ÇOK YÖNLÜ DİZİLERİN ÇOKDEĞİŞKENLİLİĞİ YÜKSELTİLMİŞ ÇARPIMLAR GÖSTERİLİMİ ARACILIĞIYLA AYRIŞTIRIMI ve UYGULAMALARI

ÖZET

Bu savın düzenlenim aşamasında, günümüzde yüksek boyutlu veri çözümleyişinde (analizinde) ve işleyişinde sıklıkla kullanılan çok yönlü dizi yapılarının ayrıştırımı için Çokdeğişkenliliği Yükseltilmiş Çarpımlar Gösterilimi (ÇYÇG) tabanlı ayrıştırım yönteminin uygulanım olanakları araştırılmıştır. Çok yönlü dizi, 1-yönlü dizi olarak tanımlayabileceğimiz yöneylerin (vektörler) ve 2-yönlü dizi olarak tanımlayabileceğimiz dizeylerin (matrisler) genelleştirilmiş durumudur. Çok yönlü dizi ayrıştırımı ise dizey ayrıştırımında olduğu gibi, N-yönlü bir diziyi her biri özüne özgü değişik nitelikleri taşıyacak biçimde öğelere ya da bileşenlere ayırmaktır. Genel olarak N-yönlü dizilerin ayrıştırımı için çokludoğrusal cebir (ing: multilinear algebra) tabanlı yöntemler kullanılmaktadır. Sav kapsamında gündeme alınan ayrıştırım yöntemleri ise sayıtçıl (istatistiksel) bir açılım yöntemi olan ÇYÇG'ni taban almaktadır. Bir N-yönlü dizi için ÇYÇG ise N sayıda destek yöneyi ve bunların dışında bir değişmez bileşen, N sayıda 1-yönlü bileşen, $N(N-1)/2$ sayıda 2-yönlü bileşen ve bu biçimde gittikçe artan yönlülüğü olan 2^N sayıda bileşenden oluşmaktadır. Kuşkusuz, uygulamalarda bu düzeyde çok bileşen ile çalışmak sayısal bir indirgeyiş sağlamayacağından bu gösterilimi kesimcil uygulayarak ilgili çok yönlü diziyi en iyi biçimde anlatabilen sayıda bileşen kullanılmaktadır.

Sav düzenleyişi kapsamında ÇYÇG ile çok yönlü dizilerin ayrıştırımı dışında ÇYÇG tabanlı 3 değişik ayrıştırım yöntemi türetilmiştir. Bunlardan ilki olan İndirgeyimcil ÇYÇG (İ-ÇYÇG), ÇYÇG'nin sayıtçıl (istatiksel) özelliklerinin yanısıra izgecil özellikler de taşıdığından melez bir yöntem olarak görülebilir. Bu yöntemde, erekteki (hedefteki) çokludizi, iki çarpanlı toplamlar yapısında ayrıştırılır. Çarpımdaki toplam yön sayısı değişmemek koşuluyla, çarpanların her birinde istenilen sayıda yön alınabilir. Ancak, ilk çarpanlarda hep eş sayıda yön bulunmalıdır. Buradaki uygulamalarda, çarpanlardan biri yöney (bir yönlü dizi) alınmış ve böylece diğer çarpanda yön sayısı 1 düşürülebilmiştir. Savda, bu biçimdeki İndirgeyimcil ÇYÇG çok yönlü dizi yaklaştırımında uygulanmış ve sonuçları uygulama bölümünde verilmiştir. Yönlerin iki çarpana dağıtımı olgusu eşsiz değildir. Üstelik, ardışık olarak yön değişimlerine yol açabilen seçimler de gündeme getirilebilmiş ve yaklaştırım niteliği olumlu etkilenebilmiştir.

