On a Supervised Classifier that can better learn from Mixed Peaks-over-Threshold

Emtair. M. Abdalla
Department of Mathematics, Faculty of Science, Sirte University
Sirte, Libya

Abstract: The traditional supervised classifier is based on the assumption of weighted Gaussian mixing components. The classifier is absolutely misleading when the underlying assumption is much violated as is the case of Peaks-Over-Threshold (POT) data. This paper, presents an alternative parametric classifier that can learn optimally from mixtures of such typically skewed components. The classifier is theoretically based on the assumption that the mixing components can be modeled by Generalized Pareto distribution with apriori known sub-group indices.

Key words: statistical training, POT-data, Optimal Classifier, Generalized Pareto distributions, apriori probability, mixing components, optimal boundary, predicted sub-grouping.

1. Criteria of Classification

Generally, the probabilistic model of a mixture of $g \geq 2$ sub-groups can be presented by the density

$$f(x) = \sum_{k \leq g} \pi_k f_{\theta_k}(x) \quad (1)$$

The model is supposed to be completely specified with its known parameter vectors θ_k, prior probabilitiesπ_k and specific distribution $f_{\theta_k}(x)$ of sub-group G_k such that $\bigcup_{k \leq g} G_k = G$.

Restriction will be to two-group mixtures. The objective is to find atwo-group optimal decision rule

$$\hat{k}(x) = \begin{cases} 1, & x \in \mathrm{G}_1 \\ 2, & x \in \mathrm{G}_2 \end{cases} \quad (2)$$

that can assign an object $x$to its optimal sub-group with loss

$$\ell(\hat{k}(x)|k) = \begin{cases} 0, & \hat{k} = k \\ \ell > 0, & \hat{k} \neq k \end{cases} \quad (3)$$

due to misclassification, through which we define discriminant functions

$$d_k(x) = \sum_{i \leq g} \pi_i \ell(\hat{k}(x)|i) f_{\theta_i}(x), k \leq 2 \quad (4)$$

The decision rule is expected to maximize the posterior probabilities, $p(k|x)$, where

$$p(k|x) = \frac{\pi_k f_{\theta_k}(x)}{f(x)}, f(x) > 0 \quad (5)$$

More details can be found in Falk et al [6]. A completely specified mixture leads to the deduction of an optimal boundary $x^{[opt]}$ and hence predicts the optimal sub-groups $\mathrm{G}_k^{[opt]}, k = 1,2$. The optimal boundary of respective mixture model is represented by the set

$$x^{[opt]} = \{x \in \mathrm{G} : d_1(x|\theta_1) = d_2(x|\theta_2)\} \quad (6)$$

which lead to the optimal sub-grouping

$$\mathrm{G}_j^{[opt]} = \begin{cases} x : x \in \mathrm{G} \wedge d_j(x) = \underset{k}{\operatorname{argmin}}\, d_k(x|\theta_k), & j = 1 \\ x : x \in \mathrm{G} \wedge d_j(x) = \underset{k}{\operatorname{argmin}}\, d_k(x|\theta_k) | \mathrm{G}_1, j = 2 \end{cases} \quad (7)$$

2. GMM-Classifier

The traditional parametric supervised Gaussian mixture modeling (GMM-) classifier is based on the assumption of Gaussian mixture

$$\pi_1 g_{\theta_1}(x) + \pi_2 g_{\theta_2}(x) \quad (8)$$

with parameter vector $\theta_k = (\mu_k, \sigma_k)$, prior probability vector $\pi = (\pi_1, \pi_2)$ and sub-group specific distributions $g_{\theta_k}(x), k = 1,2$.where

$$g_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, x \in \mathbb{R} \quad (9)$$

See Figure 1.

IJRS

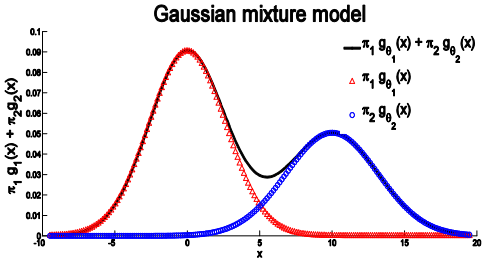

Figure 1. Gaussian mixture modelling of two components.

Figure 1., shows a mixture of two Gaussian components with parameter vectors $\theta_1 = (0,7), \theta_2 = (10,10)$ and prior probability vector $\pi = (0.6, 0.4)$.

