Classifier Design for Population and Sensor Drift

Keith Copsey and Andrew Webb

QinetiQ, St. Andrews Road, Malvern, WR14 3PS
{k.copsey,a.webb}@signal.qinetiq.com

Abstract. The basic assumption in classifier design is that the distribution from which the design sample is selected is the same as the distribution from which future objects will arise: i.e., that the training set is representative of the operating conditions. In many applications, this assumption is not valid. In this paper, we discuss sources of variation and possible approaches to handling it. We then focus on a problem in radar target recognition in which the operating sensor differs from the sensor used to gather the training data. For situations where the physical and processing models for the sensors are known, a solution based on Bayesian image restoration is proposed.

Keywords: classification; generalisation; sensor drift; population drift; Bayesian inference; target recognition.

1 Introduction

In classifier design, we often have a design or training set that is used to train a classifier; a validation set (used as part of the training process) for model selection or termination of an iterative learning rule; and an independent test set, which is used to measure the *generalisation performance* of the classifier: the ability of the classifier to generalise to future objects. These datasets are often gathered as part of the same trial and it is common that they are, in fact, different partitions of the same dataset.

In many practical situations, the operating conditions may differ from those prevailing at the time of the test data collection, particularly in a sensor data analysis problem. For example, sensor characteristics may drift with time or environmental conditions may change. These effects result in changes to the distributions from which patterns are drawn. This is referred to as *population drift* [4].

The circumstances and degree of population drift vary from problem to problem. This presents a difficulty for classifier performance evaluation since the test set may not be representative of the operating conditions and thus the generalisation performance quoted on the test set may be overly optimistic.

Designing classifiers to accommodate population drift is problem specific. In Section 2 we review some of the causes of population drift and the approaches that can be taken to mitigate against distributional changes. In section 3, a generic Bayesian approach is described and in section 4, this is applied in a target recognition example in which the probability densities of interest can be calculated using knowledge of the properties of the imaging radars.

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 744–752, 2004.

2 Population Drift

2.1 Sensor Drift

Pattern recognition techniques need to be developed for drifting sensors. An example is an electronic nose, a device that contains a number of different individual sensors whose response characteristics depend on the chemical odour present. These have applications in quality control, bioprocess monitoring and defence. All chemical sensors are affected by drift, stability problems and memory effects. Data processing techniques are required to handle these effects autonomously. This may be simple preprocessing to remove shifts in zero points of responses, and changes in sensitivity handled by a gain control, but there may be more complex effects.

2.2 Changes in Object Characteristics

In medical applications there may be drift in the patient population (changes in patient characteristics) over time. Population drift also occurs in speech recognition when a new speaker is presented. There are various approaches including analysing a standard input from a new speaker and using this to modify stored prototypes. In credit scoring, the behaviour of borrowers is influenced by short-term pressures (for example, Budget announcements by the Chancellor of the Exchequer) and classification rules will need to be changed quite frequently [5].

In radar target recognition, classifiers need to be robust to changes in vehicle equipment fit which can give rise to large changes in the radar reflectivity [1]. Within a Bayesian density estimation framework for classification, one approach is to introduce hyperpriors to model target variability. In condition monitoring, the healthy state of an engine will change with time. In object recognition in images, it is important that the classifier has some invariance to object pose (translational/rotational invariance).

In each of the examples above, it is an advantage if the classification method can be dynamically updated and does not need to be re-computed from scratch (using new sets of training data) as the conditions change.

2.3 Environmental Changes

The training conditions may only approximate the expected operating conditions and a trained classifier will need some modification [6]. The signal-to-noise ratio of the operating conditions may be different from the (controlled) training conditions and may possibly be unknown. In order to derive a classifier for noisier operating conditions, several approaches may be adopted including noise injection in the training set and modifying the training procedure to minimise a cost function appropriate for the expected operating conditions [8]. Errors-in-variables models are relevant here.

