Outlier detection with partial information: Application to emergency mapping.

Davide D'Alimonte[1]
and Dan Cornford[2]

1. NASA/Goddard Space Flight Center, Greenbelt, Maryland, USA - email: davide@ardbeg.gsfc.nasa.gov

2. Neural Computing Research Group, Aston University, Birmingham, UK - email: d.cornford@aston.ac.uk

February 15, 2007

Abstract

This paper, addresses the problem of novelty detection in the case that the observed data is a mixture of a known *'background'* process contaminated with an unknown other process, which generates the outliers, or novel observations. The framework we describe here is quite general, employing univariate classification with incomplete information, based on knowledge of the distribution (the *probability density function*, *pdf*) of the data generated by the *'background'* process. The relative proportion of this *'background'* component (the *prior 'background'* probability), the *pdf* and the *prior* probabilities of all other components are all assumed unknown. The main contribution is a new classification scheme that identifies the maximum proportion of observed data following the known *'background'* distribution. The method exploits the Kolmogorov-Smirnov test to estimate the proportions, and afterwards data are Bayes optimally separated. Results, demonstrated with synthetic data, show that this approach can produce more reliable results than a standard *novelty detection* scheme. The classification algorithm is then applied to the problem of identifying outliers in the SIC2004 data set, in order to detect the radioactive release simulated in the 'joker' data set. We propose this method as a reliable means of novelty detection in the emergency situation which can also be used to identify outliers prior to the application of a more general automatic mapping algorithm.

1 Introduction

Often when working with real data, one is faced with the situation that the observations obtained result from a mixture of different processes, which can be very challenging to model. An example of this, within the emergency mapping context, is the identification of outliers due to extreme events. This can be seen in the context of the Spatial Interpolation Comparison 2004 (SIC2004) exercise, designed to assess the reliability of state of the art automatic mapping algorithms. In order to assess the robustness of the methods a *'joker'* data set was created which simulated a release of radioactive material [1]. None of the algorithms that participated in the SIC2004 exercise coped very well with this data set [2]. The reasons for this are quite understandable; participants had been provided with 10 days of 'prior' data, which represented typical or *'background'* conditions and had not had any examples of extreme values, which were generated by a distinctly different processes from the *'background'* process.

Often, when considering emergency mapping scenarios, a proportion of the data will correspond to the typically observed *'background'* process, which given a reasonable observation system will be well observed and well characterised, in terms of its *probability density function*, *pdf*. However, in general, we will not know in advance the *pdf* of the various other processes we will encounter in an automatic mapping context, and we will also not know the proportion of the observations generated from this unknown process (or possibly processes). In this work we address this issue using a classification based approach, allowing us to identify the novel observations, their *pdf* and their *prior* probability.

Classification consists of segmenting a set of data points into different classes, each of which in this case corresponds to a different generating process. This task is optimally solved knowing the distribution (the *pdf*) and the relative proportion (the *prior probability*, P) of the data generated by each process. In this case Bayes' theorem defines the *posterior probabilities* of each data point being generated by each component process, and

afterwards data can be segmented minimizing the misclassification rate [3]. Unfortunately, in many problems it is not possible to characterize each process (i.e., to define each class conditional *pdf* and / or to know the corresponding prior), and various techniques [4] have been developed in order to provide approximate classification schemes, for instance *novelty detection* methods [5, 6].

This work investigates classification, with emphasis on novelty detection, with incomplete information when it is only possible to define the *pdf* of a single *'background'* generating process. Both the relative proportion of this known component (the *prior*), as well as the *pdf* and the *prior* of all other components are unknown. We essentially treat this as a two class classification problem, with arbitrarily complex *pdfs* for each process. Although this specific case is usually addressed with a *novelty detection* scheme, this work shows some drawbacks associated with a novelty detection approach. Instead, we propose on a new scheme based on the Kolmogorov-Smirnov test to find the greatest number of observations that follow the distribution of the known *'background'* process. This establishes the *prior* probabilities and the data are then Bayes-optimally separated. Results are demonstrated with synthetic data, and then the method is applied to the SIC2004 data sets[1].

2 Methods

The classification scheme is described in a sequential manner. Section 2.1 briefly overviews the probability density modelling. Section 2.2 describes the proposed scheme for identifying the prior probability of the known *'background'* component. Finally, Section 2.3 presents the classification algorithm.

