PCA for fuzzy data and similarity classifier in building recognition system for post-operative patient data

Pasi Luukka *

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

Abstract

In this article we propose a method which tackles a problem where data is linguistic instead of real valued numbers. The proposed method starts with representing data as fuzzy numbers. Then generalized principle component analysis (PCA) is used, which can be used to reduce the data dimensionality and also to clear out some irregularities from the data. After this, the data is defuzzified and then the similarity classifier is used to get the required classification accuracy. Here post-operative patient data set is used to build this expert system to determine based on hypothermia condition, whether patients in a post-operative recovery area should be sent to Intensive Care Unit, general hospital floor or go home. What makes this task particularly difficult is that most of the measured attributes have linguistic values (e.g. stable, moderately stable, unstable, etc.). Results are compared to existing result in literature and this system provides mean classification accuracy of 62.7% whereas second highest reported results are with linguistic hard C-mean with 53.3%.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Similarity classifier; PCA for fuzzy data; Post-operative patient data; Linguistic attributes; Medical diagnostic

1. Introduction

Data can be represented in many different forms. In simplest form data reflects whether a certain characteristic property is present or absent. Then it is usually represented with binary values such as +1 and −1, or 0 and 1. Property can also have some grades or other gradual evaluations and then the values are usually expressed by natural numbers. When considering observation and measurements the result in concrete situations is usually given by a real number or by a set of real numbers i.e. by a vector. It can also be presented as a space of functions (e.g. by Fourier coefficients) or e.g. in the image analysis it can be a suitable set of matrices. Opinions of experts or of panels of experts can be also interpreted as data. Opinions can be uttered as statements, rules or conclusions (Bandemer & Näther, 1992). The appropriate universe is then given by a suitable set of statements, rules and conclusions as possible values for the “choice” of the opinions of the experts in question. When we go from the simple representation to more complex representation of data the uncertainty of the increases.

In this paper we have tackled the situation where the problem is to determine where patients in a post-operative recovery area should be sent to next. This data set is freely available in UCI machine learning data repository (Newman, Hettich, Blake, & Merz, 1998). What makes this problem particularly interesting and difficult is that here data is represented with linguistic statements such as that patients internal temperature can be e.g. low, mid or high. Uncertainty in this case stems for linguistic attributes but also from the fact that data is collected from patients and humans are not exactly alike. So in equal or comparable situation the data collected does not show identical states if we try to measure the data too precisely. Because of this, the linguistic representation seems to be a better choice. This kind of uncertainty is typically found with data taken from living beings and reflects then the rich variability of nature.

Data analysis is almost always burdened with uncertainty of different kinds. Possibly the uncertainty is even
increased by the method of analysis applied (Bandemer & Näther, 1992). Fuzzy logic is a logic that allows vagueness, imprecision and uncertainty. Zadeh (1965) introduced the concept of fuzzy sets and the respective theory that can be regarded as the extension of the classical set theory. One of the fundamental mathematical constructs is the similarity measure. In the same way as the notion of the fuzzy subset generalizes that of the classical subset, the concept of similarity can be considered as being a multi-valued generalization of the classical notion of equivalence (Zadeh, 1971).

Here we propose a method which addresses linguistic attributes as fuzzy numbers. Post-operative patient data is represented as fuzzy numbers generalized PCA for fuzzy data (Denoeux & Masson, 2004) is used to reduce data dimensionality and to clear unwanted irregularities. After this defuzzification is carried out and the results are then inputted into similarity classifier (Luukka & Leppälampi, 2006) and classifier is giving the classification accuracy of how well the desired decision was being able to make.

Rest of the paper is organized as follows: In Section 2, we present problem and the data set used in more detail. In Section 3 proposed method is described in detail and also how different parts link together is explain. In Section 4 experimental results are presented and also comparison to other results in literature is given. In last section discussion is given.

