GENETİK ALGORİTMA VE SÜREKLİ ARTAN FONKSİYONLARLA ZAMAN SERİLERİNİN SINIFLANDIRILMASI

ÖZET

Zaman serileri uygulamalı bilimlerde sıkça karşılaşılan veri tiplerinden birisidir. Periyodik aralıklarla yapılan ölçmelerin kaydedilmesi ile elde edilen zaman serilerine, finans piyasalarında, moleküler biyolojide, istatistiksel süreç yönetiminde, astrofizikte veya özetle zamansal verilerin sözkonusu olduğu bütün alanlarda rastlanır.

Zaman serileri üzerinde uygulanan veri madenciliği teknikleri arasında sınıflandırma en yaygın olanıdır. Zaman serilerinin sınıflandırılması, farklı sınıflara karşılık gelen zaman serilerinin birbirinden ayırd edilmesi olarak nitelendirilebilir. Örnek olarak önünden gezegen geçen yıldızların parlaklığını ölçerek elde edilmiş zaman serileri ile uydusu olmayan yıldızlardan elde edilen zaman serileri arasında belirgin farklar bulunmaktadır. Yüzbinlerce yıldız arasında yörüngesinde gezegen bulunanları tespit edebilmek için sınıflandırma yöntemleri kullanılır.

En yakın komşu veya karar destek makinaları gibi uzaklık hesabına dayanan sınıflandırma yöntemleri herhangi bir değişikliğe gerek olmaksızın zaman serilerine uygulanabilmektedir. Ancak zaman serileri arasında, ölçme tekniği veya doğal nedenlerle oluşan değişimler sınıflandırma performansına doğrudan etki eder. Yukarıda verdiğimiz astrofizik uygulamasından bir örnek vermek gerekirse; birer gezegene sahip iki yıldız için elde edilecek zaman serilerinin genel örüntüsü benzer olmasına karşın, gezegenlerin yörünge hızları büyük ihtimalle farklı olacağı için örüntülerin zaman eksenlerinde kaymalar göze çarpacaktır. Örüntüleri birbirine denk getirebilmek için zaman serilerinden biri veya her ikisinin zaman eksenlerinde büzme ve/veya germe uygulamak gerekecektir. Hizalama olarak da adlandırılan bu işlem için uygulamaların çoğunda doğrusal olmayan büzme/germe'ler gerekmektedir. Hizalama yapılmadan yapılacak sınıflandırmalar çok az uygulama dışında tatmin edici sonuçlar üretmez.

Zaman serilerinin sınıflandırılma başarısını arttırmak için hizalama yapma gerekliliği 1960'ların sonlarında başlayan ses tanıma uygulamalarında yoğun olarak hissedilmiş ve halen kullanılan Dynamic Time Warping (DTW) ve türevi hizalama yöntemleri önerilmiştir. Karşılaştırılan iki zaman serisine ait mesafe matrisinde, zaman serilerinin başlangıcına denk gelen köşeden, bitişe karşılık gelen köşeye doğru çizilebilen alternatif yollar arasında, en az maliyetli yolu Dinamik Programlama (DP) ile belirleme esasında dayanan DTW, çok farklı disiplenlere uygulanarak genel kabul görmüştür.

Bir mesafe matrisi üzerinden çalıştığı için hesaplama karmaşıklığı karesel olan DTW yöntemini iyileştirmek için birçok yöntem önerilmiştir. Bunlardan en yaygın olarak kullanılan Constrained Dynamic Time Warping (CDTW) matrisin arama uzayını ana

köşegenden geçen bir band ile sınırlamaktadır. Bandın genişliğini daraltarak daha hızlı ancak yaklaşık sonuçlar elde etmek mümkün olmaktadır. Benzer şekilde bandı bir miktar genişleterek hizalama kalitesi, hızdan ödün vererek iyileştirilebilmektedir.

