MatrixNet:

Using a new multivariate technique in high energy physics

August, 2013

Author: Viktoria Doneva

Supervisor: Till Moritz Karbach

CERN openlab Summer Student Report 2013

Project Specification

This project focuses on testing and developing algorithms for multivariate data analysis, that separate signal processes from abundant backgrounds and on helping with organizing and filtering colossal amounts of raw data, gathered from the Large Hadron Collider beauty (LHCb) experiment, to find extremely rare events of interest.

Moreover, working on this project also meant trying to apply new, faster method to take the place of systems that are now used at CERN to pare down the relevant data, but require relatively extensive processing and analysis to determine relevance and usefulness.

.

Abstract

Separating signal processes from abundant backgrounds is the first crucial step in the analysis of particle collision data.

The more signal one is able to isolate, the more precise measurements of the physical parameters can be made, and the more knowledge can be gained about the theory that predicts the interactions of elementary particles.

MatrixNet (Yandex) is a new multivariate analysis (MVA) tool that has the potential of increasing over the performance of existing MVA tools. The scope of this project is to apply MatrixNet to a dataset collected by the LHCb experiment, to isolate decays of the Bs particle into a Ds particle and a kaon. With these, a central parameter of the Standard Model of particle physics, the CKM angle gamma, can be better constrained.

Table of Contents

1. Introduction: MatrixNet and CERN 5

2. MatrixNet 6

3. Implementation 7

3.1 Interface (server side) 7

3.2 Client side 9

4. Training optimization 11

5. Future work 12

6. Conclusion 12

1. Introduction: MatrixNet and CERN

Yandex (a Russian Internet company which operates the largest search engine in Russia) is one of the CERN partners in the openlab program.

MatrixNet is a machine learning tool that helps Yandex rank search results for relevancy and improves search results based on a wide range of dynamic factors related to the Web pages that match any particular query.

In the scope of this project is applying the MatrixNet search techniques to data analysis for complex physical processes and the filtering of huge data sets, with the purpose to help in identifying extreme outliers with a high level of precision, in accurately establishing statistical relevance and more generally in confirming theories and discoveries with more certainty.

It is a machine learning technology that taps into algorithms that consider tens of thousands of dynamically weighted factors to find, rank, and return search results. It is also capable of limited learning by building on previous experience.

The MatrixNet technology has been specifically designed to process massive data sets using a complex ranking formula while mitigating over-fitting (finding dependencies and relationships between data points that do not actually exist) to present results that are as relevant as possible for a machine to determine.

2. MatrixNet

One of the problems in machine learning is overfitting. When a computer uses a large number of factors on a relatively small learning, it begins to find dependencies that do not exist. A computer may deem this accidental combination of factors to be essential for a search result to be relevant to the search query.

MatrixNet is a method of machine learning whose key feature is its resistance to overfitting, which allows the Yandex' search engine to take into account a very large number of factors when it makes the decision about relevancy of search results.

MatrixNet allows generating of a very long and complex ranking formula, which considers a multitude of various factors and their combinations. Alternative machine learning methods either produce simpler formulas using a smaller number of factors or require a larger learning sample.

Another important feature of MatrixNet is that allows customization of a ranking formula for a specific class of search queries, that means that it allows adjusting specific parameters for specific classes of queries without causing a major overhaul of the whole system.

In addition, MatrixNet can automatically choose sensitivity for specific ranges of ranking factors.

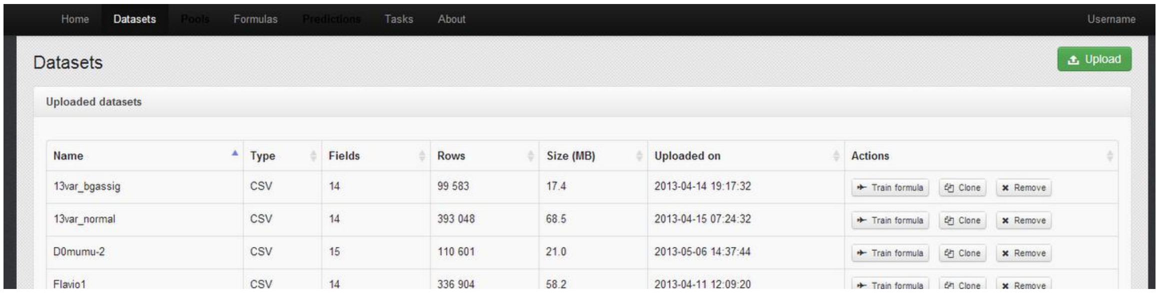
3. Implementation

3.1 Interface (server side)

The MatrixNet tool used in this project has very user friendly interface which gives access to the Yandex MatrixNet classifier. So for the implementation of MatrixNet training on particle physics data, the process went as follows:

