Modification of Gaussian mixture models for data classification in high energy physics

2015 J. Phys.: Conf. Ser. 574 012150

(http://iopscience.iop.org/1742-6596/574/1/012150)

Modification of Gaussian mixture models for data classification 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. In high energy physics, we deal with demanding task of signal separation from background. The Model Based Clustering method involves the estimation of distribution mixture parameters via the Expectation-Maximization algorithm in the training phase and application of Bayes' rule in the testing phase. Modifications of the algorithm such as weighting, missing data processing, and overtraining avoidance will be discussed. Due to the strong dependence of the algorithm on initialization, genetic optimization techniques such as mutation, elitism, parasitism, and the rank selection of individuals will be mentioned. Data pre-processing plays a significant role for the subsequent combination of final discriminants in order to improve signal separation efficiency. Moreover, the results of the top quark separation from the Tevatron collider will be compared with those of standard multivariate techniques in high energy physics. Results from this study has been used in the measurement of the inclusive top pair production cross section employing DØ Tevatron full RunII data (9.7 fb^{-1}).

1. Introduction

One of the most important step in the data analysis in experimental high energy physics (HEP) is the discrimination phase and the separation of signal from the noise, called background. A lot of machine learning techniques have been adapted for the last decade and they have become an essential part of the analysis tools. The Model Based Clustering (MBC) method and the Expectation-Maximization algorithm (EM) are well known techniques in acoustic emission, image recognition and so on, but with no utilization in the HEP. The first attempt to use MBC in this area was done in the single top analysis at Tevatron DØ experiment [1] and the method was described for example in [2].

This paper extends the first attempts and describe the modification and the genetic optimization of the MBC method. The application from the analysis of the inclusive top pair production cross section in the $t\bar{t}$ lepton+jets channel at DØ experiment is presented and it results will be used to improve the analysis that has been done on previous data sample (5.3 fb^{-1}) [3].

2. Model Based Clustering method

MBC method is a statistical classification technique based on creating mixture model of training sample for subsequent separation of testing sample. Let $\boldsymbol{x} \in \mathbb{R}^{D\times N}$ represent a set of data of dimension D with N independent and identically distributed observations and Θ represent a parameter space. According to Bayes' theorem, the *a posteriori* probability that observation $\boldsymbol{x}_i$

belongs to the class of signal can be expressed as follows

$$P(\text{signal}\,|\,\boldsymbol{x}_i) = \frac{p(\boldsymbol{x}_i\,|\,\text{signal})P(\text{signal})}{p(\boldsymbol{x}_i\,|\,\text{signal})P(\text{signal}) + p(\boldsymbol{x}_i\,|\,\text{background})P(\text{background})},$$

where $P(\cdot\,|\,\cdot)$ denotes a conditional probability and $p(\cdot\,|\,\cdot)$ is a conditional probability density function. We can express the *a priori* probabilities to signal or background as the ratio of corresponding events in training MC samples. Distributions of signal and background are considered as a Gaussian distribution mixture model (GMM)

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

where α_l denotes the weight of the l-th component, $\boldsymbol{\mu} \in \mathbb{R}^{D\times 1}$ is the vector of expected values, and $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ is the covariance matrix. In order to avoid the overtraining, the quality of testing sample classification must be taken into account for the choice of the number of mixture components. Let $\boldsymbol{z}$ be the complete data containing both observed data $\boldsymbol{x}$ and their membership to mixture components. Parameters of the GMM can be estimated via the EM algorithm, whose k-th iteration ($k \in \mathbb{N}_0$) consists of two steps:

(i) *E-step*: Calculate the auxiliary function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^k) = \mathbf{E}[l_c(\boldsymbol{\theta}\,|\,\boldsymbol{z})\,|\,\boldsymbol{x}, \boldsymbol{\theta}^k]$.
(ii) *M-step*: Find $\boldsymbol{\theta}^{k+1} \in \Theta$ maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^k)$, i.e., $Q(\boldsymbol{\theta}^{k+1}, \boldsymbol{\theta}^k) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^k),\ \forall \boldsymbol{\theta} \in \Theta$.

