Finding a Suitable Compression Algorithm for Input Data in Electron Identification Algorithms for the ALICE Transition Radiation Detector

RYAN WONG, University of Cape Town

The ALICE Experiment at CERN's Large Hadron Collider is an experiment to research nuclear matter under extreme conditions. One of the main detectors is the Transition Radiation Detector (TRD) which is designed to measure and identify electrons.
A major upgrade of the ALICE detector will increase the data rate that has to be processed online. To compensate for the increase in the data rate, online data processing will need to be redesigned. The algorithms for electron identification will be required to run on the the data stream which has been compressed. The compression algorithms require the compressed data to still provide comparable electron identification capabilities as with uncompressed algorithms. Various compression techniques, such as lossy and lossless compression, need to be tested to determine the most efficient compression algorithm while ensuring accuracy when performing electron identification. The considerations for particle identification required when compressing data are the two key data items, the average pulse height and the dependence on the drift time. The goal is to ensure the pion efficiency (fraction of pions misidentified as electrons at fixed electron efficiency) is at minimum when apply electron identification algorithms with compressed data inputs.

Additional Key Words and Phrases: Electron/pion separation, Data Compression, Lossy compression, Lossless compression, ALICE Transition Radiation Detector, Electron Identification Algorithms

1. INTRODUCTION

A Large Ion Collider Experiment (ALICE) is a heavy ion experiment at CERNs Large Hadron Collider [Lippmann et al. 2006]. The aim of the experiment is to study interacting matter at extreme energy densities. One of the main detectors for electron identification in ALICE is the Transition Radiation Detector [Kweon et al. 2009].

The Transition Radiation Detector (TRD) is part of in the central barrel of ALICE and surrounds the Time Projection Chamber (TPC). The TRD has a total of 540 chambers. Each of the TRD chambers consists of a radiator, drift chamber with pad readout and electronic readout. These components are essential for the collection of data for electron identification.

A major upgrade of the ALICE detector will increase the data rate that has to be processed online to more than 1 terabyte per second [Abelev et al. 2014]. Due to this upgrade the TRD will have to redesign its online data process to limit the data rate recorded to less than 10 gigabytes per second. The ALICE upgrade requires improved algorithms to run on a data stream which has been compressed through feature extraction.

This article focuses on finding an efficient data compression technique which will still provide comparable electron identification capabilities as with uncompressed data.

2. PARTICLE IDENTIFICATION

2.1. Particle Identification Requirements

Aamodt et al. [2008] identifies the key information required for particle identification, with the use of ALICE Transition Radiation Detector, is the average pulse height and the dependence on the drift time. The following section highlights the two main differences between data points for pions and for electrons. Pions and electrons are the two main particles detected by the TRD.

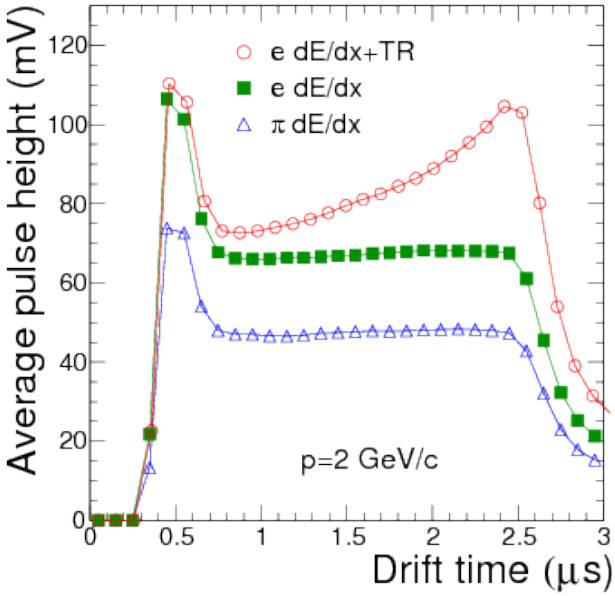

Fig. 1. Graph depicting average pulse height as a function of drift time for pions, electrons without a radiator and electrons with a radiator

2.1.1. Total amount of deposited charge

The electrons signal is typically larger than that of pions. With a momentum of 2GeV/c, in the amplification region the electron signal is about 1.5 times larger than that of pions and the region where transition radiation photons are absorbed preferably is about twice as large for electrons (See figure 1). Particle identification algorithms such as cluster counting [Ludlam et al. 1981], truncation [Andronic 2004] and likelihood method [Cherry et al. 1974] use the total amount of deposited charge for electron identification.

