Data Description in Subspaces

David M.J. Tax and Robert P.W. Duin
Pattern Recognition Group
Faculty of Applied Science, Delft University of Technology
Lorentzweg 1, 2628 CJ Delft, The Netherlands
e-mail: {davidt,bob}@ph.tn.tudelft.nl

Abstract

In this paper we investigate how the boundary of a data set can be obtained in case of (very) low sample sizes. This boundary can be used to detect if new objects resemble the data set and therefore make the subsequent classification more confident. When a large number of training objects is available it is possible to directly estimate the density. After thresholding the probability density a boundary around the data is obtained. However in the case of very low sample sizes, extrapolations have to be performed. In this paper we propose a simple method based on nearest neighbor distances which is capable to find data boundaries in these low sample sizes. It appears to be especially useful when the data is distributed in subspaces.

1. Introduction

The goal of the Data Description is to distinguish between a target set of objects and all other possible objects. It is also called One-Class classification[6], Outlier Detection[7], Novelty detection[1] or Concept learning[4]. It is mainly used to detect new objects that resemble a known set of objects. When the object does not resemble the data, it is likely an outlier or a novelty. When it is accepted by the data description, it can be classified with more confidence in a subsequent classifier.

Different methods have been developed to make a data description. Most often the probability density of a target set is modeled [8]. This requires a large number of samples to overcome the curse of dimensionality[3]. Other techniques exists, such as restricting a neural classifier to form a closed decision surface [6], various forms of vector quantization [2, 5] and recently a method based on the Support Vector Classifier, the Support Vector Data Description[9]. Because a boundary around the complete data set has to be found, all these methods still require a minimum amount of data, which can be substantial when high dimensional feature spaces are considered.

In this paper we propose a nearest neighbor method for detecting if new data resembles the target data set. It compares the distance from a test object to the training set to the nearest neighbor distances within the training set. Although it relies on the individual positions of the objects in the target set, and thereby being very noise sensitive, it is able to describe the data in the case of very low sample sizes. In section 2 we present a few simple data description methods and some of their characteristics. We will mainly focus on the detection of subspaces. In section 3 the methods are compared and in section 4 some examples for real life data is given.

2. Theory

In this section we describe two simple data description methods which can operate with very low sample sizes and therefore extrapolate from the data. Methods which require more data, like Parzen density estimation, Autoassociation Networks or Support Vector Data Descriptions will not discussed be here.

2.1. Normal distribution

When just a little amount of data is available, the most simple model is the unimodal Normal distribution. It fits this probability density model to the data:

$$p_N(\mathbf{x}) = C_d \exp\left(-\frac{1}{2}(\mathbf{x}-\mu^{tr})^T(\Sigma^{tr})^{-1}(\mathbf{x}-\mu^{tr})\right) \quad (1)$$

where C_d is a normalization constant (depending on the dimensionality d of the data). The mean μ^{tr} and the covariance matrix Σ^{tr} have to be estimated from the training set. For d dimensional data this means a total of $d + d(d+1)$ variables. Furthermore, the covariance matrix has to be inverted. In case of low sample sizes Σ^{tr} can become (nearly)

0-7695-0750-6/00 $10.00 © 2000 IEEE

singular. Often Σ^{tr} is regularized by using $\Sigma^{tr'} = \Sigma^{tr} + \lambda \mathbf{I}$. This requires a user supplied regularization parameter λ. To make an automatic method in this paper the pseudo-inverse is used.

Finally, to distinguish between target and outlier data a threshold on the probability density should be set. Accepting 95% of the data requires a threshold on the Mahanalobis distance $(\mathbf{x} - \mu^{tr})^T (\Sigma^{tr})^{-1} (\mathbf{x} - \mu^{tr})$ of:

$$\theta_N = (\chi^2_d)^{-1}(0.95) \tag{2}$$

where $(\chi^2_d)^{-1}$ is the inverse χ^2 with d degrees of freedom. This normal density method is expected to work reasonably well then the data is unimodal.