Öbür açılım yöntemlerinden biri olan Küçük Ölçeklerde ÇYÇG (KÖ-ÇYÇG) ise veri küme'sini altkesimlere ayırarak her bir altkesimde ÇYÇG yöntemini uygulayış görüşüne dayanmaktadır. KÖ-ÇYÇG yöntemi daha küçük bir uzamcıl alanda (geometrik bölgede) ÇYÇG'nin daha iyi yakınsayış sağlayacağını kanıtlar niteliktedir. Ayrıca bu ilerisürümün doğruluğu görüntü geri çatma (ing : image reconstruction) gibi gerçek yaşam uygulayışları üzerinde gösterilmiştir. Bu yöntem köklerini Sav Danışmanı olan Metin Demiralp'in önderliğindeki Bilişim Enstitüsü Bilgisayım Bilimi ve Yöntemleri Topluluğu (BEBBYT) araştırımlarında yer alan "Sıfır Oylumda ÇYÇG"

inceleyişlerinden almaktadır. Bu ise "Sıfır Oylumda YBBG (Yüksek Boyutlu Biçe Gösterilim, ing: High Dimensional Model Representation–HDMR)" inceleyişlerinin ÇYÇG uyarlanışıdır. Bu araştırımlarda, YBBG ya da ÇYÇG'nin üzerine kurulduğu uzamcıl (geometrik) bölgenin uygun bir biçimde ölçeklendiğinde ve, ölçeğin bölgeyi noktaya dönüştüreceği biçimde sıfıra götürülüşünde, değişmezliğin sonsuz baskınlığa erişeceği gösterilmiştir. Yazarın da bu doğrultuda bir yazısı bulunmaktadır. Bu baskınlık, "Sıfır Oylum Yaklaştırımı" olarak adlandırılan bir çizemin (ing: scheme) ÇYÇG uygulanışlarında kullanılabileceğini akla getirmiş ve savda bu doğrultuda da adımlar atılmıştır.

Savda kullanılan ve ÇYÇG'nden türetilmiş üçüncü yöntem olan Çokdeğişkenliliği Yükseltilmiş Çarpımlar Üçköşegencil Dizey Gösterilimi (ÇYÇGÜDG) ise her ne düzeyde ÇYÇG tabanlı olsa da özüne özgü özyineleyişli bir yapı olarak ortaya çıkmıştır. Dizeylerin dördül (kare) de olsalar dikdörtgen de olsalar, üçköşegencil (ing: tridiagonal) bir yapıya dönüştürülüşleri için geçerli olan ve işlerlik düzeyi yüksek olan bir araç olarak yapılandırılmıştır. ÇYÇÜDG'nin çok yönlü dizi ayrıştırım yöntemi olarak başarımı görüntü geri çatış ve yüz görüntüsü irdeleyiş (ing : face image retrieval) sistemi çalışmalarında gözlemlenmiştir. Savda bu doğrultuda yeterince bildirim verilmektedir.

Bu çalışma kapsamında üretilen her bir yöntem özgün olup her birinin değişik alanlardaki başarımı ilgili çok yönlü dizinin nicel ve nitel özelliklerine bağlı olduğu düzeyde kullanılan ayrıştırım yönteminin niteliğine de bağlıdır. Bu nedenle, çok yönlü dizi biçiminde anlatılabilen bir veri küme'sinin ayrıştırımının nasıl yapılacağı sorusu da aslında bu ayrıştırımın bileşenlerinin ya da tümünün ne amaçla kullanılacağı ile ilintilidir. Savda, olgunun bu yanına da özen gösterimi için ilgi çekici bir anlatım sergilenimine çabalanmıştır.

MULTI-WAY ARRAY DECOMPOSITION via ENHANCED MULTIVARIANCE PRODUCT REPRESENTATION and APPLICATIONS

SUMMARY

The basic focus of this PhD thesis is the Multi-way Array Decomposition which is frequently used for high dimensional data analysis and processing. High dimensionality can be considered in two different ways. First of all very large data sets can be categorized in high dimensionality. Secondly, matrices which have more than two directions can be considered as high dimensional and these data sets are expressed by multi-way data sets or multi-way arrays. A multi-way array can be taken as generalization of vectors which can be defined 1-way arrays and matrices that can be defined as 2-way arrays. Generally analyzing or processing a multi-way array needs a decomposition technique to get clear idea on relationships of different components or different ways. Decomposition of a multi-way array is to separate a multi-way array into components such that each of them shows different characteristics. Most of the recent researchs on the applied sciences shows that matrix decomposition based methods are insufficient for multi-dimensionality in terms of ways. The underlying reason is that the matrix based methods do not capture the correlations between the ways. This causes information loss while analyzing data set with multi-way structure and it may cause wrong inference. Thus, multi-way array decomposition is an important area for dimensionality reduction of high dimensional data sets.

