P- ISSN 1991-8941 E-ISSN 2706-6703 Journal of University of Anbar for Pure Science (JUAPS) Open Access
2009,(3), (1) :71-74

Dimensionality reduction in data from LASER applications

Qassim M. Jameel **Imad H.Aboud**

University of Anbar. College of science.

ARTICLE INFO

Received: 10 / 2 /2009
Accepted: 15 / 12 /2008
Available online: 14/6/2012
DOI: 10.37652/juaps.2009.15450

Keywords:
CATPCA,
Single linkage clustering,
LASER.

ABSTRACT

Redundant variables not only in LASER applications, but in all experimental works are disturbing statistical analysis as a result of highly correlation among them. It is not easy sometimes to identify which set of variables is redundant and which one is retained. In addition, consideration of huge sets of variables will make it difficult to point out the joint effects of any subset of variables on a certain phenomenon. It is well know that continuous variables can be transformed into a discrete (categorical) form depending on predefined intervals, thus, the categorical principal component analysis was adopted here in this paper to identify the discarded set of variables when the data contained some variability. The effect of identifying groups of retained variables was compared by observing the natural grouping of elements using single linkage clustering of elements.

Introduction

In most of applied disciplines, many variables are sometimes measured on each individual, which result a huge data set consisting of large number of variables, say p1. Using this collected data set in any statistical analysis may cause several troubles.

The dimensionality of the data set can often be reduced, without disturbing the main features of the whole data set by Principal Component Analysis (PCA) technique2. Dimensionality reduction is affected if k (<< p) of the Principal Components (PCs) convey virtually all the information inherent in the p variables3, 4. However, the constructed PCs may not be easy to interpret in terms of all the original p variables. Therefore, it is useful to reduce the number of variables as much as possible whilst capturing most of the variation of the complete data set, X.

Single-Linkage Clustering

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics.

The single linkage clustering was considered in this research work. Usually the distance between two clusters A and B is:

$$\min\{d(x,y): x \in A, y \in B\}$$

The mean distance between elements of each cluster (also called average linkage clustering)٥:

$$\frac{1}{|A|.|B|}\sum_{x \in A}\sum_{y \in B} d(x,y)$$

The Algorithm

Let's now take a deeper look at how Johnson's algorithm works in the case of single-linkage clustering. The algorithm is an agglomerative scheme that erases rows and columns in the proximity matrix as old clusters are merged into new ones.

The N*N proximity matrix is D = [d(i,j)]. The clusterings are assigned sequence numbers 0,1,......, (n-1) and L(k) is the level of the kth clustering. A cluster with sequence number m is denoted (m) and the proximity between clusters (r) and (s) is denoted d [(r),(s)].

The algorithm is composed of the following steps:

1. Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.
2. Find the least dissimilar pair of clusters in the current clustering, say pair (r), (s), according to

* Corresponding author at: University of Anbar. College of science, Iraq.E-mail address:

d[(r),(s)] = min d[(i),(j)]where the minimum is over all pairs of clusters in the current clustering.

3. Increment the sequence number : m = m +1. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)]
4. Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined in this way:
   d[(k), (r,s)] = min d[(k),(r)], d[(k),(s)]
5. If all objects are in one cluster, stop. Else, go to step 2.

CATPCA

Clustering analysis commonly falls into two main categories; clustering cases (observations) and clustering variables. In the clustering variables techniques researchers very often aimed to reduce dimensionality of the data (i.e, reducing the number of variables to the minimum such that the retained set of variables will not harm the further analysis or data investigations).

In this context, principal component analysis is a common technique used to reduce dimensionality. This technique is very sensitive to outliers and extreme values. As a result it would not give an appropriate, efficient and reliable classification of variables and/or cases. However, the categorical principal component analysis CATPCA is found to be not affected by outliers or extreme values, and therefore it assumed to give better results when adopted in situations assuming considerable data variability[٦].