2. Post-operative patient data set

Post-operative patient data set is freely available in Newman et al. (1998). Creators of this data set are Sharon Summers, School of Nursing, University of Kansas Medical Center and Linda Woolery, School of Nursing, University of Missouri. Donor of the data set is Jerzy W. Grzymala-Busse. In this data problem is to determine where patients in a post-operative recovery area should be sent to next. Because hypothermia is a significant concern after surgery, the attributes correspond roughly to body temperature measurements. There are 90 samples in this data and seven linguistic attributes:

1. L-CORE (patient’s internal temperature): (high (>37), mid (≥ 36 and ≤ 37), low (<36))
2. L-SURF (patient’s surface temperature): (high (>36.5), mid (≥ 36.5 and ≤ 35), low (<35))
3. L-O2 (oxygen saturation in %): (excellent (≥ 98), good (≥ 90 and <98), fair (≥ 80 and <90), poor (<80))
4. L-BP (last measurement of blood pressure): high (>130/90), mid (130/90 and ≥ 90/70), low (<90/70)
6. CORE-STBL (stability of patient’s core temperature): (stable, mod-stable, unstable)
7. BP-STBL (stability of patient’s blood pressure): (stable, mod-stable, unstable)

In addition there is also one numerical attribute named COMFORT, which is patient’s perceived comfort at discharge, measured as an integer between 0 and 20. Then there is also label information decision whether patient is sent to intensive care unit, prepared to go home or be sent to general hospital floor. In Table 1 linguistic values of each attribute is given.

What makes this particular classification task difficult is that data is linguistic instead of real valued numbers and what we need is tools to handle this type of data. When data is linguistic and comes from experts knowledge it is bound to be burdened with uncertainty of different kinds. In this type of medical data uncertainty, vagueness and imprecision are always present.

3. Proposed method

Here we propose a method which is capable of handling this type of data. First we use PCA for fuzzy data (Denoeux & Masson, 2004) to project the data into a lower dimension to get rid of some possible noise the data has and then defuzzify it and then use similarity classifier to get the decision of what one should do for the patient.

3.1. PCA for fuzzy data

In Denoeux and Massons article (Denoeux & Masson, 2004) principle component analysis of fuzzy data using autoassociative neural networks were considered. This method is also used here as first step to preprocess post-operative patient data. Their method extents principle component analysis to using fuzzy trapezoidal numbers as input values. These fuzzy numbers are propagated through the linear autoassociative neural network, using
fuzzy arithmetics, and the weights are adjusted to minimize a suitable error criterion, the inputs being taken as target outputs. Their method proposes to extend PCA to a wider class of data comprising real numbers, real intervals and, more generally, fuzzy numbers. Fuzzy numbers are defined as fuzzy sets of the real line whose \( \alpha \)-cuts are closed intervals (Dubois & Prade, 1988).

In their approach data is presented as \( d \) dimensional vectors of fuzzy numbers and they defined “generalized PCA” which can handle fuzzy numbers or, more closely, trapezoidal fuzzy numbers. As data is considered as fuzzy numbers also the output is given as fuzzy numbers. There the projection into a lower dimension is gained using the hidden layer in the three layer network. Next the method is described a bit more in detail.

We assume to have a collection of \( n \) objects described by \( d \) attributes taking values in the set \( F(\mathbb{R}) \) of fuzzy numbers. Data is taken in form \( x_1, \ldots, x_n \) where each \( x_r \in F(\mathbb{R}) \) is a vector of \( d \) fuzzy numbers. The objective of PCA of fuzzy data is to compress this data into a lower dimensional fuzzy data \( y_1, \ldots, y_n \), with \( y_r \in F(\mathbb{R}) \), \( r = 1, \ldots, n \), and \( t < d \). For this purpose autoassociative neural network implementation is used where PCA is generalized to the case of fuzzy data.

Let us consider the three layer network where a vector \( x \) of \( d \) fuzzy numbers is fed into the input layer. Let \( A \) be the \( t \times d \) matrix of input-to-hidden weights and \( B \) be the \( d \times t \) matrix of hidden-to-output weights. The hidden and output units are assumed to have identity transfer functions, so that output \( z \) is computed from the input vector \( x \) as

\[
z = BA^T x
\]

(1)

this network is trained in autoassociative mode, i.e., using the inputs as target outputs, and information is stored in the network by virtue of modifications made to the weight of the network. Autoassociative case may be briefly summarized as follows. Let \( E(A, B) \) be quadratic error function, defined as

\[
E(A, B) = \sum_{j=1}^{n} e(x_r, z_r)
\]

(2)

where \( e(x_r, z_r) \) denotes the reconstruction error for pattern \( r \).

