Re: connection between preselecting parts of information with high variance and PCA

Source: http://sci.tech−archive.net/Archive/sci.image.processing/2007−02/msg00218.html

- *From*: "K." <kdelac@xxxxxxxxx>
- *Date*: 24 Feb 2007 05:52:02 −0800

thanks for your answer and thanks for reading this long post :)).
you are absolutely right. the point is exactly that:

PCA before and after pre(selection) of features based on variance in
connection that PCA also selects features through measuring the
variance.

K.

On Feb 24, 1:02 pm, jg.campbell...@xxxxxxxxx wrote:

> On Feb 24, 11:28 am, "K." <kde...@xxxxxxxxx> wrote:
>
> > hi all,
> >
> > i am from the image processing / pattern recognition field and i have
> > stumbled upon an interesting mathematical problem / question. maybe it
> > will be trivial to you guys, but all the better :))
> >
> > so, the problem is related to eigenvector decomposition, i.e.
> > Karhunen−
> > Loeve's transform or as we call it − PCA (Principal Component
> > Analysis). Application area − Face Recognition. Images are rearranged
> > into vectors that represent points in n−dimensional (n being the
> > number of pixels in each image) space.

The idea is that eigenvectors of the covariance matrix of the set of images (the set of points in n−D space) will decorrelate the data. The first "most important" eigenvector (the one associated to the larges eigenvaule) captures the direction with larges variance.... well, you know the rest. You can then keep only a few of those eigenvectors with larges eigenvalues and project all the images onto that new few−D space (let's call it the k−D space, with k << n). By keeping the vectors with largest eigenvalues you are keeping a large portion of the energy of your data, or consequently you are keeping the most information. eigenvectors are linear combinations of the original dimensions, thus making PCA a simple rotation/stretch procedure. Similarity of two images (this is the basic face recognition idea) can then be determined by measuring the distance (e.g. Euclidean) between the two projections in the lower dimensional space instead of doing it in the high n−D space.

Now for the problem :) :

In my little experiment I used the preprocessed images (so the input to calculating the covariance matrix and the rest of the PCA are not pixels anymore, but some other coefficients − but this is irrelevant). My "images", or to be precise, my matrices of coefficients are the size of 128 x 128, rearranged to 1x16384 vectors (so n = 16384, original space is n−D). I performed PCA on those "images" and then performed a simple face recognition in the yielded k−space (k<<n). I got recognition rate of e.g. 40% (so i recognized about 40% of images correctly).

In the next experiment, instead of using all of the 16384 coefficients per "image", i selected only a subset of those, the ones that for my set of "images" have the larges variance. So i measured the variance of each coefficient on the same spatial coordinates across all "images", following the line of thoughts that the ones that change the most for different persons are more important for discriminating them. I kept 512 of them (i remembered their spatial locations and then selected them from each image), thus going from 1x16384 to 1x512 per "image". Now i did PCA on those 1x512 vectors and repeated the same experiment as before (i kept k−D, where k was the same as in the previous experiment) and got the recognition rate of 60%.

Some questions:

> how come the results are so different? −i know the comparison is not really fair because k << 16384 but only slightly smaller than 512 (k < 512), but nevertheless...

I have no idea. Could be pure chance . What classifier are you using? What
size of training data set? What size of test data set? Are they
separate? How many classes (subjects)? How many images per subject?

Would I be correct in thinking that 40% correct is worse than
classifying by random number? And 60% not much better?

> what is the connection between the two procedures in terms that PCA finds the directions of most variance and I preselected the coefficients with largest variances? shouldn't PCA do the same thing? is the correct conclusion that the rest of the coefficients (the remaining ~15000) are simply redundancy in terms of face recognition?

> the main question: how does preselecting the coefficients with highest variance across all images correlate to what standard PCA does? can this be compared at all since one original space is 512−D and the other 16384−D?

Thinking ... Let's say you have three features (like my squashed egg
data
set). And let us say that the long principal axis is in the x1, x2
(feature
1, feature 2) plane and diagonal. And the next p. axis is in the same
plane
and perpendicular to that. What if you pre−select x1, x2 ... Same
result. No
that doesn't really help.

PCA does two things. 1 finds directions of maximum variance. 2.
Decorrelates; i.e. if xi and xj are very highly correlated, then
little
need to include any xj when xi included.

Another attack. Have a look at how well each of the PCAs allows you to
reconstruct the (compressed) image.

comp.ai.neural−nets might have an answer; or sci.sta.math. But
summarise
your problem into a shorter post: "PCA before and after feature

selection based on variance" or something like that.

Best
regards,

Jon C.

.