Data

The considered data in this paper were the output of a research work entitled "The production of multi-element opacity targets for X-ray laser experiments" which carried out by Spindloe C. (2006/2007)[7].

Case	Element	AN	Unn.C (wt %)	Norm.C (wt %)	Atom (at %)	Error (%)
1	Si	١٤	٢٧.١٠	٢٧.٧٣	٢٠.٢٧	١.٢
2	Na	١١	٨.٢٩	٨.٤٩	٧.٥٨	٠.٧
3	Fe	٢٦	٥.٤١	٥.٥٤	٢.٠٤	٠.٧
4	Ca	٢٠	٣.٧٢	٣.٨٠	١.٩٥	٠.٤
5	Mg	١٢	٢.١٦	٢.٢٠	١.٨٦	٠.٣
6	K	١٩	٠.٦٧	٠.٦٩	٠.٣٦	٠.٤
7	Al	١٣	٠.٤٠	٠.٤١	٠.٣١	٠.٢
8	O	٨	٤٩.٩٩	٥١.١٤	٦٥.٦٣	٩.٨

Results

The use of the CATPCA revealed that three main groups of variables can be noticed.All groups are different in the variability among members of the groups as well as between groups (figure 1).

Regarding single linkage clustering of variables, figure 2 shows almost the same grouping obtained by the CATPCA.

Both figures resulted in the same conclusions regarding the retained set of variables.

In order to give a better idea about the effect of discarding variables on the grouping of elements, single linkage clustering was performed twice; once before discarding redundant variables and other one after discarding variables (figures 3 and 4). In both cases the grouping of elements never changed which means that the effect of the discarded variables can be neglected.

Discussion

CATPCA is an efficient statistical technique in the cases of nominal and ordinal set of variables. Numeric variables can be transferred to limited number of categories and treated as categorical data which will eliminate the effects of outliers and extreme values. Single linkage clustering of cases using Euclidian distance as a measure of similarity will result in a fairly similar conclusions as the CATPCA.

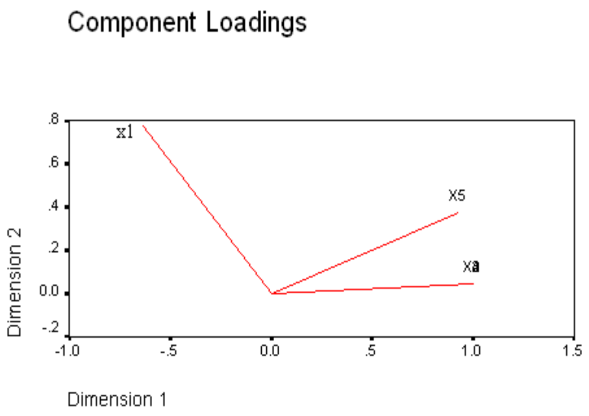
Conclusion

When redundant variables are well identified, methods of data analysis will not significantly affect after discarding such subset of variables. As obtained in this paper, dendrograms of observations (elements) clustering are not significantly different before and after discarding redundant variables.

References

1. Sharma, S. (1996). Applied Multivariate Techniques. John Wiely & Sons, Inc. New York.
2. Rencher, A. C. (1995). Methods of Multivariate Analysis. John Wiley & Sons. New York.

3. Jolliffe, I. T. (1973). Discarding Variables in a Principal Component Analysis II: Real Data. Applied Statistics, 21, 160-173.
4. Krzanowski, W. J. (1988). Princip;es of Multivariate Analysis: a user's perspective. Clarendon Press, Oxford.
5. E. B. Fowlkes & C. L. Mallows (September 1983). "A Method for Comparing Two Hierarchical Clusterings". Journal of the American Statistical Association 78 (383): 553–584.
6. Jacqueline J. Meulman, Anita J Van der Kooji and Willem J. Heiser. Principal component analysis with non-linear scaling transformations for ordinal and nominal data. http://www.sagepub.com/upm-data/5040_Kaplan_Final_Pages_Chapter_3.pdf
7. Spindloe C. The production of multi-element opacity targets for X-ray laser experiments. Central LASER Facility Annual report 2006/2007.