\[
e(x_r, z_r) = \|x_r - z_r\|^2
\]

(3)

The total error may also be expressed as a function of the global map \( W = BA \), which is the constrained to be at most of rank \( t \). It is also true that \( E(A, B) = E(CA, BC^{-1}) \) for any invertible \( t \times t \) matrix \( C \). Let \( S \) now denote the sample covariance matrix of the data and \( u_1, \ldots, u_d \) are orthonormal eigenvectors. Thus, error, \( E \) expressed as a function of global map \( W \) has a minimum of the form \( W = BA \) where

\[
A = CU^T
\]

(4)

and

\[
B = U^T C^{-1}
\]

(5)

The optimal map \( W = U_r^T U_r \) is thus the orthogonal projection \( P_r \) onto the subspace \( L \) spanned by the first eigenvectors of the data covariance matrix \( S \). PCA is recovered as a special case when \( C \) is the identity matrix. In this case the activities in the hidden layer are exactly identical to the principal components of the data. If the error function is minimized by an iterative algorithm such as backpropagation, this particular solution is however, generally not obtained, and matrix \( C \) is arbitrary. A way to resolve this ambiguity is to introduce the constraint \( A^T = B \), which given (4) and (5) translates to

\[
U_r^T C = U_r C^{-1}
\]

(6)

Since the \( (u_1, \ldots, u_t) \) form an orthogonal basis, we have \( U_r^T U_r = I_t \), where \( I_t \) is identity matrix. By left multiplying both sides of (6) by \( U_r \), we see that \( C \) verifies \( C = C^{-1} \). Hence \( C \) is now an orthogonal matrix, which implies that the hidden unit activities and the principal components are related. A simple way to impose constraint \( A^T = B \) is to rewrite the propagation equation (1) as

\[
z = BA^T x
\]

(7)

which translates in scalar notation to

\[
z_k = \sum_{j=1}^{t} B_{kj} \sum_{i=1}^{d} B_{ij} x_{i}, \quad k = 1, \ldots, d
\]

(8)

Now when using fuzzy numbers as inputs the network output may be computed by applying Zadeh’s extension principle to (8). The \( k \)th component \( z_k \) of the fuzzy output vector \( z \) for fuzzy input \( x \) is defined as

\[
\forall u \in R, \mu_{z_k}(u) = \sup_{x_1, \ldots, x_d} \min \{ \mu_{x_i}(v_i) \}
\]

(9)

the supremum being taken under the constraint

\[
u = \sum_{j=1}^{t} B_{kj} \sum_{i=1}^{d} B_{ij} v_i
\]

(10)

where addition and multiplication by a real are now the usual operations of fuzzy arithmetics (Dubois & Prade, 1988). For more about practical calculations (see Denoeux & Masson, 2004).

### 3.2. Defuzzification

As previously described methods starts by representing the data with fuzzy numbers. In Appendix A one can find how data is described by using trapezoidal fuzzy numbers. Then data is fed into the generalized PCA and it is used to compress this data into lower dimensional fuzzy data. Next we are faced with dilemma of how these lower dimensional fuzzy numbers can be inputed into the similarity classifier which requires real valued data as inputs. This is now done by defuzzification. There are several different defuzzification methods (Klir & Yuan, 1995), most common are probably Center of Area Method (COA), Bisection of Area Method (BOA) and Mean of Maxima (MOM) method.
Here Mean of Maxima was used and next presented in a bit more detailed.

Let \( \mu^{-}_M(x) \) denote the membership function of the output fuzzy number.

**Mean of Maxima Method (MOM)**

In this method the defuzzified output value \( y_{MOM} \) is defined as the average of the smallest and the largest point where the output fuzzy set \( \mu^{-} \) gets its height \( h(\mu^{-}_M) \). The formula of this method gets following form

\[
y_{MOM} = \frac{\inf M + \sup M}{2}, \tag{11}
\]

where

\[
M = \{ x \in C \mid \mu^{-}_M(x) = h(\mu^{-}_M) \}
\]

In the discrete case formula (11) takes the form

\[
y_{MOM} = \frac{\min\{x_i \mid i \in M\} + \max\{x_i \mid i \in M\}}{2}, \tag{13}
\]

where

\[
M = \{ x_i \mid \mu^{-}_M(x_i) = h(\mu^{-}_M) \}
\]

There are also other variants of this method where the smallest point of maximum or the biggest point of maximum is selected to the output of the system. These methods are called

- Smallest of Maximum (SOM)
- Largest of Maximum (LOM)

In Fig. 1 there is an example where the defuzzified output of the system is calculated. In the Figure centroid means COA method and bisector means BOA method. Now when we have defuzzified outputs of the data we are then ready to input this into the classifier.

