学 位 論 文 の 要 旨

Machine Learning Approaches for Biological and Physiological Data
（生物学的データと生理学的データのための機械学習法）

氏　名　Relator Raissa Tillada　印

Machine learning is a powerful tool in performing tasks such as classification, pattern recognition, and data mining and analysis. However, to guarantee optimal results, different machine learning techniques have been developed for different types of data for various tasks.

Supervised learning is a type of machine learning where part of the data is used to train the learning algorithm to approximate some mapping that can perform predictions on new data. One of the most common examples of supervised learning is classification. This task aims to identify to which class or category a new data sample belongs to, given a set of observed data whose classes are known. In the case where there are only two classes, usually denoted by +1 and –1, we refer to this as the binary classification task.

Predicting Drug-Protein Interactions: Drug-protein interaction prediction plays an important role in drug design and discovery. Since wet lab procedures are inherently time consuming and expensive due to the large number of candidate compounds and target genes, computational approaches became imperative and have become popular due to their promising results and practicality. The prediction problem setting can be modeled as a binary classification problem where +1 indicates that there is interaction between the drug and target protein, and –1 indicates otherwise. To improve prediction accuracy and precision, we proposed an algorithm employing both support vector machines (SVMs) and an extension of canonical correlation analysis (CCA). We introduced weighted CCA as a means of uncovering underlying relationship between similarity of drug compounds and known interactions with target proteins. By extracting the most significant features using weighted CCA and utilizing them for

training support vector machines, better prediction performance was achieved compared to methods using only SVM and classical CCA.

Enzyme Active Site Search: Prediction of active sites in enzyme proteins is highly essential not only for protein science but also for practical applications such as drug design. Because enzyme reaction mechanisms are based on the local structures of enzyme active sites, a simple measure such as the mean square deviation has been used to compare such local structures in proteins. To enhance the ability of such simple measure, we introduce parameters for the deviation, as well as regularization functions using Bregman divergences to model a new machine learning algorithm that determines the parameters of the square deviation. The proposed algorithms proved to be better than, if not comparable to, existing methods for enzyme active site search. Moreover, the algorithms presented follow a more natural form, adopting the framework of modern machine learning techniques.

Task Classification Using EEG Signals: Classification tasks in brain-computer interface research have presented several applications, biometrics and cognitive training, for instance. However, like in any other discipline, determining suitable representation of data has been challenging, and recent approaches have deviated from the familiar form of one vector for each data sample. With this in mind, we proposed a kernel between vector sets for binary task classification using EEG signals. This is motivated by recent studies where data are approximated by linear subspaces, in particular, methods that were formulated on Grassmann manifolds. The proposed kernel takes a more general approach given that it can also support input data that can be modeled as a vector sequence, and not necessarily requiring it to be a linear subspace. In addition to this, the kernel also directly computes similarity between two vector sequences, whereas known Grassmann kernels such as the Projection kernel and Binet-Cauchy kernel computes similarity between two vector sequences indirectly. We also presented a theoretical relationship between the proposed kernel and the Projection kernel. Empirical results revealed that the proposed kernel achieves promising prediction performance compared to Grassmannian methods.

Conclusion: This thesis focused on applications of binary classification using biological and physiological data. The main challenges we addressed are related to finding suitable features for input data, and kernel functions for unconventional data representation. As proposed methods showed favorable outputs relative to existing

methods, they may have good potential applications in other fields, as well as different problem settings such as multi-class classification.

機械学習はパターン認識, データマイニング, データ解析において強力なツールである. 良好な結果を得るには，様々なタスクにおける異なるタイプのデータそれぞれのために異なる機械学習技術が開発されてきた.

教師あり学習は機械学習の一つのクラスであり，入力から出力を予測するためにデータの一部が訓練アルゴリズムで用いられる．教師あり学習の中で識別は最も主要なタスクである. このタスクは新しいデータがどのクラスに属するか分類するものである. これには，クラスが既知である観測データの集合が使われる．２クラスしかない場合，通常，＋１か－１でラベルされ，２クラス分類タスクと呼んでいる.

タンパク質 - 薬剤相互作用の予測：薬剤 - タンパク質相互作用予測は薬の設計と発見において重要な役割を担う．大量に候補化合物とターゲット遺伝子がある場合，相互作用のウェット実験は時間と労力がかかりすぎる．故に，計算機的アプローチが重要視されてきた．タンパク質 - 薬剤相互作用の予測問題の設定は２クラス分類問題になる．クラスラベル＋１は相互作用があることを示し，クラスラベル－１は相互作用がないことを示す. 予測精度の向上のために，本研究ではサポートベクトルマシン（ＳＶＭ）と正準相関分析（ＣＣＡ）の拡張を組み合わせた方法を提案する．ＣＣＡの拡張として，重みつきＣＣＡを開発した．重みつきＣＣＡによって，化合物間の類似度とターゲット蛋白質との相互作用の関係を抽出する．重みつきＣＣＡを使ってもっとも重要な特徴を抽出して，その特徴をＳＶＭに利用する．実験の結果，従来のＳＶＭのみを用いた方法や，従来のＣＣＡを用いた方法より，よい予測性能を得ることができた.

酵素活性部位予測：酵素タンパク質における活性部位の予測はタンパク質科学において重要であるのみならず，産業応用上にも有用である．酵素反応のメカニズムは酵素活性部位の局所構造によっている．これまでは平均二乗誤差のような簡単な方法を使ってタンパク 質の局所構造どうしを比較してきた．この比較能力を向上させるため，偏差にパラメータを導入する．正則化関数にブレッグマンダイバージェンスを使って二乗誤差のパラメータの値を決める新しい機械学習アルゴリズムを開発した．提案法は既存の方法よりもよい探 索性能を得た．開発した方法の利点は，最近の機械学習技術の枠組みに即した自然な形式に即していることから，その枠組みを拡張させた様々な方法を導入することも可能になった.

ＥＥＧ信号によるタスク識別：脳 - 機械インターフェース（ＢＣＩ）における識別タスクは，認知科学の発展，障害者支援への応用などで注目されている．しかし，ほかのパターン認識の問題と比較して，データに適切な表現を決定することは挑戦的である．それぞれの例題を特徴ベクトルの形式で表すのが古典的な方法であったが，最近のアプローチはカーネルと呼ばれる新しい形式でデータ表現が行われるようになった．提案法は，ＥＥＧ

信号を使った２クラス分類のために，ベクトルの集合どうしのカーネルを使うものである．これは近年データ集合を線形部分空間で近似する研究が動機になっている．とくにグラスマン多様体によって定式化された方法が，ＥＥＧ信号の識別において注目されている．本研究では，ベクトル列が線形部分空間に必ずしも乗っていないとして，新たなカーネルを開発した．このカーネルはベクトル列どうしの類似度を直接計算するもので，グラスマン射影カーネルやグラスマンビネコシカーネルのように間接的にベクトル列どうしの類似度を求める方法とは対照的である．本研究では，提案カーネルとグラスマン射影カーネルの間に理論的関係を見出し，なぜ提案法のほうがよりよい予測性能を示すのか理論的に明らかにした．

結論：本論文では，生物学的データと生理学的データにおける２クラス分類の応用に注目した．本研究の主要な挑戦は入力データの適した特徴を見つけることと，カーネルによる従来にはなかったデータ表現を開発したことである．提案法は既存の方法と比べて良好な予測を行うことができるようになったことを示した．提案法は本論文で用いたデータに特化するものではなく，様々な分野の様々な分類タスクに応用可能である．