氏　　名	白石友一
学位（専攻分野）	博士（統計科学）
学位記番号	総研大甲第1149号
学位授与の日付	平成２０年３月１９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	Game-theoretical and statistical study on combination of binary classifiers for multi-class classification
論文審査委員	主　査　准教授　土谷　隆 教授　福永　健次 准教授　池田　思朗 准教授　金森　敬文(名古屋大学)

論文内容の要旨

多値判別の問題は科学の多くの分野で頻繁に見られる．二値判別の問題については，Support Vector Machine (SVM) やada-boostなど多くの有効な手法が開発されており，それらの性質についての研究も深く進んでいる．しかし，多値判別についてはいまだ決定的な手法の開発がなされておらず，現在においても多くの手法が提案され続けている．多値判別には，大きく分けて二つのアプローチがある．一つ目のアプローチは，三クラス以上のラベルを同時に扱う損失関数を考え， その損失関数を何らかの手法により直接最小化することである．このアプローチは理論的な解析が比較的容易であり，ベイズエラーへの一致性についての研究もいくつかなされている．しかし，サンプル数が多い場合には，このアプローチに立つ多くの手法は計算が困難となってしまう．もう一つの代表的なアプローチは，二値判別機を組み合わせて多値判別を行うという方法である．このアプローチは計算量が比較的抑えられること，実装が容易であることから多くの実用上の問題で用いられている．二値判別機の結果から最終的な多値判別の結論を導出する方法，すなわち二値判別機の結果を組み合わせる方法には，多数決法や，有効グラフのモデルを用いた方法，Bradley-Terryモデルを用いた方法，誤り訂正符号による方法がよく用いられる．

以上の手法においては，二値判別機の組み合わせの方法はトレーニングデータに依存せずに決定的に決まっている．この組み合わせの方法をデータから学習することにより，識別率の向上を目指した研究がいくつかなされてきた．しかし，これらの研究において提案された手法は誤判別率を大きく改善するわけではなく，「組み合わせの方法を学習するべきか，そうでないか？」という問題はこの分野で活発に議論されている．本論文ではこの問題を，ゲーム理論と統計学によるアプローチにより解決することを目指した．

まず、いくつかの二値判別機の組み合わせ法を概観し、full-model ECOC という新しい組み合わせ法の提案を行う．そして，それらの方法を数値実験による比較し，二値判別機の組み合わせ法を学習することについての問題点を議論した．

本論文の主要部分の前半においては、組み合わせの方法を学習しないことに対するある種の理論的な正当性を与えた．まず，判別の問題を，二値判別機の結果から最終的にラベルの決定を行う「決定者」と，二値判別機の結果の確率分布を定める「自然」によるゲームとして捉える．そして，組み合わせの方法を学習しないことが一種の最適性を有することを，誤り訂正符号による方法がミニマックスであることを証明することにより示す．最初に，二値判別機の出力が互いに独立であるという仮定のもとで，one-vs-allの場合に誤り訂正符号による方法がミニマックスであるということを証明した．次に，二値判別機の出力が独立であるという仮定を外した状況での解析を行った．one-vs-one

や one-vs-all などの二値判別機の学習の枠組みにより，「自然」にはどのような制約を入れるべきかが問題となる．まず，「自然」の戦略集合が複数の不等式制約の積集合で表されているとき，その制約集合とミニマックス性の関係を feasible flow によって特徴付ける定理を証明した．その定理により，誤り訂正符号による方法が自然な制約でミニマックスとなることを示した．さらに，誤り訂正符号による方法が，one-vs-one よりも one-vs-all の場合の方が少ない制約でミニマックス性を有することを示すことにより，one-vs-all と one-vs-one のどちらが優れているかという未解決問題に，一定の示唆を与えた．

後半部においては，二値判別の新しい組み合わせ法の提案を行った．最初にゲーム理論に基づいた新しい二値判別機の組み合わせ法を提案した．具体的には，誤り訂正符号による方法に拡張を加え，「自然」の範囲をデータからある程度特定したときのミニマックス戦略を求める方法を提案し，これを二次錐計画問題に定式化した．この方法は今までの多くの方法と違い，条件付確率の推定精度を考慮しつつ，どのクラスを選べば良いかについての最適確率化戦略を直接的に求めるという特徴がある．また，ゲーム理論と対応する点として，最適解の存在がゲーム理論におけるミニマックス定理により証明されること，最適値から得られる最悪の場合の誤判別率がエントロピーの概念と関係づけられることが挙げられる．次いで，二値判別機の組み合わせ法を学習する際に，既に学習に用いたデータを組み合わせ時にも用いなければならないために生じる過学習の問題に取り組んだ．この問題を，ブートストラップやクロスバリデーションを用いた stacking という手法で回避する方法を提案した．そして数値実験の結果，stacking により組み合わせ法を学習する提案手法が，学習しない方法に比べて多くの場合によりよい識別結果を与えることを示した．

論文の審査結果の要旨

[論文の概要]

