Dimension Reduction - Comparison Script for R

Giménez Gredilla, Daniel

November 2017

Contents

1 - Introduction and general data set description

The current data set is a study on **normal**, **atypical**, and reactive lymphocytes, in which 2867 numerical variables, based on colorimetric and geometric features have been measured. The aim of the study is more thoroughly explained both in the proposal and main report for this project.

The current data set has 13074 observations for 2874 variables, of which 7 (in which the response variable, *tipoCelula*, is included) are factors and 2867 are numeric predictor variables.

The response variable, *tipoCelula*, has three levels, being these **ATYPICAL_LYMPHOCYTE**, **VARIANT_LYMPHOCYTE** and **LYMPHOCYTE**. Two subsets will be generated with this data, a training set comprising 66% of all data, and a test set formed by the remaining entries. Both sets are generated by defining a random seed with the *set.seed()* function (the seed beeing 123) and applying that seed to the *sample()* function to generate an array of indexes.

2 - Goals and techniques

This script aims to assess the comparative performances of different dimension reduction techniques. They will be measured for a common accuracy metric and judged by their processing performance. Being **PCA** the most used, most assessed technique, it will be used as a kind of touchstone in respect to requirements for the other techniques.

For the effects of this project, the benchmark number of variables is that which satisfies one condition, being this that the extracted features account for an accumulated 95% of variance in the reference dimensionality reduction technique, **PCA**. Having determined this, a continued test of cumulative numbers of extracted features, between a floor of 10 and an estimative limit of 250 yields the following result.

With this number of extracted features as an objective benchmark, the following techniques will be conducted:

- **PCA**
- **ICA**
- **Factor Analysis**
- **Linear Decomposition Analysis**

The resulting features will then be divided in training and test sets, and used as input, first, for the fitting and training of an **SVM** function (*svm()*, from the *caret* package), and then as input of a *predict()* function, from the *stats* package. The predicted classes will then be cross-tested with the actual test classes. A confusion matrix and a Cohen's Kappa weighted and unweighted value will then be output, and used as that technique's entry in the final performance comparison.

Cohen's Kappa (Luengo 2009) is an alternative to **Classification Rate** that takes into account random correct hits. It was originally used to measure the degree of agreement between two subjects describing the same event. In the meantime it has been adapted for classification tasks, as it compensates for random hits in the same way that **AUC** does for **ROC**. The mathematical expression for **Cohen's Kappa** is applied to the contingency table of an event in the following way:

$$kappa = \frac{n\sum_{i=1}^{C} x_{ii} - \sum_{i=1}^{C} x_{i.}x_{.i}}{n^2 - \sum_{i=1}^{C} x_{i.}x_{.i}}$$

Where x_{ii} is the cell count in the main diagonal, n is the number of examples, C is the number of class values and $x_{.i}x_{i.}$ are the total columns and rows counts, respectively.

The value range of **Cohen's Kappa** goes from -1 (total disagreement) to 1 (perfect agreement).

Table 1: PCA observed versus predicted results

	ATYPICAL_LYMPHOCYTE	LYMPHOCYTE	VARIANT_LYMPHOCYTE
ATYPICAL_LYMPHOCYTE	3608	283	77
LYMPHOCYTE	20	97	0
VARIANT_LYMPHOCYTE	30	0	330

Table 2: PCA observed versus predicted results - percentages

	ATYPICAL_LYMPHOCYTE	LYMPHOCYTE	VARIANT_LYMPHOCYTE
ATYPICAL_LYMPHOCYTE	81.17	6.37	1.73
LYMPHOCYTE	0.45	2.18	0.00
VARIANT_LYMPHOCYTE	0.67	0.00	7.42

The reasons for chosing **Cohen's Kappa** as accuracy metric are as follows:

- The data set upon which it is going to be used has a multiclass factor response variable.
- Those labels are not ascribable to a binary synthetic class system.
- **Cohen's Kappa** yields a scalar, simple value well suited for multiclass classification.
- It is more powerful than other multiclass accuracy metrics such as **Classification Rate**, because it takes into account random hits, scoring successes separately for each class and aggregating them.

Weighted and unweighted values of **Cohen's Kappa** differ, as their name implies, in that weighted scores take into account the differential weights of several levels of disagreement between observed and predicted classes. This is a level of information that is lost in binary classification, as all disagreements between observed and predicted classes share the same level of disagreement.