In target recognition, the ambient light conditions may change. The environmental conditions, for example the clutter in which the target is embedded, will differ from the training conditions. Sea clutter is time-varying.

2.4 Sensor Change

For various reasons, it may not be possible to gather sufficient information with the operating sensor to train a classifier: it might be too expensive or too dangerous. However, measurements can be made in more controlled conditions using a different sensor and a classifier can be designed using this set of measurements. In this type of problem, a classifier needs to be designed using sensor-independent features or a means of translating operating sensor data to the training domain must be developed. This is discussed further in section 3.

2.5 Variable Priors

Prior probabilities of class membership are likely to change with time. Thus, although class conditional densities do not change, decision boundaries are altered due to varying priors. This requires high-level modelling, but often there is little data available to model the dependencies and Bayesian networks are constructed using expert opinion [2].

Costs of misclassification are also variable and unknown. The consequence is that the optimisation criterion used in training (for example, minimum cost) may be inappropriate for the operating conditions.

2.6 Conclusions

Uncertainties between training and operating conditions mean that there is a limit beyond which it is not worth pushing the development of a classification rule [4]. In some cases, population drift is amenable to treatment, but this is problem dependent.

3 Joint Density Modelling

3.1 Introduction

Let us denote measurements in the training conditions by the variable $\boldsymbol{x}$ and measurements in the operating conditions by the variable $\boldsymbol{z}$. Assuming a probabilistic relationship between the measurements of an object in the operating conditions and the measurements of the (same) object that would have been obtained in the training conditions (denoted by $p(\boldsymbol{x}|\boldsymbol{z})$), inference proceeds by considering expectations of functions $g(\boldsymbol{x})$ of $\boldsymbol{x}$, conditional on the measurements $\boldsymbol{z}$:

$$E[g(\boldsymbol{x})|\boldsymbol{z}] \;=\; \int_{\boldsymbol{x}} g(\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{x} \tag{1}$$

There are a number of special cases.

Classification. Suppose that we have designed a classifier using a design set $\mathcal{D} = \{\boldsymbol{x}_i, i = 1.\ldots, N\}$ of samples gathered under the training conditions. Denote the class posterior probabilities estimated by the classifier for a measurement $\boldsymbol{x}$ by $p(C = j|\boldsymbol{x}, \mathcal{D})$, $j = 1, \ldots, J$. Setting $g_j(\boldsymbol{x}) = p(C = j|\boldsymbol{x}, \mathcal{D})$, for $j = 1, \ldots, J$ and substituting into (1) gives

$$E[p(C = j|\boldsymbol{x}, \mathcal{D})|\boldsymbol{z}] \;=\; \int_{\boldsymbol{x}} p(C = j|\boldsymbol{x}, \mathcal{D}) p(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{x} \tag{2}$$

These expectations provide an estimate of the posterior class probabilities for the $\boldsymbol{z}$ data, based on a classifier trained using $\boldsymbol{x}$ data.

Regression. Another special case is if $g(\boldsymbol{x})$ is a regression function. The training set consists of $\{(\boldsymbol{x}_i, \theta_i), i = 1, \ldots, N\}$ and a regression function is constructed to provide an estimate of θ given $\boldsymbol{x}$. Conditioned on a value $\boldsymbol{z}$, we have (with $g(\boldsymbol{x}) = E[\theta|\boldsymbol{x}]$)

$$E\left[E[\theta|\boldsymbol{x}]|\boldsymbol{z}\right] \;=\; \int E[\boldsymbol{\theta}|\boldsymbol{x}] p(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{x} \tag{3}$$

The expectations in (3) provide an estimate of θ for the $\boldsymbol{z}$ data, based upon a regression function designed using $\boldsymbol{x}$ data.

