Pokémon Games: a Clustering proposal

Report based on Machine Learning course by Prof. Fabio Antonio Stella

Tiziano Buggio Mauro Gianfreda Luca Sala Andrea Seveso Andrea Armando Tinella

Contents

Chapter 1

Introduction

The team members of the project are Tiziano Buggio, Mauro Gianfreda, Luca Sala, Andrea Seveso and Andrea Armando Tinella.

Our goal: determine whether it is possible to cluster the Pokémon in meaningful groups only using statistics and compare the results with another pre-existent clustering based on the tournaments fights that considers also the unique abilities of each Pokémon.

The initial dataset was taken by the Kaggle user “alopez247” who merged six Pokémon game generations published in Europe from 2000 to 2013. Another dataset has been *scraped* from the site “pokemon.neoseeker.com”, to use external indices and therefore to evaluate the goodness of the cluster analysis, with respect to another clustering coming from a reliable source. The main sources of the dataset are the fan-made Pokémon-themed sites “WikiDex.com” , “Bulbapedia.com” and “Smogon.com” .

There are several reasons why this subject has been chosen, among them the most relevant are the following:

- firstly, almost all the team members are well-aware of the theme, that ensures the ability to perform an adequate analysis avoiding trivial suppositions.

- secondly, a clustering procedure seems to be adequate to the features of the data.

- finally, it is a complete dataset, which prevents missing values resulting in a more exhaustive representation.

Chapter 2

General dataset description

The first version of the dataset was composed of 721 records, ranging from 1 to 721 based on the *Pokédex order*. Each one represents a fictional creature with unique skills and abilities. Every record is composed by 23 attributes of three different types: String, Integer and Double. String attributes are the following: Name, Type_1, Type_2, isLegendary, Color, hasGender, Egg_Group_1, Egg_Group_2, hasMegaEvolution and Body_Style. There are only three Integer attributes: Number, Generation and Catch Rate. The remaining ten are Double attributes: Total, HP, Attack, Defense, Sp_Atk, Sp_Def, Speed, Pr_Male, Height_m, Weight_kg. These represent the main characteristics of each Pokémon, which are useful in determining the fight performance. The focus will be mainly on these to develop the cluster analysis.

2.1 Attributes in-depth description

Type The concept of Type is fundamental in the Pokémon's world, because it defines the strengths, weaknesses and their interactions among them. There are 18 different types.

Every Pokémon has a main type (Type_1). The 48,5% of them has also a second type (Type_2).

The most recurrent main type is "Water" assigned to the 14,56% of the Pokémon, the least frequent is "Flying" assigned to 0,42% of the Pokémon. Remarkably, "Flying" is the most assigned secondary type (24,85%).

Generation Another relevant aspect is the concept of "Generation" which represents a grouping of the Pokémon that belongs to same game and separates them based on the Pokémon they include. In each Generation a new set of Pokémon, moves and abilities are released, that didn't exist in the previous ones. The dataset considers the first six Generations.

isLegendary Is used to define if a Pokémon is legendary or not. This attribute represents the rarest and strongest Pokémon.

hasGender Defines if a Pokémon has a sexual gender.

hasMegaEvolution Specifies if a Pokémon could evolve in another, a more powerful one.

Pr_Male Related to hasGender, if a Pokémon has a sexuality, this attribute represents the percentage to find a male one. The 77 Pokémon with no gender have missing values in this field.

Fight Attributes There are six attributes that represent the characteristics used for the fight.

- HP: the health points of a Pokémon.
- Attack: the offensive power of a Pokémon.
- Defense: the defensive power of a Pokémon.
- Sp_Atk: the special offensive power of a Pokémon.
- Sp_Def: the special defensive power of a Pokémon.

- Speed: the velocity of a Pokémon. There is also another attribute, “Total”, that is the sum of the previous six.

Physical attributes

- Height_m: the height of a Pokémon in meters.
- Weight_kg: the weight of a Pokémon in kilograms.
- Body_Style: the overall physical appearance.
- Color: the main pigment of a Pokémon.

Moving on to the Knime workflow:

2.2 Descriptive Statistics

The following table represents the descriptive statistics of the original dataset, with all the attributes analyzed, even if for the cluster analysis the least relevant ones have been excluded.

