ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 433-438

Monitoring technical systems with prototype based clustering

Thorsten Bojer[1], Barbara Hammer[2], Christian Koers[1]

(1) Prognost Systems GmbH, R&D, 48432 Rheine, Germany
(2) University of Osnabrück, Department of Mathematics/
Computer Science, Albrechtstraße 28, 49069 Osnabrück, Germany,
e-mail: hammer@informatik.uni-osnabrueck.de

Abstract. We present an application of generalized relevance learning vector quantization (GRLVQ) to the supervision of piston compressors in industry. Thereby, GRLVQ constitutes a prototype-based clustering algorithm with adaptive diagonal metric based on LVQ. In the reported application, further adaptation of the distance measure is necessary in order to allow invariance with respect to small time shifts. Depending on the respective sensors, very good classification results are obtained.

1. Introduction

Neural networks constitute a well-established tool for various application areas in industry such as robotics, medicine, financial engineering, control, and monitoring [3]. Since learning vector quantization (LVQ) as proposed by Kohonen is based on prototypes, it is particularly suited for application areas where prior knowledge is available or humans should gain insight into the classification behavior. Industrial applications of LVQ as well as improvements of the basic learning algorithm are reported in Kohonen's book on self-organizing maps [6].

Being a metric based approach, LVQ is highly sensitive with respect to data representation and the induced Euclidian metric. LVQ fails if the metric is not appropriate which is very likely for high-dimensional or highly heterogeneous data, for example. Hence effort has been done in order to substitute the standard Euclidian metric by data adapted alternatives such as metrics which result in clusters of specific shapes in the case of fuzzy clustering [5] or metrics which take auxiliary information into account in the case of unsupervised processing [7]. Generalized relevance learning vector quantization (GRLVQ) as introduced in the article [4] modifies LVQ such that a very stable approach with adaptive diagonal metric is obtained. We here report an application of GRLVQ to the supervision of piston compressors in industry where heterogeneous data have to be dealt with. Thereby, we introduce a further modification of GRLVQ to account for the temporal structure of the data and to allow identification of prototypes which are invariant with respect to small shifts.

2. GRLVQ with shift invariant distance

GRLVQ learns a prototype based classification given a finite set of training examples. Denote by $X = \{(x^i, y^i) \in \mathbb{R}^n \times \{1, \ldots, C\} \mid i = 1, \ldots, m\}$ the set of

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 433-438

training examples. Denote by $\{w^1, \ldots, w^M\}$ the set of prototypes with label $c^i = c$ assigned to w^i iff prototype w^i belongs to class c. Denote by $\lambda_1, \ldots, \lambda_n$ relevance terms for the input dimensions with $\lambda_i \geq 0$ and $\sum \lambda_i = 1$. GRLVQ aims at minimizing the cost function

$$E = \sum_{i=1}^{m} \text{sgd}\left(\frac{d_J^\lambda(x^i) - d_K^\lambda(x^i)}{d_J^\lambda(x^i) + d_K^\lambda(x^i)}\right)$$

where $\text{sgd}(t) = (1 + \exp(-t))^{-1}$ denotes the logistic function. The term $d_J^\lambda(x^i) = \sum_j \lambda_j (x_j^i - w_j^J)^2$ denotes the squared weighted Euclidian distance of x^i from the nearest prototype w^J with $y^i = c^J$ and $d_K^\lambda(x^i) = \sum_j \lambda_j (x_j^i - w_j^K)^2$ denotes the squared weighted Euclidian distance of x^i from the nearest protoype w^K with $y^i \neq c^K$. E is small if the distance of the data points from the closest correct prototype, w^J, is smaller than the distance from the closest wrong prototype, w^K. Thereby, a weighted Euclidian metric involving relevance terms λ_i is used and the relevance terms λ_i are adapted during learning to achieve a good classification accuracy. The denominator of the summands appropriately scales the respective terms in order to allow stable behavior. Training consists in a stochastic gradient descent on this error function with respect to both, the prototypes w^i and relevance terms λ_i [4]:

