Journal of Physics: Conference Series **490** (2014) 012225 doi:10.1088/1742-6596/490/1/012225

Model based clustering method as a new multivariate technique in high energy physics

Michal Štěpánek, Jiří Franc, and Václav Kůs

Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Trojanova 13, 120 00 Prague 2, The Czech Republic

E-mail: michal.stepanek@fjfi.cvut.cz, francji@fnal.gov, vaclav.kus@fjfi.cvut.cz

Abstract. Analysis of experimental data has one of the most important roles in High Energy Physics. Commonly used multivariate techniques such as Boosted Decision Trees or Bayesian Neural Networks are based on learning algorithms using Monte Carlo generated samples. We implemented a new Model Based Clustering method using Bayesian statistics and a modified iterative Expectation-Maximization algorithm for weighted data that have never been applied in this area. This greatly promising method was developed especially for the data collected from the DØ experiment, which was one of two large particle physics experiments at the Tevatron proton-antiproton collider at Fermilab. We optimized and tested the proposed method in the single top search using a data sample of 9.7 fb^{-1} of integrated luminosity, which corresponds to the entire Run II DØ dataset.

1. Introduction

In this paper we describe a new multivariate analysis method for separations using modified Model Based Clustering (MBC) techniques and an Expectation-Maximization (EM) algorithm. The MBC is an analysis method which is well known in the area of acoustic emission and image reconstruction. This is the first attempt to apply this method in experimental particle physics. The MBC method separates the data into groups by creating a statistical model. We focused on the Gaussian Mixture Model (GMM), whose parameters can be relatively easily obtained by an iterative EM algorithm, which has been modified for weighted events. MBC allows us to separate given data sets without training, but we took the advantage of the available training Monte Carlo (MC) samples and applied the Bayes' rule to compute the *a posteriori* probability of membership of the observation to each data class. The method was tested on the full DØ Run II dataset of 9.7 fb^{-1} of integrated luminosity with corresponding signal and background Monte Carlo. Event selection and background modeling are described in [1]. We used the combination of variables for Bayesian Neural Networks (BNN) and Boosted Decision Trees (BDT) analysis that are presented in [2].

2. Model Based Clustering method

Let $\mathcal{S} = (\omega_1, \ldots, \omega_K)$ denote a finite set of disjoint classes with $P(\bigcup_{k=1}^{K} \omega_k) = 1$, where $P(\omega_k) > 0$ is the prior probability of the k-th class. Assume that $\boldsymbol{x} = (x_1, \ldots, x_D)$ is the observation of a D-dimensional continuous random variable $\boldsymbol{X}$. According to Bayes' theorem, the posterior probability of the k-th class, i.e. the probability that observation $\boldsymbol{x}$ belongs to the k-th class, can be expressed as

$$P(\omega_k \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x} \,|\, \omega_k) P(\omega_k)}{\sum_{k=1}^{K} p(\boldsymbol{x} \,|\, \omega_k) P(\omega_k)}, \tag{1}$$

where $P(\cdot \,|\, \cdot)$ denotes a conditional probability and $p(\cdot \,|\, \cdot)$ is a conditional probability density function. In this study, set $\mathcal{S}$ will be represented as a set of two classes, namely signal and background. We now elaborate on how the *a priori* probabilities $P(\omega_k)$ and the shape of the distribution $p(\boldsymbol{x} \,|\, \omega_k)$ for each data class can be obtained. Training MC samples from the DØ experiment are weighted, i.e. each MC event x_i is associated with a weight $\gamma(x_i)$, and the sum of weights in each class provides an estimate of the expected fraction of events in that class in the real data from detectors. Thus, we can express the *a priori* probability of each class as

$$P(\omega_k) = \frac{\sum_{\omega_k} \gamma(x_i)}{\sum_{k=1}^{K} \sum_{\omega_k} \gamma(x_i)}.$$