3. Limiting Distribution of POT-Data

The conditional distribution of a random variable $X \sim F$ over a threshold u is given by

$$F^{[u]}(x) = Pr(X \leq x | X > u)$$

$$= \frac{F(x) - F(u)}{1 - F(u)}, x \geq u$$

leading to

$$F(x) = \big(1 - F(u)\big)F^{[u]}(x) + F(u), x > u \qquad (10)$$

which is the original distribution left- truncated at u. If x_w is the right endpoint of F and $F^{[u]}(b_u + a_u x)$ has a limiting distribution as $u \longrightarrow x_w$, then according to Balkema. A & de Haan. L., [1] and Pickands. J., [9]., we can write

$$\left|F^{[u]}(x) - W_{\gamma,u,6_u}\right| \longrightarrow 0, u \longrightarrow x_w \qquad (11)$$

this implies that

$$F(x) \approx \big(1 - F(u)\big)W_{\gamma,u,6_u}(x) + F(u), x > u \qquad (12)$$

i.e.,

$$F(x) \approx W_{\gamma,\tilde{\mu},\tilde{\sigma}}(x), x \geq u \qquad (13)$$

where $W_{\gamma,\tilde{\mu},\tilde{\sigma}}(x)$ is the distribution function of generalized Pareto distribution with location $\tilde{\mu} \in \mathbb{R}$, scale $\tilde{\sigma} > 0$ and tail index γ parameters and density function $w' = W_{\gamma,\tilde{\mu},\tilde{\sigma}}$ defined as

$$w_{\gamma,\mu,\sigma}(x) = \begin{cases} \exp\left\{-\left(\frac{x-\tilde{\mu}}{\tilde{\sigma}}\right)\right\} & ,\gamma = 0 \\ \frac{1}{\tilde{\sigma}}\left[1 + \gamma\left(\frac{x-\tilde{\mu}}{\tilde{\sigma}}\right)\right]^{-\left(1+\frac{1}{\gamma}\right)} & , \begin{cases} x > \tilde{\mu} & ,\gamma > 0 \\ x < \tilde{\mu} - \frac{\tilde{\sigma}}{\gamma} & ,\gamma < 0 \end{cases} \end{cases} \qquad (14)$$

for which the location parameters always stands as the left end point of the distribution. which means that the generalized Pareto distribution, GP, can be fitted to the peak-over-threshold, POT., u, of the original distribution function F. This is the theoretical basison which will build our classifier for classifying POT-data mixtures.

4. GPMM-classifier

The two-component generalized Pareto mixture model, GPMM., is introduced as

$$\pi_1 w_{\theta_1}(x) + \pi_2 w_{\theta_2}(x) \qquad (15)$$

with parameter vectors $\theta_i = (\mu_i, \sigma_i, \gamma_i)$ and apriori probabilities $\pi_i, i = 1,2,.$
The GPMM-classifier is based on the assumption that the input data mixture can be approximated using generalized Pareto mixture modeling.

5. Training and Testing the Classifier

Technically, training means using a random sample to estimate the unknown parameters and testing is the process of using the trained model to fit the test sample. To avoid bias, a large enough training and testing samples have to be drawn from the original data mixture. The mechanism of training is as follows; firstly, the classifier separates the sub-groups with the help of known sub-group indices, secondly, estimates the parameter vector θ_k and apriori probabilities $\pi_k, k = 1,2.$, from the each sub-group sample, thirdly, it constructs the discriminant functions as in (4) and applies criterion (6) to find the optimal boundary $x^{[opt]}$ and finally, applies criterion (7) to allocate each objects to its predicted sub-group$G_k^{[opt]}, k = 1,2.$ Oncethe task of predicting sub-groups and their indicesis completed, classification error can be computed by pairwise matchingof the predicted sub-group indices with the test sample indices.

5.1 Parameter Estimation

Parameter estimation is the main task of training. The aprior probability of each sub-group is estimated from the training sample, simply, by dividing the number of the respective sub-group data by the total number of the sample mixture data. For tail index , the maximum likelihood estimator, mle, is the most reliable due to efficiency, especially for $\gamma \leq 1$, and is found by maximizing the likelihood function $L(\theta|X) =$

IJRS

$\prod_{i\leq n} f(x_i|\theta)$ or its logarithm. Iterativetechnique such as Newton-Raphson method can be used to avoid mathematical difficulty, for details ref., to [3],[14]. Pickands'. J., [10] introduced a relatively consistent estimator , pkd., with its refinement. The estimator works for all possible values of the shape parameter. The ordinary moment estimator of the shape parameter, mom, can be expressed as a probability weighted moment, pwm, for details we refer to [11]. Methods like mle, mom and others can be applied to estimate the location and scale parameters, for more information we refer to Reiss, D.,[10], and Grimshaw, D., [7], and Bartoszynski et al [12]. Once the parameters are estimated, they have to be used to construct the model that fitsthe data mixture from which we find optimal boundary and predict the optimal sub-groups.

