

PROTEİN HOMOLOJİ TESPİTİNDE BİR ÜST SINIFLANDIRMA YAKLAŞIMI

Aydın Can POLATKAN

ÖZET

Hesaplamalı biyoloji alanında sınıflandırma problemleri için makine öğrenme teknikleri sıkça ve geniş şekilde kullanılmaktadır. Bu teknikler, girdi olarak sabit uzunluklu nitelik vektörleri istemektedir. Bilindiği üzere proteinler farklı uzunluklara sahip olduklarından dolayı, tüm protein dizilimlerini sabit sayıda nitelik ile göstermek gerekir.

Bu amaçla geliştirilen etkili yöntemlerden biri protein dizilimlerinin n-peptit birleşimleridir. Yöntem n uzunluktaki her alt dizginin dizilim içerisindeki görülme yüzdesini ifade eder. Alan karmaşıklığını azaltmak amacıyla, n'nin artan değerleri için, kullanılan aminoasit alfabesi, sonuç vektörün günümüz bellek kaynaklarıyla uyumlu olmasını sağlayacak şekilde düzenli olarak küçültülmüştür.

Kullanılan bu çözümde birleşime ait bütün özellik girdileri sadece bir sınıflandırıcıya toplu olarak verilmekteydi. Bu tezde, bu özellik girdileri n-peptit birleşimlere ve küçültülen amino asit alfabelerine göre farklı gruplara ayrılıp, farklı sınıflandırıcılara verilmiştir böylece soyutlanarak daraltılan arama uzayında, gezinen birden fazla tekniğe, bir üst sınıflandırma yaklaşımı denenmiştir. Amaç doğru şekilde yakınsanan ve bizi birbirinden farklı çözüm bölgelerine ulaştıran tekniklere üstsel sınıflandırma yaklaşımı ile daha iyi sonuçlar alabilmektir. Bu yaklaşımda farklı sınıflandırıcıların çıktı değerlerini değerlendirmek üzere ortalama alma, ağırlıklı ortalama alma ve öğrenme kümesinde en başarılı olanı seçme gibi değişik durumlar karşılaştırılmıştır.

Her bir yöntem hesaplamalı biyolojinin önemli ve güncel problemlerinden biri olan uzak homoloji tespiti üzerinde test edilmiş ve sonuçlar karşılaştırmalı olarak sunulmuştur.

Sonuçlara bakıldığında eğitim kümesinde en başarılı olan sınıflandırıcının sonucunun doğru kabul edildiği durumun diğerlerine göre daha etkili olduğu gözlenmiştir. Sonuçlar arasındaki istatistiksel anlamlılığı dikkatlice incelemek için tüm yöntemler arasında öğrenci T-testleri yapılmış ve testlerin sonuçları yorumlanmıştır. Denenen bütün üst sınıflandırma yaklaşımları yalnız bir sınıflandırıcı kullanılan duruma göre daha etkili bellek kullanımına sahiptir. Destek vektör makineleriyle test edilen bu üst sınıflandırma yaklaşımının sadece uzak homoloji tespitinde değil diğer sınıflandırma problemlerinde de başarılı olacağı düşünülmektedir.

Anahtar Sözcükler: Protein Homoloji Tespiti, N-peptit Birleşimler, Destek Vektör Makineleri, Sınıflandırma, Üst Sınıflandırma.

Danışman: Hayri SEVER, Prof. Dr., Çankaya Üniversitesi, Bilgisayar Mühendisliği Bölümü

A DATA FUSION APPROACH IN PROTEIN HOMOLOGY DETECTION

Aydın Can POLATKAN

ABSTRACT

Machine learning techniques are frequently and extensively used for classifying problems in the field of computational biology. These techniques require constant length feature vectors as inputs. As far as it is known that proteins are in different lengths, therefore all proteins are needed to be represented with a constant number of features.

One of the effective methods developed for this goal is n-peptide combinations of the protein strings. These methods are represented with the availability percentage of each of the n-length substrings inside the sequence. To reduce the space complexity, for increasing values of n, amino acid alphabet is reduced regularly for the resulting feature vectors to conform available memory resources today.

In this solution, all feature inputs were given to a single classifier. In this thesis, these feature inputs are classified into specific significant groups, according to the n-peptide compositions and reduced amino alphabets. These groups are given to several different classifiers to achieve a data fusion approach with a few techniques that are wandering in the narrowed search space by abstraction. Aim is to have better results with techniques that are converging in exact and leading to different regions of a solution. In that approach, to evaluate the output values of different classifiers, various cases like averaging, weighted averaging and choosing the most successful one in the training set are compared.

Each of these methods was tested on remote homology detection problem which is one of the major and actual problems of computational biology and results are presented relatively.

As the results are considered, the case in which the output of the most successful training set is granted, observed as the more accurate one. To explore the statistical significance of differences between results, paired samples T-tests were carried out between all methods. Furthermore, all data fusion approaches tested, through out the thesis has more efficient memory usage according to the single classifier case. The data fusion approach which has been tested with support vector machines is also thought to be efficient for not only protein homology detection problems but also other problems of classification.

Keywords: Protein Homology Detection, N-peptide Compositions, Support Vector Machines, Classification, Data Fusion.

Supervisor: Hayri SEVER, Prof. Dr., Çankaya University, Department of Computer Engineering