Classification Based on Fuzzy Robust PCA Algorithms and Similarity Classifier

Pasi Luukka

Laboratory of Applied Mathematics
Lappeenranta University of Technology
P.O. Box 20, FIN-53851 Lappeenranta, Finland

Abstract

In this article classification method is proposed where data is first preprocessed using fuzzy robust principle component analysis (FRPCA) algorithms to get data into more feasible form. After this we use similarity classifier for the classification. We tested this procedure for breast cancer data and liver-disorder data. Results were quite promising and better classification accuracy was achieved than using traditional PCA and similarity classifier. Fuzzy robust principle component analysis algorithms seems to have the effect that they project these data sets into more feasible form and together with similarity classifier classification accuracy of 70.25% was achieved with liver-disordered data and 98.19% accuracy was achieved with breast cancer data. Compared to results with traditional PCA and similarity classifier about 4% higher accuracy was achieved with liver-disorder data and about 0.5% higher accuracy was achieved with breast cancer data.

Key words: Similarity classifier; Fuzzy robust PCA; Breast cancer data; Liver-disorder data; Dimension reduction;

1 Introduction

Many databases that come from the real world are admittedly coupled with noise. Noise is a random error or variance of a measured variable [1]. Data analysis is almost always burdened with uncertainty of different kinds. There are several different techniques to deal with noisy data [2].

A major problem in mining scientific data sets is that the data is often high dimensional. In many cases there is a large number of features representing the object. One problem is that the computational time for the pattern recognition algorithms can become prohibitive, when the number of dimensions grows high. This can be
a severe problem, especially when some of the features are not discriminatory. Besides the computational cost, irrelevant features may also cause a reduction in the accuracy of some algorithms.

To address this problem of high dimensionality, a common approach is to identify the most important features associated with an object, so that further processing can be simplified without compromising the quality of the final results. There are several different ways in which the dimension of a problem can be reduced. The simplest approach is to identify important attributes based on the input from domain experts. Another commonly used approach is Principal Component Analysis (PCA) [3], which defines new attributes (principal components or PCs) as mutually-orthogonal linear combinations of the original attributes. For many data sets, it is sufficient to consider only the first few PCs, thus reducing the dimension. PCA can be used also in classification as a preprocessing method. Advantages of doing this usually include the reduced dimension and hence lowered computational cost and increased classification accuracy. This is of course data dependent and these advantages are not always achieved.

Here newer principal component analysis algorithms [4] are investigated as preprocessing algorithms in task of classification, which is based on similarity classifier [5], [6]. Traditional PCA together with similarity classifier was used in [6] and PCA for fuzzy data [7] was used together with similarity classifier in [8] when data was formed from linguistics attributes. In the same way as the notion of a fuzzy subset generalizes that of the classical subset, the concept of similarity can be considered as a many-valued generalization of the classical notion of equivalence [9]. As an equivalence relation is a familiar way to classify similar objects, fuzzy similarity is an equivalence relation that can be used to classify multi-valued objects. Due to this property, it is suitable for classifying problems that can be classified based on clustering by finding similarities in objects. There is also a close link between the notion of similarity and that of distance (see for example [10] and [11]).

Data sets used in this experiment were taken from a UCI-Repository of Machine Learning Database [12]. The liver-disorder data set mainly consists of blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. We also tested our classifier to breast cancer data set [13]. There nuclear size, shape, and texture features are used to distinguish benign from malignant breast cytology. Classifier and preprocessing methods were implemented with MATLAB™-software.

2 Data Sets

Liver-disorder data set: This data was donated by R. S. Forsyth to UCI machine learning data repository [12]. The problem is to predict whether or not a male pa-
tient has a liver-disorder based on blood tests and alcohol consumption. The attribute information for liver-disorder data set is the following 1) mean corpuscular volume 2) alkaline phosphatase 3) alanine aminotransferase 4) aspartate aminotransferase,5) gamma-glutamyl transpeptidase 6) number of half-pint equivalents of alcoholic beverages drunk per day. The first five variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption.

**Breast-Cancer data set:** This data set was created by Dr. Wolberg and the purpose was to accurately diagnose breast masses based solely on Fine Needle Aspiration (FNA). He identified nine visually assessed characteristics of an FNA sample which he considered relevant to the diagnosis. The resulting data set is well-known as the Wisconsin Breast Cancer Data. The nine variables used to predict benign or malignant cases were: 1) Clump Thickness 2) Uniformity of Cell Size 3) Uniformity of Cell Shape 4) Marginal Adhesion 5) Single Epithelial Cell Size 6) Bare Nuclei 7) Bland Chromatin 8) Normal Nucleoli and 9) Mitoses.

