman, L, Random forests, Mach. Learn., 45, 5-32, (2001) · Zbl 1007.68152

[31] Biau, G, Analysis of a random forests model, J. Mach. Learn. Res., 13, 1063-1095, (2012) · Zbl 1283.62127

[32] Hall, M; Frank, E; Holmes, G; Pfahringer, B; Reutemann, P; Witten, IH, The weka data mining software: an update, SIGKDD Explor., 11, 1871-1874, (2009)

[33] Fan, R-E; Chang, K-W; Hsieh, C-J; Wang, X-R; Lin, C-J, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res., 9, 1871-1874, (2008) · Zbl 1225.68175

[34] Schölkopf, B., Smola, A.J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, London (2002)

[35] Chang, C-C; Lin, C-J, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol., 2, 27:1-27:27, (2011)

[36] Chawla, NV; Bowyer, KW; Hall, LO; Kegelmeyer, WP, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res., 16, 321-357, (2002) · Zbl 0994.68128