- Generating a dataset for training a formula, and adding a variable for classifying every entry as signal or background entry in order for the MatrixNet training to be applied

- Uploading the dataset to the MatrixNet software (Figure 1)



Datasets

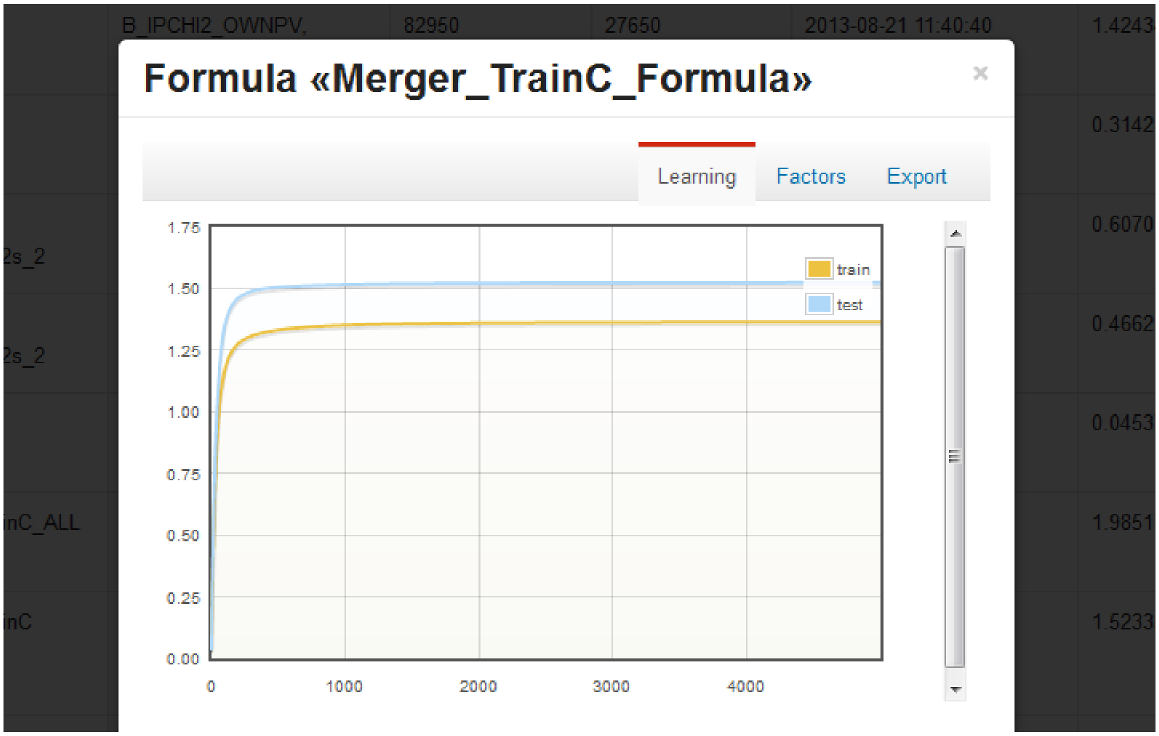
Uploaded datasets

Name	Type	Fields	Rows	Size (MB)	Uploaded on	Actions
13var_bgassig	CSV	14	99 583	17.4	2013-04-14 19:17:32	Train formula Clone Remove
13var_normal	CSV	14	393 048	68.5	2013-04-15 07:24:32	Train formula Clone Remove
D0mumu-2	CSV	15	110 601	21.0	2013-05-06 14:37:44	Train formula Clone Remove
Flavio1	CSV	14	336 904	58.2	2013-04-11 12:09:20	Train formula Clone Remove

Figure 1: Snapshot of the interface, part for uploading a data set

- Training a formula on the dataset with manually choosing the factors of interest and iterations of the training

- Analyzing the formula after it is generated (Figure 2)



Formula «Merger_TrainC_Formula»

Figure 2: Analyzing the formula

- Downloading the trained formula and applying it on a dataset

3.2 Client side

Once the formula is downloaded, it should be applied to a dataset, as previously stated. In this project I used original data gathered from the LHCb experiment. On the picture below, a mass distribution of this data set is presented.

