氏　　名	竹之内　高　志
学位（専攻分野）	博士(学術)
学位記番号	総研大甲第739号
学位授与の日付	平成１６年３月２４日
学位授与の要件	数物科学研究科　統計科学専攻 学位規則第４条第１項該当
学位論文題目	Statistical learning theory by Boosting Method
論文審査委員	主　査　教授　中野　純司 教授　江口　真透 助教授　栗木　哲 助教授　南　美穂子 助教授　村田　昇（早稲田大学）

論文内容の要旨

We deal with statistical learning theory, especially classification problems, by Boosting method. In the context of Boosting method, we can use only a set of weak learners which output statistical discriminant functions having low performance for a given set of examples. Aim of Boosting method is to construct a strong learner by combining a lot of weak learners and a typical boosting algorithm is AdaBoost. AdaBoost can be derived from a sequential minimization of the exponential loss function for a statistical discriminant function. This minimization problem is equivalent to the minimization of the extended Kullback-Leibler divergence between an empirical distribution of given examples and an extended exponential model. Statistical properties of AdaBoost have been investigated and the relationship between the exponential loss function of AdaBoost and the logistic model was revealed. In this thesis, we obtain two main results:

1. AdaBoost is extended to general U-Boost by using the statistical form of the Bregman divergence, which contains the Kullback-Leibler divergence as an example and consider a geometrical interpretation of U-Boost in terms of information geometry.

2. We propose a new Boosting algorithm η -Boost, which is a robustified version of AdaBoost.

The U-Boost is derived from a sequential minimization of the Bregman divergence between the empirical distribution and U-model. A geometric interpretation for U-Boost is given in terms of information geometry. From the Pythagorean relation associated with the Bregman divergence, we derive two special versions of U-Boost, the normalized U-Boost and the unnormalized U-Boost. We define the normalized version of U-model on the probability space and derive normalized U-Boost from this model. The normalized U-Boost corresponds to usual statistical classification methods, for example, logistic discriminant analysis. The unnormalized U-Boost is derived from an unnormalized version of U-model defined on the extended non-negative measure space and has not been seen in the previous statistical context. Especially, unnormalized U-Boost has a beautiful geometrical structure related to the Pythagorean relation and the flatness. Its algorithm is interpreted as a pile of right triangles which leads to a mild convergence property of U-Boost algorithm as seen in the EM algorithm. Based on a probabilistic assumption for a training data set, statistical discussion for consistency, efficiency and robustness of U-Boost is given.

An algorithm of AdaBoost implements the learning process by exponentially reweighting examples according to classification results. Then weight distribution is

often too sharply tuned, so that AdaBoost has a weak point on the robustness and over-learning. As a special example of U-Boost, we propose η -Boost which aims to robustify AdaBoost to avoid an over-learning. The statistical meaning of η -Boost is discussed and η -Boost is associated with a probabilistic model of mislabeling which is a contaminated logistic model. As a general U-Boost algorithm, η -Boost also has a normalized and unnormalized version. A loss function of the normalized version of η -Boost is a minus log-likelihood of a contaminated logistic model in which mislabeling probability is constant and does not depend on the input. The unnormalized version of η -Boost is a slight modification of AdaBoost and is derived from a loss function which is defined by a mixture of the exponential loss of AdaBoost and naive error loss functions. A probabilistic model of unnormalized version is also a contaminated logistic model and its mislabeling probability depends on the input. In an algorithm of unnormalized version of η -Boost, a weight distribution of AdaBoost is moderated by an uniform weight distribution and a way of combining a weak learners is adjusted by a naive error rate. As a result, η -Boost incorporates the effect of forgetfulness into AdaBoost. For both versions, a tuning parameter η is associated with a degree of the contamination of the model and we can choose it by the minimization of naive error rate. We theoretically investigated the robustness of η -Boost and confirmed it with computer experiments. Also, we applied η -Boost to real datasets and compared it with previously proposed Boosting method. The η -Boost outperformed the other method in term of robustness.

論文の審査結果の要旨

本審査委員会は竹之内高志君の博士申請論文の審査を行った.

1. 博士申請論文の概要

竹之内君の申請論文は，近年，活発に研究されている統計的学習理論における「ブースティングによる判別解析」を扱っている．その主要成果としては，ブースティングの方法の中で一般的な U-ブーストというクラスの幾何学的な構造を明らかにしたこと，そしてそのクラスの中の有用なものとして，ミスラベルに対するロバストネスの観点から η-ブーストとよばれる手法を提案したことである．

申請論文は全 7 章からなる．最初の 2 章は準備であり，第 1 章ではブースティングの歴史を述べ，第 2 章では関連する学習理論の話題が概説される．第 3 章は U-ブーストの幾何学的な構造を考察したものである．U-ブーストは，正値導関数を持つ凸-実数値関数 U から定義され，例えば，U 関数として $U(z) = \exp(z)$をとれば Freund-Schapire (1997)によるアダブーストになる．実際の U-ブーストアルゴリズムは，U-ロス関数の逐次最小化としてアダブーストアルゴリズムの一般化で定義される．Lebanon-Lafferty (2001)は，アダブーストアルゴリズムを導く指数ロスは，経験分布から指数モデルへの Kullback-Leibler ダイバージェンスと等価であることを示し，更に指数モデルに全マスが 1 である‘確率分布の要請’を課すと，ロジットブーストアルゴリズムを決める対数ロスと等価になることを示した．本章ではこれを拡張し，U-ロス関数は経験分布から U-モデルへの U-ダイバージェンスと等価になることを示した．ここで U-モデルと U-ダイバージェンスは各々指数モデルと Kullback-Leibler ダイバージェンスの自然な拡張である．そして，U-ロス関数は U-モデルを正値測度モデルと取るか，確率分布モデルと取るかにより，2 種類のバージョンが考えられることを示し，それらの間の幾何学的構造を明らかにした．第 4 章では η-ブーストを考察する．アダブーストを自然に拡張した U-ブーストクラスの中でどのような U 関数が有用であるかという問いに対して，ミスラベルに対するロバストネスの観点から η-ブーストが提案される．ミスラベルは医療診断の文脈では誤診を意味する．η-ブーストは U 関数 $U(z) = (1-\eta)\exp(z) + \eta z$ によって定義される．ここで η は $0<\eta<1$ なる定数を表す．これから導かれる U-ロス関数と U-モデルに対する統計的考察によって，η はミスラベル確率と密接な関係で結ばれることが示された．第 5 章では，人工データおよび機械学習の Web ページで公開されているいくつかの実データセットに対して，エラーレイトを中心に η-ブーストとその他の判別法の比較を行っている．特に人工的にミスラベリングを行ったデータの解析では，η-ブーストが困難な状況下でも比較的よい識別能力を保てることを示している．第 6 章では，ラベルが部分的に欠落していたり，非対称なロスを用いたりする場合へ U-ブーストを拡張することが考察される．第 7 章は簡単なまとめである．なお，本論文の内容は 1 編の和文論文および 2 編の英文論文として発表されている．

2．審議結果

申請論文は，U-ブーストの考察によって統計的学習理論に独自の統一的視点から理論的な貢献を行っていると認められる．また η-ブーストは理論的に興味深いだけでなく，現実的な問題に対しても有力な解析手法になりうると評価できる．よって博士論文審査委員会は，竹之内君の博士申請論文が，学位授与に十分値する水準にあると判定した．