EXTREME VALUE STATISTICS FOR NOVELTY DETECTION IN BIOMEDICAL SIGNAL PROCESSING

Stephen J. Roberts

Department of Engineering Science, University of Oxford, UK

NOVELTY DETECTION

Novelty, outlier or abnormality detection is a method of choice in the case where the usual assumption of supervised classification breaks down. This latter assumption is that the data set contains information regarding all classification hypotheses. This assumption will break down under two major conditions. Firstly there may be very few examples of an important class within the data set. This is often the case in medical data analysis or in condition monitoring when data from ‘normal’ system states are far easier and less costly to acquire than abnormal exemplars (patients with rare conditions, for example). In a classification system it becomes increasingly difficult to make decisions into this class as the number of examples reduces. It is argued that, in such a case, it is better to formulate a model of the ‘normal’ data and test for ‘novel’ data against this model [1, 2, 3]. Secondly the data set may contain information regarding all classes in the classification set, but the classification set itself is incomplete i.e. data from some extra class is observed at a future point. Such a classification system will classify this ‘novel’ data erroneously into one of the classes in the classification set. It is the former situation that this paper is ultimately concerned with, although the latter is clearly of importance also.

These generic issues of novelty detection have been dealt with previously in several ways. Typical methods develop a semi-parametric representation using, for example, a Gaussian mixture model (GMM) against which data are assessed. Some methods rely upon heuristic approaches, using maximum (un-normalised) Gaussian response as a measure of novelty [1, 2], creating an artificial ‘novelty’ class [4] or utilising thresholds on a density estimate of the ‘normal’ data [5]. An analytic approach to the statistics of outliers exists namely extreme value theory.

Extreme value theory (EVT) concerns the distributions of data of abnormally low or high value in the tails of some data generating distribution. As we observe more data points the statistics of the largest (or smallest) values we observe change. Knowledge of these statistics is of importance, therefore, in detecting ‘abnormal’ events which are outside our expected data extremes for outlier detection or for rejecting classification or regression on data which lies far from the expected statistics of some training data set. EVT is hence well-suited to the problem of novelty detection.

EXTREME-VALUE THEORY

Extreme-value theory forms representations for the tails of distributions. When discussing the properties of the tails of a distribution we will, for convenience, discuss the right-hand tail. The same theory applies, with minor modification, to the left-hand tail.
Consider a set $\mathcal{Z}_m = \{z_1, z_2, \ldots, z_m\}$ of $m$ i.i.d. (independent & identically distributed) random variables $z_i \in \mathbb{R}$ drawn from a distribution function $\mathcal{D}$ and define $x_m = \max(\mathcal{Z}_m)$, i.e. the largest element observed in $m$ samples of $z_i$.

Fisher & Tippett [6] showed that the limit form of the probability of observing extremum $x_m \leq x$ is given by one of only three distributions. These three distributions may also be regarded as special cases of the generalised extreme-value (GEV) distribution,

$$H_{GEV}(y_m, \alpha, \gamma) = \exp \left\{ - [1 + \gamma y_m]^{-1/\gamma} \right\}$$

(1)

for some $\gamma \in \mathbb{R}$ (the shape parameter) and where $y_m$ is the reduced variate,

$$y_m = \sigma_m^{-1}(x_m - \mu_m)$$

(2)

and $\mu_m, \sigma_m$ are norming parameters. This remarkable result is the equivalent, for maxima of random variables, to the central limit theorem for sums of random variables. As $m \to \infty$ the distribution of maxima from any $\mathcal{D}$ is in the domain of attraction of one of the three limit forms. This paper is ultimately concerned with samples assumed drawn from $\mathcal{D} = [N(0, 1)]$ (i.e. a one-sided normal distribution) whose maxima distribution converges to the so-called Gumbel form [7, 8]. The other two limit forms will not be considered further. The interested reader is pointed to (e.g.) Embrechts et al. [8] for more detailed discussions. The GEV distribution of Equation 1 gives the Gumbel distribution in the limit case of $\gamma \to 0$ [9, 7, 10]. The probability of observing some extremum, $x_m \leq x$ is hence given as:

$$P(x_m \leq x | \mu_m, \sigma_m) = \exp\{-\exp(-y_m)\}$$

(3)

As this is the cumulative distribution hence the associated density function is:

$$p(x_m = x | \mu_m, \sigma_m) =$$

$$\frac{1}{\sigma_m} \exp\{-y_m - \exp(-y_m)\}$$

(4)

For notational simplicity we will henceforth denote $P(x_m \leq x | \mu_m, \sigma_m) = P(x | m)$ and the associated density as $p(x | m)$. As we are ultimately interested in asserting whether an observed point is beyond even the expected extreme value distribution so we may use $P(x | m)$ as a novelty probability.