### 3 Classification Procedure

Here we propose a classification method where data is first preprocessed using fuzzy robust principle component analysis algorithms [4] to get the data in more feasible form and then it is classified using the similarity classifier [5],[6]. In both data sets data was splitted half. One half was used for training and one half for testing. This procedure was repeated randomly 30 times and mean classification accuracies were computed.

#### 3.1 Fuzzy Robust Principle Component Analysis

Fuzzy robust principle component analysis algorithms used here were introduced in [4]. The robust principle component algorithms which Yang & Wang proposed in [4] are basically based on Xu & Yuilles algorithms [14] were PCA learning rules are related to energy functions and they proposed objective function with the consideration of outliers. In Yang & Wang’s proposed methods they extented objective function to be fuzzy and they includes Xu & Yuilles algorithms as a crisp special cases. Next the methods are shortly presented. More thorough description can be found in [14] and [4].

Xu and Yuille [14] proposed an optimization function, subject to $u_i \in \{0, 1\}$ as
\[ E(U, w) = \sum_{i=1}^{n} u_i e(x_i) + \eta \sum_{i=1}^{n} (1 - u_i) \]  

where \( X = \{x_1, x_2, \ldots, x_n\} \) is the data set and \( U = \{u_i|i = 1, \ldots, n\} \) is the membership set. \( \eta \) is the threshold. The goal is to minimize \( E(U, w) \) with respect to \( u_i \) and \( w \). Notice here that \( u_i \) is binary variable and \( w \) is continuous variable which makes optimization hard to solve with gradient descent approach. To do this they transformed the minimization problem to maximization of the Gibbs distribution of following form:

\[ P(U, w) = \frac{\exp(-\gamma E(U, w))}{Z} \]

where \( Z \) is the partition function ensuring \( \sum_{U} \int_{w} P(U, w) = 1 \). The measure \( e(x_i) \) could be one of the following functions:

\[ e_1(x_i) = ||x_i - w^T x_i w||^2 \] \[ e_2(x_i) = ||x_i||^2 - \frac{||w^T x_i||^2}{||w||^2} = x_i^T x_i - \frac{w^T x_i x_i^T w}{w^T w} \]

The gradient descent rules for minimizing \( E_1 = \sum_{i=1}^{n} e_1(x_i) \) and \( E_2 = \sum_{i=1}^{n} e_2(x_i) \) are

\[ w^{new} = w^{old} + \alpha_t[y(x_i - u) + (y - v)x_i] \] \[ w^{new} = w^{old} + \alpha_t \left( x_i y - \frac{w^T w y^2}{w^T w} \right) \]

where \( \alpha_t \) is the learning rate, \( y = w^T x_i \), \( u = yw \) and \( v = w^T u \).

Yang & Wang proposed an objective function as:

\[ E = \sum_{i=1}^{n} u_i^{m_1} e(x_i) + \eta \sum_{i=1}^{n} (1 - u_i)^{m_1} \]

subject to \( u_i \in [0, 1] \) and \( m_1 \in [1, \infty) \). Now \( u_i \) being the membership of \( x_i \) belonging to the data cluster and \((1 - u_i)\) is the membership of \( x_i \) belonging to the noise cluster and \( m_1 \) being the so called fuzziness variable. Now \( e(x_i) \) measures the
error between $x_i$ and the class center. Notice the similar idea as in fuzzy C-means algorithm [15].

Since now $u_i$ is a continuous variable the difficulty of mixture of discrete and continuous optimization can be avoided and gradient descent approach can be used. First gradient of $E$ (7) is computed with respect to $u_i$. By setting $\delta E/\delta u_i = 0$, we get

$$u_i = \frac{1}{1 + (e(x_i)/\eta)^1/(m_1-1)}$$

(8)

Substituting this membership back and after simplification, one gets

$$E = \sum_{i=1}^{n} \left( \frac{1}{1 + (e(x_i)/\eta)^1/(m_1-1)} \right)^{m_1-1} e(x_i)$$

(9)

The gradient with respect to $w$ is

$$\frac{\delta E}{\delta w} = \left( \frac{1}{1 + (e(x_i)/\eta)^1/(m_1-1)} \right)^{m_1} \left( \frac{\delta e(x_i)}{\delta w} \right)$$

(10)

Now let

$$\beta(x_i) = \left( \frac{1}{1 + (e(x_i)/\eta)^1/(m_1-1)} \right)^{m_1}$$

where $m_1$ is called a fuzziness variable. If $m_1 = 1$, the fuzzy membership reduces to the hard membership and could be determined by the following rule:

$$u_i = \begin{cases} 1 & \text{if } e(x_i) < \eta, \\ 0 & \text{otherwise} \end{cases}$$

Now $\eta$ being a hard threshold in this situation. There is no general rule for the setting of $m_1$, but most papers set $m_1 = 2$. Yang & Wang derived following three algorithms for optimization procedure

**FRPCA1 algorithm:**

Step 1 Initially set the iteration count $t = 1$, iteration bound $T$, learning coefficient $\alpha_0 \in (0, 1]$, soft threshold $\eta$ to a small positive value and randomly initialize the weight $w$.

