Literature Review of Machine Learning Feature Selection Methods with emphasis on Medical Application on Breast Cancer

Thony Price Niklas Lindqvist

March 2018

Abstract

This document will provide a review of the past work connected to the topic of our Degree Project in Computer Science, Feature Selection Methods for Classification of Breast Cancer. It highlight different efforts of previous research that will inspire our project and be cited while making claims concerning the topic.

1 Introduction

The purpose of this litterature review is to build a foundation of knowledge for our upcomming project. The more we know about the topic beforehand the easier it will be to choose a relevant research question, refer to relevant material in out report as well as navigate the present research in the field. This document will also easily convert to a chapter about previous research.

The structure of the document:

- **Title:** Litterature title and reference.
- **Notes:** Key points from text.
- **Usage:** Potential use to us.

2 Litterature

Title: Machine learning for medical diagnosis: history, state of the art and perspective [6].
Notes: An overview of the development of intelligent data analysis in medicine from a machine learning perspective: a historical view, a state-of-the-art view, and a view on some future trends in this subfield of applied artificial intelligence.
Usage: The paper goes far back and covers the outlook for Machine Learning in the 60's to state of the art. This overview of how this dicipline evolved is good to present in our project too.

Title: Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples [8].
Notes: This report defines the term computer-aided diagnosis (CAD). The authors explain how machine learniing is applied by learning hypothesis. They propose a extension of random forest called Co-Forest to estimate the labeling confidence of undiagnosed samples and easily produce the final hypothesis. Case studies on three medical data sets and a successful application to microcalcification detection for breast cancer diagnosis show that undiagnosed samples are helpful in building CAD systems.
Usage: We use the term CAD and should refer to this report when introducing it. The paper motivates machine learning to enhance medical diagnosis which strengthen the relevance of our work. They have also implemented methods to diagnose breast cancer.

Title: Global trends in breast cancer incidence and mortality 1973–1997 [2].
Notes: Explains the global trends in breast cancer.
Usage: Use this data to motivate the importance of our work because breast cancer is very common and a leading cause of death.

Title: Beyond randomized controlled trials - Organized mammographic screening substantially reduces breast carcinoma mortality [13].
Notes: Proves the excellent results of mammography in terms of reduced mortality by 63% when screening was introduced.
Usage: Use this as motivation for screening, why it is important and thus makes our work relevant.

Title: Discovering Mammography-based Machine Learning Classifiers for Breast Cancer Diagnosis [12].
Notes: This work explores the design of mammography-based machine learning classifiers (MLC) and proposes a new method to build MLC for breast cancer diagnosis. We massively evaluated MLC configurations to classify features vectors extracted from segmented regions (pathological lesion or normal tissue) on craniocaudal (CC) and/or mediolateral oblique (MLO) mammography image views, providing BI-RADS diagnosis. Around 20,000 MLC configurations were evaluated, obtaining classifiers achieving an area under the ROC curve of 0.996 when combining features vectors extracted from CC and MLO views of the same case.
Usage: Very similar to our work, has a good introduction we can take inspiration from and along att the MLC configurations they used some might be similar to our. We might be able to compare our results with theirs. We can also take inspiration from the validation methods they've used, see if we can find the datasets they used etc. I think this has a high priority in our litterature review.

Title: Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms [10].
Notes: Feature dimension of three datasets is reduced using correlation based feature selection (CFS) algorithm. Second, classification performances of 30 machine learning algorithms are calculated for three datasets. Third, 30 classifier ensembles are constructed based on RF algorithm to assess performances of respective classifiers with the same disease data. All the experiments are carried out with leave-one-out validation strategy and the performances of the 60 algorithms are evaluated using three metrics; classification accuracy (ACC), kappa error (KE) and area under the receiver operating characteristic (ROC) curve (AUC).
Usage: In the concusion the authors writes: *The high number of the algorithms requires a more specific evaluation study to test the effect of the feature selection on classification accuracy.* wich we can use to motivate our work. Overall their work is very similar to ours and a lot of inspiration can be taken from this paper.