### 3.3. 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 (Duda & Hart, 1973).

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

\[
v_i(r) = \left( \frac{1}{|X_i|} \sum_{x \in X_i} x(f_i)^m \right)^{1/r}, \quad \forall r = 1, \ldots, t \tag{15}
\]

where power value \( m \) 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

\[
S(x, v) = \left( \frac{1}{t} \sum_{r=1}^{t} w_r \left( 1 - |x(f_i)^m - v(f_i)^m|^{p/m} \right)^{1/m} \right)^{1/p}, \tag{16}
\]

for \( x, v \in [0, 1]^t \). Here \( p \) is a parameter coming from generalized Łukasiewicz structure (Luukka, Saastamoinen, & Könnönen, 2001) (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 optimized using differential evolution algorithm (Price, Storn, & Lampinen, 2005). The similarity measure has a strong mathematical background (Klawonn & Castro, 1995; Formato, Gerla, & Scarpati, 1999) and has proven to be a very efficient measure in classification (Luukka & Leppälampi, 2006). We decide that \( x \in C_m \) if

\[
S(x, v_m) = \max_{i=1, \ldots, N} S(x, v_i). \tag{17}
\]

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 (Turunen, 1999). Secondly, the Łukasiewicz structure also has a strong connection to first-order logic (Novak, 1990) 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 (Klawonn & Castro, 1995). Good sources of information about the Łukasiewicz structure can be found in Turunen (2001), Łukasiewicz (1970)

3.4. Novelty of proposed method

Simpson (1992) listed desirable properties that a pattern classifier should possess. According to Simpson (1992) a successful pattern classifier should be able to:

1. Learn the required task quickly.
2. Learn new data without having to retrain with old data (on-line adaptation).
3. Solve non-linearly separable problems.
4. Provide the capability for soft and hard decisions regarding the degree of membership of the data within each class.
5. Offer explanations of how the data are classified, and why the data are classified as such.
6. Exhibit performance that is independent of parameter tuning.
7. Function without knowledge of the distributions of the data in each class.
8. For overlapping pattern classes, create regions in the space of the input parameters that exhibit the least possible overlap.

If we go into detail of these properties and compare the classifier to the properties listed we can see that the classifier satisfies most of the properties. The method is capable to learn the required task quickly. On-line adaptation is not included but is one area of future work. Classifier is able to solve non-linearly separable problems. It provides a partial membership for each class through the similarity concept but so far capability for both soft and hard decisions is not implemented to the classifier. It offers the explanations of how the data are classified, and why the data are classified as such. As can be seen from the results classifier can exhibit performance that is independent of parameter tuning by just setting suitable p and m values which seems to work well. In classifier algorithm distribution of the data is not needed. For overlapping pattern classes, regions in the space of the input parameters that exhibit the least possible overlap is created by finding the suitable parameter m and p values and by optimizing weights in the similarity measure for each data separately.

4. Experimental results and comparison

The post-operative patient’s data set was first represented in trapezoidal fuzzy numbers and then this data was used in principal component analysis for fuzzy data. Since there was no prior knowledge of what would be the best reduced dimension the dimension reduction was done for all dimensions from two to seven. Then the reduced data set was defuzzified using MOM method and this data was used for the data to be inputed in the classifier. In the classifier data was splitted half. One half for the training and one half for the testing. This procedure for randomly splitting data in half was done 30 times for each seven data

Table 2
Classification results for post-operative patient data set

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Max (%)</th>
<th>Mean (%)</th>
<th>Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>71.11</td>
<td>52.89</td>
<td>0.061</td>
</tr>
<tr>
<td>3</td>
<td>71.11</td>
<td>62.37</td>
<td>0.0013</td>
</tr>
<tr>
<td>4</td>
<td>66.66</td>
<td>60.00</td>
<td>0.0012</td>
</tr>
<tr>
<td>5</td>
<td>68.88</td>
<td>62.67</td>
<td>0.0012</td>
</tr>
<tr>
<td>6</td>
<td>68.88</td>
<td>59.19</td>
<td>0.0091</td>
</tr>
<tr>
<td>7</td>
<td>66.66</td>
<td>54.67</td>
<td>0.0049</td>
</tr>
</tbody>
</table>

In first column there is number of dimensions used, in second column maximum classification accuracy (in %), in third column mean accuracies are reported and for fourth column variances.

