学 位 論 文 の 要 旨

Kernel Matrix Completion
カーネル行列補完

氏 名 Rachelle Alvarez Rivero 印

We are now in the era where technology has made it possible to constantly obtain information, from people's Google search histories, to people's purchasing habits through point cards, among others. However, even with this much technological advancement, there are still instances when all that can be obtained are incomplete and noisy data. Such situation is common in biological data, where data acquisition requires extensive works of experts, and expensive tools are needed. Apparently, redoing biological experiments to fill-in the missing information or to remove the noise is inexpedient, so scientists and researchers in the field are looking for alternative approaches to resolve these issues, or to at least find a way to optimally utilize the available data for analysis. In this thesis, a novel method that can optimally exploit the data at hand for further analysis is presented. Since the integrity of data analyses and conclusions lies in the correctness of the data, it is very important that the inference accuracy is kept high and the underlying patterns in the data are maintained. These important aspects of the data are considered in the formulation of the methods introduced in this thesis.

The machine learning techniques developed in this thesis rely on the use of real-valued square matrices, called kernel matrices. Data of any type can be transformed into a kernel matrix, through the use of some functions with certain properties, called kernels. Kernel matrices are especially useful in the analysis of proteins, since proteins are represented in many ways. For instance, a protein can be represented as a string of letters from the English alphabet, corresponding to its amino-acid sequence. So much can also be inferred from a protein's cubic structure. A protein can also be described through its interaction with other proteins, and through its gene expression, which are

best represented as graphs and as vectors, respectively. Each of these different representations of proteins can now be expressed as a kernel matrix, through the use of kernels. Most machine learning algorithms for classification tasks, such as Support Vector Machines (SVMs), take in kernel matrices as inputs and hence, when performing protein classification tasks, only the proteins' kernel representations are essential, and not the raw data.

In the protein example above, the different protein representations serve as the "partial views" of the data. Intuitively, when these partial views are combined, one can gain a complete view or one big picture of the object at hand. Indeed, existing literature in protein function classification, and in some other classification studies, have shown that by combining the different views of the data, classification accuracy and performance is improved. This is where the use of kernels becomes increasingly attractive. As kernel matrices, the partial views of the data can be easily manipulated and combined, paving the way for data fusion and multiview learning.

Going back to the incomplete-data problem, a data source (e.g. gene expression data) with some lacking information produces a kernel matrix with entries in some rows and columns skipped. These rows and columns with skipped entries correspond to the missing information in the data source where the kernel matrix is derived from. Kernel matrices with skipped or "missed" entries are called incomplete. Incomplete kernel matrices cannot be fed to existing machine learning algorithms, and data analysis cannot be performed. This is a drawback of kernel methods. Early solutions for this problem include either imputing zeros in the skipped entries (zero-imputation) or imputing the mean of the remaining matrix entries (mean-imputation). Subsequent studies utilize other kernel matrices derived from supplementary data sources to infer the missing entries of a kernel matrix. However, the framework of these studies assumes that only one of the kernel matrices has missing entries, or that at least one kernel matrix must be complete. By contrast, the framework in this thesis assumes that all of the kernel matrices may have missing entries, and the kernel matrices can be completed mutually. The proposed model in this thesis is then called "Full-Covariance Mutual Kernel Matrix Completion (FC-MKMC)" method.

The proposed model FC-MKMC is based on information geometry, where the empirical kernel matrices are associated to the covariance of zero-mean Gaussians. Then, a model parameter kernel matrix is introduced that serves as the complete view of the data. To

fit this model parameter matrix to the empirical kernel matrices, FC-MKMC employs the Kullback-Leibler (KL) divergence for the distance between the model matrix and the empirical kernel matrices. The objective function is then taken as the sum of the KL-divergences and is minimized with respect to the model matrix and the missing elements.