Step 2 While $t$ is less than $T$, do steps $3 - 9$. 


Step 3 Compute $\alpha_t = \alpha_0(1 - t/T)$, set $i = 1$ and $\sigma = 0$

Step 4 While $i$ is less than $n$, do steps 5 – 8.

Step 5 Compute $y = w^T x_i$, $u = yw$ and $v = w^T u$.

Step 6 Update the weight: $w^{new} = w^{old} + \alpha T \beta(x_i) [y(x_i - u) + (y - v)x_i]$.

Step 7 Update the temporary count $\sigma = \sigma + e_1(x_i)$.

Step 8 Add 1 to $i$.

Step 9 Compute $\eta = (\sigma/n)$ and add 1 to $t$.

**FRPCA2 algorithm:** The same as FRPCA1 except steps 6 – 7:

Step 6 Update the weight: $w^{new} = w^{old} + \alpha T \beta(x_i) \left( x_iy - \frac{w}{w^Tw}y^2 \right)$.

Step 7 Update the temporary count: $\sigma = \sigma + e_2(x_i)$.

**FRPCA3 algorithm:** The same FRPCA1 except steps 6 – 7:

Step 6 Update the weight: $w^{new} = w^{old} + \alpha T \beta(x_i) \left( x_iy - wy^2 \right)$.

Step 7 Update the temporary count: $\sigma = \sigma + e(x_i)$.

In FRPCA3 this weight updating rule is called one-unit Oja’s algorithm [16] and in FRPCA3 $e(x_i)$ is replaced by $e_1(x_i)$ or $e_2(x_i)$ separately.

After the data is first preprocessed with FRPCA algorithms, the resulted new data is then classified using the similarity classifier [5], [6]. Next short description of this procedure.

### 3.2 Similarity classifier

The problem of classification is basically one of partitioning the attribute space into regions, one region for each category. Ideally, one would like to arrange this partitioning so that none of the decisions are ever wrong [17].

We would like to classify a set $X$ of objects into $N$ different classes $C_1, \ldots, C_N$ by their attributes. We suppose that $t$ is the number of different kinds of attributes $f_1, \ldots, f_t$ that we can measure from objects. We suppose that the values for the magnitude of each attribute is normalized so that it can be presented as a value between $[0, 1]$. So, the objects we want to classify are basically vectors that belong to $[0, 1]^t$.

First one must determine for each class the ideal vector $v_i = (v_i(f_1), \ldots, v_i(f_t))$ that represents class $i$ as well as possible. This vector can be user defined or calculated from some sample set $X_i$ of vectors $x = (x(f_1), \ldots, x(f_t))$ which are known to belong to class $C_i$. We can use e.g. the generalized mean for calculating $v_i$, which
is
\[ v_i(r) = \left( \frac{1}{\# X_i} \sum_{x \in X_i} x(f_r)^{m_2} \right)^{\frac{1}{m_2}}, \quad \forall r = 1, \ldots, t \] (11)

where power value \( m_2 \) (coming from the generalized mean) is fixed for all \( i, r \).

Once the ideal vectors have been determined, then the decision to which class an arbitrarily chosen \( x \in X \) belongs to is made by comparing it to each ideal vector. The comparison can be done e.g. by using similarity in the generalized Łukasiewicz structure

\[ S(x, v) = \left( \frac{1}{t} \sum_{r=1}^{t} w_r \left( 1 - |x(f_r)^p - v(f_r)^p| \right)^{m_2/p} \right)^{1/m_2}, \] (12)

for \( x, v \in [0,1]^t \). Here \( p \) is a parameter coming from generalized Łukasiewicz structure [5] (if \( p = 1 \) equation becomes again to its 'normal' form which holds in 'normal' Łukasiewicz structure or just simply Łukasiewicz structure) and \( w_d \) is a weight parameter so that different weights can be given for different attributes to emphasize their importance if it seems appropriate. In this study weights were set as one. The similarity measure has a strong mathematical background [18], [10] and has proven to be a very efficient measure in classification [6]. We decide that \( x \in C_i \) if

\[ S(x, v_i) = \max_{i=1,\ldots,N} S(x, v_i). \] (13)