In the case of GMM for weighted data, i.e., each observation $\boldsymbol{x}_i$ is associated with a weight $\gamma(\boldsymbol{x}_i)$, new values of parameters α_l^{k+1}, $\boldsymbol{\mu}_l^{k+1}$, $\boldsymbol{\Sigma}_l^{k+1}$ are given by the solution of M-step [4].

Let us mention a few words about the shape of covariance matrix $\boldsymbol{\Sigma}$. Considering the product mixture model in the form

$$p_l(\boldsymbol{x}_i\,|\,\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) = \prod_{d=1}^{D} p_l(x_{di}\,|\,[\boldsymbol{\mu}_l]_d, [\boldsymbol{\Sigma}_l]_{dd}),$$

i.e., employing solely diagonal covariance matrices, provides several advantages. First of all, choice of using diagonal covariance matrices improves the stability of the algorithm, in addition to simplifying the calculation of the matrix determinant and of the inverse matrix. Our studies suggest (see [4]) that the choice of the number of components generally has a stronger impact on the results than the form of the covariance matrix. This model is also directly applicable to incomplete data without estimating the missing values or their omission. Let us consider the product mixture model, $\mathcal{D}(\boldsymbol{x}_i)$ denote the subset of indices for the defined variables in a given vector $\boldsymbol{x}_i$, and $\mathcal{X}_d$ be the subset of vectors $\boldsymbol{x}_i$ with defined value of variable d,

$$\mathcal{D}(\boldsymbol{x}_i) = \{d \in \widehat{D}\ :\ x_{di} \text{ is defined}\}, \quad \mathcal{X}_d = \{\boldsymbol{x}_i,\ i \in \widehat{N}\ :\ d \in \mathcal{D}(\boldsymbol{x}_i)\}.$$

The formula for computation of the posterior probabilites of the missing data can be expressed in the form of

$$p(l\,|\,\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{\alpha_l \prod_{d\in\mathcal{D}(\boldsymbol{x}_i)} p_l(x_{di}\,|\,[\boldsymbol{\mu}_l]_d, [\boldsymbol{\Sigma}_l]_{dd})}{\sum_{l=1}^{M} \alpha_l \prod_{d\in\mathcal{D}(\boldsymbol{x}_i)} p_l(x_{di}\,|\,[\boldsymbol{\mu}_l]_d, [\boldsymbol{\Sigma}_l]_{dd})},$$

where $p_l(x_{di} \mid [\boldsymbol{\mu}_l]_d, [\boldsymbol{\Sigma}_l]_{dd})$ is the probability density function of a one-dimensional Gaussian distribution with mean $[\boldsymbol{\mu}_l]_d$ and standard deviation $[\boldsymbol{\Sigma}_l]_{dd}$. Analogously, the modification of the M-step is done by computation of the parameters of each dimension d only for vectors $\boldsymbol{x}_i \in \mathcal{X}_d$.

2.1. Genetic optimization of the Expectation-Maximization algorithm

Due to the fact that convergence of the EM algorithm is strongly dependent on the initial values, genetic optimization for finding maximum of the logarithmic likelihood function (i.e., the fitness function) is extremely useful. Firstly, the population $\mathcal{P}$ of P individuals $\mathcal{I}_n$, $n \in \widehat{P}$, is initialized by random starting values of the EM algorithm: α_l^0, $\boldsymbol{\mu}_l^0$, $\boldsymbol{\Sigma}_l^0$, $l \in \widehat{M}$. In the k-th iteration, new posterior probabilites $p(l|\boldsymbol{x}_i, \boldsymbol{\theta}^k)$, $i \in \widehat{N}$, are computed for each individual. Hence, the individual is associated with a $M \times N$ matrix containing current states, where the individual has evolved during the iteration process and by undergoing genetic-type operations. In this case, the gene of an individual is represented by index $i \in \widehat{N}$ associated with the probabilites of the membership to mixture components for one specific observation $\boldsymbol{x}_i$. Reproductive success is proportional to the fitness function employing rank selection, which assigns the values to individiuals according to the ranking of their fitness – the worst will have fitness 1, second worst fitness 2 etc. and the best will have fitness equal to the number of individuals in population $\mathcal{P}$.

