Tıbbi Veri Setleri İçin Akıllı Yöntem Tabanlı Öznitelik Seçme Algoritmaları

Oğuzhan CEYLAN [a], Çağıl ACAR ŞAYLAN [b], Işıl YENİDOĞAN [b], Hasan DAĞ [b]

[a] *Hesaplamalı Bilim ve Mühendislik Programı, Bilişim Enstitüsü, İstanbul Teknik Üniversitesi, İstanbul*

[b] *Enformasyon Teknolojileri Bölümü, Mühendislik ve Doğa Bilimleri Fakültesi, Kadir Has Üniversitesi, İstanbul*

Intelligent Method Based Feature Selection Algorithms For Medical Data Sets

Abstract: *In this paper we apply data mining techniques to medical data. We use a dataset consisting of the data from the outpatients of the University of Istanbul - Cerrahpasa Medical Faculty, which were treated throughout the period of 21 months between the dates March 2006 - December 2007. By using open-source data mining software WEKA, we run classification, clustering and decision tree algorithms and obtained decision rules. These decision rules were analyzed with the help of medical specialists to determine which features caused complications in the coronary arteries. Also a comparison with feature selection algorithms was performed to see if the same features could be found.*

While the number of features increase, search space of feature selection problem gets larger. If one wants to select 5 features from a set of 100 features, the selection should be performed from 1.8 10^9 different scenarios. Hence intelligent method based optimization methods need be brought to the agenda.

This work aims to develop an optimization based new feature selection algorithm instead of the cases where previously developed feature selection algorithms were insufficient. By developing this new feature selection algorithm we aim it to be used for testing the selected features.

In our previous work, we compared feature selection algorithms: Info Gain, Gain Ratio, and Correlation Based Feature Selection (CFS), and we obtained nearly the same results for both Info Gain and Gain Ratio algorithms.

This work aims to obtain better and faster results by using intelligent optimization methods. We solve optimization problem by using Harmony Search algorithm, by taking cross validation results as objective function and using Naive Bayesian classification algorithm. Our program gets the patient data prepared in Weka and uses it as input to Matlab, a commercial package developed for performing calculations using matrix operations. Optimization, classification and cross validation modules were programmed in Matlab.

High performance of the decision support system aiming to provide doctors time and source savings is important hence it is the first order concern of the human health. Because of this reason, if this newly developed feature selection algorithm passes the comparison tests it will help us to select features that increase the performance ratio. This result will positively affect public health and people working in health sector resulting in the efficient usage of the sources of the state hospitals and helping the country economy.

Key Words: Data Mining, Feature Selection Algorithms, Optimization, and Intelligent Methods

Özet: *Bu çalışmada veri madenciliğinin tıp alanında kullanımı incelenmiştir. Uygulama çalışması için İstanbul Üniversitesi Cerrahpaşa Tıp Fakültesi'nde ayakta tedavi gören hastalar arasında, Mart 2006 – Aralık 2007 tarihleri arasında 21 aylık bir sürede tedavi görmüş hastalara ait veriler bir araya getirilerek bir veri kümesi oluşturulmuştur. Bu veri kümesi üzerinde WEKA yazılımı kullanılarak sınıflama, kümeleme ve karar ağacı algoritmaları çalıştırılmış, elde edilen karar kuralları uzman desteğiyle incelenerek koroner arterlerde kalsifikasyon bulunmasında etkili olan faktörlerin neler olduğu belirlenmiş ve öznitelik seçme algoritmalarıyla aynı faktörlere ulaşılıp ulaşılamadığı belirlenmiştir.*

Öznitelik seçme probleminin çözüm uzayı öznitelik sayısı arttıkça çok fazla büyümektedir. Sözgelimi 100 öznitelikli bir veri grubundan 5 veri seçmek istiyorsak yaklaşık olarak $1{,}8x10^9$ durum arasından seçim yapılması gerekmektedir. Dolayısıyla akıllı yöntem tabanlı eniyileme yöntemleri gündeme gelmektedir.

Bu çalışmanın amacı veri madenciliğinde yaygın olarak kullanılan öznitelik seçme algoritmalarının yetersiz

kaldığı durumlarda veya seçilen özniteliklerin uygunluğunun test edilmesi gereken durumlarda kullanılması için tıbbi veri üzerinde daha önce denenmemiş bu yeni yaklaşımla en iyileme yöntem tabanlı yeni bir öznitelik seçme algoritması geliştirmektir.
Daha önceki çalışmamızda Info Gain (bilgi kazancı), Gain Ratio (kazanç oranı) ve CFS (korelasyona dayalı) öznitelik seçme algoritmaları karşılaştırılmış ve bilgi kazanç, kazanç oranı algoritmalarının hemen hemen aynı sonuçları verdikleri gözlemlenmiştir. Burada amaç eniyileme yöntemleri kullanarak, bir önceki çalışmada denenen yöntemlerden daha iyi ve hızlı sonuçlar elde edebilmektir.
Naive Bayesian sınıflandırma algoritması kullanılarak, çapraz geçerlilik ölçütü (cross validation) amaç fonksiyonu olarak alınan armoni araması (harmony search) yöntemiyle eniyileme problemi çözülmüştür. Yazılan program Weka da hazırlanmış, hasta verilerini alınıp Matlab girdi veri olarak kullanılmıştır. Eniyileme, sınıflandırma ve çapraz geçerlilik ölçütü programları Matlab yazılımı kullanılarak yazılmıştır.
Hekimlere zaman ve kaynak tasarrufu sağlamak amacıyla oluşturulacak Karar Destek Mekanizmasının, yüksek başarı oranı insan sağlığını birinci derecede ilgilendirdiği için önem taşımaktadır. Bu sebepten projenin çıktısı olarak amaçlanan yeni öznitelik seçme algoritması mukayese testlerini geçmesi durumunda başarım yüzdesini arttıracak öznitelikleri seçmemizi sağlayacaktır. Bu sonuç doğrudan kamu sağlığını ve sağlık sektöründe çalışanları olumlu etkileyecek ve devlet hastanelerindeki kaynaklarını verimli kullanılmasını sağlayarak ülke ekonomisine de katkı sağlayacaktır.

Anahtar Kelimeler: Veri Madenciliği, Eniyileme, Akıllı Yöntemler, Öznitelik Seçme Algoritmaları

Sorumlu Yazarın Adresi
Çağıl ACAR ŞAYLAN , cagil.acar@khas.edu.tr