DTW ve türevi yöntemlerin bir diğer sorunu ise, DP ile elde edilen yolun ani iniş çıkışlı olması, elde edilen hizalamaların ise pürüzlü bir yapıya sahip olmasıdır. Bunun en büyük sebebi, büzme/germe'lerin herhangi bir fonksiyon ile modellenmemesidir. Bu sorunu aşmak için Parametric Time Warping (PTW) katsayıları Newton-Rapson ile en iniyilenen ikinci dereceden bir fonksiyon önermiştir. Benzer şekilde B-spline öneren çalışmalara yapılmıştır. Ancak bu konuda önerilen çalışmaların hiçbiri Ramsay'in önerdiği ikinci dereceden diferansiyel denklem kadar büzme/germe fonksiyonlarını esnek olarak modelleyememektedir. Ramsay'ın modelinde diferansiyel denklemin ağırlıkları ile oynayarak zengin bir fonksiyon ailesini taramak mümkün olmaktadır. Ağırlıkların eniyilenmesi için Newton-Rapson kullanılmaktadır.

Zaman serilerinin sınıflandırılması ile ilgili çalışmalarda öne çıkan diğer bir konu ise önerilen hizalama yöntemlerinin test şekli ve sunuş biçimidir. Literatüre yeni bir hizalama yöntemi kazandırdığını iddia eden çalışmaların büyük çoğunluğu, önerdikleri yöntemleri sadece birkaç verikümesinde deneyerek, genel yargılara ulaşmaya çalışmaktadır. Diğer sorun ise önerilen yöntemlerin başka uygulamalarda kullanılamayacak kadar özel yazılmasıdır. Bu sebeple yöntemleri birbiri ile karşılaştırmak çoğu zaman mümkün olmamaktadır. Ayrıca çalışmaların çok büyük kısmında, hizalama ile uzaklık hesaplama kavramlarının eş anlamlı olarak kullanılması karmaşayı arttırmaktadır.

Bu doktora tezinde yukarıda anlatılan eksiklikleri gidererek zaman serilerinin sınıflandırılması konusunda literatüre katkı yapılmıştır.

- Ramsay'in modelini temel alan ancak ağırlıkları sezgisel yöntemlerle eniyileyen Signal Alignment via Genetic Algorithm (SAGA) adlı yöntem önerilmiştir.
- Önerilen yöntem UCR zaman serileri havuzunda yer alan 40 farklı verikümesi üzerinden denenmiş ve alternatifleri ile karşılaştırılmıştır.
- Zaman serilerinin sınıflandırılmasına özellikle yeni başlayanlar için faydalı olabilecek bir çerçeve yazılım hazırlanmış ve genel kullanıma ücretsiz olarak sunulmuştur.

TIME SERIES CLASSIFICATION WITH GENETIC ALGORITHM AND SMOOTH MONOTONE INCREASING FUNCTIONS

SUMMARY

A time series is a collection of measurements over time. Examples include financial data, such as stock prices, the brightness of a celestial object over a function of time, measurements of quality characteristics in samples taken from a production process at different times or horizontal position of a hand drawing a gun from a hip-mounted holster.

Time series classification has broad application to many recognition tasks such as signature verification, speech and handwriting recognition and change detection in robots' environment. Time series classification has been a topic of great interest which has two fundamental components (i) classification and (ii) alignment. In the classification component, the well known methods, such as nearest neighbor and Support Vector Machines (SVM), can be used in their original forms. In these methods, the distance between a pair of time series is calculated by using standard Euclidean metric due to its ease of implementation, space/time efficiency and its fairly well accuracy. Since a time series is a list of measurements obtained by observing an event over time, the imperfections of measurement device or differences in the examined subjects results in distortion in the time axis called time drifts or retention time difference which reduces the classification performance. In order to eliminate the distortions, one should make non-linear adjustments often called alignment.

Alignment involves the elimination of temporal variations by stretching or compressing the time axis of one or both time series. Another equivalent definition of alignment is to create a many−to−many or one−to−one mapping, or fitting *a warping function*, between a pair of time series. Although the alignment methods may follow very different strategies, they all aim to produce a mapping which is then used to correct the time drifts in the time series. The newly produced and time corrected time series are called *aligned time series*.