3.2 The Conditional Density, $p(\boldsymbol{x}|\boldsymbol{z})$

Specification of the conditional density, $p(\boldsymbol{x}|\boldsymbol{z})$ may be done in a number of ways. There may be prior knowledge of the form of the discrepancy; we have a physical model that characterises the difference between training and operating conditions. An example is a point scatterer model of a target in a radar target classification problem or a facet model of an object which we can use to simulate the image of the object under different illumination. The key to dealing with both of these examples is to first invert the operational data $\boldsymbol{z}$ to an underlying representation $\boldsymbol{\sigma}$ (e.g. the point scatterer model), and to then convert the underlying representation to measurements consistent with the training conditions (e.g. the different illumination).

4 Radar Target Classification Example

4.1 Introduction

The remainder of this paper concentrates on the development of procedures that enable a classifier designed using data gathered from one sensor to be applied to data gathered from a different sensor (provided that appropriate sensor measurement models are available). A motivating application is to use automatic target recognition (ATR) systems trained on readily available ground-based Inverse Synthetic Aperture Radar (ISAR) data to classify objects imaged by an airborne Doppler Beam Sharpened (DBS) radar seeker. It is intended that this

will be done by inverting each operational DBS image to an underlying radar cross section (RCS). A Bayesian image restoration (inverse model based) approach, which estimates the distribution of the underlying RCS, is proposed for this task. Given the distribution of the RCS we can then use the ISAR sensor measurement model to obtain the distribution of the ISAR image given the DBS image. ISAR images sampled from this model generated distribution can then be classified by an ATR system trained using ISAR data, thus providing the classification for the operational DBS image. The advantage to such a procedure is that it is easier to collect ISAR training data (by imaging targets on turntables) than it is to collect DBS data. Throughout the paper the DBS data is referred to as the $\boldsymbol{z}$ sensor data, and the ISAR data as the $\boldsymbol{x}$ sensor data.

4.2 Inverting the Data from the Operational Sensor

The Bayesian framework used to model the operational ($\boldsymbol{z}$) sensor data measurement process is illustrated by the model in Fig. 1.

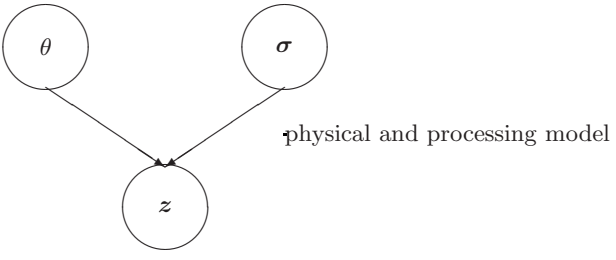

Fig. 1. Sensor model for $\boldsymbol{z}$ sensor data.

We assume that a physical and processing model is available that transforms the underlying RCS $\boldsymbol{\sigma}$ to a set of sensor measurements $\boldsymbol{z}$. The model depends on parameters θ, which may be unknown. The parameters θ will contain information on the sensor characteristics (beam shape, pulse width etc) and sensor platform dynamics (e.g. for an airborne sensor the speed, acceleration, roll).

Using Bayes' theorem, along with Fig. 1, the posterior distribution is:

$$p(\boldsymbol{\sigma}, \theta|\boldsymbol{z}) \propto p(\boldsymbol{z}|\boldsymbol{\sigma}, \theta)p(\boldsymbol{\sigma})p(\theta). \tag{4}$$

Integrating over the parameters of the sensor measurement model produces the marginal posterior distribution $p(\boldsymbol{\sigma}|\boldsymbol{z})$. The components of the posterior distribution are the prior distribution for the RCS, $p(\boldsymbol{\sigma})$, the prior distribution for θ, $p(\theta)$, and the conditional distribution for the sensor data given the measurement model parameters and the RCS, $p(\boldsymbol{z}|\boldsymbol{\sigma}, \theta)$. These distributions are examined in more detail below.

