Feature Selection Methods for Classification of Breast Cancer

NIKLAS LINDQVIST
THONY PRICE

Degree Project in Computer Science
Date: March 21, 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

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 [6]. 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 [8].

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 [13].

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. [12] 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 in a studies on breast cancer classification by Ozcift and Gulten [10].

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. We aim to answer the following:

- Does the feature selection improve the accuracy of classification compared to using all features?
- In which machine learning methods does feature selection have the greatest impact?
- What features are most important for accurate classification of breast cancer?

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 [7] where it was found that found classification improved by the use of filter methods for feature selection. Our research differ the work presented in [7] in the size of data sets data sets. In our project we will only use one data set and thus put more emphasis on breast cancer. The previous work did only investigate feature selection methods by filtering which we will extend with by implementing wrapper methods. 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 discussed by [7].

1.2 Approach

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). See summary of implementations in 1.1. The main reason for using these four methods **[insert better motivation here]** are that our knowledge in machine learning is limited and the four mentioned methods are the methods that we have previously studied. Also, the

Implementations	
Classifiers	Decision Tree Support Vector Machine Probibalistic Method Artificial Neural Network
Feature Selection by Wrapping	Recursive Feature Elimination (RFE) Sequential Feature Selector (SFS)
Feature Selection by Filter	Chi-square Entropy

Table 1.1: All classifiers should be tested with each feature selection method.

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 [11]. Secondly, we will compare computational resources of the learning phase measured in time. The Breast Cancer Wisconsin (Diagnostic) Data Set [4], contains 569 instances with 32 attributes describing the features of breast cancer. Each instance is classified as benign (357) or malignant (212).

1.3 Outline

In the Background section earlier work on the topic will be more thouroughly presented on order to clearify how this thesis relates to the field. The following chapter, Methods will detail the methods used to achive the results that are presented in chapter 4.

Chapter 2

Background

2.1 Computer Aided Diagnostics

Explain what it is

2.2 Breast Cancer

A study in Sweden by Tabár et al. [13] found breast carcinoma mortality was reduced by 63% after mammography was introduced. This clearly emphasize the benefits of screening which had resulted in a increased usage of the method to detect and diagnose breast cancer. The increasing demand for mammography image interpretation lead to a shortage of medical radiologist to perform this task and consequently non medical personnel supplement the mammography image interpretation (MII) [3]. As breast cancer continues to be the leading cause of cancer motality among women and more efficient diagnostics and pathology is high on demand the need of low-cost point-of-care is very much needed as stated by Martei et al. [9].

2.3 Feature Selection

The benefits of selecting a subset of all avaliable features are manyfold, among other it facilitates data visualization and data understanding, reduces the measurement and storage requirements and reduces training and utilization times. In cases with thousands of features it is essential to work with the data [5].

2.3.1 Filter methods

Explain

2.3.2 Wrapper methods

Explain

2.4 Related Work

Akay [1] investigated the performance of classification of a SVM with a RBF kernel using feature selection by F-score on the Wisconsin Breast Cancer Data set (WBCD). They achieved a classification accuracy of 99.51% which accordingly was among the highest scores recorded by then (2007). The highest classification rates was yielded by a model consisting of five features independently of how the test and training data was split.

Karabulut, Özel, and İbrikçi [7] made a comparative study on the effect of feature selection on classification accuracy and found up to 15.55% improvement on classification rates. The study used filter algorithms for feature selection, those were; Information Gain, Gain Ratio, Symmetrical Uncertainty, Relief-F, One-R and Chi-square. The study applied the selected features on three classification methods, Naive Bayes, Artificial Neural Network as Multilayer Perceptron, and J48 decision tree classifier on 15 different data sets including WBCD.

The results show that Multilayer perceptron benefited the most from feature selection by Chi-square, Naive Bayes by Gain Ratio and J48 by Information Gain. The study did not discuss optimal number of features.

Chapter 3

Method

3.1 Data set

The dataset used in this theises, Breast Cancer Wisconsin (Diagnostic) data set, was donated 1995 to UCI Machine Learning Repository [**Dua:2017**] by one of its creators, Nick Street. It contains 569 instances with 32 attributes describing the features of breast cancer. Each instance is classified as benign (357) or malignant (212). The 32 attribute discribes ten real-value features which are:

- **Radius:** Mean of distances from center to points on the perimeter.
- **Texture:** Standard deviation of gray-scale values.
- **Smoothness:** Local variation in radius lengths.
- **Compactness:** perimeter^2 / area - 1.
- **Concavity:** Severity of concave portions of the contour.
- **Concave points:** Number of concave portions of the contour.
- **Fractal dimension:** Coastline approximation - 1.
- **Perimeter:** Local variation in radius lengths.
- **Area**
- **Symmetry**

3.2 Implementation

Detail on what algorithms we used

3.3 Evaluation

What methods are implemented to measure the results, are we using mean, standard deviation, f1 score, anova, how many times do we run etc...