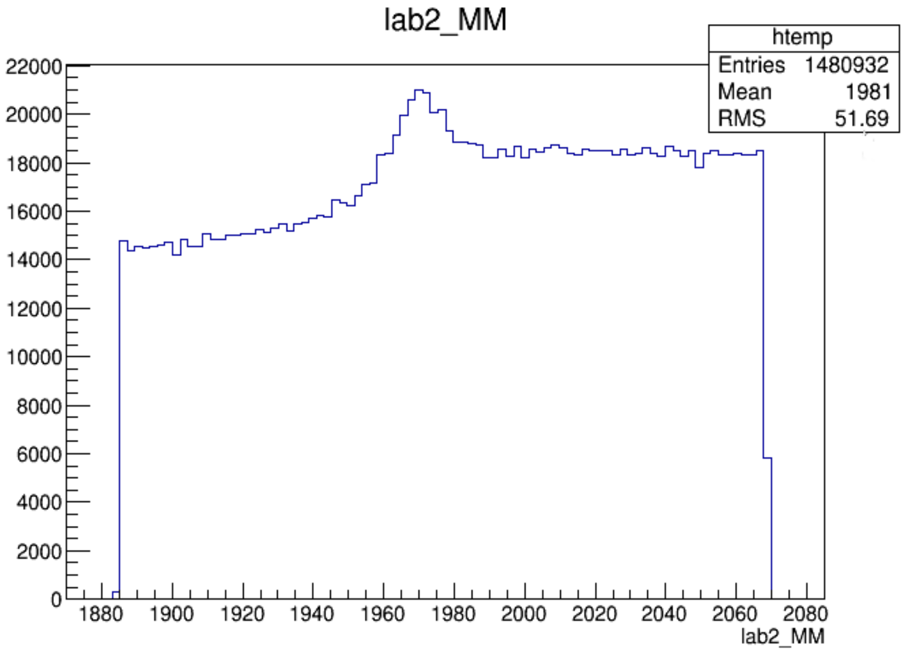


Figure 3: Mass distribution of lab2_MM

The application of the formula on the data set results in additional field in the original dataset: the MN variable which is used for determining whether a certain entry is a signal or background (values near 0 for background and 1 for signal). Below is a plot of the MN distribution after application on the data set.

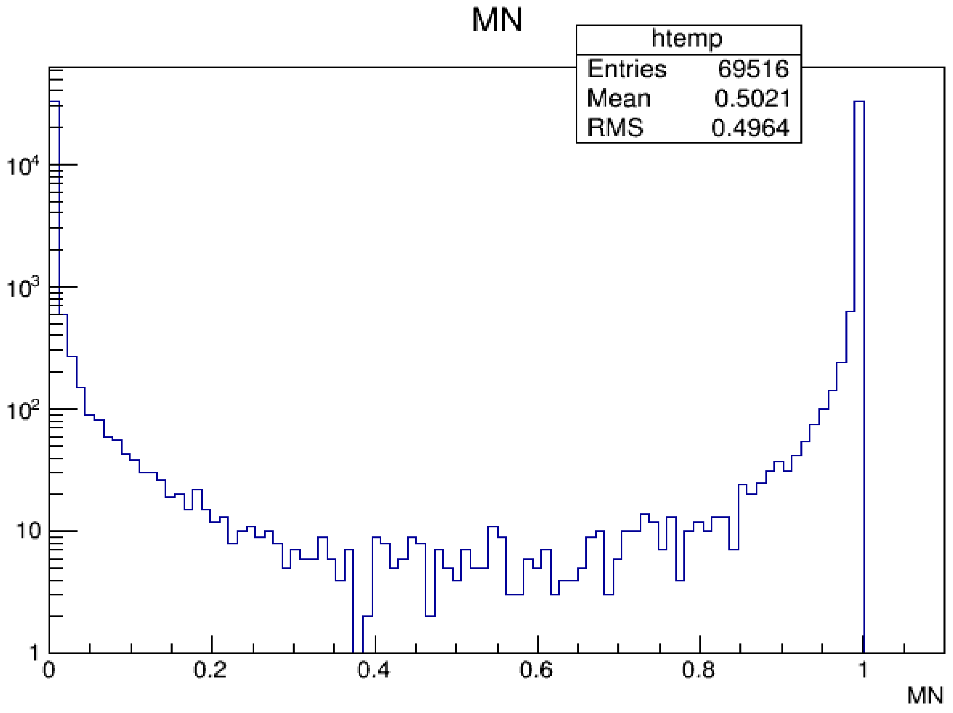


Figure 4: Distribution of MN

What is noticeable here is that MN has a continuous distribution without a clear border between estimated signal and background events. This indicates that a cut needs to be applied on MN, which can be done manually, by testing and determining which is the best value for the cut.

However this is not very scientific and reliable method. So the other option is to calculate the best cut.

