ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Learning vector quantization for heterogeneous structured data

Dietlind Zühlke[1] and Frank-Michael Schleif[2] and Tina Geweniger[3] and Sven Haase[4] and Thomas Villmann[4]

1- RWTH Aachen - Information Systems - Life Science Informatics
Ahornstr. 55, D-52056 Aachen - Germany

2- University of Bielefeld - Working Group Computational Intelligence
Universitätsstraße 25, D-33615 Bielefeld - Germany

3- University of Leipzig - Working Group Computational Intelligence
Semmelweisstrasse 10, D-04103 Leipzig - Germany

4- University of Applied Sciences Mittweida - Computational Intelligence
Technikumplatz 17, D-09648 Mittweida - Germany

Abstract. In this paper we introduce an approach to integrate heterogeneous structured data into a learning vector quantization. The total distance between two heterogeneous structured samples is defined as a weighted sum of the distances in the single structural components. The weights are adapted in every iteration of learning using gradient descend on the cost function inspired by Generalized Learning Vector Quantization. The new method was tested on a real world data set for pollen recognition using image analysis.

1 Introduction

In the area of image recognition usually only image features are used for classifying the contents of the image. Regarding the application of image recognition in the medical and biological domain classical image feature based approaches sometimes fail because of the high variability of the manifestations of a phenomenon. Given success model for such classification tasks, is the human expert. An important element of the human successes is to integrate information or features into the classification process that do not directly correspond to gray values in the image. The goal is to find a generalized possibility to incorporate such information into machine learning and technical classification procedures.

1.1 Example - Pollen recognition

One example, where the incorporation of additional features supporting the image based analysis improves the classification performance, is pollen recognition. Building on a newly developed pollen sampling hardware system we gain digital images of pollen probes. After filtering and segmenting these images there are single objects to be classified as either non-pollen objects or pollen of a specific species.

When classifying pollen objects based on image (for example microscopic views) the human expert always incorporates information about the context.

He is e.g. primed to those object classes that were already present in the probe currently under investigation. So he relies his classification of uncertain objects on those objects identified with high reliability and frequency.

To introduce a similar mechanism into our classification process we first train a "normal" classification on image features (i.a. Haralick and Zernicke Features) for a training set. By classifying this training set a threshold for reliable classification (based on the reliability value the classification delivers) is determined. All objects in the training set not exceeding this threshold in classification go into a second classification training step incorporating the distribution of the reliably classified objects on their corresponding probe.

1.2 Incorporating of heterogeneously structured data

How can this classification incorporating heterogeneously structured data look like? One obvious observation is that image features and single relative frequencies are not directly comparable to each other. We need at least a different weighting of the dimensions depending on their structural membership. Furthermore it would be preferable to use different distance metrics for the different structural components. Each one should be best suited for the data structure under investigation. The total distance is then defined as a weighted sum of the distances in every structural component.

But how to choose the weights for the single structural components? If these weights of the components are integrated into a cost function of a learning algorithm, the weights can easily be adapted with respect to minimizing the cost function in a gradient descend approach. In the following sections it is shown how these heterogeneous structures can be incorporated into Generalized Learning Vector Quantization (GLVQ). Furthermore the pollen recognition problem is used as an example.

2 Methods

First we will shortly sketch the fundamentals of Generalized Learning Vector Quantization and then describe the specific characteristics of the heterogeneous structured data incorporation.

2.1 Generalized Learning Vector Quantization

As opposed to most Learning Vector Quantizers that are motivated heuristically, Generalized Learning Vector Quantization [1] was developed to optimize a cost function that approximates the classification error. The resulting update rules are stochastic gradients on this cost function.
Cost function:

$$E = \sum_{k=1}^{n} E_k = \sum_{k=1}^{n} \left(\frac{d^+ (v_k, w_+) - d^- (v_k, w_-)}{(d^+ + d^-)} \right)$$

with n - number of examples, w_+ - Best matching unit correct class, w_- - Best matching unit incorrect class, d distance function.

Weight updates:

$$\Delta w_+ \propto -\frac{\delta E_k}{\delta w_+} = \epsilon_w \cdot \frac{d^-}{\left(d^+ + d^-\right)^2} \cdot (v - w_+)$$

and:

$$\Delta w_- \propto -\frac{\delta E_k}{\delta w_-} = -\epsilon_w \cdot \frac{d^+}{\left(d^+ + d^-\right)^2} \cdot (v - w_-)$$