Mutation is carried out using a biased mutation operation inspired by [5], which transforms the mutation operation into a guided search that can rapidly find the optimal combination of cluster assignments. It means that the probability of mutating to a cluster is related to the Mahalanobis distance from the data point to the center of the cluster, i.e., the closer a data point is to a cluster center, the higher the probability of mutating to this cluster. Genes of the individual are randomly selected with probability $p_{mut} \approx 0.1 - 0.5$.

In order to maintain the monotonic convergence property of the EM algorithm, the elitism must be employed. Elitism first copies a few best individuals to new population and the rest of evolution is accomplished by mutation. Moreover, a very common operation is parasitism, which can increase the variability of the population and thus improve performance of genetic algorithm. In every epoch, some of the worst individuals do not undergo the second selection and go directly into the new population.

2.2. Variable transformation

Variable transformation provides more suitable input for the GMM and also possibility of a combined separation via the MBC. Combination of classification methods means using their output as a new variable input, thereby creating a matrix whose columns are formed by the posterior probabilites of the membership to signal (or discriminants, in general) from different methods. Gaussianization using the Probability Integral Transformation (see [4]) can be expressed by mapping

$$x \mapsto \sqrt{2}\,\mathrm{erf}^{-1}\left(2 \cdot \int_{-\infty}^{x} f_X(t)\mathrm{d}t - 1\right),$$

where $\mathrm{erf}(\cdot)$ is the error function and $f_X(\cdot)$ is the probability density function of a random variable X. Suppose $x > 0$, the univariate Box-Cox transformation is defined by

$$x(\lambda) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \lambda \neq 0, \\ \ln(x) & \lambda = 0, \end{cases}$$

by which the variable is transformed to a normally distributed variable in terms of skewness and kurtosis. For both ways of gaussianization, worth noting that the transformation parameters can be estimated for signal or background set separately and we subsequently choose which one will be used for the gaussianization.

The goal of the decorrelation is the transformation of a set of r.v. $\boldsymbol{X} = (X_1, \ldots, X_D)^T$ with a known covariance matrix $\boldsymbol{\Sigma}$ into a set of r.v. with the identity matrix as a covariance matrix. Suppose zero mean, decorrelation can be performed by

$$\boldsymbol{Y} = \mathbb{D}^{-\frac{1}{2}}\mathbb{U}^T\boldsymbol{X},$$

where $\mathbb{D}$ is a diagonal matrix with eigenvalues of $\boldsymbol{\Sigma}$ on the diagonal and $\mathbb{U}$ is composed of the corresponding eigenvectors.

3. Application of the MBC method in high energy physics

The presented MBC method with genetic modification was used in the analysis of the inclusive top pair production cross section employing DØ Tevatron full RunII data (9.7 fb^{-1}). Our task in this analysis was to separate signal from the background in the $t\bar{t}$-lepton+jets channel. This decay family has 6 subtask according to the lepton type (electron, or muon) and to the number of jets in the decay (2,3, or 4+). The DØ collaboration prepared the final selection in lepton+jets channel, which is described in [6]. The ratio of the signal differs from 1.5 % in 2 jetbins channels to nearly 45 % in 4 jetbins channels. The required Monte Carlo samples, corresponding to real blind data sets from the detector, were generated from Alpgen+Pythia generator and split into 3 groups - training, testing and validation (called yield). Each group matched the real data in accordance with the number of events (sum of weights of entries describing the simulation of one single decay).