Minimization of the objective function to infer the missing elements of the kernel matrices involves two steps: the imputation step, where the initialized model matrix is fixed, and the objective function is minimized with respect to the set of missing elements; and the model update step, where the inferred elements are fixed, and the objective function is minimized with respect to the model parameter kernel matrix. By repeating these two steps in order, the estimates for the missing elements are improved.

The imputation and model update steps correspond to the expectation step and maximization step, respectively, of Expectation-Maximization (EM) algorithm. With the EM algorithm, the likelihood of the model parameter is maximized. This statistical framework of FC-MKMC enables it to handle the missing-data problem in a principled way.

Since FC-MKMC is full-covariance, its number of degrees of freedom is large, and is at risk of overfitting. Hence, two variants of FC-MKMC are introduced to provide model flexibility control. These parametric models are the PCA-MKMC and FA-MKMC. PCA-MKMC and FA-MKMC are based from notions of probabilistic PCA and factor analysis models, respectively. In these models, the covariances are expressed in a restricted form, in which the flexibility can be controlled through the choice of the subspace dimension.

As revealed by empirical results on experiments in real-world data such as yeast proteins, the proposed models have improved kernel estimation accuracies over traditional completion methods. The proposed models have also improved the classification performance, as verified from tests on binary classification tasks such as membrane protein prediction, and yeast protein functional classification prediction. In the latter experiment, it was also shown that the proposed parametric models have better classification performance than the full-covariance variant.

In conclusion, this thesis has contributed to learning from incomplete data by fully utilizing the available data. The practicality of the proposed models, as well as favorable

results from experiments, are good indicators that the models presented in this thesis may also have reliable applications in other fields aside from yeast protein classification tasks.

いま人類は絶え間なく情報を手に入れられる時代にいる．人は，グーグル検索履歴や，ポイントカードなどから得られる購買傾向まで手に入れられる．しかしながら，このような目覚ましい技術発展にあっても，それぞれの情報は不完全のままであることがある．例えば，生物学的データの場合，データ取得には専門家による膨大な労力を要し，高価な実験資源が必要となる，もしくは，場合によってはデータ取得が不可能な時もある．欠損情報を埋めるためには再度実験室の測定を行うのは現実的ではないので，別の解決方法を探さなくてはならない．もしくは，不完全なデータのまま解析を行う方法を見つける必要がある．本研究では，利用できる不完全なデータのみから解析を可能にする新しい方法を開発した．データ統合とそれに伴う結論はデータの正しさに由来するので，高い精度で欠損値を推定することは重要である．これらの重要な観点を考慮に入れて，欠損値の補完を行うアルゴリズムを定式化した．

本論文で提示する機械学習技術は，カーネル行列と呼ばれる，実数値正方行列に依存する．本技術は，すべてのタイプのデータをカーネル行列に変換することを前提とする．カーネル行列に変換するときはカーネル関数を使うことができ，カーネル関数がデータタイプの特性を抽出するように設計する．タンパク質を解析するときは，カーネル行列による表現は特に使い勝手が良い．なぜなら，タンパク質は多くの表現方法があるからである．たとえば，タンパク質のアミノ酸配列は 20 種類のアルファベットの系列で表される．タンパク質の中には，立体構造が決定されているものもある．タンパク質間の相互作用の有無もタンパク質の性質を示す重要な情報となる．異なる条件下での発現情報を表す遺伝子発現データはマイクロアレイ技術で取得できる．それらは文字列，グラフ，ベクトルなど様々な形式で表現される．これらを統一的な方法で表す手段がカーネル行列になる．サポートベクトルマシンなど多くの機械学習アルゴリズムはカーネル行列を入力データとして受け取る．カーネル行列は，解析精度を決める重要な要素となっている．