2.1 Modelling the probability density function

We employ Gaussian Mixture Models (GMMs) to provide flexible models for the known *'background'* and observed *pdfs*. GMM allow us to efficiently represent more complex density functions through a linear combination of simpler distributions:

$$p(x) = \sum_{j=1}^{M} p(x|j)\, P(j), \tag{1}$$

where the j *kernel* function, $p(x|j)$, is

$$p(x|j) = \frac{1}{(2\pi\sigma_j^2)^{1/2}} \exp\left\{ -\frac{(x-\mu_j)^2}{2\sigma_j^2} \right\}, \tag{2}$$

and the probabilities $P(j)$, which combine the individual kernels, are the *mixing coefficients* with $0 \leq P(j) \leq 1$ and $\sum_{j=1}^{M} P(j) = 1$. The GMM can be efficiently trained through an Expectation Maximization (EM) algorithm [7].

Although the component density functions are very simple, combining a number of them in a mixture representation results in an overall more complex density model, and provided that the number of training points is sufficiently high, can represent any continuous data distribution [3]. We do not discuss in detail the issues surrounding the use of GMMs; the interested reader can consult [3].

2.2 Identifying the prior of the known process

Consider the case where a data set consist of observations drawn from different processes, but it is only possible to define the distribution of one process, which we here call the *'background'* process. In many applications the term *'background'* process might not be appropriate, and it could be simply though of as any process whose *pdf* might be known in advance, for example where some subset of the data from the process has been labelled. In this paper we indicate by α this *known 'background'* process and group all the remaining (unknown) processes into a unique class which we label β. Our task is to identify the maximum proportion of "α-data", i.e., the number of observations generated by the "α-process". This is achieved using

the Kolmogorov-Smirnov test (Section 2.2.1) to implement the algorithm presented in Section 2.2.2.

2.2.1 The Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test [8] (here after also indicated as KS) allows one to verify whether a dataset follows some univariate *cumulative distribution function* (*cdf*). Indicate by cdf_α the *cdf* of the set of n_α labelled α data, by $p(x|\alpha)$ the corresponding *pdf*, and by cdf^*_α the cumulative distribution of a set of n^*_α data. The Kolmogorov-Smirnov test verifies whether the n^*_α data follow the cdf_α by measuring the maximum value, $D_{\rm KS}$, of the absolute difference between the two *cdf*s,

$$D_{\rm KS} = \max_{-\infty<x<+\infty} \left| cdf^*_\alpha(x) - cdf_\alpha(x) \right| . \tag{3}$$

The null hypothesis that the two *cdf*s are the same is checked against the *P-value*,

$$\text{Probability}\left(D_{\rm KS} > \text{Observed}\right) = Q_{\rm KS}\left(\left[\sqrt{n_e} + 0.12 + 0.11/\sqrt{n_e}\right] D_{\rm KS}\right), \tag{4}$$

where $n_e = \frac{n_\alpha n^*_\alpha}{n_\alpha + n^*_\alpha}$ and the $Q_{\rm KS}$ distribution is

$$Q_{\rm KS}(\lambda) = \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\lambda^2}. \tag{5}$$

2.2.2 Constructing $P(\alpha)$ with KS

Given N data points, the maximum proportion, n_α/N, possibly drawn from the "α-process" is defined as follows:

1. set the *subscript* variable $i = 1$ and $n_i = i$;
2. sample n_i points from $p(x|\alpha)$[1];
3. select from the N points to be classified, the n_i points having the minimum distance from those generated in step 2;

[1]Notice how this step directly exploits the generative nature of the GMM algorithm.

4. using Equation (4), compute the *P-value* K_i;

5. set $n_i = n_i + \Delta_n$, $i = i + 1$ and go to step 2 **or** stop when the *P-value* become less than a confidence level, and set $n_\alpha = n_i$.

The smaller Δ_n (i.e., the number of points to be added from one iteration to the next), the more accurate, but also time consuming, is the algorithm. For instance, setting $\Delta_n = 1$ and using the SIC2004 data set presented in Section 3.2 (200 data points), the algorithm takes less than a second to define $P(\alpha)$ with an AMD Athlon 3500 processor.

[Fig. 1 about here.]