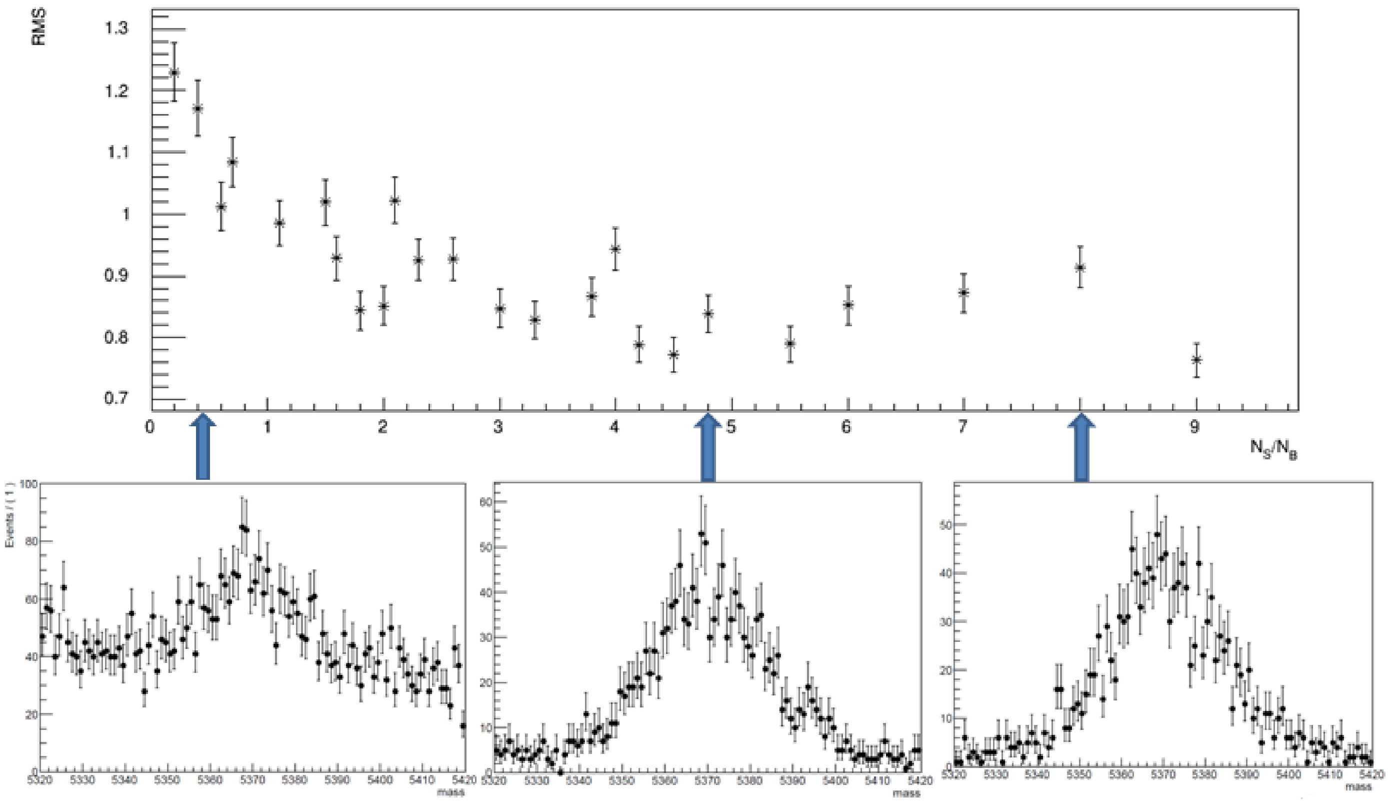
4. Training optimization

For the purpose of calculating the optimal value for the cut on the MN variable, usage of the fitter for training was again required. This process went as follows:

- Generating sets of toys with varying S/B ratio: For this project I used sets of 400 toy data sets for each value of signal/background (the number of signal entries stays the same, only the background events vary in number)

- Each sets has different value of gamma and different gamma error that are of great importance

Below is a graph of all the gamma and gamma error values I acquired in dependence of the signal/background ratio, and also few plots of the mass distribution of some of the sets of toys generated for providing better visualisation.



Using this methodology, the value for which the gamma error has the lowest value can be computed, and that is the value that needs to be used for the optimal MN cut.

5. Future work

Future analysis may result in improvements in the accuracy of the calculations used in this methodology.

One suggestion for determining the correct value for the cut on MN even more precisely is significantly increasing the number of toys produced, on which the analysis are done. Also the factors used in the training can be further optimized and sorted by relevance, process which required further testing of the entire method as explained.

6. Conclusion

This topic is of great importance and interest in the particle physics analysis. Using this method can be very helpful in terms of speeding up discoveries and/or improving the confidence level of findings.

It supports early analysis on the data in contrast to the extensive calculations used now.

This technology allows physicists to perform filtering of huge datasets in order to find extremely rare events. By achieving a high level of precision in identifying such events, scientists can confirm or refute physical models and theories.