提出論文は、多値識別問題を多数の２値識別器の結果を組み合わせることにより解く手法に関する研究をまとめたもので、全８章８４頁からなる。

第１章は、研究の背景や課題を述べている。多値識別問題を２値識別器の組み合わせによって解く場合、組み合わせ部分の学習、すなわち、組み合わせ法をデータに依存して決定することがよいか否かは、この分野の未解決問題となっており、その問題が本論文の中心課題であることが説明されている。第２章では、多値識別問題を１つの最適化問題に定式化して直接解くアプローチが概説されている。第３章では、データが特定の２クラスの内どちらに属するかを判定する one-vs-one や特定のクラスに属するか否かを判定する one-vs-all などの問題に対する２値判別器を適切に組み合せて多値識別問題を解く手法として、多数決法や Bradley-Terry 法、ECOC (Error Correcting Output Code) 法、Hamming Decoding 法が説明されている。これらの手法では、判別器の組み合わせ法はデータに依存しない。第 4 章では、組み合わせ法をデータから学習するアプローチを説明し、その拡張として、full model 法と modified ECOC 法を提案している。

本論文の主要な結果を述べた第 5 章から第８章は大きく２つに分かれており、第 5 章、第 6 章は組み合わせ法を学習しない ECOC 法のゲーム理論的解析を、第７章、第８章はデータから組み合わせ法を学習するための方法を論じている。

ゲーム理論的解析においては、２値識別機の組み合わせ問題を、２値識別機の出力ベクトルとクラスラベルとの真の同時確率分布を選択する "nature" と、２値識別機の出力ベクトルからクラスへの決定関数を選択する "combiner" との間のゲームとして定式化し、いくつかの組み合わせ手法の minmax 性を論じている。まず、第５章は、ECOC に対する既存の理論解析でよく用いられる、２値識別器群が互いに独立であるという仮定のもとで、one-vs-all により２値識別器を構成した場合には、ECOC による組み合わせが minmax であることを証明している。第 6 章はこれを発展させて独立性の仮定を排除し、one-vs-one と one-vs-all のそれぞれに対して"nature" の取りえる確率分布に自然な制約を与え、それらの制約の下では ECOC が minmax であることを証明している。また、制約条件を比較することにより、より弱い制約下で ECOC を minmax にする one-vs-all がよりよい２値識別器構成法であることを理論的に示唆している。さらに、実際のデータを用いて制約条件の妥当性を検討している。

第５章、第６章の議論は、"nature" の取る確率分布に関する情報として、緩い制約条件のみを用いることが前提であったが、一方、この確率分布を積極的にデータから推定することによる組み合わせ法を論じたのが第７章、第８章である。第７章は、full model 法に

よる確率分布の推定値の信頼区間を用いて、"nature"の取りえる確率分布の集合をデータに依存して定め、その集合での minmax 解を組み合わせ法として用いるロバスト組み合わせ法を提案している。また、解を得るための最適化問題が２次錐計画に変換でき、既存の最適化ソフトウエアによって容易に数値解が得られることを述べている。第８章は ２値識別器の出力を重みつき線形和によって組み合わせる方法を考察している。このときの問題点として、線形和の重み係数を決める目的関数に、２値識別器の学習に用いたものと同じデータを用いると、データに対する過適合を起こしやすいことが論じられている。この問題を解決するために、クロスーバリデーションやブートストラップを用いた stacking と呼ばれる手法によってデータの重複使用を回避する方法を提案している。実データによる数値実験の結果、組み合わせ法を学習する提案手法が、学習しない方法に比べて多くのケースでよりよい識別結果を与えることを示している。

第９章は論文のまとめである。

[論文の評価]

第５章，第６章で論じられたゲーム理論的考察は、出願者が提案した独自の枠組みであり、興味深いものである。特に第５章では、一般的な制約のもとでの minmax 解を、ネットワークフロー問題で使われる feasibility という概念を用いて特徴付ける定理を証明しており、最適化の観点からも十分に評価できる。さらに、多値識別に用いるための２値識別器の構成法として、one-vs-one と one-vs-all のどちらが優れているかという、この分野の未解決問題に対し、一定の理論的示唆を与えた点にも意義がある。

第７章のロバスト組み合わせ法においては、第５章、第６章での問題意識をさらに発展させて、データを考慮した一定の妥当性を持つ仮定の下で、 minimax 性を有する組み合わせ方を求める問題を、２次錐計画として定式化する点が理論的に興味深く、また実用的な速度での計算を可能にしている。また、第８章で提案された方法は、組み合わせ法の学習によって多くのケースで識別誤差を減少させる結果を示しており、組み合わせ法の学習が有利である可能性を示した点が評価できる。

以上より、本論文は博士（統計科学）を与えるに十分な内容を有すると判定した。

[その他]

第５章の内容をまとめた論文が国際学術雑誌 Neurocomputing に、また、第６章の内容をまとめた論文が電子情報通信学会論文誌Ｄに掲載されることが決定済みである。