Let $\boldsymbol{x} \in \mathbb{R}^{D \times N}$ represent a set of data of dimension D with N independent and identically distributed (i.i.d.) observations. Let $p_1(\boldsymbol{x}|\theta_1), \ldots, p_M(\boldsymbol{x}|\theta_M)$ be parametric probability density functions (pdf) of the same type, $\theta_l \in \Theta$, $l \in \{1, \ldots, M\}$, where M denotes the number of mixture components, $M \in \mathbb{N}$, $M \leq N$, and where $\Theta \subset \mathbb{R}^s$ is a parameter space. Then the *distribution mixture* is any convex combination in the form of

$$p(\boldsymbol{x} \,|\, \theta) = \sum_{l=1}^{M} \alpha_l p_l(\boldsymbol{x} \,|\, \theta_l), \qquad \sum_{l=1}^{M} \alpha_l = 1, \ \alpha_l \geq 0,$$

where α_l denotes the *weight* of the l-th component. The parameters of the distribution mixture can be estimated using the Maximum Likelihood Estimation (MLE) method. In the case of i.i.d. random variables, the logarithmic likelihood function has this form

$$l(\theta \,|\, \boldsymbol{x}) \overset{i.i.d.}{=} \ln \prod_{i=1}^{N} p(x_i \,|\, \theta) = \sum_{i=1}^{N} \ln \sum_{l=1}^{M} \alpha_l p_l(x_i \,|\, \theta_l),$$

where

$$\theta = (\alpha_1, \ldots, \alpha_M, \theta_1, \ldots, \theta_M) \subset \mathbb{R}^M \times \mathbb{R}^{s \times M}.$$

We will maximize the conditional expected value of the so-called *complete data*

$$\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T,$$

which consists of the observable data, $\boldsymbol{x}$, and of the missing data $\boldsymbol{y}$ denoting membership of the data $\boldsymbol{x}$ to the l-th component, i.e.

$$(y_i)_l = \begin{cases} 1, & \text{if } x_i \text{ belongs to the } l\text{-th component,} \\ 0, & \text{otherwise,} \end{cases}$$

$i \in \{1, \ldots, N\}, l \in \{1, \ldots, M\}, \boldsymbol{y} \in \mathbb{R}^{M \times N}, \boldsymbol{x} \in \mathbb{R}^{D \times N}$. For this purpose, the complete logarithmic likelihood function is defined as the logarithm of the probability of the complete data,

$$l_c(\theta \,|\, \boldsymbol{z}) = \ln p(\boldsymbol{z} \,|\, \theta).$$

2.1. EM algorithm for weighted Gaussian Mixture Model

We will define an auxiliary function $Q(\theta, \vartheta)$ as the conditional expected value of the complete data,

$$Q(\theta, \vartheta) = \mathbf{E}[l_c(\theta \,|\, \boldsymbol{z}) \,|\, \boldsymbol{x}, \vartheta],$$

where θ denotes a new (unknown) value of the parameter of the distribution mixture and ϑ denotes an old (known) parameter. This function is maximized using the EM algorithm (see [3]), whose k-th iteration ($k \in \mathbb{N}_0$) consists of two steps:

(i) *E-step*: Calculate the auxiliary function $Q(\theta, \theta^k)$.

(ii) *M-step*: Find $\theta^{k+1} \in \Theta$, which maximizes $Q(\theta, \theta^k)$, i.e. $Q(\theta^{k+1}, \theta^k) \geq Q(\theta, \theta^k),\ \forall \theta \in \Theta$.