上記のタンパク質の例は，異なるタンパク質の表現それぞれがデータをある視点から映し出したものとみることができる．様々な角度から見て得られた対象の情報を統合することによって，対象全体の実態をつかむことができる．実際に，タンパク質機能予測の文献においても，画像分類などそれ以外の応用における研究でも，異なる視点のデータを統合することで予測精度を向上できることが報告されている．カーネル行列によって数学的に整備された枠組みにおいてデータを統合することが容易になり，この性質からカーネル行列を介したデータ統合や多視点学習の技術が発展してきた．

不完全データ問題に戻ると，遺伝子発現データのような情報源のうちいくつかの対象に対する情報が欠けているとき，一部の行や列が欠損したカーネル行列を生むことになる．欠損した要素を伴う行や列はオリジナルの情報源において欠損が生じた対象に対応する．欠損要素を伴うカーネル行列は不完全であるという．不完全カーネル行列は既存の機械学習技術には直接使うことができないため，そのままでは解析できない．これはカーネル法の短所である．この問題に対する簡単な解として，欠損要素にゼロを代入する方法が考えられる．もしくは，平均値を代入する方法も考えることができる．これに対して，ほかの情報源から導出されるほかのカーネル行列を使って不完全カーネル行列の欠損値を推定する方法が開発された．しかし，これらの研究は，欠損値を含むカーネル

行列は一つだけであり，そのほかのカーネルは完全であることを仮定している．これに対して，本論文では，この仮定を緩和して，すべてのカーネル行列が不完全であってもよいとしている．そして，不完全カーネル行列を相互に利用して補完する方法を開発した．本論文で提案するモデルを「フル共分散相互カーネル行列補完（FC-MKMC)」法と呼ぶことにする．

提案モデル FC-MKMC は情報幾何学に基づいており，平均値ゼロの正規分布の共分散行列に経験カーネル行列を関連付けることで情報幾何学的アプローチをとれるようにしたものである．モデルカーネル行列は，入力した不完全なカーネル行列を統合したものを表している．FC-MKMC はカルバックライブラダイバージェンスと呼ばれる確率分布間がどれだけ離れているか表す指標を，モデル行列と経験カーネル行列との距離を表すのに用いた．目的関数は各々の経験カーネル行列とモデル行列との距離の和とし，これを最小化するモデル行列と欠損値を見つけるアルゴリズムを開発した．

カーネル行列の欠損値を推定するための目的関数の最小化は 2 ステップを伴う : 補完ステップではモデル行列を固定し，目的関数が最小化される欠損要素を見つける．モデル更新ステップでは，欠損要素の推定値を固定して，目的関数が最小化されるモデル行列を見つける．この 2 ステップを繰り返すことで，欠損要素の推定を実現する．

この 2 ステップは，統計学で用いられてきた期待値最大化アルゴリズムの期待値ステップと最大化ステップに対応している．期待値最大化アルゴリズムとは，確率モデルのパラメータの尤度を最大化するための方法である．この数学的な等価性は，FC-MKMC 以外の選択肢を考えるためのヒントとなった．

FC-MKMC はカーネル行列を正規分布の共分散行列に関連付けたが，そのモデルの自由度はあまりに大きかった．自由度が高いゆえに，過整合の危険のあるモデルとなっていた．本研究では，モデルの自由度を調整できるように，２つのバリアント PCA-MKMC および FA-MKMC を考案した．それぞれ確率的主成分分析および因子分析モデルに立脚したものである．これらのモデルは共分散行列をある制限した形式で表すことで，自由度を使用状況に応じて設定できるようにしたものとなっている．

酵母タンパク質の実データを用いた実験によって，提案モデルは従来の方法より精度よくカーネル行列を補完し，かつ，2 クラス分類問題での汎化能力も高いことが実証された．自由度をガットマンカイザー法などで設定した PCA-MKMC や FA-MKMC は，FC-MKMC より性能が向上した．

結論として，本論文では，不完全データからの学習を可能にする新しい方法論を示した．実データを用いた検証によって，提案モデルの実用性を示した．酵母タンパク質分類問題に限らず，多くの応用において信頼できる方法論になっていると期待される．