Support objects for domain approximation

Alexander Ypma and Robert P. W. Duin
Pattern Recognition Group
Faculty of Applied Sciences, Delft University of Technology
Lorentzweg 1, 2628 CJ, Delft The Netherlands
email: *ypma@ph.tn.tudelft.nl*

Abstract

We propose a novel algorithm for extracting samples from a data set supporting the extremal points in the set. Since the density of the data set is not taken into account, the method could enable adaptation to novel (e.g. machine wear) data. Knowledge about the clustering structure of the data can aid in determination of the complexity of the solution. The algorithm is evaluated on its computational feasibility and performance with progressively more dissimilar data.

1 Introduction

Automatic recognition of machine wear and failure calls for methods that can deal with small sample sizes in high-dimensional spaces, undersampled fault classes and dynamically changing environments. Since normal machine behaviour is typically determined in a few calibration measurements of extremal operating conditions (e.g. when putting the machine into practice), an accurate but parsimonious description of the borders of the domain in the feature space indicating normal behaviour is expected to emerge. Failure detection can be performed by approximating the normal domain and rejecting samples not matching with this description. Ultimately, machine wear will be visible as a gradual shift of the borders of the admissible domain up to a point where a fault can be diagnosed (and this domain will consequently be labelled as faulty).

Conventional classification methods assume well-defined, static classes, while the problem of tailoring the complexity of the solution to a certain problem (e.g. choosing the number of hidden units in a neural network) is error-prone in cases with small sample sizes in high-dimensional spaces (*curse of dimensionality*). Recently proposed *support vector* methods circumvent these problems by editing the data set in a clever way, and expressing the solution in terms of the dot-product (i.e. *similarity*) of a new sample with the remaining (supporting) samples [5]. In addition one could use an arbitrary (non-analytic) similarity measure by looking upon the $m \times m$ distance matrix $D = d(\mathbf{x}_i, \mathbf{x}_j), i, j = 1, \ldots, m$ of a data set of size m as a set of m samples in an m-dimensional space [1]. When the rows in the distance matrix are considered as samples, and the columns as features, reducing the number of samples (data editing) results in a matrix with for each remaining sample (the *support objects*) the distances to each of the original objects. Inspired by these ideas, we propose a novel method for the extraction of objects in a data set that describe the domain of the set in a parsimonious manner, without taking the density of the set into account.

2 Domain approximation with support objects

In Vapnik's support vector classifier, the *optimal separating hyperplane* in the feature space specified by the nonlinear kernel $k(\mathbf{x}, \mathbf{y})$ for two separable classes of data $\{\mathbf{z}_i, i = 1, \ldots, l\}$ is given by [4, 5]

$$f(\mathbf{z}) = sgn(\sum_{i=1}^{l} \alpha_i y_i k(\mathbf{z}, \mathbf{z}_i) + b) \tag{1}$$

The coefficients α_i can be computed by a quadratic minimization procedure, and turn out to be nonzero only for the samples near the classification border (i.e. on the *margin*), the *support vectors* (dark objects in figure 1a). Hence, the solution can be found by taking only a limited subset of the available samples into account, whereas it is independent of the dimensionality of the data.

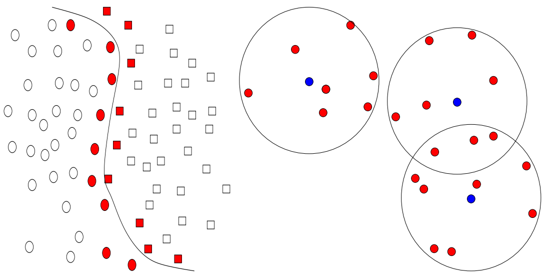

Figure 1: Support vectors in classification (l) and domain approximation (r)

The domain approximation problem can be stated in terms of equation (1) if we set the labels y_i to 1 for samples in the domain and 0 otherwise, set the threshold b to zero and choose as kernel function the rectangular kernel

$$k_r(\mathbf{x}, \mathbf{y}) = \begin{cases} 1, & \|\mathbf{x} - \mathbf{y}\| \leq r \\ 0, & otherwise \end{cases} \tag{2}$$

Since here there is no notion of a margin that is to be maximized and the labeling is different, the support vector learning algorithm determining the supporting samples in a dataset cannot be used any more. Using the fact that we approach a one-class problem, we develope an algorithm to minimize the radius r of the kernel $k_r(\cdot)$ under the constraint that each training sample is decided by $f(\mathbf{z})$ as being *in the approximation domain.*

3 Domain approximation algorithm

Consider a data set $X = \{\mathbf{x}_i\}, i = 1, \ldots, m$. For *domain approximation*, every object in the support set $J = \{\mathbf{y}_j \in X\}, j = 1, \ldots, k, k \leq m$ is now given a *receptive field* in $\mathbb{R}^m$ of radius r. A sample $\mathbf{x}_i$ in X lies in the receptive field R_p of a support object $\mathbf{y}_p$ when $p = arg\ min_j\ d(\mathbf{x}_i, \mathbf{y}_j)$ and $d(\mathbf{x}_i, \mathbf{y}_j) \leq r$. We took the Euclidean distance as distance measure $d(\mathbf{x}, \mathbf{y})$.