Chapter 4

Results and analysis

4.1 Classification impovements

Did classification improve, how, why

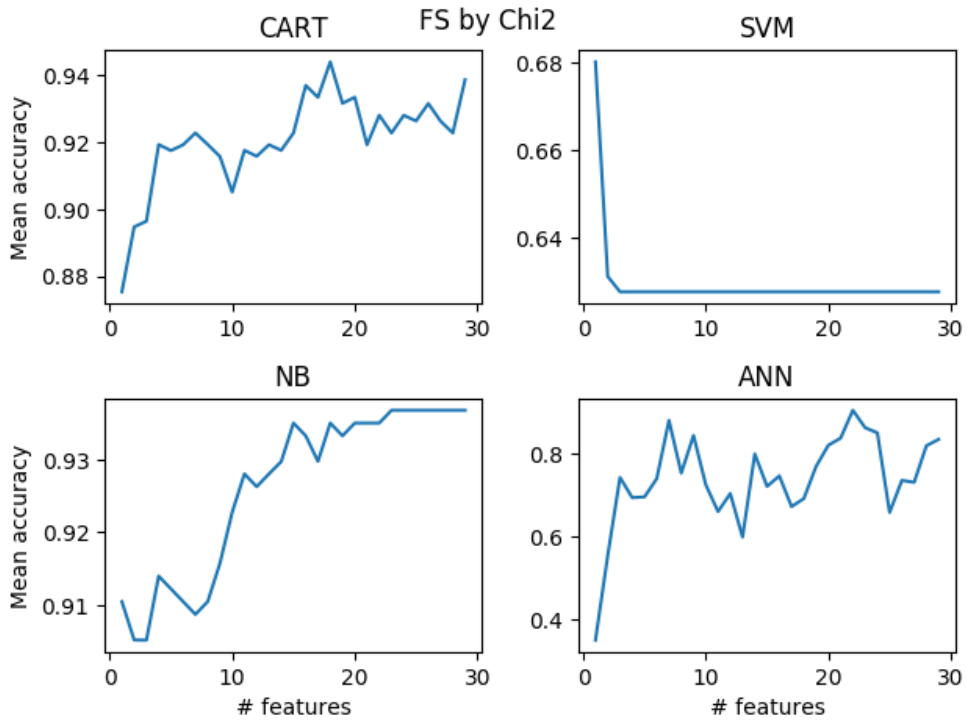

Figure 4.1: Performance by selecting features by Chi2

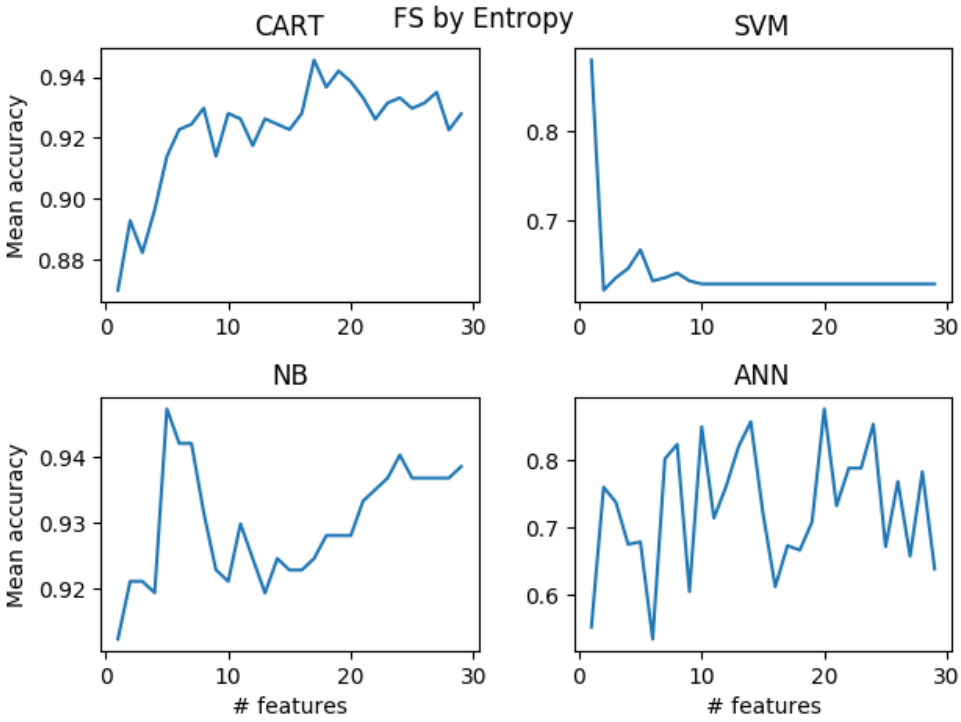

Figure 4.2: Performance by selecting features by Entropy

4.2 Best features

Could we tell which features contributes most to correct classification, how, why those

4.3 Source of errors

What can have caused faulty results, can our results be trusted?

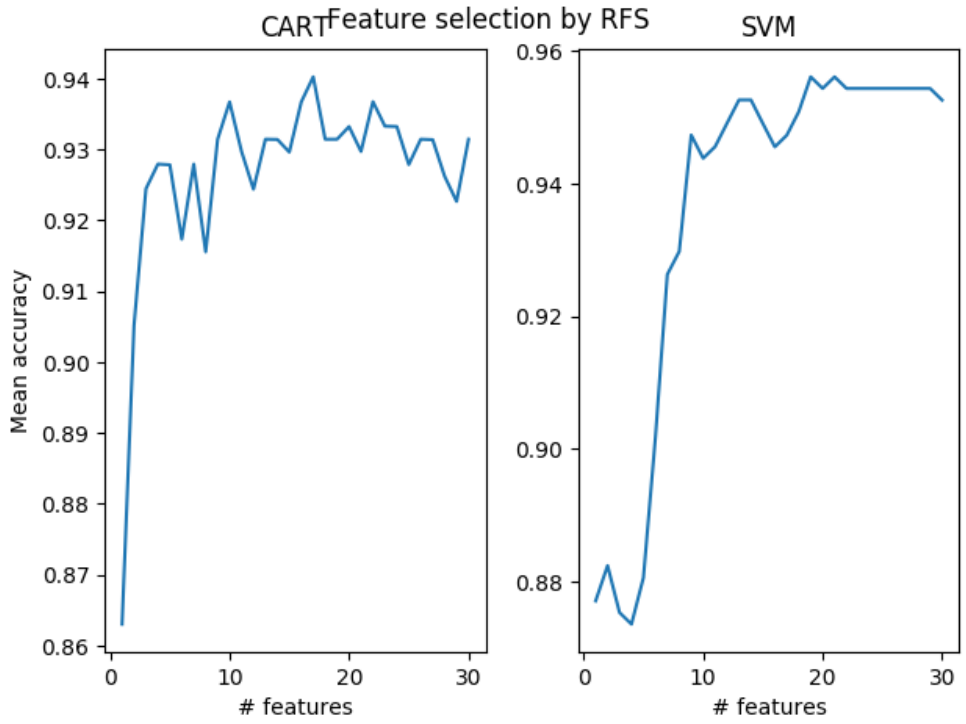

Figure 4.3: Performance by selecting features by RFS

Chapter 5

Discussion

Let's discuss here

Chapter 6

Conclusion

I guess this will be the last thing we write

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.

[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] A.M. Culpan. "Radiographer involvement in mammography image interpretation: A survey of United Kingdom practice". In: *Radiography* 22.4 (2016), pp. 306–312.

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

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

[6] 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.

[7] 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), pp. 323–327.

[8] 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 (Nov. 2007), pp. 1088–1098.

[9] Yehoda M. Martei et al. "Breast Cancer in Low- and Middle-Income Countries: Why We Need Pathology Capability to Solve This Challenge". In: *Clinics in Laboratory Medicine* 38.1 (2018), pp. 161–173.

[10] 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.

[11] Nancy Chinchor Ph.D. “MUC-4 Evaluation metrics”. In: *Fourth message understanding conference (MUC-4)*. McLean, Virginia, June 1992.

[12] Raúl Ramos-Pollán et al. “Discovering Mammography-based Machine Learning Classifiers for Breast Cancer Diagnosis”. In: *Journal of Medical Systems* 36 (Aug. 2012), pp. 2259–2269.

[13] László Tabár et al. “Beyond randomized controlled trials - Organized mammographic screening substantially reduces breast carcinoma mortality”. In: 91 (June 2001), pp. 1724–31.

Appendix A

Unnecessary Appended Material