The form of the prior distribution for $\boldsymbol{\sigma}$ will depend on the representation of $\boldsymbol{\sigma}$. For our example, $\boldsymbol{\sigma}$ is represented as a two-dimensional grid of values, $\boldsymbol{\sigma} = \{\sigma_{i,j}, 1 \leq i, j \leq d\}$, where $d \times d$ is the dimensionality of the grid. The prior

distribution for $\boldsymbol{\sigma}$ is taken to be multivariate Gaussian (having concatenated the rows of the grid into a single vector), although this would not be appropriate in many situations. Ideally, the prior distribution should reflect the RCS grids that would be expected for the range of targets to be identified (i.e. determined using expert knowledge).

A prior distribution $p(\theta)$ is required for the imaging model parameters θ. This will depend on the exact form of θ, which is determined by the physical model transforming the RCS to the sensor measurements. In many cases each of the variables that comprise θ (e.g. pulse width) will be known to within a tolerance, so independent Gaussian distributions would be appropriate.

The form of the conditional distribution $p(\boldsymbol{z}|\boldsymbol{\sigma}, \theta)$ depends on the physical and processing model for the sensor data generation and the noise in that physical model. Typically, the model will be assumed to consist of a (known) deterministic function of $\boldsymbol{\sigma}$ and θ, together with additive noise. For our example, the operational sensor measurement process is taken to consist of the application of a 3×3 point spread function (PSF) to the underlying RCS grid, followed by the addition of independent Gaussian noise to each pixel (although the presented Bayesian algorithm would be unchanged by the addition of multivariate Gaussian noise on the whole image). Edge effects from the application of the PSF are dealt with by adding a temporary boundary layer of zeros to the RCS grid. The additive noise for each pixel is drawn independently from a zero mean Gaussian distribution, $N(0, \psi_z^2)$. The variable θ is therefore given by the noise variance ψ_z^2 and the matrix psf_z defining the PSF. The documented example assumes that the elements of θ are known (and correct), corresponding to a point mass prior distribution for $p(\theta)$. The resulting sensor measurement distribution is:

$$p(z|\boldsymbol{\sigma}, \theta) = \prod_{i=1}^{d} \prod_{j=1}^{d} N(z_{i,j}; \mathrm{psf}(\boldsymbol{\sigma}, psf_z)_{i,j}, \psi_z^2) \tag{5}$$

where $z = \{z_{i,j}, 1 \leq i, j, \leq d\}$ are the pixels of the measurement z, and the (i, j)-th element of the image created by applying the point spread matrix psf to the RCS grid $\boldsymbol{\sigma}$ is represented by $\mathrm{psf}(\boldsymbol{\sigma}, psf)_{i,j}$.

It turns out that for the example in this paper, the distribution $p(\sigma|\boldsymbol{z})$ can be expressed analytically as a multivariate Gaussian distribution. This would not be the case for the majority of sensor measurement models and prior distributions. Indeed, calculation of the normalisation constant for the posterior distribution is usually not tractable, and for most physical and processing models, statistics of interest (such as the mean and covariance) will not be available analytically. In such cases, rather than making simplifications to allow analytic inference on the posterior distribution, a full Bayesian approach to the problem is maintained by drawing samples from the posterior. All inferences can then be made through consideration of these samples. In most circumstances it will not be possible to sample directly from the posterior distribution, in which case a Markov chain Monte Carlo (MCMC) algorithm [3, 7] is used.

4.3 The Conditional Density $p(\boldsymbol{x}|\boldsymbol{z})$

The density $p(\boldsymbol{x}|\boldsymbol{z})$ used in the equations of Section 3 can be expressed as:

$$p(\boldsymbol{x}|\boldsymbol{z}) = \int_{\boldsymbol{\sigma}} p(\boldsymbol{x}, \boldsymbol{\sigma}|\boldsymbol{z})\, d\boldsymbol{\sigma} = \int_{\boldsymbol{\sigma}} p(\boldsymbol{x}|\boldsymbol{\sigma}) p(\boldsymbol{\sigma}|\boldsymbol{z})\, d\boldsymbol{\sigma} \tag{6}$$

The density $p(\boldsymbol{x}|\boldsymbol{\sigma})$ represents the physical and processing model for the training sensor measurements, under the assumption that the imaging model parameters are known. For our example, the same measurement process as for the operational sensor is used (see (5)), but with a different matrix psf_x defining the PSF, and a potentially different additive noise variance ψ_x^2.