2.1.2. Transition radiation peak

At longer drift times the electron signal increases, since the transition radiation photons are mostly absorbed within the drift region, while the signal of pions remains constant (See figure 1). The signal's time information can be used to aid the identification of electrons. The two likelihood methods (LQ), LQX method [Andronic 2004] and 2-dim LQ method, make use of the time information to help with electron identification.

The electron/pion discrimination performance of a detector is measured with the pion efficiency ϵ_π. The pion efficiency is the fraction of pions that is misidentified at a fixed electron efficiency ϵ_e. The electron efficiency is the fraction of electrons that is identified correctly. The goal of ALICE TRD is to have a pion efficiency of 1% for 90% electron efficiency [Wilk 2010].

2.2. Particle Identification with Neural Networks

The current algorithm for particle identification with TRD is make use of artificial neural networks. Artificial neural networks are statistical learning algorithms that are used to estimate or approximate functions that can depend on a large number of inputs and unknowns [McCulloch and Pitts 1943]. They are systems of interconnected

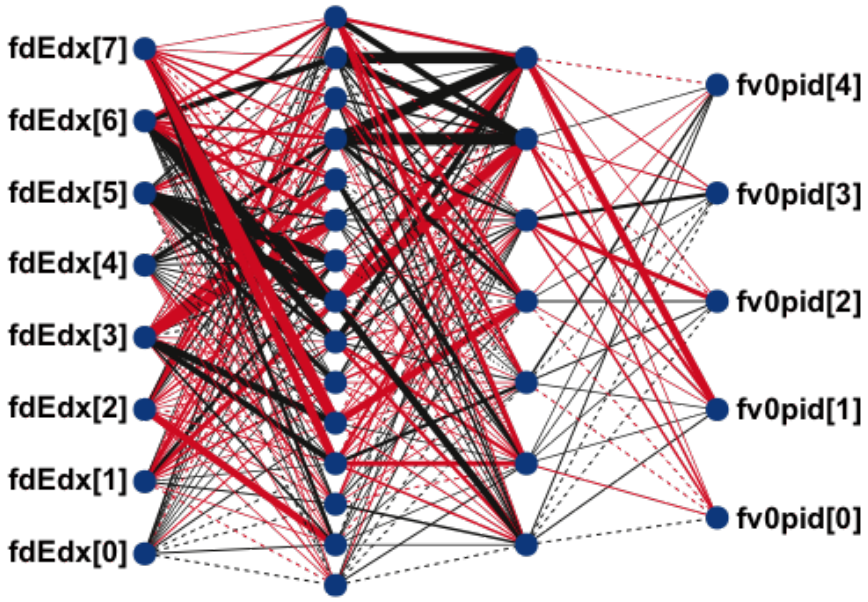

Fig. 2. Reference network as used in AliRoot v4-17-Rev12 for Layer 1 and particles of 2 GeV/c

neurons and are able to use machine learning and pattern recognition for computation on input values.
Bellotti et al. [1993] performed the first successful approach for electron identification using neural networks. Bellotti used a network which was a three-layered multilayer perceptron which used the deposited charge as input. He showed that artificial neural networks could increase the performance by a factor of up to three compared to the likelihood methoods.
The follow section highlights the neural network algorithm currently implemented discussed by Wilk [2010].

2.2.1. Neural Networks Using Deposited Charge Information

A standard method for particle identification for the TRD is with the use of artificial neural networks with the deposited charges as input parameters. The network is a feed-forward network which composes of the input layer, two hidden layers and the output layer (See figure 2). Several factors specify the topology of the network. The number of input neurons is limited to eight due to the compromise between discrimination power and needed storage space. The number of output neurons is two as only electrons and pions are taken into account. The number of hidden layers and hidden neurons is determined by several tests. Fifteen and seven hidden neurons, was found to give the best performance [Wilk 2010].