Figure 1 shows an example of the dependency of the KS test result, K_i, on the number of sample points, n_i. For i small, there is a significant chance of finding among the N data to be classified, a set of n_i points that follow the $p(x|\alpha)$ distribution. In this case the corresponding *P-value* (see Equation (4)) will be close to 1. Once all the data distributed according to $p(x|\alpha)$ have been identified, the *P-value* drops to 0. A threshold confidence level of 0.99 was used in this study to define n_α, although this will require some knowledge of the cost-loss function of the specific application to choose optimally. The maximum proportion of samples possible drawn from the "α-process" is hereafter denoted as $\tilde{P}(\alpha) = n_\alpha/N$, where the tilde notation is used to highlight the difference with respect to the mixing coefficient, $P(\alpha)$, of the component from which the α samples are drawn. Notice that $\tilde{P}(\alpha) \geq P(\alpha)$. Analogously, $\tilde{P}(\beta) = n_\beta/N$ is the minimum proportion of data *not* generated by the "α-process", with $n_\beta = N - n_\alpha$.

The Kolmogorov-Smirnov test is used to check whether the unlabelled data closest to the points sampled from the $p(x|\alpha)$ (see points 2 and 3 of the iterative procedure described previously) follow the $p(x|\alpha)$ distribution; this could have been assessed applying another test, for instance the χ^2 test. Here, we are interested in improving data classification in those cases where the distributions of data generated by different processes overlap significantly – that is when the *novelty detection* scheme may produce

sub-optimal results, as shown in Section 3. Notice that the *cdf*s on which is based the Kolmogorov-Smirnov test approaches 0, or 1, in correspondence to the tails of the two distributions. Hence, applying the Kolmogorov-Smirnov test permits us to assess the two distributions similarity giving more relevance to data points close to the median, and less relevance to the data points in the tails.

2.3 Classification scheme

As stated in Section 2.2, assume that samples are drawn from two processes, α and β, and only the distribution of the points generated by the "α-process" is known. Now suppose that the prior $P(\alpha)$ has been defined with the method described in Section 2.2. On this basis, we suggest the following classification algorithm.

A data point to be classified, say x, is optimally attributed to the class α when

$$P(\alpha|x) \geq P(\beta|x), \tag{6}$$

where $P(\alpha|x)$ and $P(\beta|x)$ are the *posterior probabilities*. Equation (6) can be rewritten using Bayes' theorem as

$$\begin{aligned} p(x|\alpha)\tilde{P}(\alpha) &\geq p(x|\beta)\tilde{P}(\beta) \\ &\geq p(x|\beta)(1-\tilde{P}(\alpha)), \end{aligned} \tag{7}$$

where we used $\tilde{P}(\alpha) + \tilde{P}(\beta) = 1$. Thus, $\tilde{P}(\beta)$ does not need to be derived directly from the data. Instead, we need to model $p(x|\alpha)$ and $p(x|\beta)$. Since we do not know if there are some other unlabelled "α-data", it is not possible to model $p(x|\beta)$ after having removed all the labelled "α-data". As an alternative, we model the unconditional distribution

$$\begin{aligned} p(x) &= p(x|\alpha)\tilde{P}(\alpha) + p(x|\beta)\tilde{P}(\beta) \\ &= p(x|\alpha)\tilde{P}(\alpha) + p(x|\beta)(1-\tilde{P}(\alpha)), \end{aligned} \tag{8}$$

and then derive the unknown conditional distribution

$$p(x|\beta) = \frac{p(x) - p(x|\alpha)\tilde{P}(\alpha)}{1-\tilde{P}(\alpha)}. \tag{9}$$

This leads to the following classification rule

$$p(x|\alpha)\tilde{P}(\alpha) \geq \frac{p(x) - p(x|\alpha)\tilde{P}(\alpha)}{1 - \tilde{P}(\alpha)}(1 - \tilde{P}(\alpha))$$
$$\geq p(x) - p(x|\alpha)\tilde{P}(\alpha). \tag{10}$$

Finally,

$$p(x|\alpha)\tilde{P}(\alpha) \geq \frac{1}{2}p(x), \tag{11}$$

which means that the point x is attributed to the "α-process" when $p(x|\alpha)\tilde{P}(\alpha)$ accounts for at least half of the $p(x)$ density, quite meaningful also from the intuitive point of view.