Alignment methods can be used to improve the performance of a classifier by integrating the alignment method with the distance calculation. In this setup, whenever a distance calculation between a pair of time series is requested by the classifier, the time series are first aligned, then the Euclidean distance of the aligned time series is returned to the classifier. By using such an approach, the alignment becomes an integral part of the distance calculation so that it is usually considered as *a distance measure*.

The methods proposed in the studies on alignment techniques are usually considered as new alternatives to the standard Euclidean distance metric or other distance measures based on different alignment methods. For instance, Dynamic Time Warping (DTW),

one of the first alignment techniques emerging from the spoken word recognition field, is perceived as a new distance definition. DTW is indeed a generalization of the Minkowski distance which can handle time series of different lengths. However, many other distance measures, such as the Euclidean distance, can not be applied to time series of different lengths. In such cases, the time series first need to be "aligned" by re−interpolating them to equal length. This implies that a distance measure does not always behave as an alignment technique but rather works in tandem with alignment methods. Therefore, we prefer to distinguish alignment methods from distance measures even if the alignment is a variant of some distance measure.

A drawback of the DTW and its variants is the non-smooth alignments because of not using a mechanism controlling the smoothness of the warping function. Therefore the generated paths contain sharp corners leading to stair-case patterns in the alignments. Although the slope constraints have some control over the shape of the warping path, they do not impose smoothness and are not enough to control the curvature.

In order to generate smooth alignments, a new alignment method called Signal Alignment via Genetic Algorithm (SAGA) is introduced. SAGA is built upon a second order ordinary differential equation (ODE) whose solution generate smooth monotone increasing functions. The optimal weights of ODE was determined by using the Genetic Algorithm (GA). SAGA was tested on UCR time series repository and achieved a superior performance over DTW in terms of accuracy. As it is expected, SAGA also produced very smooth warping functions.

Another tricky part of time series classification is the application specific adjustments either in alignment or data processing steps. For instance, multi−dimensional time series are often converted to one dimension by summing, averaging or using other dimensional reduction algorithms because of the fact that majority of the alignment algorithms are designed to work only with one−dimensional time series. Amplitude normalization, baseline correction or measuring the quality of the alignment are the other minor tasks to be handled in times series classification. As a result, many alignment techniques proposed in the literature are entangled with the application specific tunings which make them unusable for standalone alignment purposes. For the same reason, the performance of the proposed alignment methods can be evaluated on limited number of application domains. The studies about time series classification have also a reasonable tendency to focus only on the field of study. Therefore, there are only a few studies trying to test their methods on datasets from different application domains.

In order to overcome the difficulties highlighted above, we contribute to the time series classification by proposing a framework which is built upon a clear distinction between classification and alignment so that one can freely change alignment method and observe its performance without dealing with classification. Likewise, different classification techniques can be tested by keeping alignment method fixed. New classification or alignment algorithms can also be integrated to the framework by implementing the corresponding interfaces. The framework has also the benefit of facilitating parallel computing resources if available. Additionally, the framework analyses the classification performance based on statistical tests.

By using the proposed framework, we implemented one classification and three alignment methods and tested on 40 different datasets kindly provided by Eamonn Keogh of University of California, Riverside. The most significant outcome of our experiments is that using an alignment method dramatically improves the classification performance on nearly every dataset except a few. Our second finding is that the performance of alignment techniques heavily relies on the characteristics of investigated dataset as we present such examples in the experiments. We also tested parallel programing feature of the framework by utilizing the facilities in High Performance Computing Center of Turkey. The framework has been designed as an open source project, so that researchers can easily implement their own algorithms.

The contribution of this theses can be summarized in the following items:

- A new alignment method called SAGA is introduced that combines smooth monotone increasing functions with genetic algorithm.
- SAGA was thoroughly tested on 40 different datasets presented in UCR time series repository.
- A time series classification framework is created to provide an easy to use and flexible platform that will be helpful especially for novices.