$$w^J += \epsilon \cdot \text{sgd}' \cdot \frac{d_K^\lambda(x^i)}{(d_J^\lambda(x^i) + d_K^\lambda(x^i))^2} \cdot \Lambda \cdot (x^i - w^J)$$

$$w^K -= \epsilon \cdot \text{sgd}' \cdot \frac{d_J^\lambda(x^i)}{(d_J^\lambda(x^i) + d_K^\lambda(x^i))^2} \cdot \Lambda \cdot (x^i - w^K)$$

$$\lambda_l -= \epsilon' \cdot \text{sgd}' \cdot \frac{d_K^\lambda(x^i) \cdot (x_l^i - w_l^J)^2 - d_J^\lambda(x^i) \cdot (x_l^i - w_l^K)^2}{(d_J^\lambda(x^i) + d_K^\lambda(x^i))^2}$$

Thereby, ϵ and ϵ' are learning rates, Λ is the diagonal matrix with relevance terms, and the relevance terms are in addition normalized after each step. As pointed out in [4], the derivatives with respect to the prototypes contain the Hebb terms of standard LVQ and additional scaling factors which account for the stability of this approach. Similarly, gradients with respect to the relevance terms contain plausible Hebb terms as introduced in [1] and additional stabilizing factors. Note that this approach can be done online, allowing to deal with huge data sets as well as new training data in a simple way.

The cost function E is usually multimodal. Hence the algorithm is sensitive with respect to initialization of the prototypes and can easily get stuck in local optima. Therefore, we use a dynamic approach. Thereby, one prototype for each class is initialized in the cluster center. Additional prototypes are added dynamically during training if the cost function has reached a plateau and a significant number of errors still exists. The new prototypes are initialized in the center of the misclassified points of the respective class. Monitoring a test error during this procedure is advisable in order to avoid overfitting.

Data points we will deal with consist of several discretized periodic time series which are measured by several sensors within one rotation of the piston rod of a reciprocating compressor. Thereby, the same technical defect might

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 433-438

express to different patterns over several rotations because characteristic features might be subject to slight time-shifts or time-scales. Hence the adaptive diagonal metric is to be further modified. Assume data are contained in a vector space of the form $\mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_m}$ and enumerate the components of a data point by $x = (x_1, \ldots, x_m)$ where $x_i = (x_{i1}, \ldots, x_{in_i})$ describes one time series, i.e. one dimension of the sensor data for one rotation. Invariance with respect to small shifts within x_i can be achieved if we use the following local distance measure for the comparison of two time-series:

$$D_L^i : \mathbb{R}^{n_i} \times \mathbb{R}^{n_i} \to \mathbb{R}, \quad D_L^i(x, y) = \sum_{j=1}^{n_i} 2(L+1)^2 \sum_{k=-L}^{L} \frac{(x_{j+k} - y_j)^2}{1 + 2|k|}$$

where indices are taken modulo n_i. Note that this measure is no longer a metric since it is not symmetric. One could easily use the symmetric function $D_L^i(x,y) + D_L^i(y,x)$, instead; however, this is not necessary for our application. L determines the size of a local window within which values are compared. Note that *all* values of x are compared with one value of y within this window and the distances are linearly weighted with decreasing values towards the borders of the window. Hence time series x and y where characteristic features are subject to a small time shift within the window size have still a small distance from each other. The term $2(L+1)^2$ is introduced for normalization. We now substitute the term $d_J^\lambda(x^i)$ in E by the value

$$\sum_{j=1}^{m} \lambda_j D_L^j(x^i, w^J)$$

where w^J is the closest prototype with the same label as x^i according to this modified distance, and we substitute $d_K^\lambda(x^i)$ by the value

$$\sum_{j=1}^{m} \lambda_j D_L^j(x^i, w^K)$$

where w^K is the closest prototype with a different label than x^i. Prototypes and relevance terms λ_j are adapted with stochastic gradient descent on E. Naturally, alternative optimization methods could be used. Due to the fact that few prototypes are necessary and hence E is a comparably simple function in our application, a simple gradient descent is sufficient for our application.