There are several reasons for why Łukasiewicz structure is chosen in defining memberships of objects. One reason is that in the Łukasiewicz structure, it holds that the mean of many fuzzy similarities is still a fuzzy similarity [19]. Secondly, the Łukasiewicz structure also has a strong connection to first-order logic [20] which is a well studied area in modern mathematics. Thirdly, it also holds the fact that any pseudo-metric induces fuzzy similarity on a given non-empty set \( X \) with respect to the Łukasiewicz conjunction [18]. Good sources of information about the Łukasiewicz structure can be found in [21] and [22].

4 Experimental results and comparison

Liver-disorder data set: Classification results of liver-disorder data set are collected in Table 1. In Figure 1 the effect of \( m \) and \( p \) values in similarity measure are studied with respect to classification accuracy. From Figure 1 one can see that the area where suitable \( m \) and \( p \) values should be chosen is between \( m_2 = (0.8] \) and \( p = [1.10] \). In Figure 2 one can see how reducing the dimension with robust fuzzy principle component analysis algorithms effects the classification accuracy. As can be seen from the Figure 2 best results was found with all three cases when projection was made with FRPCA algorithms and six dimensions were used. Parameter which effected most to classification accuracy in FRPCA algorithms was fuzziness.
variable $m_1$. When other parameters in FRPCA algorithms were varied there was no significant difference in mean classification accuracies. The effects of changing the fuzziness variable $m_1$ is studied in Figure 3. Generally when $m_1 = 1$ we are considering crisp case and we make crisp division in objective function and larger the $m_1$ value the fuzzier the division. As can be seen from the figure, most suitable $m_1$ values seemed to be around $m_1 = 1.75$ with all three cases. Best combination for classification of liver disorder data set seemed to be FRPCA3 and similarity classifier. There mean accuracy was 70.25% and variance 0.0110. This is about 7% higher accuracy when we compare the results achieved when original data was used and about 4% higher classification accuracy compared to results, where first traditional PCA was used and then classified with similarity classifier. When results with FRPCA3 are compared to FRPCA1 and FRPCA2 one can see that they are about 2% better with FRPCA3.

[Table 1 about here.]

[Fig. 1 about here.]

[Fig. 2 about here.]

[Fig. 3 about here.]

**Breast cancer data set:** In Table 2 results with breast cancer data set are reported. In Figure 4 effects of parameter values $p$ and $m_2$ in similarity measure to classification accuracy is studied. As can be seen from the Figure 4 $p$ value should be chosen between $p = (0 \ 10]$. Mean value $m_2$ does not seem to effect so much to classification accuracy and any value between $m_2 = (0 \ 20]$ can be chosen as long as $p$ value is suitable.

In Figure 5 one can see how reducing the dimension with robust fuzzy principle component analysis algorithms effects the classification accuracy. As can be seen from the Figure 5 better results could be found with lower dimension and with FRPCA1 only two dimensions was needed for the highest mean classification accuracy of 98.19% variance being 0.0176. With FRPCA2 and FRPCA3 dimension where highest classification accuracies where reached was 6 in both cases and mean accuracies were 98.13% with FRPCA2 and 98.16% with FRPCA3. There was no significant differences in classification accuracies using similarity classifier and three FRPCA algorithms with this data set. When compared to original data 0.7% higher mean classification accuracy was achieved when data was first preprocessed with FRPCA1. When mean accuracies were compared with pairwise tw-tailed t-test it indicated that means are significantly different in 95% confidence interval. Suitable value for fuzziness variable $m_1$ was also studied and results from this can be seen in Figure 6. Best accuracies were reached when $m_1$ values was chosen between $m_1 = [1.5 \ 1.8]$.

[Table 2 about here.]
5 Discussion

In this article classification method was studied were data is first preprocessed using fuzzy robust principle component analysis (FRPCA) algorithms to get data into more feasible form and then classified using similarity classifier. Fuzzy robust principle component analysis algorithms used here seem to be more robust and better classification accuracies were achieved with chosen data sets which were breast cancer data and liver-disorder data. Results were compared to results achieved using traditional PCA and similarity classifier and also to results where original data was used and classified using similarity classifier. From comparison one can see that better results were achieved using these FRPCA algorithms. With liver-disorder data set accuracy of 70.25% was reach and with breast cancer data set accuracy of 98.19% was achieved. Noticable was also that only two dimensions were needed in order to achieve this 98.16% mean classification accuracy. When traditional PCA was used nine dimensions where needed. When achieved classification accuracies are compared to results with traditional PCA and similarity classifier about 4% higher accuracy was achieved with liver-disorder data and about 0.5% higher accuracy was achieved with breast cancer data.