Whilst maximum-likelihood estimation of the norming parameters is required for arbitrary distributions, if the analytic form of the observations distribution, $\mathcal{D}$, is known then $\sigma_m$ and $\mu_m$ may be estimated in closed form. These forms are asymptotically correct as $m \to \infty$. As was shown, however, in [3] they parameterise extreme-value distributions which fit experimental data well even for small $m$. The calculation of the norming parameters is a standard problem in EVT. As in [3] we build models for arbitrary observation densities using a Gaussian mixture, hence we consider extremes relative to the centres of the Gaussian components. These relative measures are transformed as in [3] to a one-sided normal distribution, $[N(0, 1)]$. The asymptotic form of the location parameter for $\mathcal{D} = [N(0, 1)]$ is given as [3],

$$\mu_m = (2 \ln m)^{1/2} - \frac{\ln \ln m + \ln 2 \pi}{2(2 \ln m)^{3/2}}$$

(5)

and the scale parameter as,

$$\sigma_m = (2 \ln m)^{-1/2}.$$  

(6)

Figure 1 shows the good agreement between the theoretical distributions (lines) and experimental estimates from Monte-Carlo simulations (points). Note that $P(x | m)$ offers a good measure of novelty.

**Gaussian Mixture Models**

The Gaussian Mixture Model (GMM) is widely used in data density estimation and to form the latent (hidden) space of radial-basis function networks. The choice of the Gaussian kernel is not an arbitrary one. Gaussian mixture methods and radial-basis systems using Gaussian bases both have the property of
universal approximation, are analytically attractive and are hence favoured in the literature. The form of the GMM is simple, consisting of a set of \( k = 1..K \) components each of which has standard Gaussian form and is uniquely parameterised by its mean and covariance matrix. For such a mixture Bayes’ theorem implies that the data likelihood, \( p(x) \), may be written as mixture of component (conditional) likelihoods of the form

\[
p(x) = \sum_k \pi(k)p(x|k) \quad (7)
\]

where \( \pi(k) \) are the component priors. If we assume that the statistics of outlying events are dominated by the component closest (in the Mahalanobis sense) then we are interested in the resultant probability of data lying interior to some distance \( h \) from the component mean. Integrating the (one-sided) density function gives

\[
P(h(x)) = \int_{c_{k*}}^{h(x)} p(w|k^*)dw \quad (8)
\]

where \( c_{k*} \) is the mean of the closest component and \( w \) is a dummy variable of integration. The more complex case in which \( x \) is multidimensional is easily dealt with by evaluating the above Equation using the Mahalanobis distance defined in its standard form:

\[
h(x)_{k*} = \sqrt{(x - c_{k*})^T C_{k*}^{-1} (x - c_{k*})} \quad (9)
\]

where \( C \) is the covariance matrix and \( c \) is the centroid. The Mahalanobis distance, evaluated with respect to a Gaussian component, has a density function given by:

\[
p(h(x)) = \frac{2}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} h(x)^2 \right) \quad (10)
\]

Combining Equations 8 and 10 and using the standard form for the Gaussian integral gives:

\[
P(h(x)) = \text{erf} \left( \frac{h(x)}{\sqrt{2}} \right) \quad (11)
\]

where \( \text{erf}(\cdot) \) is the error function.

To evaluate the resultant EVT probability at some point \( x \) it is noted that the theory (and parametric form) of the EVT distribution derived for the \( |N(0,1)| \) distribution (Equation 3) may be used directly as \( h(x) \) is expected (with respect to a Gaussian component) to be distributed as \( |N(0,1)| \). Furthermore, if a total of \( m \) points are observed and used to fit the model so an estimate of the effective number of points ‘seen’ by each component is simply \( m_k = m \pi(k) \). The EVT probability measure is hence written as:

\[
P(x|m) = \sum_k P(h(x)_k|m_k)p(k|x) \quad (12)
\]
where \( p(k|x) \) is the posterior probability of the \( k \)-th Gaussian in the mixture given \( x \). As we are, by definition, interested in extremal points to any Gaussian in the mixture so these posteriors are close to zero for all but one of the components (the one lying closest, in the Mahanolobis sense, to \( x \)), \( k^* \) say, for which \( p(k^*|x) \approx 1 \). The above Equation may hence be simplified as:

\[
P(x|m) \approx P(h(x)_{k^*}|m_{k^*})
\]  

(13)

All the results shown here are obtained using the Expectation-Maximisation (EM) algorithm [11] to estimate a maximum-likelihood parameter set for the means, covariance matrices and component priors in the mixture model. Details of the standard algorithmic approach taken to implement the EM procedure may be found in [12] for example. The number, \( K \), of the components in the GMM is obtained by penalised likelihood methods, as detailed in [13] for example. Once the mixture model is obtained all EVT probabilities are evaluated using Equation 13 above.

**RESULTS**

**Epilepsy data set**

The first data set consists of a single channel of EEG from an epileptic subject. The signal was parameterised over one-second segments using information from the singular spectrum of a phase-space embedding (see [14] for details). A four component Gaussian mixture model was adapted to the parameter space over a 10-minute section of the data in which no epileptic activity was observed. Figure 2 shows the resultant novelty probabilities from (upper trace) the error-function approach and (middle trace) the EVT approach. It is clear that both methods highlight, in particular, the four regions of epileptic activity in this record. The EVT method, however, gives a considerably more interpretable trace as background (non-epileptic) activity has uniformly low probability. The lower plot of the figure shows the activity highlighted by the first region of novelty in the EVT trace (between 63 and 71 seconds into the recording). The characteristic 3Hz ‘spike and wave’ activity is very clear.