One could choose the set of k support objects J such that corresponding radius $r(J)$ is minimized, while all original objects are captured by some support object's receptive field (fig. 1b, dark objects in circle center), i.e.

$$J = arg\ min_{S_k} r(S_k) \tag{3}$$

where $|S_k| = k, S_k \in 2^X$ i.e. S_k is a subset of X of length k and the corresponding receptive field radius $r(S_k)$ is given by

$$r(S_k) = max_i d(\mathbf{x}_i, \mathbf{y}_j), \mathbf{x}_i \in R_j, \mathbf{y}_j \in S_k \tag{4}$$

We propose a variant of the *k-means clustering* algorithm adapted for the domain approximation problem (referred to as the *kcenters* algorithm) to cope with the subset selection problem that is still present. In the basic approach, different trials are repeated with different random choices for the initial subset. During the algorithm, each support object represents the center of the samples in its receptive field. If a better center can be found within its receptive field (i.e. such that the radius can be decreased, or *relaxed*), swap the former and the latter objects, until the subset can't be improved any more. Ultimately, the best subset over all trials is retained. Moreover, the support set size k is also a parameter to be optimized, for which we use a *successive approximation* scheme: the support set size is increased from one to at most the number of samples, and the optimal subset at size k is used as initial subset for size $k + 1$. The additional support object is initially taken to be the sample most remote to its receptive field center. As convergence criterion we choose the relative improvement in radius when increasing the support set from size k to $k + 1$.

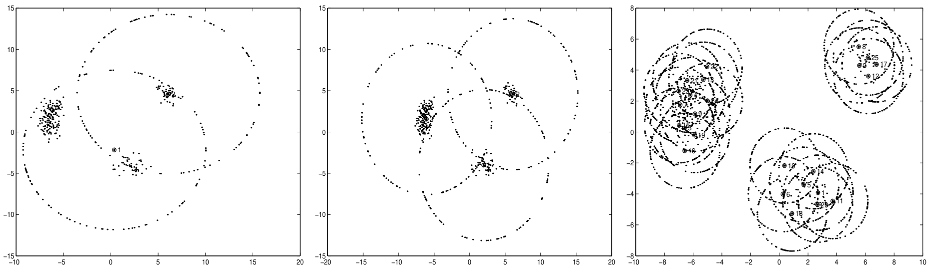

Figure 2: Trade-off between tolerance (left) and complexity (right)

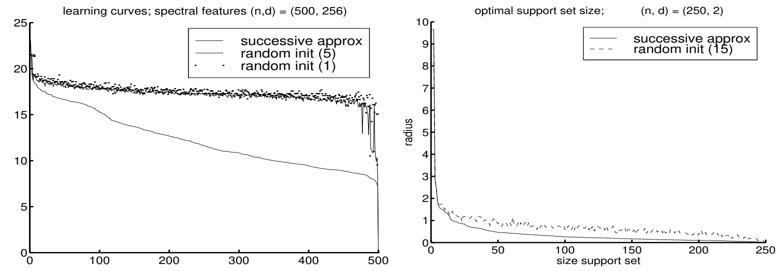

Figure 3: Successive approximation vs. random initialization

Note that there is a trade-off between complexity of the solution (support set size) and generalization capability (figure 2). Estimation of an appropriate support set size can be aided by using the data clustering structure. Moreover, the method assumes there are no serious outliers present in the data. For the application at hand (small number of well-defined calibration measurements) this assumption seems to be acceptable. Otherwise, an elementary outlier detection algorithm could be used prior to domain approximation.

4 Evaluation of the algorithm

The condition of mechanical machinery can be monitored by measuring the vibration behaviour of its rotating parts [2]. Measurements were taken from a submersible pump that was operating in three states: normal, presence of imbalance and presence of bearing failure (in the last case also at three different operating speeds). Nonlinear mapping of the data with multidimensional scaling revealed that the data consist of three clusters, where the cluster denoting bearing failure is composed of three subclusters. In [6] we showed that taking a few principal components of the acceleration spectrum already suffices for accurate recognition. This can be understood from the fact that a vibration signal $\mathbf{x}$ can be modeled as a harmonic series $\mathbf{s}$ plus additive white noise ϵ. Now the principal eigenvectors of the correlation matrix $R_{xx} = E[\mathbf{x}\mathbf{x}^T]$ span the signal subspace [3], which is probably low-dimensional. Hence, a 256-points acceleration spectrum was obtained and normalized w.r.t. mean and standard-deviation.