Title: Support vector machines combined with feature selection for breast cancer diagnosis [1].
Notes: From abstract: In this paper, breast cancer diagnosis based on a SVM-based method combined with feature selection has been proposed. Experiments have been conducted on different training-test partitions of the Wisconsin breast cancer dataset (WBCD), which is commonly used among researchers who use machine learning methods for breast cancer diagnosis. The performance of the method is evaluated using classification accuracy, sensitivity, specificity, positive and negative predictive values, receiver operating characteristic (ROC) curves and confusion matrix. The results show that the highest classification accuracy (99.51%) is obtained for the SVM model that contains five features, and this is very promising compared to the previously reported results
Usage: Super similar to ours, we should explain their results thuroughly and how our work will differ/build upon theirs, like applying more FS methods.

Title: A comparative study on the effect of feature selection on classification accuracy [7].
Notes: From abstract: Feature selection provides classifiers to be fast, cost-effective, and more accurate. In this paper the effect of feature selection on the accuracy of NaiveBayes, Artificial Neural Network as Multilayer Perceptron, and J48 decision tree classifiers is presented. These classifiers are compared with fifteen real datasets which are pre-processed with feature selection methods. Up to 15.55% improvement in classification accuracy is observed, and Multilayer Perceptron appears to be the most sensitive classifier to feature selection.
Usage: Vary similar, I think we should put emphasis on this report in our literature review and explain their results how we can biuld upon those and how our work will differ.

Title: MUC-4 Evaluation metrics [11].
Notes: -
Usage: This is a report from a conference where the F1 score measurement was introduced, we refer to this when we explain what F1 score is and hy we use it.

Title: UCI Machine Learning Repository [4].
Notes: -
Usage: Source of the dataset.

Title: Radiographer involvement in mammography image interpretation: A survey of United Kingdom practice [3].
Notes: A peper on the methods used by radiologist and what problems exist in this practices.
Usage: This we refer to when motivating why ML methods is necessary in terms of making diagnosis more efficient.

Title: Breast Cancer in Low- and Middle-Income Countries: Why We Need Pathology Capability to Solve This Challenge [9].
Notes: See title.
Usage: Again, strengthens the motivation of our work.

Title: An introduction to variable and feature selection [5].
Notes: A thurough explanation what feature selection is and how it works.
Usage: -

2.1 Conclusion

Here is a conclusion!

References

[1] Akay, M. F. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications 36*, 2, Part 2 (2009), 3240 – 3247.

[2] Althuis, M. D., Dozier, J. M., Anderson, W. F., Devesa, S. S., and Brinton, L. A. Global trends in breast cancer incidence and mortality 1973–1997. *International Journal of Epidemiology 34*, 2 (2005), 405–412.

[3] Culpan, A. Radiographer involvement in mammography image interpretation: A survey of united kingdom practice. *Radiography 22*, 4 (2016), 306 – 312.

[4] Dheeru, D., and Karra Taniskidou, E. UCI machine learning repository, 2017.

[5] Guyon, I., and Elisseeff, A. An introduction to variable and feature selection. *An Introduction to Variable and Feature Selection 3* (2003), 1157–1182.

[6] Igor, K. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine 23* (2001), 89–109.

[7] Karabulut, E. M., Özel, S. A., and İbrikçi, T. A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology 1* (2012), 323 – 327.

[8] Li, M., and Zhou, Z. H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 37* (Nov 2007), 1088–1098.

[9] Martei, Y. M., Pace, L. E., Brock, J. E., and Shulman, L. N. Breast cancer in low- and middle-income countries: Why we need pathology capability to solve this challenge. *Clinics in Laboratory Medicine 38*, 1 (2018), 161 – 173.

[10] Ozcift, A., and Gulten, A. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *Computer Methods and Programs in Biomedicine 104*, 3 (2011), 443 – 451.

[11] Ph.D., N. C. Muc-4 evaluation metrics. In *Fourth message understanding conference (MUC-4)* (McLean, Virginia, June 1992).

[12] Ramos-Pollán, R., Guevara-López, M. A., Suárez-Ortega, C., Díaz-Herrero, G., Franco-Valiente, J. M., Rubio-del Solar, M., González-de Posada, N., Vaz, M. A. P., Loureiro, J., and Ramos, I. Discovering mammography-based machine learning classifiers for breast cancer diagnosis. *Journal of Medical Systems 36* (Aug 2012), 2259–2269.

[13] Tabár, L., Vitak, B., H Chen, H., F Yen, M., Duffy, S., and Smith, R. Beyond randomized controlled trials - organized mammographic screening substantially reduces breast carcinoma mortality. 1724–31.