For our documented example, the density $p(\boldsymbol{x}|\boldsymbol{z})$ in (6) turns out to be multivariate Gaussian. Thus, samples can be drawn easily from this distribution and used to approximate the expectations of Section 3. Given samples $\{\boldsymbol{x}^{(s)}, s = 1, \ldots, N\}$ from $p(\boldsymbol{x}|\boldsymbol{z})$, the posterior classification probabilities defined in (2) become:

$$E[p(C = j|\boldsymbol{x}, D)|\boldsymbol{z}] \approx \frac{1}{N} \sum_{s=1}^{N} p(C = j|\boldsymbol{x}^{(s)}, D) \tag{7}$$

More complicated operational and training sensor measurement models and prior distributions require the use of MCMC samples from $p(\boldsymbol{\sigma}, \theta|\boldsymbol{z})$. In particular, we can sample from $p(\boldsymbol{x}|\boldsymbol{z})$ by passing the RCS samples through the physical and processing model for sensor data $\boldsymbol{x}$ (i.e. by sampling from $p(\boldsymbol{x}|\boldsymbol{\sigma})$ for each MCMC sample for $\boldsymbol{\sigma}$).

4.4 Description of Experiment

A two class problem has been used to illustrate the approach, with the targets defined on underlying $d \times d$ RCS grids, where $d = 5$. The PSF matrices for the sensors were taken to be:

$$psf_z = \begin{pmatrix} 0.3\ 0.3\ 0.3 \\ 0.3\ 1.0\ 0.3 \\ 0.3\ 0.3\ 0.3 \end{pmatrix} \qquad psf_x = \begin{pmatrix} 0.1\ 0.1\ 0.1 \\ 0.1\ 1.0\ 0.1 \\ 0.1\ 0.1\ 0.1 \end{pmatrix} \tag{8}$$

The different PSF widths mimic the effects of different sensor resolutions, with the data from the operational sensor being more distorted than that from the training sensor. The standard deviations of the additive noise applied to the sensor data were $\psi_x = \psi_z = 0.1$.

The sensor data were created by generating underlying RCS grids and then simulating the sensor measurement processes. The RCS grids were generated by sampling from multivariate Gaussian distributions (after concatenation of the rows of the grid). For both classes, the Gaussian covariance matrix was diagonal, with the same standard deviation $\psi_{\sigma} = 0.5$ across all components. The means were dependent on the class and are displayed graphically in Fig. 2.

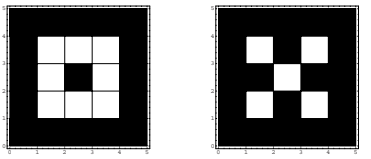

Fig. 2. Mean values for the target RCS grids (class 0 to the left, class 1 to the right). Dark pixels correspond to value 0, light pixels to value 1.

A multivariate Gaussian classifier (outputting posterior class probabilities) was selected for the $\boldsymbol{x}$ sensor data classifier. The classifier was trained by generating $n_{tr} = 100$ $\boldsymbol{x}$ sensor data measurements from each class. The test/operational data were obtained by generating $n_{te} = 1000$ $\boldsymbol{z}$ sensor data measurements from each class. Thus, the test data came from a different sensor to that used in the design phase of the classifier.