2.2.2. Neural Networks with Preprocessed Variables

The following are variables required for particle identification:

— the number of clusters above a high threshold N_p
— the number of time bins above a low threshold N_{thresh}
— the time bins above a low threshold T_{max}
— the deposited charge of the second largest cluster Q_2
— the integrated charge below the low threshold Q_{sum}

3. DATA COMPRESSION

Data compression involves encoding information using fewer bits than the original data [Mahdi et al. 2012]. It is either lossy or lossless. Lossy and lossless data compression differ in that lossy data compression involves removing unnecessary information while lossless data compression eliminates statistically redundant information.

3.1. Considerations in Data Compression

Jain and Lakhtaria [2013] highlight several factors that are needed to be considered with data compression as different systems have different compression requirements.

— *Compression techniques* need to be considered when compressing data such as lossy or lossless data compression. Data that is needed to be accurate requires lossless data compression while lossy algorithms may be used for approximations on data.
— *Compression ratio* is the percentage resulting from dividing the compression size by the original file size. It determines how well data is compressed in terms of size. The compression ratio is used to draw comparisons between data compression techniques.
— *Compression time* may vary depending on the algorithm used for data compression. Real-time data compression may require data compression to be faster as opposed to smaller in compression size.

Considerations on the context of the data need to be taken into account when applying data compression algorithms.

3.2. Types of Data Compression

3.2.1. Lossless Data Compression

Lossless data compression is a class of data compression algorithms. The algorithms allow for the exact original data to be reconstructed from the compressed data. Lossless data compression is important for obtaining the original data when applying decompression without having any assumptions made on the data.

Run Length Encoding Technique. Run Length Encoding (RLE) is a simple form of data compression [Porwal et al. 2013]. This technique looks at the runs of the data. The run is the length of a repeated consecutive data element. The compressed data stores the data element as a pair with the consecutive number of counts of that element.
RLE can be efficiently and easily applied. A disadvantage with RLE is that it may not be an optimal solution for data with very few runs or data with each consecutive element being unique. In the worst case the size of the output data is twice as large as the input data.

Huffman Coding. Huffman [1952] developed the Huffman Coding Algorithm which involves converting characters in a data file into binary code. Huffman Coding involves a pre-analysis of data. The general idea behind Huffman Coding is to convert the most common characters in a file to have the shortest binary code and the least common have the longest binary code.
Huffman Coding is an optimal algorithm for symbol-by-symbol coding, where symbols are unrelated. Huffman Coding will produce a worst case scenario when the probability of a symbol exceeds 50%.

Shannon Fano Coding. Fano [1963] developed the Shannon Fano Coding which is a compression technique to construct prefix code based on a set of symbols and their probabilities. It guarantees that all code word lengths are within one bit of their theoretical ideal unlike Huffman Coding. Shannon Fano Coding orders elements from most probable to least and recursively divides into them two sets whose probabilities are as close as possible to being equal.

Shannon Fano Coding does not always give an optimal prefix code. When the probabilities near the inverses of powers of 2 the Shannon Fano algorithm is most efficient.

Arithmetic Encoding. Arithmetic encoding is similar to Huffman coding except it differs as rather than separating the input into component symbols and replacing each with a code, arithmetic coding encodes the entire message into a single number and a fraction between 0 and 1 [Rodionov and Volkov 2007]. It provides extremely high coding efficiency and ensures lossless data compression.

3.2.2. Lossy Data Compression

Lossy compression uses inexact approximations for representing data that is being encoded [Witten et al. 1994]. It is mostly used in compression of multimedia such as audio, video and images. Lossy compression suffers from generation loss which causes a file to lose quality with multiple compressions and decompressions.

Lossy Transform Codecs. Lossy transform codecs cuts samples of data in a file into small segments then transforms it into a new basis space which is then quantized. The entropy code is generated from the quantized values.