6. Applications

Application 1.Using different methods to train the classifier estimate the parameters and learn from the data sample.See Table 1.

Estimator	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}$
mle	91.0088	1.9596	0.0002
mom	91.0088	1.9497	0.0052
pkd	91.0088	1.9892	0.0255
pwm	91.0088	2.0856	0.0641

Table 1: application of different estimators toPOT- data.

Table 1. shows application of the estimation methods to estimate the location, scale and shape parameters of POT-data.

Application 2.A comparison between the GPMM-classifier and the GMM-classifier in modeling a mixture ofPOT- components. The GPMM-classifier fitted the mixture using the model

$$0.71\, w_{-0.05,\ 89.89,\ 1.46}(x) + 0.29\, w_{0.08,92.48,\ 2.35}(x)$$

whereas the GMM-classifier fitting is given by

$$0.7 g_{91.31,\ 1.46}(x) + 0.3\, g_{95.01,2.95}\,(\mathrm{x})$$

See Figure 2.

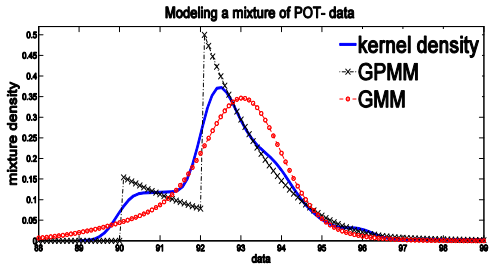

Figure 2. Modeling a mixture of POT-components.

Figure 2. shows a modeling of a mixture of two POT-components using generalized Pareto modeling, GPMM, and Gaussian mixture modeling GMM.

Application 3. Using theGPMM-classifier to estimate the mixing components, find the optimal boundary and predict the sub-groups of the mixture

$$0.5\, w_{0,2,0}(x) + 0.5\, w_{0,4,0}(x)$$

See Figure 3.

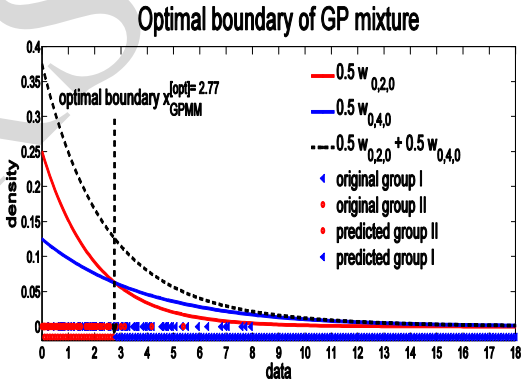

Figure 3. theoretical optimal boundary and sub-grouping of the underlying mixture.

Figure 3. shows the mixture modeling, its optimal boundary and predicted sub-groups modeled by GPMM-classifier.

Application 4. Shows a superiority of the proposed GPMM-classifier to the traditional GMM-classifier in modeling a sample mixture of POT-data generated from the Gaussian mixture

$$0.40\, g_{90,5}(x|x>91) + 0.60\, g_{90,0.8}(x|x>92)$$

The GPMM-classifier estimated the real POT-data using the mixture

IJRS

$$0.6\, w_{-0.27,91,1.8}(x) + \ 0.4\, w_{-0.51,92,1.3}(x)$$

whereas the GMM-classifier estimatedmixture model is

$$0.52\, g_{92.5,1.3}(x) + 0.48\, g_{94,2.7}(x)$$

See Figure 4.

