審 査 の 結 果 の 要 旨

論文提出者氏名　　襲田　勉

創薬における新薬開発工程において、臨床試験に至る前の基礎研究で労力がかかるステップに、活性化合物を発掘して化合物最適化を行う化合物スクリーニングがある。化合物スクリーニングを効率的に実現するため、コンピュータで統計処理を行って効率的に探索するというバーチャルスクリーニングへの期待が高まっているが、化合物の探索空間には大多数の非活性化合物データとともにごく少数の活性化合物データが散在することをはじめとして多様な特徴があり、化合物スクリーニングに適応した新たな方法が必要となっている。

本論文では、化合物の探索における課題を洗い出し、解決手段を与えて、活性化合物をより効果的に探索する方法の確立を目指す。具体的には、その洗い出した課題に着眼して、多様な化合物の構造を適切にモデリングして化合物空間で適用範囲の広い記述子を提案するとともに、能動学習の手法を対象の表現空間に適用するにあたって着目した課題を解決する学習法を与えることによって、広大な探索空間の中に疎に存在する所望の活性化合物を効率的に探索する方法を示している。

本論文は、第1章の導入、第2章の問題定義に続き、第3章で能動学習、第4章で既存のバーチャルスクリーニングに関してまとめ、能動学習法に関する成果を第5章から第7章で、それら能動学習の探索可能な範囲を広めるためのモデリングと学習への展開に関する成果を第8,9章に記している。

第5章では、薬剤候補化合物である正例が非活性化合物である負例の中に疎に散在する探索空間を対象として、正例と負例を識別する仮説を早期に構築する能動学習法を提案している。この能動学習法は、稀な正例データを効率的に探索するため、エントロピーと学習データ・予測データの密度比に基づいてクエリ点を決めるノイズ耐性のある学習法となっており、その有効性を実験データセットで検証している。

化合物の探索空間に関して洗い出した課題の1つに、この探索空間が化合物データ数の多さによって爆発的に広がるだけでなく、その構造を変換する記述子数にも起因して特徴量の数の点でも爆発する点があげられる。また、探索初期段階において、ラベル付けされたデータの数に比して特徴量の数はかなり多い。そのような状況において、機械学習・統計解析の観点からは過学習を避けるための特徴量の絞り込みが必要である。能動学習においてラベル付けされたデータが増大するにつれ、

特徴量も増大する状況では、１つのデータを選び出すことにより分離境界を探索しつつ特徴量も探索する方が、分類精度を向上させると考えられる。第6章において、これらの議論をもとに、絞り込まれた特徴量において算出する不確かさだけではなく、残りの個々の特徴量が有する多様さを考慮した能動学習を提案し、それを創薬データセットと汎用的データセットに適用することにより有効性を明らかにしている。

負例が探索空間のほとんどを占めるインバランスデータセットにおいて、主要なクラスの予測における正解率を上げることにより、データ全体の予測の正解率は高くなるが、マイナークラスの予測における正解率は低下してしまう。化合物スクリーニングにおいて、マイナークラスは活性化合物のクラスであるものの、活性化合物は非活性化合物と比較してはるかに高い価値を持つ。そのため、全体の正解率を高くする目的は、必ずしも、活性化合物を探索する目的とは合致しない。さらに、探索初期段階において探索すべきマイナークラスのデータが、ほとんど存在しないこともある。そこで第7章においては、活性に寄与する構造を統計的に解析することにより、種類の異なる活性化合物から探索すべき種類の活性化合物を推測する方法とともに、学習アルゴリズムが全体の正解率を高くするような仮説を作成したとしても、選んだデータに対するラベル付けを行うときに判明する誤差に基づいて、その仮説の修正を行う能動学習を提案し、その有効性を創薬化合物データセットで確かめている。

第8,9章では、まず能動学習による探索可能範囲を広げることを目標に、統計的解析を行う前提となる記述子がまだ開発されていない種類の化合物に対し、その構造を表現する記述子を構築する課題に取り組んでいる。成果として、アミノ酸の構造を記述する基本構造と、2つの基本構造の間のアミノ酸のつながりを表す中間記述子との組み合わせにより、多様なペプチドを記述する方式を提案している。また、標的タンパクに関する情報が少ない状況において、バーチャルスクリーニングが適用可能な範囲を広げる成果も与えている。これは、活性ペプチドから代謝上安定した構造をもつ低分子化合物に変換する方法が標的タンパク質の立体構造に基づくことに着目して、ペプチドのSARルールを統計的に解析して、低分子化合物を選ぶための評価関数とするリガンドベースのアプローチを提案したものである。

以上をまとめるに、本論文は化合物の新たな表現方式と、それら表現データが構成する空間に適した能動学習で効果的に探索する方法を与えており、コンピュータ科学を広めかつ深めることに貢献する成果を与えている。

審査委員会は、平成２４年８月２１日に論文提出者に対し、学位請求論文の内容及び専攻分野に関する学識について口頭による試験を行った結果、本人は博士(情報理工学)の学位を受けるに十分な学識と研究を指導する能力を有するものと認め、合格と判定した。