It can be proved that in the case of a finite mixture model for weighted data [4, 5], the E-step has the form of

$$Q(\theta, \theta^k) = \sum_{l=1}^{M} \sum_{i=1}^{N} \ln [\alpha_l] p(l \,|\, x_i, \theta^k) \gamma(x_i) + \sum_{l=1}^{M} \sum_{i=1}^{N} \ln [p_l(x_i \,|\, \theta_l)] p(l \,|\, x_i, \theta^k) \gamma(x_i),$$

$$p(l \,|\, x_i, \theta^k) = \frac{p_l(x_i \,|\, \theta_l^k) \alpha_l^k}{\sum_{l=1}^{M} p_l(x_i \,|\, \theta_l^k) \alpha_l^k},$$

where $p_l(x_i \,|\, \theta_l^k)$ denotes the probability that event $x_i \in \mathbb{R}^{D \times 1}$ belongs to the l-th component.

In the case of GMM, i.e. considering the pdf of the l-th component in the form of

$$p_l(x_i \,|\, \theta_l) = \frac{1}{(\sqrt{2\pi})^D \sqrt{|\mathbb{C}_l|}} e^{-\frac{1}{2}(x_i - \mu_l)^T \mathbb{C}_l^{-1} (x_i - \mu_l)},$$

with the vector of expected values $\mu \in \mathbb{R}^{D \times 1}$ and the covariance matrix $\mathbb{C} \in \mathbb{R}^{D \times D}$, then the M-step can be expressed as

$$\alpha_l^{k+1} = \frac{1}{\sum_{i=1}^{N} \gamma(x_i)} \sum_{i=1}^{N} p(l \,|\, x_i, \theta^k) \gamma(x_i), \qquad \mu_l^{k+1} = \frac{\sum_{i=1}^{N} p(l \,|\, x_i, \theta^k) \gamma(x_i) x_i}{\sum_{i=1}^{N} p(l \,|\, x_i, \theta^k) \gamma(x_i)},$$

$$\mathbb{C}_l^{k+1} = \frac{\sum_{i=1}^{N} p(l \,|\, x_i, \theta^k) \gamma(x_i) (x_i - \mu_l^{k+1})(x_i - \mu_l^{k+1})^T}{\sum_{i=1}^{N} p(l \,|\, x_i, \theta^k) \gamma(x_i)}.$$

2.2. Computational aspects of the EM algorithm

In general, the classification accuracy on the training data set can be improved by increasing the number of components, but this does not guarantee a corresponding improvement on test data. This reflects a feature of many supervised classification methods whereby the training data can be overfitted, i.e. the classifier can be overtrained.

In addition, it is crucial to choose appropriate initialization parameters. Convergence of the EM algorithm to a local optimum may produce different results for different runs. We set the initial weight of each component to $\alpha_l^0 = \frac{1}{M}$, while the initial expected values μ_l^0 are set equal to the sample means, and the initial covariance matrices $\mathbb{C}_l^0$ are diagonal matrices containing the sample variance on the diagonal. This choice makes it more likely for the algorithm to converge to a better local maximum.

Furthermore, it has been observed that, when the size of the data set is too small, the variances can become close to zero, which corresponds to subpopulation distributions given by Dirac delta functions. It should also be mentioned that our choice of using diagonal covariance matrices has improved the stability of the algorithm, in addition to simplifying the calculation of the matrix determinant and of the inverse matrix. However, our studies suggest that the choice of the number of components generally has a stronger impact on the results than the form of the covariance matrix.

3. Analysis of single top channel Monte Carlo from the DØ experiment

The top quark, which is the most massive elementary particle ($m_t = 173.5 \pm 0.6 \pm 0.8$ GeV), was discovered at the Tevatron proton-antiproton collider by the DØ and CDF collaborations in 1995 [6, 7]. The top quark can be produced in pairs via the strong interaction, or individually, i.e. without the corresponding anti-particle, via the electroweak interaction ("single top channel"). Single top production was discovered at DØ and CDF in 2009 [8, 9]. There are three modes of single top quark production: s-channel, t-channel, and associated tW production. The tW channel is highly suppressed (i.e. the predicted cross section is very small) at the Tevatron due to the presence of two massive particles in the final state, namely the top quark and the W boson, and is omitted in the most of the studies. The process $q\bar{q} \longrightarrow tb+X$ is referred to as "s-channel" single top production, or tb, and the processes $q'b \longrightarrow tq + X$ and $q'g \longrightarrow tq\bar{b} + X$ are referred to as "t-channel" single top production, or tqb. In this study, the event selection required either two or three jets, and the electron and muon channels were merged. In order to improve the signal purity, only events containing either one or two b-tags were used in the analysis. Since the kinematic properties of single top events are not very different from those of background (i.e. no individual variable provides enough discrimination between signal and background), the use of multivariate analysis (MVA) techniques is expected to improve the results.