Within the Bayesian algorithm, the mean of the Gaussian prior distribution for $\boldsymbol{\sigma}$ was set to be zero around the outer layer, and 0.5 within the central 3×3 grid. The covariance matrix was set to be diagonal with the standard deviations for each grid location set to $\varsigma_\sigma = 0.5$. We note that the prior distribution incorporates expert knowledge that the values in the outer layer of the RCS grid (for both targets) are likely to be smaller than the inner values.

For each test data $\boldsymbol{z}$ measurement, 500 samples were drawn from the distribution $p(\boldsymbol{x}|\boldsymbol{z})$, and used in (7). Class decisions were made by selecting the class with the maximum expected posterior class probability.

4.5 Experimental Results

To assess the performance of the Bayesian approach, three additional sets of classification results have been obtained (all based on Gaussian classifiers).

C1) A classifier applied directly to the $\boldsymbol{z}$ sensor data. $n_{tr} = 100$ $\boldsymbol{z}$ sensor data measurements from each class were made available for training.
C2) A classifier applied directly to the $\boldsymbol{x}$ sensor data. Test data for the $\boldsymbol{x}$ sensor measurements were made available.
C3) The classifier for the $\boldsymbol{x}$ sensor data applied directly to the $\boldsymbol{z}$ sensor data.

Classifiers C1 and C2 rely on data that is not available under the proposed scenario, so provide an indication of the the performance that could be expected in idealised situations, rather than a baseline performance.

Figure 3 displays the classification rates obtained for the Bayesian approach and the three additional classifiers. The poorer performance of classifier C3 relative to the other classifiers indicates that, if ignored, the change in sensor between operational and training conditions does (as would be expected) reduce the classifier performance. There is little to choose between classifiers C1, C2 and the Bayesian approach, indicating that (given appropriate sensor measurement models) we have provided a mechanism for dealing with situations where the operating sensor differs from the sensor used to gather the training data.

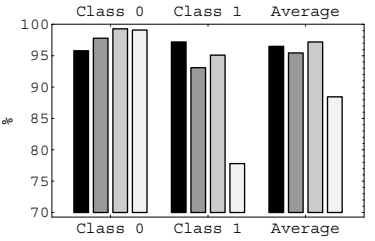

Fig. 3. Classification rates. From left to right, within each set of bars, the results are for the Bayesian image restoration based approach, classifiers C1, C2 and C3 respectively.

5 Conclusions

The basic assumption in classifier design is that the distribution from which the design sample is selected is the same as the distribution from which future objects will arise. This paper examines the validity of that assumption and considers a problem in radar target recognition in which the operating sensor differs from the sensor used to gather the training data. A Bayesian image restoration based solution is proposed for situations where the physical and processing models for the sensors are known. The approach is illustrated on a simplified problem.

Acknowledgements

This research was sponsored by the UK MOD Corporate Research Programme.

References

1. K.D. Copsey and A.R. Webb. Bayesian approach to mixture models for discrimination. *Advances in Pattern Recognition, Springer Lecture Notes in Computer Science*, 1876:491–500, August 2000.
2. K.D. Copsey and A.R. Webb. Bayesian networks for incorporation of contextual information in target recognition systems. *Advances in Pattern Recognition, Springer Lecture Notes in Computer Science*, 2396:709–717, August 2002.
3. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in Practice*. Chapman and Hall, 1996.
4. D.J. Hand. *Construction and Assessment of Classification Rules*. John Wiley, Chichester, 1997.
5. M.G. Kelly, D.J. Hand, and N.M. Adams. The impact of changing populations on classifier performance. *Proc. of the 5th ACM SIGKDD Conf, San Diego*, pages 367–371, 1999.
6. M.A. Kraaijveld. A Parzen classifier with an improved robustness against deviations between training and test data. *Pattern Recognition Letters*, 17:679–689, 1996.
7. A.R. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, Chichester, 2nd edition, August 2002.
8. A.R. Webb and P.N. Garner. A basis function approach to position estimation using microwave arrays. *Applied Statistics*, 48(2):197–209, 1999.