The results of the separation for muon 4+ jets channel are presented in the figures 1 and 2. The former shows that the output discriminants from the blind data has the same distribution as the validation MC sample, the ROC curve for this channel, the optimal cut computed as the maximum of the ratio of true positive S and $\sqrt{S+B}$ (where B stands for false positive), and finally, the check of the overtraining. It is clear from the last subplot that the output discriminants of the signal and the background have the same distribution both in the training sample and in the testing sample.

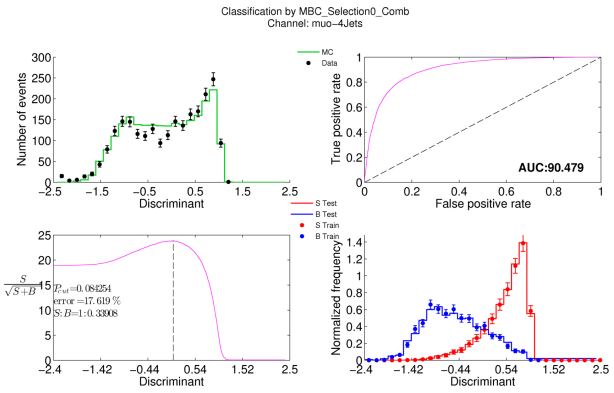

Figure 1. Control plots of results from muon 4+ Jets channel

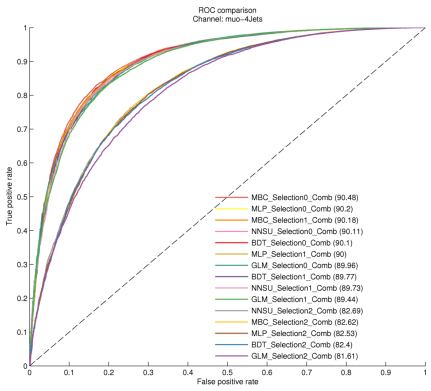

Figure 2. Results muon 4+ Jets channel: MBC, MBCcomb, ROC comparison.

The figure 2 shows the comparison of the quality of various methods ranked by the area under the ROC curve. Different variables selections were used during the analysis. It can be observed that the results of the MBC method is comparable with other well known standard multivariate techniques in high energy physics. The MBC method had even the best result in the presented channel (muon, 4+jets) displayed in the figure 2.

4. Discussion

The genetic modification of the MBC method and its application in high energy physics was presented. Although the field of high energy physics is one of the most difficult for machine learning and separation, the MBC method proved its quality and its results are at least comparable with other classification multivariate techniques. There are still some pending work and the MBC method can be improved via modification of the fitness function in terms of focus on testing sample separation efficiency, for instance.

Acknowledgments

This work has been supported by the MSMT (CZ) grant INGO II INFRA LG12020 and CTU (CZ) grant SGS12/197/OHK4/3T/14.

References

[1] 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
[2] Štěpánek M, Franc J, and Kůs V 2014, Model based clustering method as a new multivariate technique in high energy physics, Journal of Physics: Conference Series 490 - 012225.
[3] Abazov V M 2011 *et al*, D0 Collaboration, Measurement of the top quark pair production cross section in the lepton + jets channel in proton-antiproton collisions at $\sqrt{s} = 1.96$ TeV *Phys Rev. D.* Vol. 84, Issue 1.
[4] Štěpánek M 2014 Application of statistical methods for separation of signal in the production of top quark at the DØ experiment (CTU FNSPE)
[5] Wicker J E 2006 Applications of modern statistical methods to analysis of data in physical science (University of Tennessee)
[6] Abazov V M 2014 *et al*, D0 Collaboration, Measurement of differential $t\bar{t}$ production cross sections in $p\bar{p}$ collisions *ARXIV:1401.5785*