Lossy Predictive Codecs. Lossy predictive codecs use previous decoded data to predict the current data sample.

In some cases, lossy data compression can compress a file into a much smaller compressed file, as opposed to lossless compression, while still keeping the general content of the original data.

4. DATA COMPRESSION IN ALICE TRD

4.1. Importance of Data Compression in ALICE TRD

The amount of data collected by the ALICE TRD is expected to be extremely large. Sharma et al. [2014] highlights the many advantages of data compression. The advantages include reducing the complexity and cost of the data storage. The aim for the data compression for ALICE TRD is to limit the data rate and processing to record less than 10 gigabytes per second.
While the data compression needs to be as efficient as possible, careful considerations need to be made to the compression algorithms to ensure that the compression does not affect the accuracy of the electron identification algorithms.

4.2. Approaches to Data Compression in ALICE TRD

Analysis on lossless and lossy compression approaches are required for the data generated by the TRD chambers in the ALICE experiment.

4.2.1. Lossless compression of TRD data

Although Jain and Lakhtaria [2013] found that Huffman Coding (from comparison between RLE, Huffman Coding and Shannon Fano) produces optimal results for language related data, Huffman Coding would not produce optimal results for ALICE Data as the ALICE TRD data is not a set of unrelated symbols and as mentioned in section 3.2.1 Huffman coding is an optimal algorithm for symbol-by-symbol coding. Porwal et al. [2013] found that arithmetic encoding is significantly better than Huffman Coding. Arithmetic coding showed to give a better compression ratio and compression time over Huffman Coding.
The lossless compression technique for TRD may require the use of a probability model for the data fields. By looking at patterns with sample positioning of values in correlation with its signal from the TRD, the data can be grouped with a probability

distribution for entropy coding. Additionally exploitation on the correlation between consecutive samples can be achieved using a prediction scheme. A similar approach has been investigated with the ALICE TPC experiment and could be applied to the TRD [Nicolaucig et al. 2002].
Wulff [2009] studied the effects of zero suppression on the resolution and the need for a good data compression by noise suppression and to minimise the signal loss. Zero suppression is the removal of redundant zeroes from data. Wulff found that possibly dangers in data compression include losing important information such as if deposited charges are rejected then position reconstruction might produce errors.
Other alternative models for data compression can also be applied such as using the time information such as the space correlation and order time correlation.

4.2.2. Lossy compression of TRD data

A lossy compression technique could be explored to possibly even further increase the compression ratio. Lossy compression needs to preserve the data for the average pulse height and the dependence on the drift time as they are the main interest for the experiment. Experimentation with TPC data have currently been implemented using area of bunches (signals coming from each pad), center of mass of bunches and the correlation between bunch area and sample values [Nicolaucig et al. 2002]. A similar approach can be looked into to be applied to the TRD data.

4.3. Compression Analysis

The compression algorithms are to be tested on the simulations done with the use AliRoot [Brun et al. 2003]. AliRoot is the ALICE offline-line framework for simulation, reconstruction and analysis.
As mentioned in section 3.1 the effectiveness of compression algorithms are measured using the compression ratio and the data rate.
The accuracy of the TRD is measured with pion efficiency as discussed in section 2.1. Accuracy in the data compression is measured by the comparison between the pion efficiency for electron identification with compressed data and the original data.

5. CONCLUSIONS

Investigation of lossless and lossy compression approaches for the data generated by the TRD chamber in the ALICE experiment shows that several considerations need to be taken when compressing the TRD data. The two key data items for data compression for the TRD are the average pulse height and the dependence on the drift time. The compression algorithms need to keep the overall computational complexity feasible for real-time implementation while still allow electron identification algorithms to produce the same results as if the original data was used in electron identification.

REFERENCES

Kenneth Aamodt, A Abrahantes Quintana, R Achenbach, S Acounis, D Adamová, C Adler, M Aggarwal, F Agnese, G Aglieri Rinella, Z Ahammed, and others. 2008. The ALICE experiment at the CERN LHC. *Journal of Instrumentation* 3, 08 (2008), S08002.