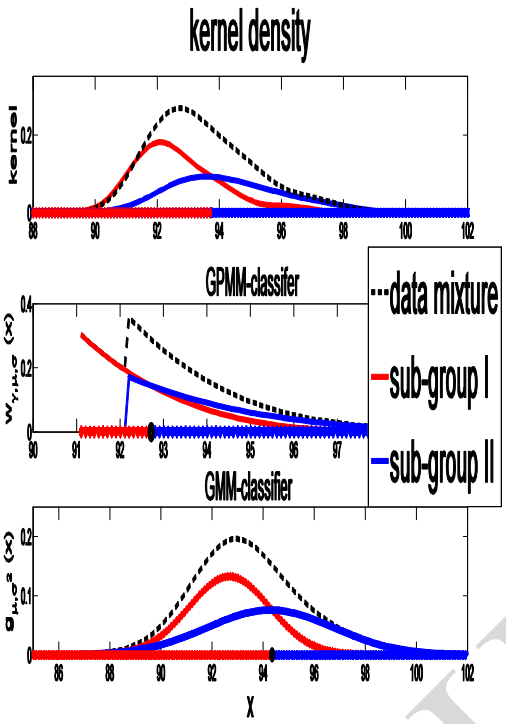

Figure 4. results of kernel density estimator, GPMM-classifier and GMM-classifier.

Figure 4. above: kernel density modeling of the original POT-data mixture with original optimal boundary, and predicted sub-groups. Middle: GPMM-classifier modeling. Below: GMM-classifier modeling. The three models show the optimal boundary, estimated mixture components and predicted sub-groups.

The superiority of the GPMM-classifier relative to the GMM-classifier.Figure 5. shows bootstrap of classification error as computed from 500 repeated training and testing samples of size 81 drawn from 163 sample points of a given POT-data mixture. The two figure show a superiority of the GPMM-classifier to the traditional GMM-classifier, where the minimum, model and maximum values of the GPMM-classifier are less than the corresponding ones of the GMM-classifier.

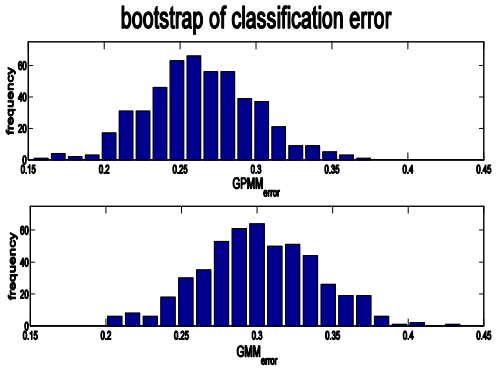

Figure 5. bootstrap of classification error.

7. Discussion and Conclusion

Unlike the traditional Gaussian classifier, the proposed classifier is based on the assumption of generalized Pareto mixing components and hence takes into account the properties of POT-data such as skewness and possible heavy tailedness. This merit makes the classifier superior to the GMM-classifier in modelling samples drawn from POT-data mixtures.

8. Reference

[1] Balkema, A.A. and de Haan, L. (1974). Residual life time at great age. Ann. Probab.2, 792–804.
[2] Bernardo, M. and Smith, M. (1994). Bayesian Theory. John Wiley & Sons, Chichester.
[3] Conte, S. D. and de Boor, C. (1972). Elementary Numerical Analysis. An Algorithmic Approach. McGraw Hill, Tokyo.
[4] Davidson & Smith (1990). Models for Exceedances over high Thresholds. Journal of the Royal Statistical Society B,52, pp. 393-442.
[5] Drees, H.(1993). Refined Estimation of the Extreme Value Index. PhD Thesis, University of Siegen.
[6] Falk, M., Marohn, F. & Tewes, B. (2002). Foundation of Statistical Analysis and Application with SAS. Birkhäuser, Berlin.
[7] Franc & Hlaváĉ (2004). Statistical Pattern Recognition Toolbox for Matlab, ser's Guide. Technical University, Prague.
[8] Grimshaw, S. D. (1993). Computing the Maximum Likelihood Estimates for the Generalized Pareto Distribution. Technometrics, 35, pp. 185-191.
[9] Hasti, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning. Springer, Berlin.
[10] Pickands . J(1975). Statistical inference using extreme value order statistics. Ann. Statist. 3, 119–131.
[11] Reiss, R.-D. and Thomas, M. (2007). Statistical Analysis of Extreme Values. Birkhäuser, Berlin.
[12] Bartoszynski, R and Niewaidomska-Bugaj, M. (2008). Probability and Statistical Inference. Jon Wily & Sons, Inc, Publication.
[13] Schlesinger & Hlaváĉ (2001). Ten Lectures on Statistical and Structural Pattern Recognition. Kluwer Academic Publishers.
[14] Stoer, J. (1972). Einfürung in die Numerische Mathematik I. Springer-Verlag, Berlin.

IJRS