Feature Selection Methods for Classification of Breast Cancer

NIKLAS LINDQVIST
THONY PRICE

Degree Project in Computer Science
Date: March 11, 2018
Supervisor: Pawel Herman
Examiner: Orjan Ekeberg
Swedish title: Attributurvalsmetoder för klassificering av bröstcancer
School of Computer Science and Communication

Abstract

» This will be written later on...

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Sammanfattning

» This will be written later on...

Träutensilierna i ett tryckeri äro ingalunda en oviktig faktor, för trevnadens, ordningens och ekonomiens upprätthållande, och dock är det icke sällan som sorgliga erfarenheter göras på grund af det oförstånd med hvilket kaster, formbräden och regaler tillverkas och försäljas Kaster som äro dåligt hopkomna och af otillräckligt.

Contents

1 Introduction **1**
1.1 Research Question 2

2 Methods **3**

Bibliography **4**

A Unnecessary Appended Material **6**

Chapter 1

Introduction

Hospitals today are well equipped with data collection devices to do monitoring and data can be collected and shared in information systems. Collected data is a foundation for learning, both for medical personnel as well as for machines. As machine learning algorithms from the very beginning have been used to analyze medical data sets, machine learning is a well studied field within medical diagnosis [5]. Computer aided diagnosis (CAD) makes use of machine learning techniques that learn a hypothesis, a statistical prediction about a patient's diagnose, from a large set of previously diagnosed examples in order to assist medical experts in making more accurate diagnostics more efficiently [7].

Breast cancer is a disease of major concern and is the leading cause of cancer deaths among women [2]. At present there are no effective ways to prevent breast cancer. However, efficient diagnosis in an early stage can increase the chance of full recovery. This makes early detection and diagnosis an important issue and screening mammography is the primary imaging modality for early detection of breast cancer [4].

Multiple studies of CAD on breast cancer have been conducted, primarily focusing on classifying mammography data as malignant or not, such those of Ramos-Pollán et al. [10] and Akay [1]. The act of feature selection, removing redundant or irrelevant features from a dataset, can provide classifiers to be faster, more cost-effective and accurate [8]. It is also explicitly mentioned as a topic in need of more research a studies on breast cancer classification by Ozcift and Gulten [8].

1.1 Research Question

In our thesis we will study the impact of different feature selection methods on the classification rate of malignant breast cancer by different machine learning methods. Does the feature selection improve the accuracy of classification compared to using all features? Furthermore, in which machine learning methods does feature selection have the greatest impact? Our hypothesis is that overall the feature selection will improve the classification rate of all the machine learning methods used in this research scope. This hypothesis is based on previous research on this topic by Karabulut, Özel, and İbrikçi [6] where it was found that found classification improved by the use of filter methods for feature selection.

Our research differ the work presented in [6] in the size of data sets data sets. In our project we will only use one data set. Also, feature selection methods used in our research will not be restrained to filter methods, as in [6]. Lasty, the research scope in this thesis includes a study of the effect of different feature selection methods on Support Vector Machines (SVM) which was not investigated by [6].

Chapter 2

Methods

Trials will be conducted with feature selection Wrapper methods and feature selection Filter methods. The result of the feature selection methods will be used in a Decision Tree (DT), Support Vector Machine (SVM), Probabilistic Method (PM) and a Artificial Neural Network (ANN). The main reason for using these four methods are that our knowledge in machine learning is limited and the four mentioned methods are the methods that we have previously studied. Also, the methods are well studied and there are several conducted studies which can be used for comparison.

A comparison of the classification rate on the machine learning methods without any feature selection and with feature selection will be conducted in order to establish the importance of feature selection in different machine learning approaches when classifying breast cancer. The evaluation criteria will primary be F1 score that conveys the balance between recall and precision of classification performance [9]. Secondly, we will compare computational resources of the learning phase measured in time. The Breast Cancer Wisconsin (Diagnostic) Data Set [3], contains 569 instances with 32 attributes describing the features of breast cancer. Each instance is classified as benign or malignant.

Bibliography

[1] Mehmet Fatih Akay. "Support vector machines combined with feature selection for breast cancer diagnosis". In: *Expert Systems with Applications* 36.2, Part 2 (2009), pp. 3240–3247. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2008.01.009. URL: http://www.sciencedirect.com/science/article/pii/S0957417408000912.

[2] Michelle D Althuis et al. "Global trends in breast cancer incidence and mortality 1973–1997". In: *International Journal of Epidemiology* 34.2 (2005), pp. 405–412.

[3] Dua Dheeru and Efi Karra Taniskidou. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[4] P.C. Gøtzsche et al. "Beyond randomized controlled trials: Organized mammographic screening substantially reduces breast carcinoma mortality [2] (multiple letters)". In: 94 (Jan. 2002), pp. 578–583.

[5] Kononenko Igor. "Machine learning for medical diagnosis: history, state of the art and perspective". In: *Artificial Intelligence in Medicine* 23 (1 2001), pp. 89–109.

[6] Esra Mahsereci Karabulut, Selma Ayşe Özel, and Turgay İbrikçi. "A comparative study on the effect of feature selection on classification accuracy". In: *Procedia Technology* 1 (2012). First World Conference on Innovation and Computer Sciences (INSODE 2011), pp. 323–327. ISSN: 2212-0173. DOI: https://doi.org/10.1016/j.protcy.2012.02.068. URL: http://www.sciencedirect.com/science/article/pii/S2212017312000692.

[7] M. Li and Z. H. Zhou. "Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples". In: *IEEE Transactions on Systems, Man, and Cybernetics -*

Part A: Systems and Humans 37.6 (Nov. 2007), pp. 1088–1098. ISSN: 1083-4427. DOI: 10.1109/TSMCA.2007.904745.

[8] Akin Ozcift and Arif Gulten. "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms". In: *Computer Methods and Programs in Biomedicine* 104.3 (2011), pp. 443–451. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2011.03.018. URL: http://www.sciencedirect.com/science/article/pii/S0169260711000836.

[9] Nancy Chinchor Ph.D. "MUC-4 EVALUATION METRICS". In: *FOURTH MESSAGE UNDERSTANDING CONFERENCE (MUC-4), Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*. 1992. URL: http://www.aclweb.org/anthology/M92-1002.

[10] Raúl Ramos-Pollán et al. "Discovering Mammography-based Machine Learning Classifiers for Breast Cancer Diagnosis". In: *Journal of Medical Systems* 36.4 (Aug. 2012), pp. 2259–2269. ISSN: 1573-689X. DOI: 10.1007/s10916-011-9693-2. URL: https://doi.org/10.1007/s10916-011-9693-2.

Appendix A

Unnecessary Appended Material