2.2. Nearest Neighbor Method

Instead of estimating complete probability densities, an indication of the resemblance can be obtained by comparing distances. This method is based on the comparison between the local density of the test object and the nearest neighbor in the training set. The local density is estimated by the nearest neighbour information[3]:

$$p_n(\mathbf{x}) = \frac{k_n/N}{V_n} \tag{3}$$

where k_n are the n nearest neighbors of $\mathbf{x}$, V_n is the volume of the cell containing $\mathbf{x}$ and the n nearest neighbours and N is the number of training objects. When $n = 1$ and objects with lower density as the nearest neighbor object are rejected, this method boils down to a comparison between two distances. Define the distance d_1 between the test object $\mathbf{x}$ and it's nearest neighbor in the training set $\mathrm{NN}^{tr}(\mathbf{x})$ as

$$d_1 = \|\mathbf{x} - \mathrm{NN}^{tr}(\mathbf{x})\| \tag{4}$$

Secondly, define the distance between this nearest neighbor $\mathrm{NN}^{tr}(\mathbf{x})$ and it's nearest neighbor in the training set $\mathrm{NN}^{tr}(\mathrm{NN}^{tr}(\mathbf{x}))$:

$$d_2 = \|\mathrm{NN}^{tr}(\mathbf{x}) - \mathrm{NN}^{tr}(\mathrm{NN}^{tr}(\mathbf{x}))\| \tag{5}$$

When the first distance is much larger than the second distance, the object will be regarded as an outlier. We use the quotient between the distances as indication of the validity of the object:

$$\rho_{NN}(\mathbf{x}) = \frac{d_1}{d_2} \tag{6}$$

where $\mathrm{NN}^{tr}\mathbf{x}$ is the nearest neighbor of $\mathbf{x}$ in the training set. In all cases the L_2-norm is used, the threshold is set to $\theta_{NN} = 1.0$ (that is, the new object should be at least as close as the nearest neighbor in the training set). The fact that the local density of $\mathbf{x}$ should be larger or equal to the density of the nearest neigbour, makes this method is therefore very useful for distributions with relatively fast decaying probabilities.

3. Comparisons on artificial data

Two quantities are important for comparing data description methods, (1) the covered volume of the description and (2) the covered part of target distribution. The volume should be as small as possible, while the covered part of target distribution should be maximal. Especially for high dimensional feature spaces both quantities are very hard to estimate. For a 1D situation the difference between lower and upper boundary can be used (both methods are unimodal). For 2D and 3D data a grid of objects in a box around the data can be created. For higher dimensionalities, the grid becomes unacceptably large and random objects have to draw from the box.

We will start by comparing the methods in three simple data distributions, the uniform distribution, Normal and t-distribution. For uniform data the Nearest Neighbor method is expected to work well (it contains a sharp decaying probability on it's boundary). The t-distribution is an example of a distribution containing a lot of remote objects which will deteriorate the performance of the Nearest Neighbor method very much.

Covered surface in 1D For several samples of the uniform distribution between -1 and 1 a Nearest Neighbor method and Gauss is trained. In figure 1 the width of the Nearest Neighbor method is plotted versus the Gaussian description for 100 1D samples. For sample sizes larger than 2 the Nearest Neighbor method gives smaller descriptions. This is visible in the figure by the fact that almost all ratios are below $y = x$. Only for data set size of two objects the normal density is a bit tighter. Note that for larger sample sizes, the Nearest Neighbor method finds on average a width of 1.0, while the Gaussian in general overestimates the width of the description. For a sample size of just 2 the width of

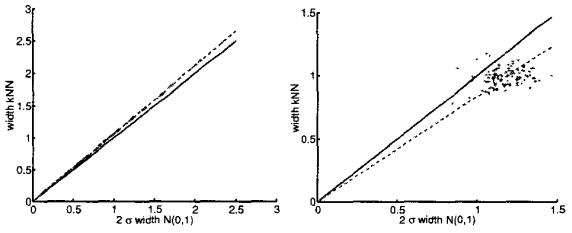

Figure 1. Covered surface of Nearest Neighbor method vs Normal distr. for 1D uniform distribution, left 2 and right 25 training objects. 100 runs

the Nearest Neighbor method and Normal distributions can be computed exactly: $\Delta_{NN} = \frac{3}{2}d$ and $\Delta_N = \sqrt{2}d$ where

d is the distance between the two points. r_{NN}/r_N corresponds to the slope in figure 1.