Variable Principal normalization.

Fig. 1. Distribution of the variables as grouped by CATPCA.

```
Dendrogram using Single Linkage

                          Rescaled Distance Cluster Combine

    C A S E      0         5        10        15        20        25
  Label     Num  +---------+---------+---------+---------+---------+

  X2          2   ⇩×⇩⇩⇩⇘
  X3          3   ⇩⇙  ▭⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇘
  X4          4   ⇩⇩⇩⇩⇩⇙                                ⇔
  X1          1   ⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩×⇩⇩⇩⇩⇩⇩⇩⇙
  X5          5   ⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇙
```

Fig. 2. Dendrogram using Single Linkage.

```
                          Rescaled Distance Cluster Combine

    C A S E      0         5        10        15        20        25
  Label     Num  +---------+---------+---------+---------+---------+

  Case 5      5   ⇩⇘
  Case 7      7   ⇩⇳⇩⇘
  Case 2      2   ⇩⇙ ⇔
  Case 4      4   ⇩×⇩✓⇩⇩⇩⇩⇩⇘
  Case 6      6   ⇩⇙ ⇔    ▭⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇘
  Case 3      3   ⇩⇩⇩⇙    ⇔                             ⇔
  Case 1      1   ⇩⇩⇩⇩⇩⇩⇩⇩⇩⇙                             ⇔
  Case 8      8   ⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇙
```

Fig.3. Dendrogram of clustering cases using Single Linkage (all variables).

```
                            Rescaled Distance Cluster Combine

    C A S E      0         5        10        15        20        25
  Label     Num  +---------+---------+---------+---------+---------+

  Case 5      5   ⇩⇘
  Case 7      7   ⇩⇳⇩⇩⇩⇘
  Case 2      2    ⇩⬃   ⬄
  Case 4      4    ⇩✖⇩⇩⇩✓⇩⬂
  Case 6      6    ⇩⬃   ⬄ ▫⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⬂
  Case 3      3    ⇩⇩⇩⇩⇩⬃ ⬄                              ⬄
  Case 1      1    ⇩⇩⇩⇩⇩⇩⇩⬃                              ⬄
  Case 8      8    ⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⬃
```

Fig.4. Dendrogram of clustering cases using Single Linkage (only retained variables).

تقليص ابعاد البيانات المأخوذة عن التطبيقات الليزرية

عماد هجول عبود   قاسم محمد جميل

الخلاصة

تعتبر المتغيرات الفائضة ليس فقط في التطبيقات الليزرية و انما في كل الاعمال التجريبية من المزعجات التي تعترض سبل التحليل الاحصائي كنتيجة للارتباطات العالية التي يمكن تاشيرها بين هذه المتغيرات. في بعض الاحيان لا يكون سهلا اعتبار أي مجموعة جزئية من المتغيرات فائضة و أي مجموعة يمكن اعتبارها لاغراض البحث. اضافة الى ذلك فان الابقاء على مجموعات كبيرة من البيانات سوف يجعل من الصعب ايجاد تفسيرات دقيقة لمساهمة كل متغير عندما يشترك تاثيره مع متغير او اكثر من مجموعة المتغيرات المعتمدة. و لان البيانات المستمرة يمكن تحويلها الى بيانات متقطعية (حقلية)، لذا فقد تم تبني طريقة المكونات الاساسية الحقلية في هذا البحث لتمثيل مجموعة المتغيرات الفائضة عندما تنطوي البيانات على قدر من التغاير. لقد تم اختبار تاثير المتغيرات المستبعدة على طريقة تجميع المشاهدات باستخدام احد اساليب التحليل العنقودي.