3. Supervision of piston compressors

Supervision of piston compressors and early fault detection are desirable for several reasons: severe damage can be prevented, idle times of the machines can be minimized, and necessary repairs can be better organized. Large scale industrial piston compressors vary from machine to machine with respect to sensor data. Hence exact modeling of the process would be very time consuming. Adaptive machine learning tools offer a valuable alternative. They can in particular be trained for each machine with low cost and be further adapted to

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 433-438

take the sign of wear of a machine into account. Further demands within this context consist in the fact that machine learning tools should be understandable for experts and it should be possible to assign a confidence to each classification of a given data point. LVQ and variants offer insight into the classification through the prototypes which constitute typical representatives for the several classes and which can be interpreted by experts. A confidence level of the classification can be obtained by standard statistical methods. Alternatively, the distance of the data point from the closest prototype compared to the distance of the data point from the next prototype with different class can serve as confidence level. Note that GRLVQ aims at optimizing this confidence level for the given training set by the choice of E, whereas LVQ tries to optimize the classification accuracy directly. Of course, we could use alternative classification tools such as the SVM for the given classification task. However, one condition for our application is that the classifier can easily be interpreted by humans. In contrast to the SVM, this is easily achieved by prototype based classification. Moreover, LVQ likely shows very good generalization performance being, like SVM, a large margin optimizer [2].

A typical piston compressor is equipped with several sensors which measure pressure and oscillation at a rotation of the piston rod. Data are here monitored in 36 segments of a full rotation and the mean and absolute value is stored for each segment. Another time series is provided by the position of the piston rod. Additional values which are used for classification are global quantities like the various temperatures and characteristic values of the pV-graph. All values are discretized into 5 segments which are determined from the data statistics for each machine. The borders of the segments are chosen in such a way that standard values of the machine roughly fall into segment 3, segments 1 and 2 indicate values which are smaller than the standard value, segments 4 and 5 indicate values which are larger than the standard ones.

Depending on the number of sensors, this preprocessing yields heterogeneous and high dimensional data which decomposes into around 25 time series with 36 values each, 20 analysis yield one additional value, and around 40 additional global features describe the machine. The different input dimensions have a different impact on the fault classification. It can, for example, be expected that the pressure is more important for accurate failure detection than the temperature of the system. Hence metric based approaches like LVQ with the standard Euclidian metric will likely fail because they accumulate noise of data dimensions which are of minor importance for accurate classification. Adaptation of the metric is advisable. We use GRLVQ for the classification because it shares the benefits of LVQ: classification can be interpreted based on the prototypes, and the distance of a data point from the prototypes allows to determine a confidence level of the classification. GRLVQ allows to automatically adapt the metric and hence to overcome the problems caused by the large input dimension. Thereby, the possibility of insight into the classification is preserved since GRLVQ restricts to a diagonal metric such that the factors λ_i can be interpreted as relevance terms for the classification. In addition, GRLVQ shows better stability than LVQ since it constitutes a stochastic gradient descent on the cost function E. In addition, we have to substitute the Euclidian metric for the comparison of time series by the shift-invariant version as described above, because relevant features of the same failure can be found at

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 433-438

expression 1	44444 5554444	444 554
expression 2	44444 55544	444 555
expression 3	444444 444554	4444 4

Table 1: Different expressions of the same failure (sticking valve) within three rotations of the piston rod. Characteristic features of the time series (entries 4 and 5 indicating too large values) are subject to small shifts. The shown time series represent the discretized maximum and mean oscillation of the cylinder. Segments with value 3 (= normal value) are left blank.

slightly shifted positions of the time series. This effect is partially caused by the discretization of all values. As a consequence of this discretization, prototypes are more expressive for humans and easier to process with machine learning tools; however the process is not continuous and makes a slight shift invariance of the metric necessary. See Tab. 1 for an example of the expression of the same failure in two time series and three rotations of the piston rod.