Multi-way array decomposition based on Enhanced Multivariance Product Representation (EMPR) has been kept at the focus in this thesis study. Generally multi-linear algebra based methods are used for multi-way array decomposition. However for this study a statistical expansion, EMPR, based methods have been chosen for decomposing multi-way arrays. EMPR is a generalization of High Dimensional Model Representation (HDMR). HDMR is proposed by Sobol to represent multivarite functions with the less variate function components and it's usage has been pervaded for different applications. The basic property of HDMR is that the determination of the components can be accomplished if the geometry is orthogonal. However even if the geometry is orthogonal HDMR can be still insufficient for the representation. This kind of situations occur if the target function structure is far from the additivity because HDMR is purely an additive expansion. That is the reason why EMPR has been arosen.

The main difference between EMPR and HDMR is the existence of the support terms of EMPR. Support terms participate in HDMR expansion as multiplication of one-way arrays. However this participation is done in such a way that all terms have the same number of ways. The efficiency of support terms has been shown for multi-way arrays by comparing HDMR and EMPR approximation results on synthetic multi-way data sets. The results are supported for the conjecture that HDMR works better on multi-way arrays which have additive structure.

EMPR for an N-way array comprises 2^N additive terms each of which is a product of an EMPR component with a sufficient number of compatible support arrays. Amongst the components, one is a scalar which is multiplied by all support arrays to form the first additive term of EMPR while N number of components are one-way arrays which are multiplied by the all support arrays except the one depending on the same independent variable of the relevant component to form the univariate terms. The number of the bivariate terms is $N(N-1)/2$ each of which contains a bivariate component as the factor and other additive terms contain components with increasing number of ways. Despite EMPR is an expansion composed of a finite number of additive terms, this number may be extremely big in the practical applicational sense when the number of the independent variables grows. Then, certainly this kind of expansion is needed to be truncated such that it approximates the target multi-way array in the best approximation quality. This truncation is necessary because in practica;ity a decomposition technique with so many omponents does not provide any efficient numerical reduction on multi-way data.

Another necessity for the usage of EMPR on multi-way array is to find the optimal support terms to catch the best approximation with truncation. During the research of this necessity different kind of methods have been arosen for different kind of applications. In this manner three different EMPR based decomposition techniques have been developed apart from EMPR itself for multi-way arrays and they are reported in this thesis.

First of these methods is Reductive Enhanced Multivariance Product Representation (R-EMPR), R-EMPR can be regarded as a hybrid method because of it's spectral features as well as statistical structure. With Reductive Multilinear Array Decomposition (RMAD) it is possible to decompose an N-way array into a 1-way array and $(N-1)$-way array by using spectral features of N-way array. However R-EMPR offers a different utilization of RMAD. R-EMPR uses the 1-way outputs of RMAD as support terms. This kind of support array determination gives the flexibilty to algorithm for choosing different combination of RMAD outputs. For example, support arrays can be changed in accordance with the order of reduced ways. Also it is possible to design a tree algorithm to choose for best support array determination. R-EMPR is applied on the approximation problem of multi-way arrays. The multi-way arrays are chosen from the real chemical experiments and the results are reported on the applications section.

The second method, Small Scale EMPR, is based on a philosophy such that dividing the data set into small sub-datasets and applying EMPR on each piece. Small Scale philosophy was first used for HDMR and quite remarkable results have been produced. However small scale HDMR was developed for multivariate functions, thus continuous entries.