3 Results and Discussion

This section shows some results from the proposed schemes and from a standard *novelty detection* approach. The latter consists in fitting the labelled "α-data" with a Gaussian distribution and then using a confidence level of 0.1 to identify *novel data* (i.e., those data to be attributed to the β-class).

3.1 Synthetic data

Synthetic data are drawn from the following mixture model

$$p(x) = p(x|\alpha)P(\alpha) + p(x|\beta)P(\beta). \tag{12}$$

To make the simulation more general, each component process, $p(x|\alpha)$ and $p(x|\beta)$, is a GMM with two kernels (see Table 1) that generates a bimodal distribution. According to our hypothesis, only the $p(x|\alpha)$ is assumed to be known once data are sampled.

The percentage of samples to be attributed to the "α–process" is presented in Figures 2 for classification schemes based on *i.* the posterior probabilities computed through the data generating probability densities and prior probabilities (this classification scheme is hereafter also identified as *virtual* classification, being based on parameters supposed not to be known,

and is used only for result benchmarking); *ii.* posterior probabilities derived through the proposed KS classification scheme; and *iii.* the novelty detection scheme.

Results from different tests (see Table 1) are shown as a function of the similarity between $p(x|\alpha)$ and $p(x|\beta)$, measured through the KL divergence

$$\mathrm{KL}(p(x|\alpha)\|p(x|\beta)) = -\int\int p(x|\alpha)\ln\left(\frac{p(x|\beta)}{p(x|\alpha)}\right)\mathrm{d}x\mathrm{d}y. \qquad (13)$$

[Fig. 2 about here.]

When $p(x|\alpha)$ and $p(x|\beta)$ are mostly dissimilar, that is for large value of the KL distance, both the KS and the novelty detection schemes properly estimate the number of samples attributed to the "α–process" by the virtual classification. As an example, the first row panels of Figure 3 show the original distributions from which data have been sampled, the estimates from the KS scheme, and novel data (first second and third column, respectively).

Reduction of the distance between $p(x|\alpha)$ and $p(x|\beta)$ leads the novelty detection scheme to over-estimate the portion of "α–data", while the KS approach is still in agreement with the findings of the virtual classification, as detailed in Panels 3(d), 3(e) and 3(f). This result highlights a major drawback of the novelty detection scheme – data not novel with respect to $p(x|\alpha)$ do not necessary follow the $p(x|\alpha)$ distribution!

[Fig. 3 about here.]

When the distance between $p(x|\alpha)$ and $p(x|\beta)$ becomes very small, also the KS scheme starts to attribute to the "α–process" more data than the virtual classification (detail are given in the last row panels of Figure 3). This result is fully justified by the task addressed in this work, that is to find the maximum proportion of data possibly generated by the "α–process". The same can not be said for to the novelty detection scheme, where finding this maximum proportion is not the aim of a principled approach.

3.2 Outlier identification for automatic mapping: SIC2004

The SIC2004 exercise raises some interesting challenges. Firstly there is considerable prior data made available at 200 observation sites over 10 'typical' days. In this paper we ignore the spatial nature of this data (equivalent to wrongly regarding the generating process as homogeneous) and assume that these samples can be aggregated to define a stationary, homogeneous background radioactivity probability density which defines our known *'background'* process, $p(x|\alpha)$. This is modelled using a GMM chosen to have 5 components, which is flexible enough to represent the quite complex distribution but is less prone to over-fitting than a more complex model. It is trained using a k-means based initialisation followed by an EM algorithm to find a maximum likelihood solution [9]. The GMM fits the SIC2004 background data well.

[Fig. 4 about here.]

We now model the data for the two test data sets used in SIC2004, the 'normal' data that arises from typical conditions (not shown) and the *'joker'* data set that contains the simulated release of radioactive material. The *pdf* of the background data, $p(x|\alpha)$, along with the *pdf* of the observations from the *'joker'* data set, $p(x)$ are shown in Figure 4. It is clear that the *'joker'* data set is largely similar in the main body of the *pdf* to the *'background'* data, with the exception of a very small peak around 1200 nSv/hr due to the contaminant release (not shown on the plot). It is not surprising that the algorithm determines that there are only three outliers in this case, these being shown in both spatial and data scale in Figure 5. In the current implementation of the algorithm we use a P-value of 0.99 to assess whether an observation is likely to have come from $p(x|\alpha)$, and in this data set the estimate of $p(\alpha)$ at 0.86 – 0.96, the range resulting from the fact that this is a Monte Carlo based algorithm requiring simulation from $p(x|\alpha)$. In operational practice it might make sense to exploit multiple runs to provide a more stable estimate of $p(\alpha)$. Note that all runs provide identical outlier

detection results, detecting the release sites shown in Figure 5, as might be expected in this rather simple data set; none of the other observations is ever classified as an outlier.

