Knowledge Discovery in Biological Data

Streszczenie rozprawy doktorskiej

Złożonej przez

Indrajit Saha

ubiegającego się o stopień doktora filozofii w
Instytucie Podstaw Informatyki,
Polskiej Akademii Nauk

Promotor:

Dr hab. Dariusz Plewczyński
Interdyscyplinarne Centrum Modelowania
Matematycznego i Komputerowego (ICM),
Uniwersytet Warszawski

Sierpień, 2015

Streszczenie

W ciągu ostatniej dekady, znaczny postęp w dziedzinie biologii molekularnej i technik sekwencjonowania następnej generacji doprowadził do gwałtownego wzrostu ilości informacji biologicznej w publicznych i prywatnych repozytoriach wiedzy. Jednym z trudniejszych problemów, które stoją przed środowiskiem naukowym, jest wydobycie użytecznej informacji z różnych źródeł masowych danych, proces odkrywania wiedzy i eksploracji statystycznej i wizualnej danych. Bioinformatyka jest nauką dokonującą odkryć przy wykorzystaniu metod obliczeniowych. Dodatkowo sam proces integracji ogromnych ilości heterogenicznych danych, przechowywanych w rozproszonych repozytoriach na całym świecie, często obejmuje zadania grupowania obiektów, ich klasyfikacji, przewidywania i identyfikacji sygnatur.

Analiza skupień jest popularna metoda uczenia bez nadzoru, która dzieli dostępną przestrzeń cech na regiony w oparciu o miarę podobieństwa (lub zróżnicowania). Otrzymane w ten sposób klastry danych mogą być zarówno binarne jak i rozmyte. Tradycyjne metody analizy skupień, takie jak metoda k-średnich, czy metoda c-średnich, mogą tworzyć modele nie optymalne z uwagi na obeność lokalnych minimów w problemie klasyfikacyjnym. Dodatkowo metody te optymalizują zwykle pojedyńczą miarę dobroci tworzenia skupień, co nie zawsze wyczerpuje złożoność oczekiwań użytkownika metody, szczególnie w przypadku nauk biologicznych. Dlatego też użyteczne może być równoległe użycie wielu miar dobroci rozwiązania optymalizacyjnego. W tym celu stworzono wielokryterialną metodę analizy skupień przy użyciu np. zasady Pareto. Jednak w kolejnym kroku zaobserwowano, że użycie pojedynczej metody analizy skupień również przynosi pewne błędy wynikające ze słabości metod teoretycznych. W przypadku tego samego zbioru danych, różne algorytmy, lub nawet ten sam algorytm ale użyty wielokrotnie z innymi wartościami parametrów, otrzymują różne wyniki końcowe. Dlatego też zaproponowano użycie zespołów algorytmów analizy skupień, które wydają się osiągać lepsze wyniki w porównaniu do pojedycznych metod. W kolejnym kroku zadać można pytanie czy wyniki metod uczenia bez nadzoru mogą być użyte do trenowania metod uczenia nadzorowanego. W tym przypadku takich metod uczenia z nauczycielem można zadać podobne pytania, związane z tworzeniem zespołów algorytmów uczących się i oceną ich efektywności.

W mojej rozprawie doktorskiej poruszam oba zagadnienia badawcze, tj. budowanie zespołów metod uczących się bez nadzoru i nadzorowanych, łącznie obu podejść do bardziej zaawansowanej i automatycznej statystycznej analizy danych biologicznych. Zadanie stojące przez bioinformatyką, zostało przeze mnie realizowane przy użyciu statycznych metod uczenia oraz dużych biologicznych baz danych. Opracowałem metodę ustalającą automatycznie optymalną liczbę skupień w procedurze klastrowania analizując dane mikromacierzowe. Wielokryterialna analiza skupień została przeze mnie zastosowana do analizy obrazów mózgu z rezonansu magnetycznego. Następnie analizowałem bazę cech fizyko-chemicznych aminokwasow (AAindex), przewidywałem miejsca modyfikacji posttranslacyjnej białek, jak również dane o wiązaniu się krótkich peptydów do białek HLA (Human Leukocyte Antigen Class-II protein), czy tworzeniu się kompleksów białko-białko.

Zastosowania praktyczne zostały poparte analizami teoretycznymi, wnoszącymi większe zrozumienie problemów nauk obliczeniowych, informatyki teoretycznej, jak również lepsze poznanie natury zjawisk biologicznych, obserwowanych narzędziami współczesnej biologii molekularnej i genetyki.

Abstract

Over the past few decades, major advances in the fields of molecular biology and genomic technology have opened the window for the scientific community in order to generate massive amount of biological data. Such data are stored in various repositories in distributed manner all around the globe. Recently, mining such repositories in order to extract useful information/knowledge has become an attractive and challenging task. In this regard, computational methods are used to make biological discoveries. Hence, it is either called Computational Biology or Bioinformatics. Generally, the objective of the computational biological researcher is to discover new biological insights and create a global perspective from which unifying principles in biology can be derived. Therefore, for this huge amount of biological data, stored in repositories distributed, often needs mining tasks like clustering, classification, prediction and frequent pattern identification.

Clustering is a popular unsupervised pattern classification technique which partitions the input space into K regions based on some similarity/dissimilarity metric where the value of K may or may not be known a priori. Clustering can either be hard or fuzzy. Conventional partitional clustering algorithms, such as K-means and Fuzzy C-means often get stuck at local optima. Moreover, they optimize a single cluster validity index and thus may not be able to capture different characteristics of data sets. Hence, it is more useful to simultaneously optimize several cluster quality measures that can capture the different data characteristics. In order to achieve this, the problem of clustering a dataset has been posed as one of multiobjective (MO) optimization in literature. However, these clustering techniques have its own strengths and weaknesses. For a particular dataset, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. Recently, cluster ensembles have emerged as an effective solution that is able to overcome these limitations, and improve the robustness as well as the quality of clustering results. Therefore, it is quite significant, if ensemble of different clustering solutions is used to create a final clustering result. On the other hand, in supervised classification, a set of patterns having known class information are used as training samples by which the classifier learns the decision rule with some suitable algorithm. The learned classifier then categorizes the unknown input pattern to one of the classes with the help of the learned decision rule. Similarly, like ensemble based clustering, ensemble of classifiers has also drawn the attention of researchers nowadays. Hence, these facts motivated us to develop various unsupervised and supervised techniques in this dissertation.

This dissertation deals with the development of new clustering and classification techniques for various biological problems. For the clustering purpose, improved differential evolution based automatic fuzzy clustering technique is developed to find the number of clusters automatically from gene expression datasets. Subsequently, the multiobjective clustering technique is developed for Magnetic Resonance (MR) brain image segmentation. Thereafter, consensus clustering approach is developed with the use of recently proposed differential evolution based clustering technique and other state-of-the-art methods to analyse the AAindex database which contains various physicochemical and biochemical properties of amino acids. For developing the classification techniques, a new ensemble based classifier is developed for the binding prediction of peptide to the Human Leukocyte Antigen (HLA) Class-II protein. Subsequently, another ensemble based predictor is developed for automatic prediction of protein post translational modification sites. Thereafter, protein-protein interaction prediction is evaluated using state-of-the-art classifiers for Yeast and Human.