This is the first study that implements Small Scale philosophy on multi-way arrays. The basic idea of Small Scale EMPR is that, on smaller geometries, EMPR (or HDMR) gets better convergence for constant approximation (Zero degree truncation of EMPR expansion). According to this idea a multi-way array has been divided into small pieces and on each sub-array EMPR has been applied at most first order approximation. After combining the components of sub-array decompositions then the entire approximation has been calculated cumulatively. The algorithm's performance has been shown on real-life applications such as image reconstruction in this thesis' relevant sections.

Third of the developed methods is Tridiagonal Matrix Enhanced Multivariance Products Representation (TMEMPR). TMEMPR has specific features compared with the other EMPR based decomposition techniques. First of all, TMEMPR is always first order approximation because it has been constructed for matrices which are bivariate entities in elements. Thus the matrix EMPR contains just four additive terms. However its consecutive use in the decomposition of the residual matrices increases the number of the additive terms. This recursive scheme stops at the smallest edge of the rectangularity if the matrix has an empty null space adjoint to the smallest edge. The resulting decomposition's kernel matrix is however in a tridiagonal format and therefore it has been named as Tridiagonal Matrix Enhanced Multivariance Products Representation.

Despite TMEMPR has been designed for only two-way arrays its utilization can be extended for multi-way array decomposition and it can be done by using planar unfoldings of three-way arrays. In this sense it has been focused on the decomposition of the planarly unfolded three-way arrays with an application on images.

TMEMPR is constructed on the EMPR method base however it's main properties are quite different. For example while EMPR terms go to the number of way, TMEMPR terms can go to the number of small dimensionality of the array. Another difference between EMPR and TMEMPR is the determination of the components, EMPR components are found via statistical perspective while TMEMPR uses linear algebraic methods like matrix-vector multiplications to find the components. In this theisi we also have studied how to use this new method on the multi-way arrays. To this end a known transformation technique from multi-way arrays to matrices has been used. We first unfold the multi-way array and then use TMEMPR, after the truncation matrix is folded back into the multi-way array we obtain the approximation. With various application results of the algorithm on multi-way arrays have been collected by taking the certain three-way data sets which correspond to RGB images. TMEMPR method has also been used for face image retrieval system for greyscale image database. The results show the efficiency of TMEMPR is competitive with Singular Value Decomposition (SVD).

Support arrays play a fundamental role in the approximation quality of EMPR truncations and therefore TMEMPR truncations. In the case of continuity the target is a multivariate function and the supports become univariate functions. The construction of the support entries in both continuous and discrete cases is a serious problem and is quite nonlinear as we have mentioned a little bit in this study. One of the most recent research studies in the Demiralp Group (Group for Science and Methods of Computing) focuses on the support function construction at the zero volume limit where the all volume element subintervals are taken to zero. This study is in progress and it is expected that certain splinewise structures seem to be constructed. Until know we have behaved pragmatically and chosen the support functions mostly as directional fluctuations.

In the case of TMEMPR, the effect of the support arrays on the tridiagonal kernel matrices is such that the tridiagonality can be taken away towards diagonal form. This happens by changing the dominancy of the diagonal elements of the kernel matrix. In other words, the lower and upper diagonals of the tridiagonal kernel matrix can be reduced through appropriate choices of support vectors.

In this study, different divide-and-conquer algorithms have been designed for decomposition of multi-way arrays which are explained above. The main aim is to work with less terms and getting more accuracy for different real life data sets. Designed methods efficiency have been showed on different data sets which come from different application fields, however a point to be noted is the structure of data sets, which is as effective as the structure of these designed algorithm. As an example a statistical expansion may not be appropriate to represent a multi-way data set which have dominant spectral properties. Therefore these important details are also considered in this study while developing the decomposition algorithms.

Each of the methods that developed under this study is original and their performance is related to the field that they are used, the quantitative and qualitative characteristics of the multi-way array and the main properties of the decomposition technique. Thus the question that how is it possible to decompose and reduce a multi-way data, is related to how the components will be used and how the all decomposed array will be used. Specified examples have been taken for each case to explain these ideas clearly and the some of the algorithms' performances are showed on tables and some of them illustrated on images.