2.2 Heterogeneous Learning Vector Quantization

Under the assumption that the total distance of two samples in the heterogeneous structured feature space can be represented as the sum of weighted distances in every single structural component (with internal homogeneous weight), we get a new cost function:

$$E = \sum_{k=1}^{n} E_k = \sum_{k=1}^{n} \sum_{j=1}^{D} \alpha_j \cdot E_{jk} = \sum_{k=1}^{n} \sum_{j=1}^{D} \alpha_j \cdot \left(\frac{d_j^+ \left(v_k, w_+\right) - d_j^- \left(v_k, w_-\right)}{\left(d_j^+ + d_j^-\right)} \right)$$

with n - number of examples, D - number of structural components, α_j - weighting parameter of structural component j with $0 \leq \alpha_j \leq 1$ and $\sum_{j=0}^{D} \alpha_j = 1$, w_+ - Best matching unit correct class, w_- - Best matching unit incorrect class, $v_k = \left([v_k]_1, \ldots, [v_k]_D\right)$ - kth sample with the $[v_k]_j$ being different structural components of v_k, $d_j = d\left([v_k]_j, [w_+]_j\right)$ different (not necessary comparable) distance measures (metrics).

The weight updates are obtained as derivatives of the cost function:

$$\Delta [w_+]_j \propto -\frac{\delta E_k}{\delta [w_+]_j} = -\frac{\delta E_{jk}}{\delta [w_+]_j} = -\alpha_j \cdot \epsilon_w \cdot \frac{d_j^-}{\left(d_j^+ + d_j^-\right)^2} \cdot \frac{\delta d_j^+}{\delta [w_+]_j}$$

and:

$$\Delta [w_-]_j \propto -\frac{\delta E_k}{\delta [w_-]_j} = -\frac{\delta E_{jk}}{\delta [w_-]_j} = \alpha_j \cdot \epsilon_w \cdot \frac{d_j^+}{\left(d_j^+ + d_j^-\right)^2} \cdot \frac{\delta d_j^-}{\delta [w_-]_j}$$

For different metrics the derivatives $\frac{\delta d_j^+}{\delta [w_+]_j}$ and $\frac{\delta d_j^-}{\delta [w_-]_j}$ differ accordingly e.g.:

- Euclidean metric:

$$\frac{\delta d_j^-}{\delta [w_-]_j} = -2 \cdot \left([v_k]_j - [w_-]_j\right)$$

- Divergences (z.B. Cauchy-Schwarz-Divergence, see [2] for more detail):

$$\begin{aligned} D_{CS}\left(v, w\right) &= \frac{1}{2} \cdot \log\left(\int v^2\left(x\right) dx \cdot \int w^2\left(x\right) dx \right) \\ &- \log\left(\int v\left(x\right) \cdot w\left(x\right) dx \right) \end{aligned}$$

with integral becoming sums (because of discrete data) :

$$\begin{aligned} D_{CS}(v,w) &= \frac{1}{2}\cdot\log\left(\sum v_i^2\cdot\sum w_i^2\right) \\ &- \log\left(\sum v_i\cdot w_i\right) \end{aligned}$$

and thus

$$\frac{\delta D_{CS}([v]_j,[w_+]_j)}{\delta[w_+]_j} = \frac{[v]_j}{\sum_j v_i^2} - \frac{[w_+]_j}{\sum_j v_i\cdot w_{+_i}}$$

The adaptation scheme of the weighting parameters alpha again follows from the respective derivative of the cost function as before:

$$\Delta\alpha_j \propto -\frac{\delta E}{\delta\alpha_j} = -\frac{\delta E_k}{\delta\alpha_k} = -\epsilon_\alpha\cdot E_{jk} = -\epsilon_\alpha\cdot\left(\frac{d_j^+(v_k,w_+) - d_j^-(v_k,w_-)}{\left(d_j^+ + d_j^-\right)}\right)$$

2.3 Heterogeneous-LVQ for pollen recognition

2.3.1 Basic data set

In the pollen recognition problem, we have a data set of pollen objects. We calculated a set of 63 image features (e.g. Haralick contrast, shape of object, Zernicke features, see [3] for details) for every segmented pollen object. This basic data set includes 4856 samples from 12 pollen classes and 191 probes. The objects were labeled with their pollen classes by an experienced pollen counting expert using the digital microscopy images.

2.3.2 Preparation of data set for HLVQ

The basic data set was classified using a hierarchical linear discrimination approach as published in [3] and in the classification process a reliability for every classified object was calculated. We than divided the data set into (1) pollen recognized with a reliability of at least 80% and (2) pollen recognized with less then 80% reliability.

From group (1) we calculated for every probe present in this group the relative frequency for all pollen species. The feature values of the objects in group (2) were extended by the relative frequencies for all pollen species of the corresponding probe as an additional structural component. Thus we have a feature vector $v = (v_1, v_2)$ with v_1 being a vector of image features (dimension of 63) and v_2 being a vector of relative frequencies (dimension of 12) of the corresponding probe, the feature vector came from.