4.1 Successive approximation algorithm

We compared the successive approximation algorithm to the *kcenters* algorithm with random initialization (1 and 5 trials) and arbitrary choice of support set size. Using the spectral features data set, we monitored the final radius of the surrounding spheres as a function of the number of spheres. In fig. 3a the successive approximation algorithm (the bottom line) can be seen to yield smaller final radius (vertical axis), while the radius is monotonically decreasing with

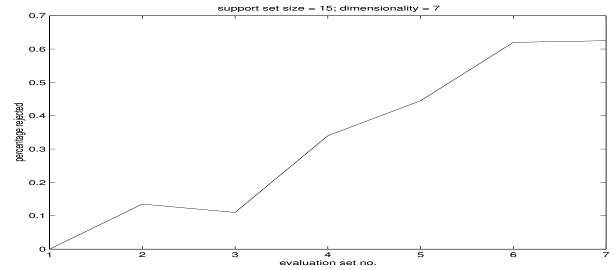

Figure 4: Rejection of novel data with successive approximation

support set size (horizontal axis). The radius obtained with 1 trial random initialization fluctuates randomly around the 5 trial random initialization results. It is clear that the stabilizing effect of more trials is at the expense of increased computing time, whereas successive approximation needs for each support set size only one run of the algorithm. Moreover, using more spheres leads to significant improvement with successive approximation, whereas the random case reaches a plateau. This illustrates that there is a relatively homogeneous distribution of distances in high-dimensional spaces, since the sphere radius can constantly be relaxed. With increasing support set size, the probability that a random guess for the initial subset is adequate decreases, hence the radius will be less improved using random initialization. When using data consisting of the first two principal components of the previous data set, again (fig. 3b) the successive approximation algorithm proved superior, but the difference is much smaller. Due to the three clusters in the data, combined with small dimensionality of the space, a sudden decrease in the radius occurs with three spheres (to a very small value). Moreover, the clusters are already represented adequately by the spheres, resulting in only marginal improvement in radius using more spheres.

4.2 Generalization to novel data

We analyzed the capability of the algorithm for incorporating gradually more dissimilar data by generating new samples out of the existing data. The offset of a new sample from its corresponding original sample is Gaussian distributed with zero mean and standard deviation s times the mean signed difference between the point under consideration and its nearest neighbours in the original set. For s running from 0 (original set) to 5, different validation sets were contructed. To track generalization, an independent test set of the same origin as the original set was used. Note that the use of multidimensional spheres can give rise to excessive tolerance in singular directions when the data lies in some (possibly nonlinear) subspace, hence we used the principal components of the spectral features data set in this experiment. It was observed that increasing feature size and support set size decreased generalization drastically (i.e. test samples and validation data for moderate values of s were frequently rejected).

Results for reasonable values for dimensionality (7) and number of spheres (15) are shown in fig. 4. The first set is the train set, the second set the test set, and sets 3 to 7 are validation sets with $s = 1, \ldots, 5$. Since the domain is fitted as tight as possible, a test set already shows some rejections. Then the rejection rate increases with the amount of novelty, up to a point where always a certain fraction of the new samples lies somewhere on the original domain (the distribution of the offset has zero mean).

5 Discussion

We proposed a novel algorithm for the extraction of objects describing the domain of a data set using multidimensional spheres with minimal radius. The algorithm was demonstrated to be an adequate and efficient way for approximating the domain of a dataset. For more reliable detection of novelty the radii should not be uniform over the set, but vary with local resolution. Moreover, a predefined tolerance specifying the admissible amount of novelty should be incorporated in the method, since the proper rejection threshold will typically vary with each practical application. Since the density of points is never used in the algorithm, but the description is made on the basis of border points, the algorithm bears potential for on-line learning in problems with small sample sizes.

6 Acknowledgment

This work was partially supported by the Dutch Foundation for Applied Sciences (STW), project no. DTN-44.3584.

References

[1] R. P. W. Duin. Relational discriminant analysis and its large sample size problem. *submitted to ICPR'98*, 1998.

[2] J. S. Mitchell. *An introduction to machinery analysis and monitoring - 2nd ed.* PennWell Publ. Comp., 1993.

[3] P. Pajunen, J. Joutsensalo, J. Karhunen, and K. Saarinen. Maximum likelihood estimation of equispaced sinusoids in rotating machine fault detection. In *Proc. of the ICSPAT'95, Boston, USA*, pages 1164–1168, 1995.

[4] B. Schölkopf. *Support vector learning.* PhD thesis, TU Berlin, 1997.

[5] V. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, New York, 1995.

[6] A. Ypma, R. Ligteringen, E. E. E. Frietman, and R. P. W. Duin. Recognition of bearing failures using wavelets and neural networks. In *Proc. of TFTS'97*, pages 69–72. University of Warwick, Coventry (UK), 1997.