[Fig. 5 about here.]

The current implementation does not exploit any spatial structure in the data, but in the context of automatic mapping, this is an omission. Further work is required to model the spatial and spatio-temporal context, possibly using mixtures of space-time Gaussian process models or using some for of post processing to identify clusters of outliers. Space-time clusters of outliers are more likely to represent anomalous observations, rather than instrumental error which might be expected to be uncorrelated in both space and time. In this section we have briefly shown the application of our new novelty detection algorithm to the SIC2004 data set, however more work is required to apply the methods to larger and more challenging data sets. This work will be undertaken as part of the INTAMAP project [http://www.intamap.org].

3.3 Why not seek $P(\alpha)$ directly through the EM algorithm?

It may seem that the two-class problem addressed in this work could be solved by directly modelling the data distribution through a mixture model constraining one kernel to the known probability density $p(x|\alpha)$,

$$p(x) = p(x|\alpha)P(\alpha) + \sum_{k=1}^{C} p(x|k)P(k), \tag{14}$$

and

$$P(\alpha) + \sum_{k=1}^{C} P(k) = 1. \tag{15}$$

From this perspective, prior probabilities, $P(\alpha)$ and $P(k)$, and the *location* and *scale* parameters of each kernel, $p(x|k)$, could be defined through a standard EM learning procedure [7], keeping $p(x|\alpha)$ fixed. This would seem to allow the computation of the posterior probabilities, $P(\alpha|x)$ and $P(k|x)$, providing the desired optimal classification, however this is not the case.

The objective of the EM algorithm is to find the maximum-likelihood estimate of the mixture model parameters. Thus, the prior for the known process, $P(\alpha)$, found through the EM algorithm does not necessarily represent the maximum portion of data that follows the $p(x|\alpha)$ distribution. Indeed it was found through numerical experiments (results not presented here) that in most case the data likelihood becomes higher for $P(\alpha)$ going to zero, this corresponding to the situation where the distribution of the entire data set modelled only by $\sum_{k=1}^{C} p(x|k)P(k)$.

4 Summary and Conclusions

This work addressed the problem of data classification with incomplete information assuming that we only know the distribution of data drawn from a single *'background'* generating process. We define the corresponding prior probability by identifying the largest number of observations that follow this known *'background'* distribution. The proposed classification scheme is of benefit when it is necessary to find observations that are more likely to have been drawn (or do not come) from the known *'background'* process while, at the same time, and there is no need to discriminate among the other processes.

The effectiveness of the KS method in determining the prior probability of the known process was demonstrated with simulated data and then applied to the real problem of detecting the simulated nuclear release in the SIC2004 data. This requirement for classification with incomplete information may arise in many other contexts besides the application presented here, ranging from medical diagnosis to remote sensing data analysis.

The theory and the results presented in this work refers to the univariate case. Although not investigated here, an analogous approach could be applied to higher dimensional data through the generalization of the Kolmogorov-Smirnov test (an extension to the two dimensional case is given in [8]). Since it may become difficult to model data density in a high di-

mensional space, feature extraction procedures could be used to reduce the data dimensionality: the proposed classification scheme could then be implemented on the basis of a lower dimensional features instead of higher dimensional data. In this regard, the *NeuroScale* [10] algorithm is particularly suited for extracting features that preserve the original data structure. For this reason, the *NeuroScale* model can also be successfully applied to data visualization (see for instance, [11]). We expect to pursue this further in future work.

As discussed in Section 3.2 further improvements might be expected in our ability to identify outliers by taking the spatio-temporal context into account. This could be done as an augmentation of the existing model which might be rather complex to implement, or more manageably as part of a post-processing step to identify clusters of outliers that might result from a real emergency rather than a simple instrument malfunction. Another obvious enhancement might be to use more general mixture models, for example gamma mixture models might be more appropriate for variables such as dose rates that are physically constrained to be positive.

