氏名	Adrian Pino Angulo
学位の種類	博士（応用情報科学）
学位記番号	博情第　51 号
学位授与年月日	平成　31 年　3 月　22 日
学位授与の要件	学位規則第 4 条第 1 項該当（課程博士）
論文題目	A New Balance for Efficiency and Accuracy of Feature Selection for High-dimensional Dataset
論文審査委員	（主査）教授　申　吉浩 （副査）教授　竹村　匡正 （副査）准教授　大島　裕明

学位論文の要旨

Machine Learning has been one of the hottest trends for the last ten years. Supervised classification as a sub-field of machine learning, is increasingly gaining popularity among researchers due to its versatility and power of application at any field where data is available. Among the most common examples of supervised learning we can find: micro-array problem classification, cancer diagnosis and network intruder detection. Supervised classification is a central issue in machine learning and consists on finding a classification function l :$\mathbf{D} \rightarrow v(c)$ that is able to classify an arbitrary instance with unknown class from $v(c) \in C$. l is built from analyzing the relation between instances in **D**. The performance of supervised classifiers is often measured in three directions: efficiency, representation complexity and accuracy. The efficiency refers to the time required to learn the classification function l; while the representation complexity often refers to the number of bits used to represent the classification function. All these three factors can be strongly affected when there exist features in **D** that do not contain useful information to predict the class variable. Feature selection methods are able to identify and remove unneeded, irrelevant and redundant features from data that do not contribute to the improvement of the accuracy of a predictive model. Feature selection allows us to build models as good or with better accuracy whilst requiring less data. The process of selecting features is composed of two basic components: an evaluation function and a search engine. The evaluation function is a metric that evaluates quantitatively how good are a set of features to discriminate among class labels. On the other hand, the search engine is in charge of generating all the potential sets to be evaluated. Feature selection algorithms can be divided into three broad categories: wrapper, filter and embedded methods. To evaluate a feature set F, wrapper methods use some accuracy score of a classifier after being trained in the dataset projected by F. Wrapper methods are very low in efficiency since training and testing the inferred function is required for each evaluation. Conversely, filters make use of explanatory analysis on data to assign a score to each feature set. Filters are usually less computationally expensive than wrappers, but they output a feature set that is not tuned to a specific type of predictive model. Embedded methods learn which features best contribute to the accuracy of the model while the model is being created. The most common type of embedded feature selection

methods are regularization or penalization methods. Filter-based feature selection can be also classified as: *feature ranking*, *pairwise evaluation* and *consistency-based* algorithms. The *feature ranking* methods evaluate relevance of individual features using statistical measures. That is, features are ranked using their individual relevance score and then the top features are selected. Although the ranking feature algorithms are usually simple and fast, they have two serious drawbacks that may affect the performance of super- vised classifiers. First, redundant features are likely to be selected. Second, they usually can not detect interacting features. Oppositely to the *feature ranking* algorithms, pairwise

evaluation methods can detect and eliminate relevant features, but also are able to remove redundant features by computing the correlation between features. Consistency-based algorithms can detect interacting features by collectively evaluating relevance (correlation) of a feature set to the class. Although exhaustive search of all possible feature sets is computationally too expensive, the result can be expected to be accurate. In this paper, we propose several feature selection algorithms for high-dimensional data that can efficiently find very accurate solutions when compared with other benchmarking algorithms. Our contribution is as follows.

• We first, propose four new feature selection algorithms based on consistency mea- sures, which are improvements of the current state-of-the-art algorithms: *Steepest-Descent-Consistency-Constrained* (SDCC), the *Linear-Consistency-Constrained* (LCC), *Super Linear-Consistency Constrained* (SLCC), respectively.

• Second, we propose a rule-based feature selection algorithm, namely, *Probabilis- tic Attribute Value Integration for Class Distinction* (PAVICD), which can detect interacting features and is extremely fast.

• Third, we propose a new version of the pairwise-evaluation-based algorithms, the *Fast Correlation based Filter* (FCBF) and the *Correlation-based Feature Selection* (CFS).

• Lastly, we propose an improvement of the hybrid feature selection algorithm, namely *Genetic Bee Colony for Feature Selection* (GBC).

All the proposed algorithms are tested in terms of accuracy, number of selected features and running time required. Results of the experiments in high-dimensional data exhibits that in most of the datasets our proposed algorithms are faster and more accurate than the original algorithms.

論文審査の結果の要旨

平成３１年２月１５日、論文内容およびこれに関連する事項について試問を行った結果、研究内容・公表実績・学力のいずれにおいても、本研究科における博士論文としての水準を満足するものと判定した。

本研究は、機械学習研究の中心的な領域の一つである、特徴選択に関わるものである。当研究領域は研究の歴史も長く、非常に多くのアルゴリズムが発表されている。その一方で、近年、ビッグデータに象徴される大規模データが利用可能になるに従い、大規模データに対して高速かつ正確に特徴選択を行う、新しい次元のアルゴリズムの開発が重要になっている。特に、数万から百万個の特徴を含む高次元データセットから、数個から数十個の特徴を選択できることが求められている。本研究は、従来提案されている中から、良好な性能を有する複数の特徴選択アルゴリズムに対し、効率と正確性の両面から根本的な改善を加えるものであり、改善には研究的にみて斬新な手法を用いているとともに、提案されたアルゴリズムは直ちに実用に供することができるなど、実用的な価値も極めて高い。提案アルゴリズムは、最急降下法に基づく SDCC に対し、特徴の探索範囲を動的に制御することにより、効率・正確性の両方を改善した FSDCC 、現在最も高速なアルゴリズムの一つである LCC に対し、性能を左右するハイパーパラメータの最適化を擬似焼きなまし法 (simulated annealing) により自動化した SA-based LCC、最も広く用いられている MRMR と CFS のアルゴリズム上の計算冗長性を排除し著しく効率を高める MRMR+及び CFS+、ミツバチの挙動モデルに基づく遺伝子分析において広くも利用されている GBC を効率・正確性両面で改善した GBC+である。結果は、３編の原著論文で発表され、うち、一篇はインパクトファクターが付与された雑誌から出版されている。また、国際発表論文の一編は Best Student Paper Award を受賞している。

論文は、新規の考え方・手法を提案しているが、60 篇を超える既存研究の詳細な検討に基づいて、提案手法の新規性について具体的に記述している。提案手法の有効性については、多くのデータセットを用いた綿密な実験を実施し、その結果を多面的な指標を用いて評価し、効果を実証している。いずれの提案も、従来研究の重要な問題点を有効に改善するものであり、学術的に有用であるほか、直ちに実用に供し得るものである。

論述も明解であり、かつ、必ずしも当該領域の専門家でなくても理解が可能なように構成されており、公開が前提の博士論文の要件を満足している。

尚、Adrian Pino Angulo 氏は、日本国文科省の国費留学生である。