Regarding the main fight attributes (Attack,Defense, Sp_Atk, Sp_Def, Speed) one can notice that they share a similar mean that ranges from 65.714 to 75.014. They also have a close Standard Deviation which ranges from 27.016 to 29.297. This happens despite the five chosen attributes have a very different maximum, but a similar minimum:

	Min	Mean	Max	Std. Dev.
Total	180	417.9459	720	109.6637
HP	1	68.38	255	25.8483
Attack	5	75.0139	165	28.9845
Defense	5	70.8086	230	29.2966
Sp_Atk	10	68.7379	154	28.788
Sp_Def	20	69.2913	230	27.0159
Speed	5	65.7143	160	27.2779
Generation	1	3.3232	6	1.6699
Pr_Male	0.0	0.5534	1	0.2
Height_m	0,1	1.145	14.5	1.0444
Weight_kg	0,1	56.7734	950	89.0957
Catch_Rate	3	100.2469	255	76.5735

Chapter 3

Cluster analysis

3.1 Data preparation

With the most relevant statistics looked through,we have proceeded preparing the dataset to obtain the best possible clustering solution. Firstly, the data have been normalized using the "min-max" option setting the min to 0.0 and the max to 1.0. Then all the attributes that were not relevant for the clustering purpose have been removed, namely all the string attributes and the double attribute "Pr_Male" due to the presence of several missing values. Secondly, the linear correlation between the remaining attributes has been analyzed and it has been found that no particularly high correlations incurred among the data, except "Total", since it was simply the sum of all the other features. The attribute "Catch_Rate" was the only one to have an inverse correlation with all the others, which means that a high value of this attribute implied low values in all the remaining ones and the other way around. It has been decided to remove Total and also the attribute Generation, because it was clearly meaningless for our goal.

Later, it has been decided to apply some random clustering algorithms with the purpose to find outliers. It was found that applying Hierarchical clustering setting the "Euclidean" distance function, different types of linkage and varying the number of clusters, four Pokémon (namely Schukle, Wailord, Chansey and Blissey) remained always isolated.

This can be explained with the excessively different value of their features compared to all the others, so it has been decided to discard them. Finally, the data have been plotted to see the distributions.

3.2 Algorithm comparison

It has been decided to compare different clustering models to find out which ones are the most appropriate, starting from a cluster number equal to 11, the same of the dataset obtained from the scraping.

Density-Based Algorithm Density-Based Spatial Clustering of Applications with Noise (DBSCAN): despite that this clustering model didn't need to set the number of clusters, the results of DBSCAN analysis were poor.

Even changing the "Epsilon" and "Minimum Points" parameters, this clustering analysis showed a single cluster and many values labeled as 'Noise', probably as a result of a significant difference between cluster densities.

Prototype-Based Algorithm

- Fuzzy C-Means (FCM): it has been decided to reject the FCM results because many values belonged to more than one cluster, so the results were not clear.
- K-means: appeared to be the clearest method. Even if cluster dimension was different, density was quite similar.

- Expectation Maximization (EM): It seemed to work well despite it's a probabilistic assignment.
- The Partitioning Around Medoid method has not been applied because the distribution of the data was not supposed to have a center.
- Self-Organizing Map (SOM): has shown good results, but it has been applied to the same data on which it was developed.

In the end it has been decided to consider only one of the Prototype methods, the K-means.

This, because the K-means assigns clusters based on a certain characteristic, EM considers probability, instead.

Hierarchical

- Single-link: Dendrogram results were different from expected because of the aggregation of one (or a few) records at a time.
- Average-link: the Dendrogram solution seemed to be better than the previous one.
- Complete-link: even if it seemed to be the best, it still aggregated values with a higher distance compared to the previous two models.

Further, also the Ward's Method will be considered, referred to Hierarchical Clustering.

So far, it has been decided to do the comparison between the different types of Hierarchical Clustering and the K-means algorithm.

3.3 Internal measures

The following step is the **cluster validity** i.e. evaluating the overall goodness and effectiveness of the different clustering algorithms.

Considering the self conception of clustering and the ubiquitous risk of aggregating not significant objects, the validity of the clustering models has an high complexity.

One of the types of indices used in clustering evaluation are the internal ones. The main objective of these measures is to define the degree of cohesion and separation of the defined clusters.

Cophenetic

Average	Single	Complete	Ward
0.716	0.391	0.577	0.623

The Cophenetic index is very useful to evaluate the best linkage in the Hierarchical models, i.e. the model which fits better to the data.

As expected, the single linkage had the worst result of all of them; therefore it has been decided to exclude it from the analysis; the other three models had very good results instead.

It must be considered, nonetheless, that a high output of the Hierarchical Average is not sufficient to accept its use. The Hierarchical Complete and Ward, lastly, had an acceptable result after all, with the Ward one slightly better.