3 - Dimension Reduction Techniques

3.1 - PCA

PCA is the most used unsupervised, linear dimension reduction technique currently available. It is also the best, in the mean-square error sense (Fodor 2002). Its central idea is the construction of a set of features from a number of initial variables (Jolliffe 2002). The number of new features will be less than the initial variables, while retaining as much as possible of the initial variation. This is achieved by linear transformations of the original data, and then establishing a descending order of the new features attending to the amount of variation retained or explained by each of them.

For this project, 210 are retained and use in the fitting, training and prediction of classes. The following section depicts the results of this protocol.

After fitting and predicting, the hit and accuracy values are extracted and represented in **Table 1** and **Table 2**, in absolute and percentage values, respectively.

PCA yields a weighted Cohen's Kappa value of 0.77, which will be a benchmark for other techniques.

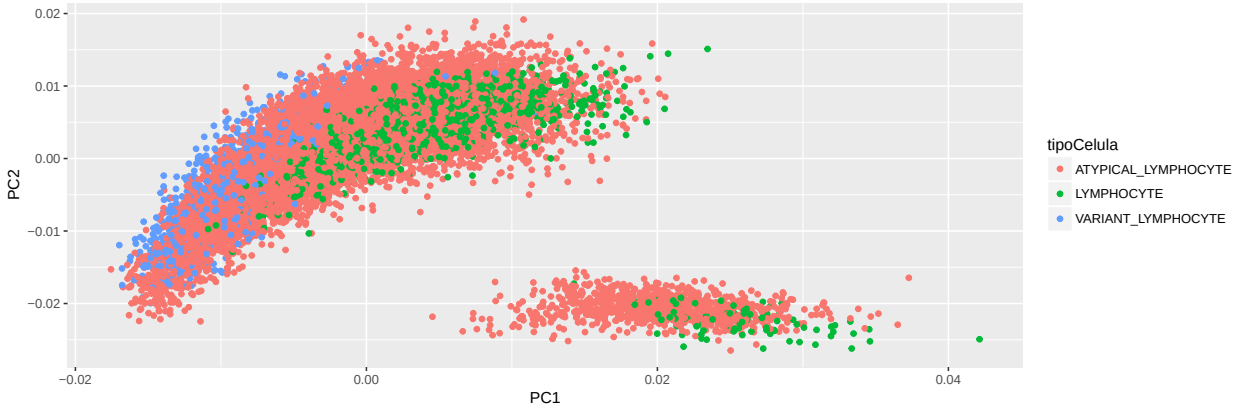
As a visual, informative measure, it is interesting to have the pca fitted components plotted, coloured by *tipoCelula* label.

Table 3: ICA observed versus predicted results

	ATYPICAL_LYMPHOCYTE	LYMPHOCYTE	VARIANT_LYMPHOCYTE
ATYPICAL_LYMPHOCYTE	3608	283	77
LYMPHOCYTE	20	97	0
VARIANT_LYMPHOCYTE	30	0	330

Table 4: ICA observed versus predicted results - percentages

	ATYPICAL_LYMPHOCYTE	LYMPHOCYTE	VARIANT_LYMPHOCYTE
ATYPICAL_LYMPHOCYTE	81.17	6.37	1.73
LYMPHOCYTE	0.45	2.18	0.00
VARIANT_LYMPHOCYTE	0.67	0.00	7.42



Points in space are grouped by factors and colored by response label.

3.2 - ICA

Independent Component Analysis (*ICA*) is a statistical method for transforming an observed multidimensional random vector into components that are statisticaly as independent from each other as possible, this is, a tendency to **redundancy reduction** (Tobergte and Curtis 2013). In its linear approach, as with other dimension reduction algorithms, its goal is to take a zero-mean, m-dimensional variable, and by means of a linear transformation, find its n-dimensional transform, such that $n \leq m$, this transformation having some suitable properties. The vectors obtained from this transformation are neither orthogonal nor ranked in order.

Feature extraction is a prominent application of *ICA*. It is originally motivated by results in neuroscience that suggest that the same cited principle of redundancy reduction is applied by the brain for the early processing of sensory data.