4. Experiments

Data is taken from two 4-cylinder piston compressors with 20 sensors. Data points decompose into 28 time series and 52 additional values. 14 different failures have been observed yielding a total number of 65 training patterns for failures. In addition, 8 data points expressing normal states of the machines have been added. Data is randomly separated into training and test set. The learning task is to predict all classes correctly. In all runs we used the shift invariant metric with time window length $L = 5$ to compare time series. We used dynamic prototype generation, iteratively adding prototypes for classes which contain misclassified points. Training was stopped when a predefined classification accuracy was achieved on the training set (95% for GRLVQ). In this way, we obtained one or two prototypes for each class during training. We trained standard LVQ, standard LVQ with relevance factors set by hand according to expert knowledge, and GRLVQ as presented in this paper. The learning rate for the prototypes was set to 0.1 in all runs, the learning rate for the relevance terms was set to 0.01. Results of 15 runs are reported in Table 2.

We also used LVQ with the standard Euclidian without relevance adaptation and without shift invariance, which achieved a performance of at most 60% on the test sets (results not reported in Tab. 2). LVQ with shift invariant metric achieves a classification accuracy of only about 70% due to the high dimensional data which accumulates noise. If expert knowledge is taken into account and sensors are weighted according to their relevance by experts prior to training, the classification accuracy increases by more than 10%. The result of GRLVQ is even better and, in particular, automatically obtained during training. Thereby, the relevance factors λ_i found by GRLVQ mirror the prior knowledge of the experts: they emphasize signals corresponding to the pressure with high values of λ_i, whereas relevance terms e. g. corresponding to the

LVQ		LVQ + expert metric		GRLVQ	
train (min-max)	test (min-max)	train (min-max)	test (min-max)	train (min-max)	test (min-max)
machine 1:					
69.6 (64.8-71.9)	66.7 (65.3-69.8)	91.6 (89.1-92.4)	81.6 (75.2-83.4)	98.2 (96.3-100)	97.2 (93.5-100)
machine 2:					
72.3 (68.4-74.3)	65.3 (62.5-67.2)	92.1 (88.3-97.2)	84.5 (74.2-86.3)	99.1 (98.4-100)	97.7 (97.6-100)

Table 2: Training and test set accuracy (% correct classification) for two piston compressors and several prototype based learning algorithms in 15 runs.

temperature vanish. Hence results obtained via training the relevance terms yield very good accuracy compared to classification based on relevance terms set by experts because the relevance terms are more precisely tuned for the specific machine and respective classification task in automated relevance adaptation.

5. Discussion

We have presented an application of prototype based neural clustering to monitor technical systems. Thereby, clustering based on the standard Euclidian metric fails due to heterogeneous and high dimensional data involving several time series. Hence we have used an extension of LVQ with adaptive metric, GRLVQ, and we have further extended the basic ingredients of the metric by components which are invariant with respect to small time shifts. The method allows a good classification accuracy of more than 95% for two data sets taken from two piston compressors. The proposed method is part of an industrial system developed for online and offline supervision of piston compressors. The neural classifier greatly reduces the necessity of human interaction in control, while it preserves the possibility of human insight into the classification. Further experiments will extend the method to adaptive relevance terms within the time windows of time series, and adaptive quantization of the measured sensor data, which is so far performed based on data statistics.

References

[1] T.Bojer, B.Hammer, D.Schunk, K.Tluk von Toschanowitz, Relevance determination in learning vector quantization, in: M.Verleysen(ed.), European Symposium on Artificial Neural Networks'2001, D-facto publications, pp.271-276, 2001.

[2] K.Crammer, R.Gilad-Bachrach, A.Navot, and N.Tishby. Margin analysis of the LVQ algorithm, NIPS'2002.

[3] F.Fogelman Soulié, P.Gallinari, Industrial Applications of Neural Networks, World Scientific, 1998.

[4] B.Hammer, T.Villmann, Generalized Relevance Learning Vector Quantization. Neural Networks, 15:1059-1068, 2002.

[5] F.Höppner, F.Klawonn, R.Kruse, T.Runkler, Fuzzy Cluster Analysis, Wiley, 2000.

[6] T.Kohonen, Self-Organizing Maps, Springer, 1997.

[7] J.Sinkkonen, S.Kaski, Clustering based on conditional distributions in an auxiliary space. Neural Computation, 14:217-239, 2002.