The dashed line is the minimum least squares estimate over 100 instantiations. It gives an indication of the relative covered surfaces by the two methods. A slope smaller than 1.0 indicates that on average the Nearest Neighbor method method covers less space than the normal density. It also means that a part of the uniform distribution is not captured. We can conclude that for the 1D uniform data the Nearest Neighbor method gives tighter descriptions while rejecting a small part of the target distribution.

Covered probability In the left plot of figure 2 the covered space by both methods is shown, and on the right the total covered probability. The Nearest Neighbor method is

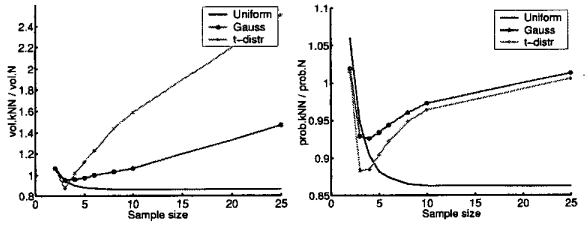

Figure 2. Left) covered surface, right) covered probability for 1D uniform distribution

very economical in the covered space for the uniform distribution, but for sample sizes larger than 5 it becomes inefficient, especially for the normal and t-distribution. The right plot shows that the savings in surface result in a 10% decrease in the integrated probability which shows that the Nearest Neighbor method is more efficient in case of sample sizes less than 5 objects even when outliers are available.

2D data in figure 3 the methods are compared for all 2D data sets. For both the target and outlier class the acceptation rate is measured (using a grid for the outlier class). In the figure the ratio $\frac{\text{NN accepted}}{\text{N accepted}}$ is shown. A ratio lower than 1.0 for the target set indicates that the Nearest Neighbor method accepts less target objects than the Normal. On the other hand, a ratio lower than 1.0 for the outlier set means a better outlier rejection. The figure shows that for all data distributions the Nearest Neighbor method only covers 60% of the target set covered by the Normal model. On the other hand the outlier acceptance of the Nearest Neighbor method is smaller in comparison with the normal model, especially for the uniform distribution.

Data in a subspace To simulate data in a subspace, a normal distribution in 2D is taken, which is embedded in n-D. The target class has a standard deviation of 1.0 in the first

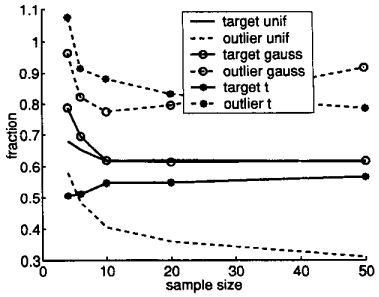

Figure 3. Ration between Nearest Neighbor method and Normal acceptance rates for target and outlier data from uniform, 2D Normal and 2D t-distribution

two features, and 0.1 in the remaining features. To detect how tight the subspace is covered by the outlier methods, the outlier class contains two classes with the same standard deviations, but their means are at ± 1 for the low variance features. A scatter plot of the second and third feature is shown in the left plot of figure 4. Note that taking a PCA for feature reduction will remove all important structure Again

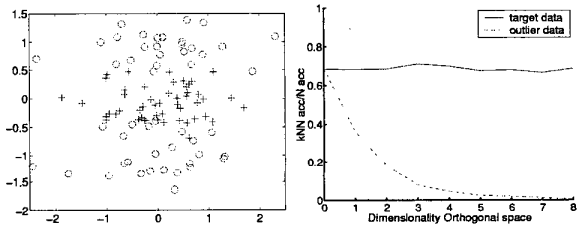

Figure 4. Performance of the Nearest Neighbor method and N for pancake data for different dimensionalities

the ratio $\frac{\text{NN accepted}}{\text{N accepted}}$ is measured for the target and outlier data. It appears that the normal distr. accepts about 85% of the target set, but also 85% of the outliers. The Nearest Neighbor method on the other hand just accepts about 60% of the target data, but reject almost all outliers, especially in high dimensional spaces. By increasing the Nearest Neighbor method threshold from 1.0 to 2.0 the target acceptation increases to 80%, but the outlier acceptance to 30%.

4. Real experiments

In this section we investigate the properties from the previous section in real problems. We focus on a machine diag-

nostics problem: the characterization of a submersible water pump[9]. Both target objects (measurements on a normal operating pump) and negative examples (measurements on a damaged pump) are available.