Acknowledgements

This work was partially supported by the BBSRC contract 92/EGM17737. The SIC2004 data was obtained from [http://www.ai-geostats.org]. This work is partially funded by the European Commission, under the Sixth Framework Programme, by the Contract N. 033811 with DG INFSO, action Line IST-2005-2.5.12 ICT for Environmental Risk Management. The views expressed herein are those of the authors and are not necessarily those of the European Commission.

References

[1] EC EUR 21595 EN. Automatic mapping algorithms for routine and emergency monitoring data. Report on the Spatial Interpolation Comparison (SIC2004) exercise, Dubios, G. (Ed.). 2005.

[2] G. Dubois and S. Galmarini. Introduction to the spatial interpolation comparison (SIC) 2004 exercise and presentation of the datasets. *Applied GIS*, 1(2):1–11, 2005.

[3] C. M. Bishop. *Neural networks for pattern recognition.* Clarendon, 1995.

[4] M. Markou and S. Singh. Novelty detection: a review. part 1: statistical approaches. *Signal Processing*, 83:2481–2497, 2003.

[5] C. M. Bishop. Novelty detection and neural network validation. *IEEE Proc.Vision and Image & Sig. Proc.*, 141:217–222, 1994.

[6] S. Roberts. Novelty detection using extreme value statistics. *IEE Proceedings on Vision Image ' Signal Processing*, 146(3):124–129, 1999.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, B39(1):1–38, 1977.

[8] W. H. Press, S.A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: the art of scientific computing.* Cambridge University Press, 1992.

[9] I. T. Nabney. *Netlab: Algorithms for pattern recognition.* Springer Edition, 2001.

[10] D. Lowe and M. E. Tipping. Neuroscale: novel topographic feature extraction using RBF networks. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, pages 543–549, London, UK, 1997.

[11] D. D'Alimonte, D. Lowe, I. T. Nabney, V. Mersinias, and C. P. Smith. Milva: An interactive tool for the exploration of multidimensional microarray data. *Bioinformatics*, 21(22):4192–4193, November 2005.

List of Tables

List of Figures

	π	μ	σ
α_1	0.6	-1.5+Δ	0.05
α_2	0.4	-1.0+Δ	0.03
β_1	0.4	-0.25	0.05
β_1	0.6	0.5	0.1

Table 1:

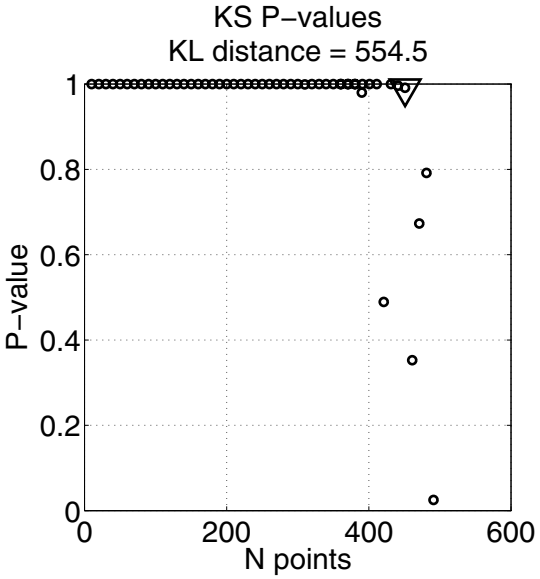

KS P−values
KL distance = 554.5
P−value
N points
0
0.2
0.4
0.6
0.8
1
0
200
400
600

Figure 1:

