氏　　名	川喜田　雅則
学位（専攻分野）	博士（統計科学）
学位記番号	総研大甲第947号
学位授与の日付	平成１８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 method for local learning in statistical classification
論文審査委員	主　査　教授　栗木　哲 教授　江口　真透 助教授　南　美穂子 助教授　藤澤　洋徳 助教授　池田　思朗 教授　村田　昇（早稲田大学）

論文内容の要旨

The main objective is to study boosting methods in statistical classification. Several ensemble learning methods including boosting have attracted many researchers' interests in the last decade. In particular, it has been reported that the boosting methods perform well in many practical classification problems. The boosting algorithm constructs an accurate classifier by combining several base classifiers, which are often at most slightly more accurate than random guess. While many researchers have studied the boosting methods, their success has still some mysterious aspects. More intensive theoretical studies are required to clarify such mysteries.

We describe a survey on several ensemble learning methods. We set up a statistical classification problem and make some notations to develop discussion from learning theories. Some theoretical preliminaries for analyzing the performance of classification methods are also overviewed. Then, we survey some existing ensemble learning methods. In particular, we review theoretical properties of boosting methods, which have been clarified by several researchers.

The application of AdaBoost with decision stumps to shark bycatch data from the Eastern Pacific Ocean tuna purse-seine fishery is described. Generalized additive models (GAMs) are one of the most widely-used tools for analyzing fisheries data. It is well known that AdaBoost is closely connected to logistic GAMs when appropriate base classifiers are used. We compared results of AdaBoost to those obtained from GAMs. Compared to the logistic GAM, the prediction performance of AdaBoost was more stable, even with correlated features. Standard deviations of the test error were often considerably smaller for AdaBoost than for the logistic GAM. In addition, AdaBoost score plots, graphical displays of the contribution of each feature to the discriminant function, were also more stable than score plots of the logistic GAM, particularly in regions of sparse data. AsymBoost, a variant of AdaBoost developed for binary classification of a skewed response variable, was also shown to be effective at reducing the false negative ratio without substantially increasing the overall test error. In the analysis of shark bycatch data, we observed that there existed several spatially local structures. This indicates that the classification rule that varies depending on location may improve the prediction performance.

Boosting with decision stumps, however, may not capture complicated structures in general since decision stumps are considerably simple classifiers. Use of more complicated base classifiers possibly improves the approximation ability of boosting.

However, several literatures have pointed out that the use of complicated base classifiers may increase the generalization error of boosting. In addition, it is difficult to find what types of base classifiers are appropriate to each problem without any prior knowledge.

To overcome these difficulties, we propose a new method, the local boosting, that is a localized version of boosting method based on the idea similar to but not the same as the local likelihood. Application of the local likelihood may improve the approximation ability considerably but also increases the computational cost, which makes the algorithm infeasible. The local boosting, however, includes a simple device for computational feasibility. We show that the local boosting has the Bayes risk consistency in the framework of PAC learning. It is seen that the estimation error increases compared to the ordinary boosting with simple base classifiers when we use the ordinary boosting with more complicated base classifiers or when we use the local boosting. However, the increase caused by the local boosting is not steep. The approximation error of local boosting is guaranteed to be smaller than that of the ordinary boosting, which is often much smaller with appropriate bandwidth. Therefore, when same base classifiers are used, the local boosting attains the Bayes risk consistency in wider situations than the ordinary boosting by controlling the trade-off between estimation error and approximation error. Several simulations confirm the theoretical results and the effectiveness of the local boosting over the ordinary boosting in both binary and multi-class classifications.

In addition, we mention the relationship between the kernel classification rule and the local boosting when the training data are used as kernel centers. While the local boosting is derived by localizing the ordinary boosting, it is also derived by extending the kernel classification rule. Both methods are related to the reproducing kernel Hilbert space induced by some kernel functions. These relationships enabled us to derive another types of the upperbound of their generalization error. These bounds indicate that the estimation error of the local boosting may be larger than that of the kernel classification rule. Conversely, the local boosting has more powerful approximation ability than the kernel classification rule. However, these topics are still under investigation, which are expected to be studied further in the future.

論文の審査結果の要旨

統計的学習理論におけるパタン識別手法として近年提案されたものとして，ブースティング法（アダブースト）が知られている．申請論文は，このアダブーストに関連して，(i) データ解析手法としての観点からの従来法との比較，および（ii）局所化（局所アダブースト）の提案とその性質の理論的解明を行ったものである．

申請論文は全5章から構成されている．第1章では序説として，論文全体の流れと目的が説明されている．第2章ではアンサンブル学習に関する数学的準備とブースティングの具体的なアルゴリズムが定式化されている．特にVC次元の定義と関連する不等式，ならびにアダブーストの基本的な統計的性質についての概説がなされている．これらは第4章の数学的準備と位置づけられる．本申請論文の主要な結果は，第3章と第4章で与えられる．第3章では，マグロ漁におけるサメの混獲のデータを題材として，アダブーストと従来法であるロジスティック一般化加法モデルのデータ解析の比較を行っている．アダブーストによって構成される識別関数が，ロジスティック一般化加法モデルと同様，説明変数の関数（スコア）の和に分解されることに着目し，両者のスコアの挙動の解明を通して，アダブーストがデータ解析手法として安定した性質をもつことを確認している．また，スコアの図示化の方法（スコアプロット）を提案し，データ解析における利用法を提案している．第4章では，局所尤度の考えをアダブーストに適用した局所アダブーストが提案されている．ここでいう局所化は，カーネル中心をランダムに与えるというものであり，従来の局所尤度とは異なる局所化の工夫がされたものである．また，提案手法の経験リスク関数の挙動の漸近解析が行われている．一般に，ある識別アルゴリズムの経験リスク関数が漸近的にベイズリスクに収束するときベイズリスク一致であると言い，アルゴリズムの好ましい性質とされる．本章では，提案した局所アダブーストがベイズリスク一致であることを証明している．さらにその証明の中で，リスクの確率的な上界（最悪評価）を導出している．これは，通常のアダブーストにおいて知られている上界を著しく改良するものである．なお得られた上界をより詳しくみると，局所アダブーストにおいては，一般的にトレードオフの関係にあると言われる推定誤差と近似誤差のうちの推定誤差をやや犠牲にして近似誤差を改良することによって汎化誤差を改良することも分かる．数値実験によると，通常のアダブーストで学習できない例題に対しても局所アダブーストが良好に働くことが分かるが，このことはリスクの漸近評価に関する理論結果をサポートするものと考えられる．第5章では，論文全体の結論と課題，今後の展望が述べられている．

なお本申請論文の内容は英文論文2編（1編は ”An introduction to the predictive technique AdaBoost with a comparison to generalized additive models,” *Fisheries Research*, Vol. 76, 328-343, 2005 として出版済，1編は投稿中）にまとめられている．

審査委員会の審査結論は以下の通りである．アダブーストという統計科学の観点からも重要な手法に関して，(i)「スコアプロット」を用いたデータ解析法を提案し，その比較を通して従来法との違いを明らかにしたこと，ならびに（ii）アダブーストの局所化（局所アダブースト）を新規に提案し，その理論的性質を数学的に厳密な形で解明したことは高く評価できる．以上から本申請論文は学位授与に値する．