ICA is a generative model (it describes how the observed data are generated by describing the components), and it seeks the minimization of mutual information between the transformed variables. It depends on the supposition of nongaussianity for the data; gaussian data is independent and of mean zero, it has no skewness and as such can only be estimated up to an orthogonal transformation (Hyvärinen and Oja 2000).

Applying it to the current data set, the results are as follows:

After fitting and predicting, the hit and accuracy values are extracted and represented in **Table 3** and **Table 4**, in absolute and percentage values, respectively.

Curiously, **ICA** yields results similar to **PCA**, with a weighted Cohen's Kappa of 0.77. Even though this

Table 5: Factor Analysis observed versus predicted results

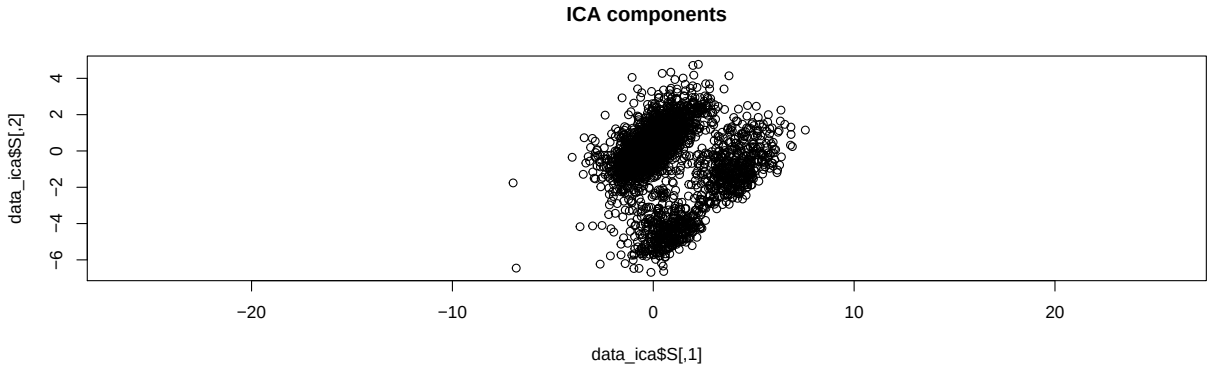
	ATYPICAL_LYMPHOCYTE	LYMPHOCYTE	VARIANT_LYMPHOCYTE
ATYPICAL_LYMPHOCYTE	3579	231	66
LYMPHOCYTE	43	149	0
VARIANT_LYMPHOCYTE	36	0	341

Table 6: Factor Analysis observed versus predicted results - percentages

	ATYPICAL_LYMPHOCYTE	LYMPHOCYTE	VARIANT_LYMPHOCYTE
ATYPICAL_LYMPHOCYTE	80.52	5.20	1.48
LYMPHOCYTE	0.97	3.35	0.00
VARIANT_LYMPHOCYTE	0.81	0.00	7.67

puts it at eye level with **PCA**, the processing power needed for this technique is noticeably larger, and thus, the hollistic assessing of **ICA** still has it behind **PCA**.

In this plot the ICA components can be visualized.



3.3 - Factor Analysis

The basic idea underlying Factor Analysis is that p observed random variables, $\mathbf{x}$, can be expressed, except for an error term, as linear functions of $m(< p)$ hypothetical (random) variables or *common factors* (Jolliffe 2002). The aim of Factor Analysis is to group variables that share a "common theme" under the same grouping, such that the dimensionality of the dataset is decreased.

Factor Analysis has been applied in psychology to identify groups of inter-related variables, as those components of intelligence that can be placed under a single factor g or *general intelligence*, grouping factors such as *broad visual perception* (it includes all the intelligence variables related to visual tasks), or *broad auditory percention* (same as before, but with auditory tasks). This is interpreted as someone with a high g having good *broad auditory and visual perceptions*, and g sinthetically explaining the behaviour of the factors and variables "contained" within itself.

This technique is applied here to the given data set.

After fitting and predicting, the hit and accuracy values are extracted and represented in **Table 5** and **Table 6**, in absolute and percentage values, respectively.

