Evaluation of Clustering Patterns using Singular Value Decomposition (SVD): A Case Study of Metabolic Syndrome

Josephine M. Namayanja, University of Maryland, Baltimore County (UMBC), USA

ABSTRACT

Computational techniques, such as Simple K, have been used for exploratory analysis in applications ranging from data mining research, machine learning, and computational biology. The medical domain has benefited from these applications, and in this regard, the authors analyze patterns in individuals of selected age groups linked with the possibility of Metabolic Syndrome (MetS), a disorder affecting approximately 45% of the elderly. The study identifies groups of individuals behaving in two defined categories, that is, those diagnosed with MetS (MetS Positive) and those who are not (MetS Negative), comparing the pattern definition. The paper compares the cluster formation in patterns when using a data reduction technique referred to as Singular Value Decomposition (SVD) versus eliminating its application in clustering. Data reduction techniques like SVD have proved to be very useful in projecting only what is considered to be key relations in the data by suppressing the less important ones. With the existence of high dimensionality, the importance of SVD can be highly effective. By applying two internal measures to validate the cluster quality, findings in this study prove interesting in context to both approaches.

Keywords: K-Means Clustering, Metabolic Syndrome, Silhouette Coefficient, Singular Value Decomposition, Sum of Square Error

INTRODUCTION

In a realistic sense humans by nature are different, however it might seem that if individuals are affected by similar factors – either internal or external, they tend to exhibit similar characteristics. What if these characteristics vary from one individual to another or between groups of individuals? This study identifies groups of clusters in individuals that are diagnosed with MetS (MetS Positive) and is furthered with a comparison of those do not fit into the diagnosis (MetS Negative). In order to understand the patterns of the clusters formed, there are two techniques applied – 1) Simple K – Means clustering and 2) Singular Value Decomposition (SVD) as a pre-processing task performed prior to the former. Because the process of collecting medical data can be quite tasking on the part of health professionals due to the limitless number of features that surround its domain, it presents even more of a problem for those, such
as researchers and analysts, who try to expand the use of this type of data. Such datasets are highly dimensional and thus contain just about too many attributes for a single record. This often makes it difficult for knowledge workers to not only make a viable selection but to also represent this information in a manner that supports both visualization and clarity. Therefore (Thomassian, Castelli, & Li, 1998) commend that computational research presents data reduction techniques such as SVD and Principle Component Analysis (PCA) that can be applied to manage this curse of dimensionality.

In the first approach (also referred to as Non-SVD) of this study, a Simple K-means clustering algorithm is used as an exploratory analysis technique to group individuals of selected age groups with similar characteristics into \( k \) clusters where \( k \) refers to the ideal number of clusters (Bunn & Ostrovsky, 2007) with an aim of maximizing intra cluster similarity and minimizing inter cluster similarity. While in the second approach SVD (also referred to as SVD-based clustering in this study) is initially applied to create a factorization of a given matrix \( X \) and hence reduce the dimensionality of its underlying data structure to produce a more appropriate subspace of the data based on some ranking system (Phillips, Watson, & Wynne, 2008). This ranking system describes the linear independence of a data matrix which is determined by the dependency of rows/columns on each other. Further still, this technique can be applied to any type of data set \( D \) represented as any size of matrix but the linearity may differ depending on the distribution of the data. (Martin, 2005) adds that it also provides a method for mathematically discovering correlations within data and in this study the use of SVD is followed by clustering the reduced dataset into \( k \) clusters. The purpose of both approaches sums up into a comparison of clusters formed at both iterations to determine the performance of the clustering algorithm depending on the quality of clusters formed. As the technique of clustering has previously been applied to numerous applications such as information retrieval, machine learning, data mining research (Bunn & Ostrovsky, 2007) and computational biology (Frahling & Sohler, 2006), SVD has taken phase in face recognition, medical imaging as stated in (Hsu et al., 2007; Lee et al., 2004; Lui et al., 2009; Ma et al., 2009; Zanderigo et al., 2009) and is also used as a complimentary technique on Latent Semantic Indexing (LSI) and clustering which is the case in this study. More so, computational methods such as Markov Chains and SVD itself have been looked at previously in alliance with several aspects of the health domain from brain functionality (Brockwell, Kass, & Schwartz, 2007) and Alzheimer’s disease (Li et al., 2005) to cardio (Brinegar et al., 2009) plus gene related studies (Alshalalfa, Alhaji, & Rokne, 2008). With the objective of this study in mind, the comparison of cluster quality is validated using two internal measures which include: 1) Sum of Square Error (SSE) which looks at the distance (Euclidean distance) of the data point \( x \) from the cluster centroid \( c \) and 2) Silhouette Coefficient (Sil) which considers the distances of individual data points to other points internally within a cluster and externally to other clusters formed (Frahling & Sohler, 2006). This foreword on the study is extended with a motivation that describes Metabolic Syndrome in depth and samples on real statistical based examples of its effects.

**Motivation**

With close to 72 million obese Americans (Ascribe Newswire: Health, 2008), the prevalence of Metabolic Syndrome (Met S) is alarming. An estimation of 45% of the population highlights that the older generation (Sherman, 2009; Howe, 2008; The Consumer’s Medical Journal, 2009; Miner, 2010) within the age of 60 and 70 years plus, has a diagnosed rate of approximately 60% and 40% respectively (Sherman, 2009; Howe, 2008). Having no particular definition for MetS, previous studies define the latter also called Insulin resistance syndrome (Nutrition Action Health Letter, 2009) or Syndrome X, as a group of conditions when combined, put an individual at an increasing risk of having heart disease, and diabetes (Sherman, 2009; The Consumer’s Medical Journal, 2009; Miner, 2010; Uemura et al., 2009; Brzycki, 2010; National Institutes
A Gravitational Search Algorithm Approach for Optimizing Closed-Loop Logistics Network
www.igi-global.com/chapter/gravitational-search-algorithm-approach-optimizing/69899?camid=4v1a