The data set for the HLVQ was build from those pollen objects in group (2) with the extended features. Additionally we removed pollen from probes with less than five pollen recognized reliably or less than two classes already reliably recognized. The resulting data set was class-wise divided into training and test set, resulting in 459 pollen objects for training (set TR) and 452 for testing (set TE).

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

2.3.3 The whole procedure of pollen recognition

The whole training procedure is as follows:

1. First training phase: Training of a basic classification ([3]) based only on the image feature values of all 4856 pollen samples; test this classification on all pollen samples; calculating a reliability for every object.
2. Second training phase. Preparation of HLVQ data set, training of HLVQ using the HLVQ training set TR.
3. Test phase: Classify the HLVQ test pollen samples TE with the HLVQ.

In a working system the input is a set of all pollen objects from one probe. First the image features are calculated for all those pollen objects. They are classified using the basic classification from the first training phase. Again the set is split into (1) objects with reliability of at least 80% and (2) objects with reliability of less then 80%. From group (1) the relative frequencies are calculated for this probe, that are added to the feature values of the objects in group (2) as second structural component. The extended feature vectors in group (2) are than classified using the HLVQ from the second training phase.

3 Experiments

3.1 Experimental setting

We made three different experiments and compared the error rate of the resulting classification with the error rate of the first classification step on the data. For the first experiment, we split up image features and relative frequency features and used the normal euclidean distance for both structural components and adapted the weights of this structural components. Using the CS divergence for the relative frequency features and the euclidean distance for the image features in the given HLVQ was the second experimental setting.

3.2 Results

The basic classification (using a multi-step method introduced in [3]) achieved with a single unadapted metric a correct classification rate of 110 by 452 samples on the testing set, thus about 24%.

The settings described above were all tested with 10 different HLVQs each using two different annealing strategies for ϵ_α, each one with five different initializations. The HLVQs contained 100 prototypes (nearly 10 per class) that were adapted over 120000 learning steps (thus 10000 per class). Both HLVQs achieved much higher recognition rates than the basic classification. Table 1 shows the mean and standard deviation of the number of correctly classified pollen that were obtained in both experimental settings.

For both test modalities we first assessed the mean value of the distances in the single structural components. We used these values for normalizing the differences before weighting them with the α_is. Thus we can identify the relevance of single structural components from the values of the α_is.

Experimental setting	absolute mean	relative mean	standard deviation
	of correctly classified pollen		
HLVQ euclidean–euclidean	$386,5$	$85,5\%$	$0,8\%$
HLVQ euclidean–CS-Divergence	$317,5$	$70,2\%$	$1,2\%$

Table 1: Number of correctly classified pollen for testing set in pollen recognition using different experimental settings.

In the first experimental setting (euclidean–euclidean) the weighting of the image feature component v_1 was in the mean about $\alpha_1 = 0.42$ and thus the weighting of the relative frequency component v_2 at $\alpha_2 = 0.58$. That means, that the image feature component was weighted less than the relative frequency component.

For the second experimental setting (euclidean–CS-Divergence) the weighting of the image feature component v_1 always converged to $\alpha_1 = 1$, i.e. $\alpha_2 = 0$. That means, that the relative frequency features are not considered at all for the classification. We assume that this is the reason for a worse recognition rate in this test modality. In our opinion the HLVQ has a tendency to avoid high variability in the distances. Therefore it may decrease the weighting of the relative frequencies measured by CS-Divergence, which has such a high variability in the dot products.

4 Conclusion and perspectives

In this paper we introduced a general concept to incorporate structured, heterogeneous data into a learning process for classification. Introducing this concept into the cost function of GLVQ we developed a method performing very good on the problem of pollen object recognition. Using adequate metrics the method finds a better influence weighting for the different structural components and thus performs better for classification. We were surprised at the highly positive influence the relative frequencies had on the correct classification rate.

The incorporation of relative frequencies can also support image features in other domains (e.g. pathology imaging). It is also possible to use this general concept in other learning methods. When incorporating categorical data it is necessary to use a median version of the learning method under investigation.

References

[1] A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems* 8. Proceedings of the 1995 Conference, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

[2] T. Villmann and S. Haase, Mathematical Aspects of Divergence Based Vector Quantization Using Fréchet-Derivatives. Technical Report, Dep. of Mathematics/Natural Sciences, and Computer Sciences, University of Applied Sciences Mittweida, Germany, November 2009.

[3] D. Zühlke and M. Häusler and U. Heimann. PollenMonitor - a system for automatic determination of pollen concentration in ambient air. *The 4th European Symposium on Aerobiology* 12-16 August 2008 University of Turku, Finland.