Table 3
Classification results comparison for post-operative patient data set

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LERS</td>
<td>48</td>
</tr>
<tr>
<td>HCM1</td>
<td>50</td>
</tr>
<tr>
<td>HCM2</td>
<td>51.1</td>
</tr>
<tr>
<td>LHCM</td>
<td>53.3</td>
</tr>
<tr>
<td>Proposed method</td>
<td>62.7</td>
</tr>
</tbody>
</table>

In first column there is method and in second column classification accuracy.
sets. Weights were optimized with differential evolution algorithm. This was done first time for the similarity classifier in Luukka and Sampo (2004) where one can found more information about it. Also best area for $p$ and $m$ values were found using grid search technique. Classification results using grid search technique can be seen in Fig. 2.

In Table 2 there is reduced dimensions in the first column. Classification accuracies for maximum classification accuracy and for the mean classification accuracy are in second and third column. In the last column variances are reported for each dimension. As can be seen from the table the highest mean accuracy of 62.67% was achieved.

Fig. 3. linguistic attributes presented as fuzzy trapezoidal functions. (a) Patient’s internal temperature (L-CORE), (b) patient’s surface temperature (L-SURF), (c) patient’s oxygen saturation in percentile (L-O2), (d) patient’s last measurement of blood pressure (L-BP), (e) stability of patient’s surface temperature (SURF-STBL), (f) stability of patient’s core temperature (CORE-STBL) and (g) stability of patient’s blood pressure (BP-STBL).
with dimension five. Also variance was only 0.0012 with 99% confidence interval using Student’s t distribution. Classification results from dimension five are also plotted in Fig. 2 with respect to parameter $p$ and $m$ value changes. As can be seen from the figure best accuracies were found when $p$ value was around $p = 8$ and mean value $m$ around $m = -4.8$. Suitable good area for these parameters can be considered to be $p = [5, 10]$ and $m = [-5, -4]$.

Classification results are compared to results presented in literature in Table 3. There results are compared with earlier methods such as Linguistic Hard C-Means (LHCM) (Auephanwiriyakul & Theera-Umpon, 2004), regular Hard C-Mean (HCM) with two variants and to learning system learning from examples based on rough sets (LERS) (Woolery, Grzymala-Busse, Summers, & Budihardjo, 1991). In regular HCM numeric data set was used (HCM1). There linguistic values were mapped as a numeric value e.g. L-COREs value low, was simply $34$, etc. (see more about that in Auephanwiriyakul & Theera-Umpon (2004)). Centroid from fuzzy numbers were used as numerical values in the second case (HCM2).

As can be seen from the results comparison Table 3 method based on pca for fuzzy data and similarity classifier managed with highest accuracy of 62.7% where as second highest mean accuracy was achieved with Linguistic Hard C-Mean with accuracy of 53.3%.

5. Discussion

In this paper a method is proposed which can be used for decision making system for cases where data is linguistic and we need means to work with such data set. The method starts with representing the data in fuzzy numbers. Then dimension reduction is carried out for the data. Third step in this process is defuzzification and then data is input into similarity classifier to get the classification decision. To get the best possible results one needs to find the correct reduced dimension, find the suitable parameter values and weights for the similarity measure. This all was carried out for post-operative patient data.

As can be seen from the comparison part, here proposed method provides most accurate results and it clearly has potential for being used as decision making system for post-operative patient data. Results compare well to existing result in literature and this system provides mean classification accuracy of 62.7% whereas as second highest reported results are with linguistic hard C-mean with 53.3%.

For the future work this method is also to be tested for other problems were collected data is highly heterogenous and has linguistic and missing values.

Appendix A

See Fig. 3.

References