Table 1. Results of the separation: signal tb vs. all background in 2-bTag 2-Jets. NcS – the number of signal components, NcB – the number of background components, $Err_{\text{S-}*}$ – the error in the signal set, Err_* – the error in the whole set.

NcS	NcB	$\text{AUC}_{\text{ROC-TS}}$	$\text{AUC}_{\text{ROC-STS}}$	$\text{AUC}_{\text{ROC-YS}}$	Err_{TS}	Err_{STS}	Err_{YS}	$Err_{\text{S-TS}}$	$Err_{\text{S-STS}}$	$Err_{\text{S-YS}}$
1	350	61.622	67.355	58.788	5.539	5.398	6.934	99.876	99.877	99.893
350	350	70.589	80.319	66.596	19.420	14.359	23.709	59.872	50.340	59.146
20	110	65.946	68.895	62.498	8.844	8.228	10.758	89.829	89.218	89.598
290	110	71.384	76.938	67.086	25.166	22.675	30.199	47.783	41.123	47.231
350	110	71.339	78.143	67.412	27.462	25.075	32.946	44.185	34.771	42.952
20	80	62.563	63.998	59.420	5.710	5.734	7.178	99.592	99.654	99.666
290	80	71.065	76.800	67.083	29.017	26.652	34.336	42.577	35.555	41.885
170	20	69.699	73.134	65.759	43.201	42.165	49.338	26.716	22.829	26.256
1	1	70.694	70.984	66.917	11.902	11.774	14.503	80.507	80.683	80.207

The MBC method was tested on single top Monte Carlo data, corresponding to the full DØ Run II data set. Details about the Monte Carlo data sets and the official DØ analysis using BDT, BNN, and the Matrix Element (ME) method are given in [10, 11]. The MBC method was applied to 12 sub-tasks $\{\text{signal} : tb, tqb, tb + tqbg\} \times \{\text{1-bTag, 2-bTag}\} \times \{\text{2-Jets, 3-Jets}\}$, each corresponding to a subset containing a number of variables between 20 and 39. We focused on the 2-bTag 2-Jets channel, which is the most sensitive to s-channel single top production with the *a priori* signal to background ratio 1:18. All Monte Carlo events were divided into three samples: a "small training sample" (STS) containing $\frac{1}{4}$ of all MC events, a "testing sample" (TS) containing $\frac{1}{4}$ of all MC events, and a "yield sample" (YS) containing the remaining $\frac{1}{2}$ of all MC events, which is used for the final comparison with the real data. The area under the ROC curve (AUC) varied between 0.65 and 0.8 depending on the analysis channel. More detailed results are given in table 1 with regard to the tb 2-bTag 2-Jets channel.

Figure 1 shows that the number of events with a high probability of being classified as signal increases with the number of components. This suggests that better results can possibly be obtained for signal as opposed to background when a higher number of components is used.