![Figure 2: Novelty probability for epileptic seizure activity using error-function of Eqn. 11 (upper plot) and EVT (middle plot). The lower trace shows the first period of activity flagged as novel by the EVT approach. All x-axis times are in seconds from the start of the recording.](image)

**Anaesthesia data set**

The next results presented come from a single channel of EEG recorded during anaesthetic uptake. Much of the problem associated with automated depth of anaesthesia assessment lies in the fact that different anaesthetic agents have differing affects on the EEG [15] thus making a global assessment of anaesthetic depth per se difficult. If, however, we take an approach whereby we build a model of non-anaesthetised EEG activity we may monitor the departure from this of unseen data. During an operative procedure, for example, one would hence consider movement towards this waking state as a cause for concern. From a training set of seven non-anaesthetised subjects a model for waking EEG activity was made. A single channel of EEG was parameterised using a 5th order autoregressive (AR)
model and the reflection coefficients [16] from this model were used as signal descriptors over successive one-second sections of EEG (sampled at 128Hz). A six component Gaussian mixture model was inferred from this data, with the number of components being set via a penalised likelihood measure [13]. Unseen data (no part of which was in the training set) in which a subject is taken from waking to anaesthetised states was then presented to the wake-EEG model and the resultant EVT probability calculated. Figure 3 shows the time course of $1 - P_{EVT}$ for this data. The clear transition from wake to anaesthetised is evident.

![Figure 3: Novelty probability for depth of anaesthesia data. This may be interpreted as a wakefulness probability in this example.](image)

Vigilance data

Whenever people perform repetitive, boring or long-term tasks a loss in concentration can occur. These lapses in vigilance may have serious consequences under certain circumstances. Problematically, from the perspective of analysis, different subjects have very different responses to dealing with tasks when tired. It is therefore appropriate to model a short reference section of ‘normal’ (i.e. awake & alert) and thence screen unseen data with respect to this section. The features used in the analysis were AR reflection coefficients from a single EEG channel and the blink rate evaluated via electrical activity of the eyes. All features were evaluated on a five-second basis. Figure 4 shows the EVT probability [top] along with a tracking task performance measure [middle] (larger values indicate worse performance) and the human-scored ‘alertness’ labels [bottom] over a one-hour section of data. Higher values of this ‘alertness’ label indicates the human scorer considered the EEG to be more associated with alert brain activity. All traces are smoothed using a flat FIR filter of length 60s for comparison purposes only. The first three minutes of another recording (on the same subject) formed the ‘training’ set of ‘normal’ activity. All subsequent analysis screens for novelty against this data. The plots show $\log(1 - P_{EVT})$ and are truncated at $10^{-20}$. Very low regions of the novelty trace, i.e. truncated at $10^{-20}$, occur in the region between 15 and 45 minutes into this one hour recording. This is the region in which the subject’s performance on a boring tracking task was at its poorest. It is noted that predictive methods, i.e. regression from the features onto the task performance measure, are very poor in this instance. Note also that the human-scored labels correlate poorly with the task performance measure (bottom and middle traces).

MRI image

The magnetic resonance data presented in this paper consists of a $100 \times 100$ region of a slice through a tumour patient’s brain. Data was collected using both T2 and proton density (PD) spin sequences, which are used directly to form a two-dimensional feature space. A five component Gaussian mixture was fitted to the entire data set using the Bayesian approach detailed in [13]. The data was thence screened for novelty based on the extreme-value statistic. Figure 5 shows the superposition of one of the original MR images (T2) with the novelty probability. The bright white region near the
Figure 4: Vigilance data: the top trace of the plot is \( 1 - P_{EVT} \) on a log scale. The middle and bottom traces are the tracking performance measure and the smoothed human-scored label set. Very low regions of the novelty trace are truncated at \( 10^{-20} \) and occur in the region between 15 and 45 minutes into the one hour recording. This is the region in which the subject’s performance on a boring tracking task was at its poorest.

The middle bottom of the image is the sole region of very high novelty (with probability close to unity) and this accurately picks out the tumour location in the image.

CONCLUSIONS

The issue of determining whether some datum is indicative of abnormal or novel system behaviour is paramount in areas such as medical data analysis where the acquisition of large quantities of abnormal data is difficult. Several approaches to screening data for such novel segments have been proposed previously. Extreme-value theory and the associated statistics provide, however, a complete and principled approach to the problem which avoids the need for heuristic methods. EVT forms an elegant methodology for any problem in which the chance of finding extremes of some quantity is required.

References


**Acknowledgements**

The author would like to thank Helen Stone and Oreste Ortolani for some of the data used in this paper. Part of this work is supported by BAE SYSTEMS to whom the author is most grateful.