References


List of Figures

1  a) Mean classification results and b) variances with liver-disorder data plotted with respect to $p$ and mean values $m_2$ in similarity measure. 12

2  a) and b) Mean classification accuracies and variances studied with respect to reduced dimension and comparison of FRPCA algorithms with liver-disorder data set. Fuzziness variable is set here as $m_1 = 1.75$ 13

3  Study of the effect of fuzziness parameter $m_1$ to classification accuracy with liver-disorder data set. 14

4  a) Mean classification results and b) variances with breast cancer data set plotted with respect to $p$ and mean values $m_2$ in similarity measure. 15

5  a) and b) Mean classification accuracies and variances studied with respect to reduced dimension and comparison of FRPCA algorithms with breast cancer data set. Fuzziness variable is set here as $m_1 = 1.75$ 16

6  Study of the effect of fuzziness parameter $m_1$ to classification accuracy with breast cancer data set. 17
Fig. 1. a) Mean classification results and b) variances with liver-disorder data plotted with respect to $p$ and mean values $m_2$ in similarity measure.
Fig. 2. a) and b) Mean classification accuracies and variances studied with respect to reduced dimension and comparison of FRPCA algorithms with liver-disorder data set. Fuzziness variable is set here as $m_1 = 1.75$. 

![Image of graphs](image-url)
Fig. 3. Study of the effect of fuzziness parameter $m_1$ to classification accuracy with liver-disorder data set.
Fig. 4. a) Mean classification results and b) variances with breast cancer data set plotted with respect to $p$ and mean values $m_2$ in similarity measure.
Fig. 5. a) and b) Mean classification accuracies and variances studied with respect to reduced dimension and comparison of FRPCA algorithms with breast cancer data set. Fuzziness variable is set here as $m_1 = 1.75$.
Fig. 6. Study of the effect of fuzziness parameter $m_1$ to classification accuracy with breast cancer data set.
List of Tables

1. Classification results for liver-disorder data set. In first column used method is listed in second column there is number of dimensions used, in third column mean classification accuracy (in %) is given and in fourth column variances are reported

2. Classification results for breast cancer data set. In first column used method is listed in second column there is number of dimensions used, in third column mean classification accuracy (in %) is given and in fourth column variances are reported
Table 1
Classification results for liver-disorder data set. In first column used method is listed in second column there is number of dimensions used, in third column mean classification accuracy (in %) is given and in fourth column variances are reported

<table>
<thead>
<tr>
<th>Method</th>
<th>Dimen.</th>
<th>Mean</th>
<th>Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver\textsubscript{original}</td>
<td>6</td>
<td>63.09</td>
<td>0.0055</td>
</tr>
<tr>
<td>Liver\textsubscript{PCA}</td>
<td>6</td>
<td>66.50</td>
<td>0.0073</td>
</tr>
<tr>
<td>Liver\textsubscript{FRPCA}1</td>
<td>6</td>
<td>68.25</td>
<td>0.0080</td>
</tr>
<tr>
<td>Liver\textsubscript{FRPCA}2</td>
<td>6</td>
<td>67.88</td>
<td>0.0081</td>
</tr>
<tr>
<td>Liver\textsubscript{FRPCA}3</td>
<td>6</td>
<td>70.25</td>
<td>0.0110</td>
</tr>
</tbody>
</table>
Table 2
Classification results for breast cancer data set. In first column used method is listed in second column there is number of dimensions used, in third column mean classification accuracy (in \%) is given and in fourth column variances are reported

<table>
<thead>
<tr>
<th>Method</th>
<th>Dimen.</th>
<th>Mean</th>
<th>Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>$BC_{original}$</td>
<td>9</td>
<td>97.49</td>
<td>0.0350</td>
</tr>
<tr>
<td>$BC_{PCA}$</td>
<td>9</td>
<td>97.72</td>
<td>0.0027</td>
</tr>
<tr>
<td>$BC_{FRPCA1}$</td>
<td>2</td>
<td>98.19</td>
<td>0.0176</td>
</tr>
<tr>
<td>$BC_{FRPCA2}$</td>
<td>6</td>
<td>98.13</td>
<td>0.0060</td>
</tr>
<tr>
<td>$BC_{FRPCA3}$</td>
<td>6</td>
<td>98.16</td>
<td>0.0070</td>
</tr>
</tbody>
</table>