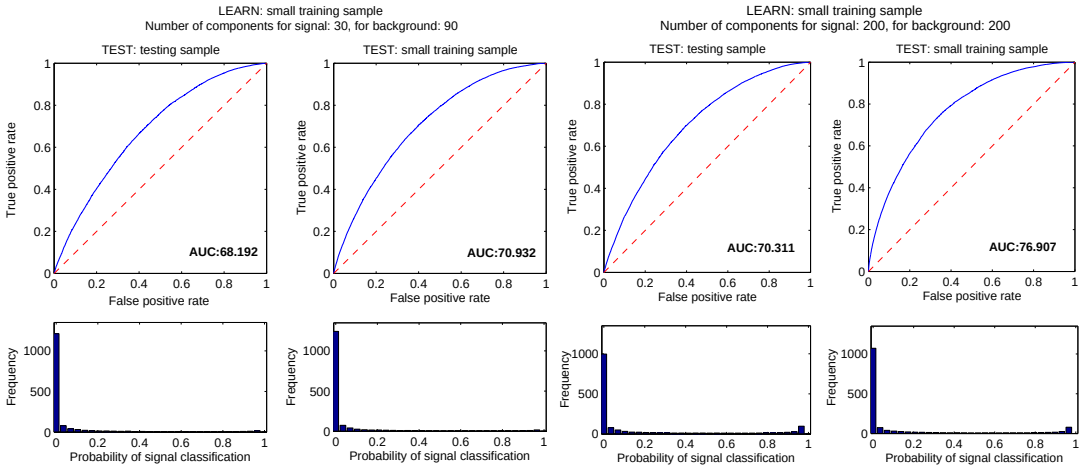

Figure 1. ROC curves and histograms of component weights for signal tb vs. all background in 2-bTag 2-Jets with two different settings of the number of components.

4. Discussion

Our results show that MBC makes it possible to separate signal from background in high energy physics applications. We typically doubled the signal to background ratio for all sub-tasks. Using 290 signal and 110 background components of the distribution mixture, we obtained the best AUC value on the testing sample of ~ 71.4, thereby improving the signal tb to background ratio in the 2-bTag 2-Jets from 1:18 to 1:8 with the *a posteriori* probability threshold $P_t(\text{signal}\,|\,\boldsymbol{x}) = 0.5$ in (1). As expected, the MBC method provides better results when the difference between the signal and the background distribution is more pronounced. Unfortunately, in single top channels the patterns of signal and background are nearly the same. In order to improve the results, we plan to implement a transformation of the input variables using a combination of ϕ-divergences (see [12]) with principal component analysis. Further studies are necessary in order to identify the optimal number of components in each channel.

Acknowledgments

This work has been supported by the MSMT (CZ) grant INGO II INFRA LG12020.

References

[1] Tsai Y T *et al* 2012 single top quark production in 9.7 fb^{-1} of data event selection and background modeling (DØ note 6365)
[2] Bala S *et al* 2012 Measurement of the single top production cross section in 9.7 fb^{-1} of data (DØ note 6375)
[3] McLachlan G and Peel D 2000 Finite mixture models *J. Wiley & Sons*
[4] Bilmes J 1997 A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models *Technical Report* TR-97-021, ICSI
[5] Štěpánek M 2013 Aplikace separačních metod na reálná data z urychlovačů částic (CTU FNSPE)
[6] Abe F 1995 *et al* Observation of top quark production in $p\bar{p}$ collisions with the collider detector at Fermilab *Phys. Rev.* **74**, No. 14, pp. 2626-31
[7] Abachi S 1995 *et al* Search for High mass top quark production in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV *Phys. Rev.* **74**, No. 13, pp. 2422-26
[8] Abazov V M 2009 *et al* Observation of single top-quark production *Phys. Rev.* **103**, No. 9, pp. 092001-7
[9] Aaltonen T 2009 *et al* Observation of electroweak single top-quark production *Phys. Rev.* **103**, No. 9, pp. 092002
[10] DØ Collaboration 2013 Evidence for s-channel single top quark production in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV *Preprint* arXiv:1308.6690
[11] Tsai Y T 2013 Measurement of electroweak top quark production at DØ (University of Rochester)
[12] Kůs V *et al* 2008 Extensions of the parametric families of divergences used in statistical inference *Kybernetika* **44**, No. 1