On the pump several vibration sensors are mounted. From the recorded time series subsamples are taken and Autoregressive Model features are calculated (see for extended explanation [9]). The AR features set contains 96 training objects in a 64 dimensional feature space. A scatter plot of the first 2 principal components is shown in the upper left plot of figure 5.

As a comparison we also included other outlier detection methods, the Parzen density estimator and the Support Vector Data Description (SVDD). The method is inspired by the Support Vector Classifier by Vapnik [11] but is adapted to give a boundary around a data set instead of a classifier between classes. For more information see [10].

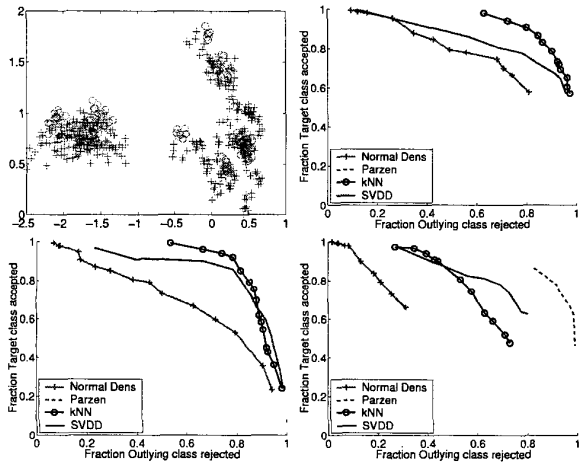

Figure 5. AR feature set of Delft pump data. Upper right are the original 64 features, lower left are 10 and lower right are 2 features used.

In the other plots in figure 5 the target acceptance is plotted versus the outlier rejection by adjusting the threshold regulating the target acceptance rate. For higher dimensionalities the Nearest Neighbor method rejects almost all of the outlier class and still accepts most of the target class. When the dimensionality of the data is reduced below the intrinsic dimensionality, the Parzen density is feasible and performs best. Because the data distribution is multimodal, the Normal density does never perform very well.

Finally, reducing the dimensionality of this dataset using PCA does not improve performance for the outlier detection. Only the Parzen density is comparable with the Nearest Neighbor method in the full 64-dim. feature space.

5. Conclusions

In this paper we showed that for very small sample sizes (less than 5 samples per feature) the Nearest Neighbor method can be very useful to detect if new data resembles old data. Although an acceptance of 100% of the target set can be guaranteed, a minimal volume (which means a minimal acceptance of outliers) is shown. This is especially apparent if data is located in a subspace.

For moderate sample sizes model based methods work pretty good (if the model reasonably fit the data) or else use the SVDD. For higher sample sizes the Parzen density estimation works best.

6 Acknowledgments

This work was partly supported by the Foundation for Applied Sciences (STW) and the Dutch Organization for Scientific Research (NWO).

References

[1] C. Bishop. Novelty detection and neural network validation. *IEE Proceedings on Vision, Image and Signal Processing. Special Issue on Applications of Neural Networks*, 141(4):217–222, May 1994.

[2] G. Carpenter, S. Grossberg, and D. Rosen. ART 2-A: an adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks*, 4(4):493–504, 1991.

[3] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[4] N. Japkowicz. *Concept-Learning in the absence of counter-examples: an autoassociation-based approach to classification*. PhD thesis, New Brunswick Rutgers, The State University of New Jersey, 1999.

[5] T. Kohonen. *Self-organizing maps*. Springer-Verlag, Heidelberg, Germany, 1995.

[6] M. Moya, M. Koch, and L. Hostetler. One-class classifier networks for target recognition applications. In *Proceedings world congress on neural networks*, pages 797–801, Portland, OR, 1993. International Neural Network Society, INNS.

[7] G. Ritter and M. Gallegos. Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters*, 18:525–539, April 1997.

[8] L. Tarassenko, P. Hayton, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Proceedings of the Fourth International IEE Conference on Artificial Neural Networks*, volume 409, pages 442–447, 1995.

[9] D. Tax and R. Duin. Data domain description using support vectors. In M. Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks 1999*, pages 251–256. D.Facto, Brussel, April 1999.

[10] D. Tax and R. Duin. Support vector domain description. *To appear in Pattern Recognition Letters*, 1999.

[11] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.