This technique gives a weighted Cohen's Kappa Value of 0.79. This value is surprising, given that it puts the **Factor Analysis** technique's accuracy over the **PCA**, which is a peculiar result. This can be a specific case in which **Factor Analysis** is better in a metric sense, while it is clearly more costly in processing power.

Table 7: LDA observed versus predicted results

	ATYPICAL_LYMPHOCYTE	LYMPHOCYTE	VARIANT_LYMPHOCYTE
ATYPICAL_LYMPHOCYTE	3607	66	14
LYMPHOCYTE	36	314	0
VARIANT_LYMPHOCYTE	15	0	393

As a visual, informative measure, it is interesting to have the **Factor Analysis** output fitted components plotted, coloured by *tipoCelula* label.

All this said, what **Factor Analysis** does (and what makes it unique) is create a matrix of loadings, this is, it creates columns with synthetic factors that "rely" on and group up subsets of the original variables. These new features are semantic groupings of the original variables, giving this technique an added value as it reveals hidden relations between variables.

Given these new relationships, it is up to the analyst to name them in a semantically comprising manner (let it be said, as an example, that two theoretical variables "speed" and "agility" were revealed to be bound; the wrapping extracted feature could be named, maybe, "mobility"). In this project these complex relationships won't be named, but given more time, the traits that bind these variables together could be analysed and these new features named.

3.4 - LDA

Linear Discriminant Analysis, or **LDA**, is a generalization of *Fisher's Linear Discriminant*. It is a well-known technique for feature extraction, and it has been widely used for such uses as facial recognition, image retrieval or microarray data classification. **LDA** focuses on the response variable classes. It projects the data onto a lower-dimensional vector space such that the ratio of the between-class distance to the within-class distance is maximized, thus achieving maximum discrimination.

Mathematically, given a data matrix, classical **LDA** aims to find a transformation that maps each column a_i of A, for $1 \leq i \leq n$ in the N-dimensional space to a vector b_i in the l-dimensional space. It creates clusters, such that the quality of each cluster is high if it is well-separated from other clusters and tightly grouped (Klecka 1980).

This technique is applied here.

After fitting and predicting, the hit and accuracy values are extracted and represented in **Table 7** and **Table 8**, in absolute and percentage values, respectively.

This technique gives a weighted Cohen's Kappa Value of 0.79. It seems a high value is extracted from a data set with well defined class groupings and it may be that, for this application specifically, **LDA** is a good feature reduction technique.

Table 8: LDA observed versus predicted results - percentages

	ATYPICAL_LYMPHOCYTE	LYMPHOCYTE	VARIANT_LYMPHOCYTE
ATYPICAL_LYMPHOCYTE	81.15	1.48	0.31
LYMPHOCYTE	0.81	7.06	0.00
VARIANT_LYMPHOCYTE	0.34	0.00	8.84

4 - Conclussions

For the time being, **no definitive conclussions** can be extracted from this data. The protocols still need to be refined for the final report of this project, so that possible artifacts and misleading parameters are avoided. If any partial conclussions have to be extracted from this data, it is that the precise area in which these techniques are applied is of an utmost importance to the final utility of each of them, and that no generalisation (e.g. "*PCA is always the best solution*") can be made without an exhaustive eapplication.

5 - References

Fodor, Imola. 2002. "A Survey of Dimension Reduction Techniques." doi:10.1.1.8.5098.

Hyvärinen, Aapo, and Erkki Oja. 2000. "Independent Component Analysis: Algorithms and Applications." *Neural Networks* 13 (45): 411–30. doi:10.1016/S0893-6080(00)00026-5.

Jolliffe, I T. 2002. "Principal Component Analysis, Second Edition." *Encyclopedia of Statistics in Behavioral Science* 30 (3): 487. doi:10.2307/1270093.

Klecka, William. 1980. "Discriminant Analysis." *Advances in Neural Information Processing Systems* 17 (60): 1569–76. doi:10.4135/9781412983938.

Luengo, Æ J Æ. 2009. "A study of statistical techniques and performance measures for genetics-based machine learning : accuracy and interpretability," 959–77. doi:10.1007/s00500-008-0392-y.

Tobergte, David R., and Shirley Curtis. 2013. "Independent Component Analysis by Minimization of Mutual Information." *Journal of Chemical Information and Modeling* 53 (9): 1689–99. doi:10.1017/CBO9781107415324.004.