B Abelev, ALICE collaboration, and others. 2014. Upgrade of the ALICE experiment: letter of intent. *Journal of Physics G: Nuclear and Particle Physics* 41, 8 (2014), 087001.

A Andronic. 2004. Electron identification performance with ALICE TRD prototypes. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 522, 1 (2004), 40–44.

R Bellotti, M Castellano, C De Marzo, N Giglietto, G Pasquariello, and P Spinelli. 1993. A comparison between a neural network and the likelihood method to evaluate the performance of a transition radiation detector. *Computer physics communications* 78, 1 (1993), 17–22.

R Brun, P Buncic, F Carminati, A Morsch, F Rademakers, and K Safarik. 2003. Computing in ALICE. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 502, 2 (2003), 339–346.

M.L. Cherry, D. Mueller, and T.A. Prince. 1974. The efficient identification of relativistic particles by transition radiation. *Nucl.Instrum.Meth.* 115 (1974), 141–150. DOI:http://dx.doi.org/10.1016/0029-554X(74)90439-X

R. Fano. 1963. A heuristic discussion of probabilistic decoding. *Information Theory, IEEE Transactions on* 9, 2 (April 1963), 64–74. DOI:http://dx.doi.org/10.1109/TIT.1963.1057827

D.A. Huffman. 1952. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE* 40, 9 (Sept 1952), 1098–1101. DOI:http://dx.doi.org/10.1109/JRPROC.1952.273898

Amit Jain and Kamaljit I Lakhtaria. 2013. A Comparative Study of Lossless Compression Algorithm on Text Data. In *International Conference on Advances in Engineering and Technology. Elsevier India*. 536–543.

MinJung Kweon and others. 2009. The Transition Radiation Detector for ALICE at LHC. *arXiv preprint arXiv:0907.3380* (2009).

C Lippmann and others. 2006. The ALICE Transition Radiation Detector. In *Proc. of the SNIC Conference, SLAC, April 3Ä6*.

T. Ludlam, E. Platner, V. Polychronakos, M. Deutschmann, W. Struczinski, and others. 1981. Particle Identification by Electron Cluster Detection of Transition Radiation Photons. *Nucl.Instrum.Meth.* 180 (1981), 413. DOI:http://dx.doi.org/10.1016/0029-554X(81)90081-1

Omar Adil Mahdi, Mazin Abed Mohammed, and Ahmed Jasim Mohamed. 2012. Implementing a novel approach an convert audio compression to text coding via hybrid technique. *International Journal of Computer Science Issues* 9, 6 (2012), 53–59.

WarrenS. McCulloch and Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5, 4 (1943), 115–133. DOI:http://dx.doi.org/10.1007/BF02478259

Aldo Nicolaucig, Marco Mattavelli, and S Carrato. 2002. Compression of TPC data in the ALICE experiment. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 487, 3 (2002), 542–556.

Shrusti Porwal, Yashi Chaudhary, Jitendra Joshi, and Manish Jain. 2013. Data compression methodologies for lossless data and comparison between algorithms. *IJESIT* 2, 2 (2013), 142–7.

Anatoly Rodionov and Sergey Volkov. 2007. P-adic arithmetic coding. *CoRR* abs/0704.0834 (2007). http://arxiv.org/abs/0704.0834

Neha Sharma, Jasmeet Kaur, and Navmeet Kaur. 2014. A Review on various Lossless Text Data Compression Techniques. (2014).

Alexander Wilk. 2010. *Particle identification using artificial neural networks with the ALICE transition radiation detector*. Ph.D. Dissertation. Munster (Westfalen), Univ., Diss., 2010.

Ian H Witten, Timothy C. Bell, Alistair Moffat, Craig G. Nevill-Manning, Tony C. Smith, and Harold Thimbleby. 1994. Semantic and generative models for lossy text compression. *Comput. J.* 37, 2 (1994), 83–87.

Elke Svenja Wulff. 2009. *Position Resolution and Zero Suppression of the ALICE TRD*. Ph.D. Dissertation. Diplomarbeit, Westfälische Wilhelms-Universität Münster.