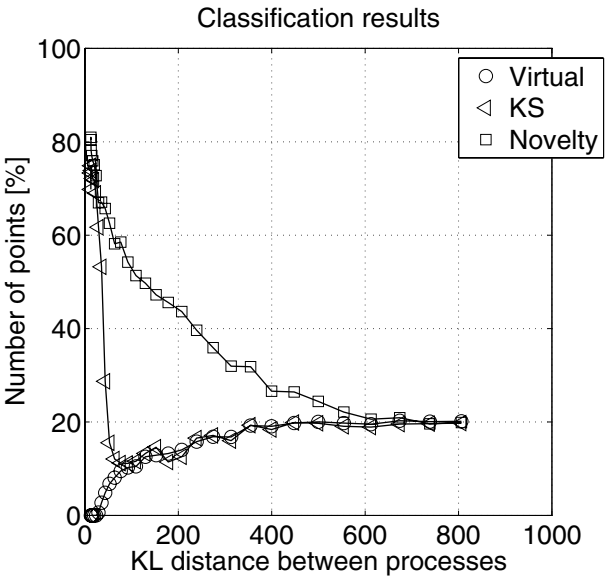

Classification results
Virtual
KS
Novelty
Number of points [%]
KL distance between processes
0
20
40
60
80
100
0
200
400
600
800
1000

Figure 2:

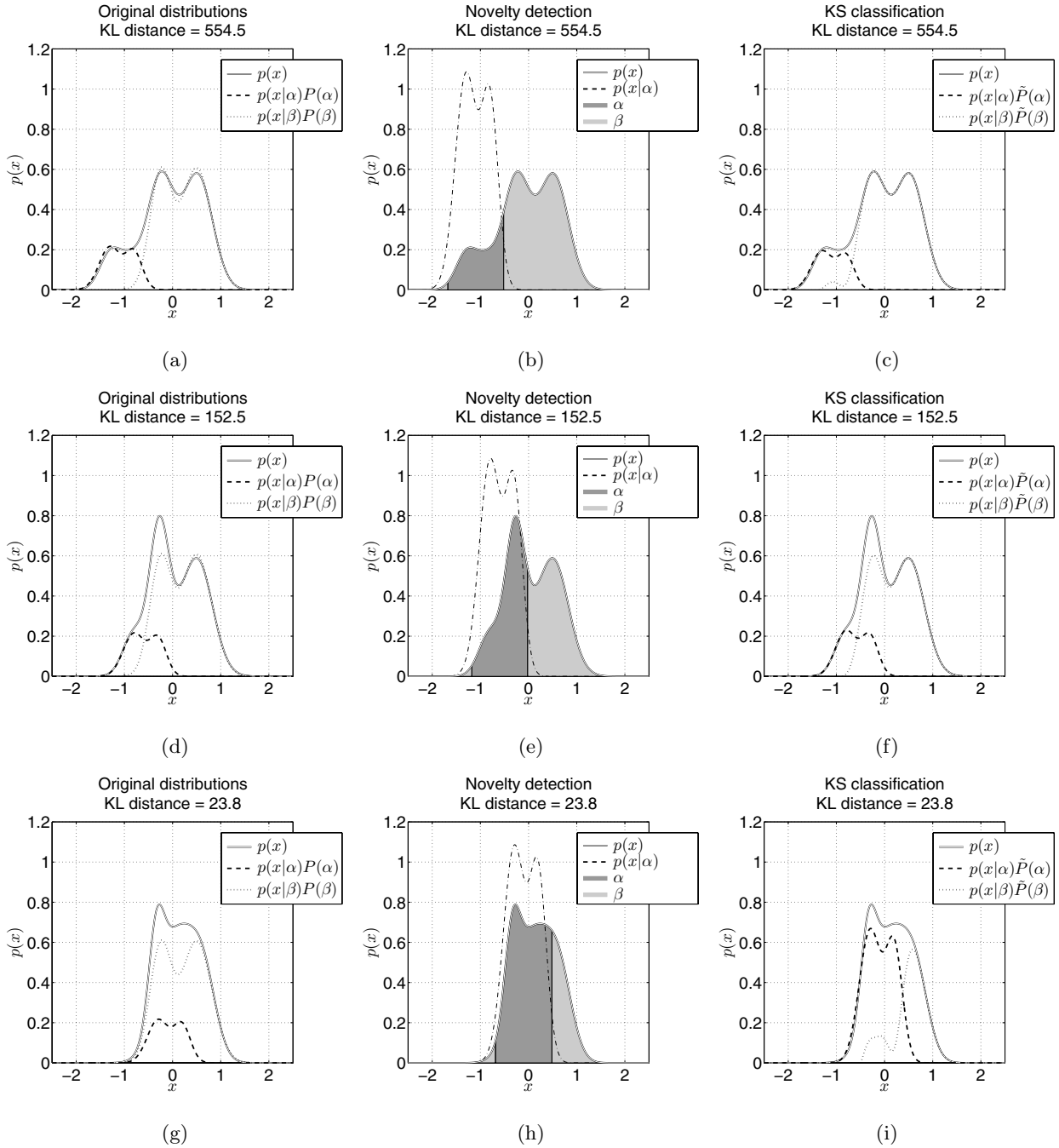

Original distributions
KL distance = 554.5
Novelty detection
KL distance = 554.5
KS classification
KL distance = 554.5
Original distributions
KL distance = 152.5
Novelty detection
KL distance = 152.5
KS classification
KL distance = 152.5
Original distributions
KL distance = 23.8
Novelty detection
KL distance = 23.8
KS classification
KL distance = 23.8
$p(x)$
$p(x|\alpha)P(\alpha)$
$p(x|\beta)P(\beta)$
$p(x|\alpha)$
α
β
$p(x|\alpha)\tilde{P}(\alpha)$
$p(x|\beta)\tilde{P}(\beta)$
x
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

Figure 3:

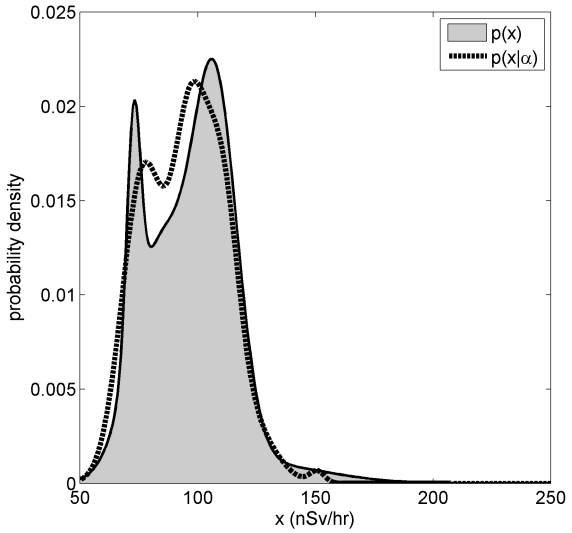
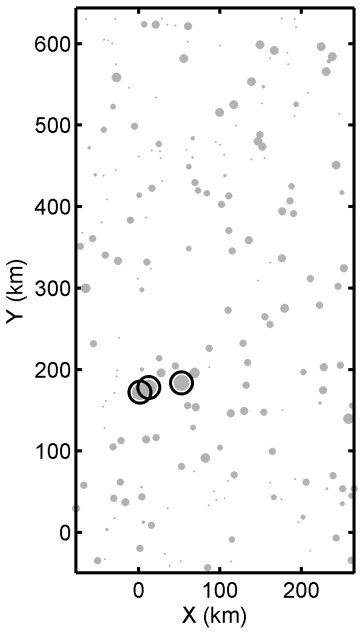
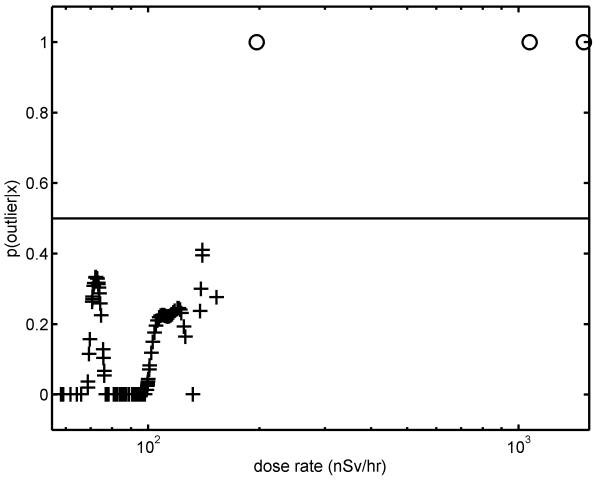

0.025
0.02
0.015
0.01
0.005
0
probability density
50
100
150
200
250
x (nSv/hr)
p(x)
p(x|α)

Figure 4:


600
500
400
300
200
100
0
Y (km)
0
100
200
X (km)

(a)


1
0.8
0.6
0.4
0.2
0
p(outlier|x)
10^2
10^3
dose rate (nSv/hr)

(b)

Figure 5: