Particle Passage Detector Dataset Regression Problem

Giovanni Pellegrino
Politecnico di Torino
Student id: s331438
s331438@studenti.polito.it

***Abstract*—In this report it is introduced various possible approaches to the Particle Passage Detector regression problem. In particular, the proposed approaches leverage a purely analytical method and another based on knowledge of the physical domain. The first method, which examines the correlation among the pads, achieves satisfactory results, while the second method, estimating the Signal-to-Noise Ratio (SNR), proves too approximate and leads to overfitting due to the failure to identify the noisy pads of the RSD.**

I. Problem overview

The proposed competition addresses a regression problem using the *Particle Passage Detector Dataset*, which consists of 90 features, 5 for each of the 18 pads:

- *pmax*, the magnitude of the positive peak of the signal, in mV
- *negpmax*, the magnitude of the negative peak of the signal, in mV
- *rms*, the root mean square (RMS) value of the signal
- *area*, the area under the signal
- *tmax*, the delay (in ns) from a reference time when the positive peak of the signal occurs

Notably, 6 pads (30 features) include noise as they are not physically connected to the *Resistive Silicon Detector* (RSD). This sensor is widely employed in machine learning applications, such as 4D particle tracking [1].

The goal of this project is to build a data science pipeline that predicts for each event, given as inputs the characteristics of the signals measured by each pad, the target (x,y) coordinates where the particle of interest passed.

The dataset is partitioned into two distinct sets:

- a *development* set, including 385,500 events
- an *evaluation* set, with of 128,500 events.

Into the dataset there are no missing values (NaN), so we don't need to make any special adjustments.

The problem is well-balanced: for each pair of coordinates (x,y), there are 100 samples.

The dataset contains various outliers for each feature, suggesting that, in general, all pads have a certain level of noise. It will be crucial to select a model that is robust to noise to make the most accurate predictions possible. However, it's worth noting that the feature *tmax* for pads 0, 7, 12, 15, 16, 17 does not include outliers, indicating a potential correlation between these pads.

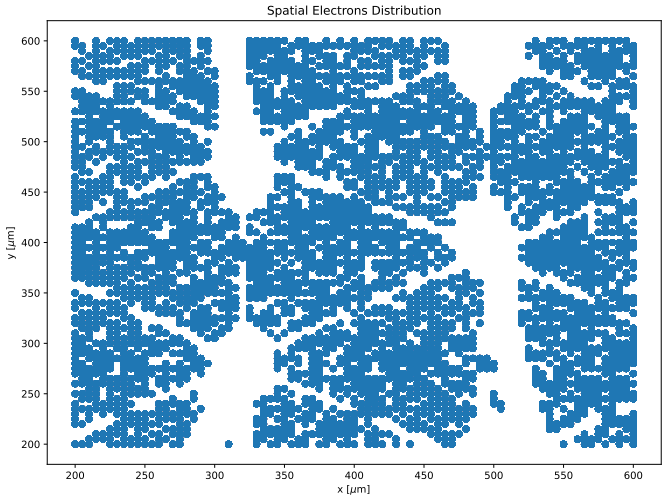

Fig. 1. This is the scatter plot of the detected electrons position: the area is a 600x600 square (in μm) and it resembles the asterisk-shape of the pads

In Figure 1, the *Spatial Electrons Distribution* is shown, and several observations can be made:

- the distribution is not uniform along the y and x-axis
- it is evident that electrons pass closer to 5 of the 12 pads (the plot resembles the asterisk-shaped pads), and it is expected that values for each event may vary in magnitude
- since the entire detecting area is not covered, the regressor may face challenges in accurately predicting positions outside this zone.

For the sake of simplicity, these three points will not be considered in either preprocessing or model selection phase and this could potentially lead to an approximate regressor.

II. Proposed approach

A. Preprocessing

In this phase, it is important to select the appropriate pads that do not capture real measurements to minimize noise in each measurement. This helps significantly reduce the number of features. Since each pad provides 5 measurements, identifying and excluding the 6 noisy pads is beneficial, resulting in a reduction of features from 90 to 60.

Two specific methods were tested: one analytical and the other leveraging signal processing theory.

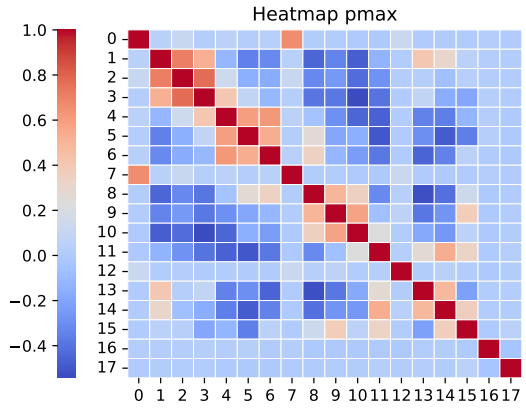

Fig. 2. Correlation heatmap of the feature pmax: a clear contrast can be seen in rows/columns 0, 7, 12, 15, 16, 17 which indicates the potential pads affected to noise

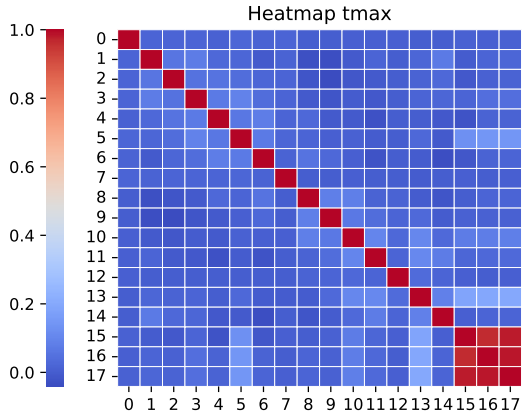

Fig. 3. Correlation heatmap of the feature tmax: the anomaly high correlation is spotted for the pads 15, 16, 17

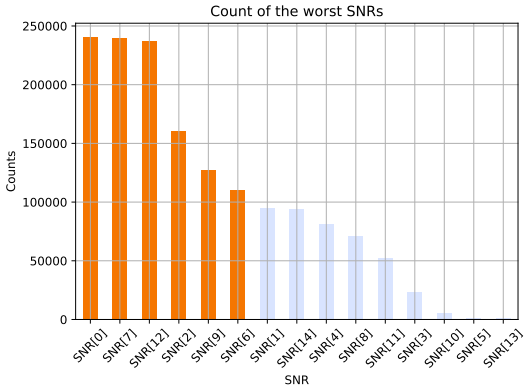

Fig. 4. Bar plot of the worst SNR: the first 6 highlighted bars are the pads where a weak signal has been founded the most (0, 2, 7, 9, 12, 14)

- *analytical approach* (correlation preprocess): the focus was on evaluating the correlation between different features using a *heatmap*, a graph which uses a color gradient to depict changes and magnitudes of a third variable in a two-dimensional graph. This choice is motivated by the awareness that the features exhibit a correlation with the position, and the goal is to specifically examine the nature of this correlation among them.
  What can be particularly observed from the *heatmap* of *pmax* (Figure 2) and that of *tmax* (Figure 3) is noteworthy. Regarding pmax, a low correlation index are clearly visible on particular rows/columns (except for the value on the diagonal indicating the correlation with itself, which is maximum). These rows and columns correspond to pads 0, 7, 12, 15, 16, 17. To highlight the behavior of pads 15, 16, and 17, it can be observed in Figure 3 that the last 3 pads exhibit a high correlation among themselves, which might indicate a *systematic error* [2] in the instrument for those three pads when not connected. However, the assertion of the presence of systematic error is incorrect, as anticipated in the Problem Overview, since no outliers were found in those columns of the parameter *tmax*. This indicates more probably the presence of white Gaussian noise[1] rather than systematic error.
- *signal processing approach* (SNR preprocess): in this case, an approximate calculation of the *Signal-to-Noise Ratio* [3] was performed, and the six pads with lower values were excluded. This measure provides information regarding the signal quality relative to the background noise level. It is possible to assume the passage of an electron as a signal with finite energy and finite power. The *Signal-to-Noise Ratio* is typically written as *SNR* and computed as

$$SNR_{dB} = 10 * \log_{10}\left(\frac{P_s}{P_N}\right) \tag{1}$$

where SNR_{dB} is the *SNR* in decibel scale, P_s is the power of the signal and P_N is the power of the noise. The logarithmic scale was chosen because the values of *SNR* may be very high and difficult to compare with other values of *SNR*.
To evaluate P_s, it is possible to compute it as

$$P_s = \frac{E_s}{T} \tag{2}$$

where E_s is the energy of the signal or the area under the signal and T is the period of the signal. Plugging (2) in (1)

$$SNR_{dB} = 10 * \log_{10}\left(\frac{E_s}{TP_N}\right) \tag{3}$$

$$= 10 * \log_{10}\left(\frac{E_s}{P_N}\right) - 10 * \log_{10}(T) \tag{4}$$

Equation (4) is obtained by applying logarithmic properties to equation (3). We can assume that the period of

[1]Noise that is present in all frequencies

the signal is the same for each measurement, given that the start and end of the entire event are set by the RSD. In our case, the second term, which would represent the intercept, can be eliminated[2] and equation (4) becomes

$$SNR_{dB} \approx 10 * \log_{10}\left(\frac{E_s}{P_N}\right) \tag{5}$$

Given the limited signal information, we can represent it approximately as a *discrete signal* using two key points: *pmax* and *negpmax*. However, it's important to note that this method is a simplified representation of the continuous signal and deviates significantly from more accurate approximation methods [4]. Using this approach, we can calculate the energy of the (real) signal as the sum of the squares of *pmax* and *negpmax* [5]

$$E_s = \sum_{i=1}^{N} x_i^2 = pmax^2 + negpmax^2 \tag{6}$$

Regarding to the power of white noise, according to our initial assumptions, we can treat it as an ergodic process [6]. From this, we derive that its power is equal to its variance.

$$P_N = var(random(0,1)) \tag{7}$$

Taking 2/3 of the data as representative of the entire dataset, as depicted in Figure 4, yields a slightly different result: the columns where noise is present are 0, 2, 6, 7, 9 and 12.

The two methods share pads 0, 7, 12, while differing for pads 2, 6, 9, and 15, 16, 17. It was decided to use two datasets with the respective columns dropped, reducing the dimensions from 90 to 60. To further reduce dimensionality, *PCA* was performed on the *SNR* preprocess (*PCA on SNR preprocess*), retaining something above 90% of the information (Figure 5). However, using PCA could potentially worsen the regressor performance due to information loss.

B. Model selection

Given the noisy nature of the data and the intricate relationship among them, a *random forest* was chosen as regression algorithm. This algorithm leverages multiple decision trees (trained on various subsets of data and features) to make predictions. In this way, no normalization is needed since it works on one feature at a time.

C. Hyperparameters tuning

There are two main sets of hyperparameters to be tuned:

- *SNR* dataset features for *PCA*
- *random forest* parameters

PCA was performed with a *Standard Scaler* since the data involve different physical quantities and have varying value ranges. To prevent overfitting on the dataset, tuning was performed on the parameters of the *random forest*. Concerning

[2]We are not interested on the exact value but it is sufficient that all the SNR computed are consistent in order to compare the values among them

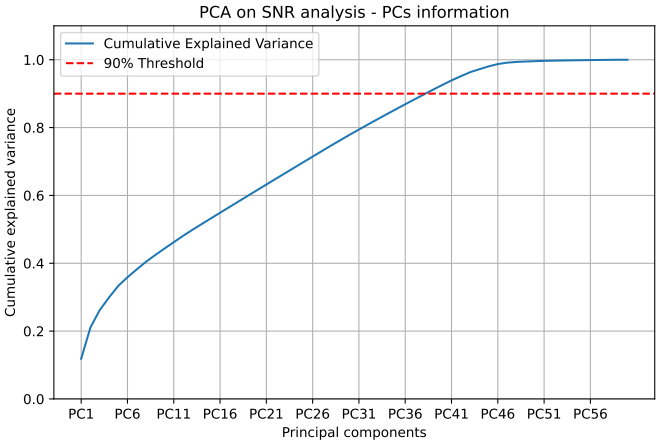

Fig. 5. This graph shows that the 90% of information (cumulative explained variance) is stored in the first 38 PCs. To stay above the threshold, one extra PC was picked. This drastic reduction of dimensionality (from 90 to 60 and then 39) may have throwback in our regressor

the maximum depth of the tree (*max_depth*), as the random forest is an ensemble of decision trees, for each dataset used a single decision tree was constructed and its *max_depth* was derived (Table I).

Preprocess	max_depth
Correlation	29
SNR	31
PCA on SNR	35

TABLE I
MAX_DEPTH DERIVED FROM THE DECISION TREE FOR EACH PREPROCESS METHOD

Less emphasis was placed on the number of decision trees in the random forest (number of estimators) as these improve the model up to a certain point [7]. The number of features to consider when looking for the best split (*max_features*) was chosen to be the square root of the number of features. In this way, the regressor has a higher level of interpretability and avoids overfitting.

The parameters chosen to tune were:

- *min_samples_split* = [2,4,8]: the minimum number of samples required to split an internal node
- *min_samples_leaf* = [2,3,5]: the minimum number of samples required to be at a leaf node

A grid search was run for all three methods to acknowledge the best combination of parameters and which regressor performed better (lowest *average Euclidean distance* between the actual positions of electrons on the *RSD* detector and the predicted ones).

III. RESULTS

The result are summerized in Table II. It is easy to see that the best configuration is given by the *analytical approach*: the method which uses SNR analysis takes too many approximated assumptions and uses noisy pads, PCA adjust the SNR result but it is still way worse than the *correlation preprocess*

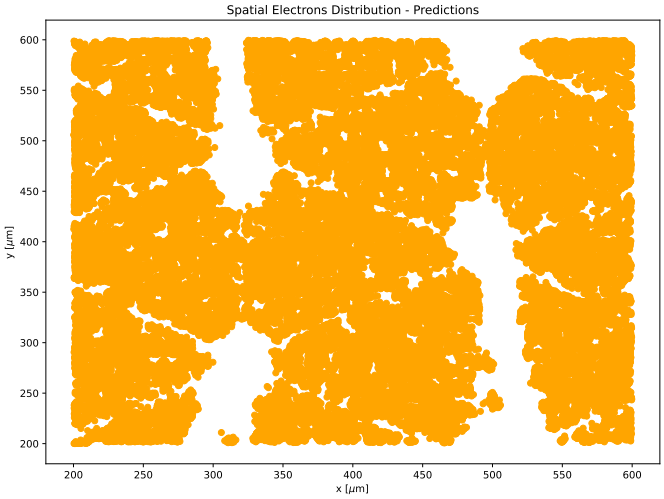

Fig. 6. This is the scatter plot of the predicted electrons position using 20% of the dataset as validation (the other 80% was used to train the regressor): it is possible to resemble the asterisk-shape of the 5 pads but some points are positioned where they should not be.

method. Unlikely as expected, PCA method performed better than SNR method due to the fact that contains less noisy information.

Looking on the features importance it is possible to note that *rms* and *tmax* contribute little to the total features importance: training a model without these features, a better public score is obtained, reaching 4.881μm of error.

Preprocess	min_samples_leaf	min_samples_split	Public Score
Correlation	2	4	5.112 μm
SNR	2	4	163.566 μm
PCA on SNR	2	2	114.967 μm

TABLE II
RESULT OF THE RANDOM FOREST REGRESSOR ON THREE DIFFERENT TYPE OF PREPROCESSING

IV. DISCUSSION

Knowing very little about the nature of the signal, using signal processing techniques has led to an overfitted regressor with a high level of noise. It has been demonstrated that in this problem, a good regressor can be obtained by discarding noisy pads (after checking correlation levels), in this case, 0, 7, 12, 15, 16, 17. Furthermore, even just by increasing the number of estimators (*n_estimators* = 200) of the random forest and performing another grid search (obtaining as best parameters *min_samples_leaf* = 2, *min_samples_split* = 2), a public score of 4.818 μm is achieved (predictions of this regressor is shown in Figure 6).

Other techniques can be implemented to improve the regressor, such as conducting another grid search on other parameters of the random forest, including the number of estimators. However, the result obtained is more than satisfactory for the purposes of analysis and predicting the position of electrons, as the error is lower than 10 times than the actual position.

REFERENCES

[1] M. Mandurrino, R. Arcidiacono, M. Boscardin, N. Cartiglia, G.-F. D. Betta, M. Ferrero, F. Ficorella, L. Pancheri, G. Paternoster, F. Siviero, V. Sola, A. Staiano, and A. Vignati, "Analysis and numerical design of resistive ac-coupled silicon detectors (rsd) for 4d particle tracking," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 2020.

[2] J. Heinrich and L. Lyons, "Systematic errors," *Annual Review of Nuclear and Particle Science*, vol. 57, no. 1, pp. 145–169, 2007.

[3] D. H. Johnson, "Signal-to-noise ratio," *Scholarpedia*, vol. 1, no. 12, p. 2088, 2006. revision #126771.

[4] A. Oppenheim and D. Johnson, "Discrete representation of signals," *Proceedings of the IEEE*, vol. 60, no. 6, pp. 681–691, 1972.

[5] A. V. Oppenheim, *Digital Signal Processing*. Englewood Cliffs: Prentice-Hall, 1975. Print.

[6] S. L. Miller and D. G. Childers, *Probability and Random Processes: With Applications to Signal Processing and Communications*. Amsterdam: Elsevier, 2nd ed., 2012. Print.

[7] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," in *Machine Learning and Data Mining in Pattern Recognition*, (Berlin, Heidelberg), pp. 154–168, Springer, 2012.