

氏 名 小森 理

学位（専攻分野） 博士（統計科学）

学位記番号 総研大甲第 1335 号

学位授与の日付 平成 22 年 3 月 24 日

学位授与の要件 複合科学研究科 統計科学専攻
学位規則第 6 条第 1 項該当

学位論文題目 Boosting Methods for Maximization of the Area under
the ROC Curve and their Applications to Clinical Data

論文審査委員 主 査 教授 栗木 哲
准教授 藤澤 洋徳
教授 西井 龍映（九州大学）

With the advent of information age, huge amount of data has been collected in laboratories and hospitals. It includes not only clinical data such as age, laboratory test values, the size of internal organ; but also genomic data such as gene expression patterns, single nucleotide polymorphism (SNP) and proteome. Based on the information, we want to predict as accurately as possible the condition of the subject (diseased or non-diseased), who comes to a hospital and has gone through some clinical tests. However, it is often difficult to analyze these variety of medical data within a traditional statistical framework. Moreover, there exist criteria that are suitable for medical and clinical sciences. Hence, we have tried to develop a new statistical method that can deal with these data and provide us with a useful information for the discrimination, based on a criterion that is widely used by medical doctors or clinical researchers.

In medical and biological sciences, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) have gained in popularity. The ROC curve originated from the signal detection theory, where the performance of the radar operator who monitors enemy warplanes is measured or compared using the curve. It is also applied in psychology, and now is used in a variety of discrimination problems. Its appealing points are that the false positive rate (FPR) and the true positive rate (TPR) are both measured in the ROC curve, and that the curve is independent of the population prevalence of disease. FPR and 1-TPR express different aspects of the classification performance, so it is important to report the values separately, when evaluating the goodness of the classification. The independence also is suitable for quantifying the inherent accuracy of classification, and this property makes the AUC different from other accuracy measures such as the error rate, the relative risk or the odds ratio.

In this thesis, we have developed a new statistical method that is designed to optimize the AUC based on a boosting technique, which is widely used in the machine learning community. The method can deal with both usual low dimensional settings as well as high dimensional settings. The main concept of boosting is that a strong classifier (score function) is constructed by combining many various "weak classifiers". The weak classifier means that its discriminant ability is slightly better than random guessing. The method includes an implicit procedure of marker selection in its boosting algorithm, and produces a score function after an appropriate number of iterations. The resulting score plots are shown to be useful for understanding how each marker is associated with the outcome variable, say, the status of the subjects (non-diseased or diseased). Hence, our method put importance on the classification accuracy as well as the interpretation of the result. We also have extended this AUC-based boosting method to pAUCBoost, which focuses on the partial area under

the ROC curve (pAUC) that is often more relevant in some clinical or medical situations.

In Chapter 1, we review other accuracy measures than the AUC and pAUC, which are also important in clinical evaluation of markers; we investigate the properties and consider why the AUC and pAUC are getting popular in recent years. In Chapter 2, we also review the status of progress and development in machine learning community, and characterize the property of boosting from an objective viewpoint. We propose a new statistical method, termed AUCBoost, in Chapter 3 and discuss the statistical properties and demonstrate its utility. In Chapter 4, we focus on PSA data analysis. This is a collaborative research with medical doctors in Keio University Hospital. PSA is an abbreviation of prostate specific antigen, and is a primary marker for prostate cancers. The subject with PSA larger than 4 ng/ml is usually recommend to undergo biopsy; however, the value is affected by the age and the size of the prostate gland and other clinical covariates. Hence, we consider a optimal combination of these markers as well as the association to the prostate cancer, using AUCBoost. As a result, we present a ``nomogram'', by which medical doctors determine whether they perform biopsy in consideration of PSA, age, the volume of prostate gland and the number of biopsy undergone. The point of this nomogram is that the cutoff points are determined so that the sensitivity is at least 95 percent. This idea is quite different from existing nomograms that are based on a probability of having the cancer, and much more suitable for practical medical diagnosis. In Chapter 5, we extend AUCBoost to pAUCBoost, which focuses on the partial area under the ROC curve. We show that pAUCBoost is preferable to AUCBoost in some clinical situations. In Chapter 6, we mention ongoing and future work that I am engaged in now. Finally, we close this thesis with acknowledgements to all persons who supported me during my hard and pleasant doctor course.

博士論文の審査結果の要旨

統計的判別手法の性能を記述する方法として、ROC (Receiver Operating Characteristic) 曲線がよく知られている。とくにその曲線の下側面積 (AUC, the Area Under the ROC Curve) は、統計的判別手法の性能を要約する量として、医学診断分野などの応用分野で広く用いられる。本研究の主たる目的は、AUC を評価関数 (効用関数) とする判別方式を、機械学習の標準的な手法であるブースティング法により構成し、得られた手法の性能を評価し、さらにはそれを実問題へ適用するというものである。

提出された論文は、6章と1つの Appendix からなり、英語で執筆されている。第1章は序章である。第2章は、機械学習による統計的判別の概説にあてられている。続く第3章から第5章までは本論文の主要部である。第3章では、AUC を評価関数とし、その値をできるだけ大きくするような判別方式をブースティング法によって構成する方法 (AUC ブースト) を提案している。より正確には、評価関数の微分可能性を保障しブースティングアルゴリズムを構成するために、AUC の代わりに近似 AUC を用いている。どちらの評価関数に対しても最大化によって得られる判別関数は極限の意味では変わらないことを証明することによって、この近似が正当であることを示している。また数値実験と公開データの解析によって、従来の方法との比較を行っている。第4章では、前立腺がん診断で主要なマーカーである PSA 検査値に関する実データ解析が行われている。AUC を評価関数にとることによって、PS ノモグラムとよばれる診断アルゴリズムを提案している。得られた診断法は、4つのマーカーの値を組み合わせたものであり、従来法よりも精度が高く、また医師が医療の現場で使いやすいものとなっている。第5章は、第3章の改良として、AUC の代わりに pAUC (partial AUC) を評価関数とするブースティング判別法 (pAUC ブースト) を提案している。ここで pAUC とは ROC 曲線の一部の下側面積である。AUC と pAUC のどちらを用いるかは、問題に応じて使い分けられるべきものである。さらに本手法を、乳がん細胞の遺伝子発現のデータの予後予測の解析問題に適用し、従来法との比較検討を行っている。第6章は現在進行中の研究と将来的な展望である。

審査委員会の審査結果は以下の通りである。論文を通して評価関数として用いている AUC および pAUC は医学診断で広く用いられているものであるので、本論文で提案する方法の実際的な価値は高い。第3章と第5章で提案しているブースティング法 (AUC ブースト, pAUC ブースト) は、評価関数を AUC (pAUC) という非凸ロス関数とするものであり、その解析においては従来の指数ロス関数や対数ロス関数などの凸ロス関数に対する一般論を適用することができない。そのために、第3章では別のテクニカルな方法によって近似 AUC の利用の正当性を示しているが、このことは評価できる。また第4章では、医療の現場に足を運んだ現場主義によって、最終的に医療の現場で実用に耐える方法を提案しているが、これは統計科学の研究のあり方として好ましいものである。なお第3章の内容は、申請者の単著による査読付き英文論文 (雑誌名: *Annals of the Institute of Statistical Mathematics*) として既に採択されている。以上から審査委員会は、申請論文は学位授与に値すると判定した。