Connectivity

Hier. Complete	Hier. Ward	Hier. Average	K-means
444.527	353.938	104.135	351.413

The connectivity index has noticeable results when it is very low, ideally 0. Comparing the connectivity index of the different models, the Hierarchical Complete had an outstandingly high value. This output would have forced to do a cutoff of this model, which is the opposite to what appeared from the Dendrogram.

The Hierarchical Average, instead, had a very low connectivity index, significantly lower than the other models, with the remaining two models in the upper-middle position.

Silhouette

Hier. Ward	K-means	Hier. Average	Hier. Complete
0.132	0.195	0.169	0.094

The Hierarchical Complete model had a bad performance in the Silhouette Index as well, with a value lightly under the 10% of the optimal value. The remaining three models had better outputs instead, in particular K-means.

The results of the different measures were similar also when different number of clusters were considered. The overall performance of the models, however, was relatively scarce (considering, for example, that the Silhouette maximum result of 1 was very far from the actual results). The final results questioned the reliability in the groups creation.

All in all, Hierarchical complete model was a candidate to be excluded, whereas the three remaining ones to be maintained.

3.4 External measures

Another way of evaluating the quality of the different clusters models is to use the external criteria. In this evaluation, the different models are compared with a pre-existent and defined cluster: in this case, the tier list of the competitive site "Smogon.com". For each of the three indices used (Rand, Jaccard and Fowlkes&Mallows), the best value is 1 and the worst is 0.

	Rand	Jaccard	Fow.Mal
Hier.Complete	0.741	0.207	0.351
Hier. Ward	0.743	0.132	0.258
Hier. Average	0.519	0.218	0.385
K-means	0.748	0.138	0.269

Considering 11 as the number of cluster, overall the best result was obtained by the Hierarchical Average model. It had in fact the lowest Rand measure, but the multiplicity of cluster made preferable a higher output in the Jaccard or Fowlkes&Mallows measures. The Hierarchical Average scored the best outputs in these two measures, although the remaining three clustering models obtained a nearly comparable output in these indices.

Modifying the number of clusters in a neighborhood of 11, the results suggested a similar solution, with some other model different from the Hierarchical Average being the best, but with an irrelevant difference.

However, an important fact must be taken in consideration in the definition of the pre-existent cluster. The source of the cluster, the competitive Pokémon-fight site "Smogon.com", is influenced by a number of elements apart from the raw statistics of each Pokémon, like trends of the moment, moves of each Pokémon, types influenced by a *rock-paper-scissors* dynamic, personal preferences of the players, strategies etc. All these aspects were not taken in consideration in the initial hypothesis of clustering, which could be the main reason for non-satisfying results obtained by external indices.

As a final consideration, with both external and internal measures taken into account it seems appropriate to suggest the Hierarchical Average method as a clustering solution, with 11 clusters as it has been initially planned.

3.5 Monte Carlo Procedure

It has been decided to use Monte Carlo Procedure to evaluate if the proposed solution is effectively better than a random solution, using the internal index 'Silhouette'.This is a statistical test, very useful in order to avoid the *apophenia* phenomenon. Therefore it requires a level of significance, the alpha parameter, a value typically set to 0.01. Furthermore, it requires also the number of clusters to create with the algorithm, and the number of samples to develop the procedure.

Therefore, it has been set:
Method: hierarchical average;
Number of clusters: 11;
Alpha level: 0.01;
Number of samples: 1000.

Results: the statistics obtained is 0.17. It can be concluded that the solution is a valid one, in respect with a alpha level equal to 0.01. It is a valid solution referring to a quantile of value 0.05.

3.6 The Fundamental Problem

At this point of the analysis, we are interested to see if the proposed cluster solution can be further improved, modifying the number of clusters to be created: in fact, we started from the assumption that the groups are 11, as reported by the scraped site. The question is: by varying the number of clusters, is it possible to get better results?

Optimal number of clusters choice - "The Fundamental Problem"- has been set:

minimum number of clusters: 4;

maximum number of clusters: 14.

Results:

Connectivity and Silhouette measures suggested a number of clusters equal to 4;

Dunn measure suggested a number of clusters greater than 10.

However the values of the Dunn index are very low, so the difference by changing the number of clusters is not very relevant. Therefore, it can be argued that the optimal number of clusters to consider in the solution is 4.

We now repeat the Monte Carlo Procedure, to see if, with 4 clusters, the validity of the solution persists and if, eventually, improves.

Results: compared to a quantile of 0.05, the same as earlier, the statistic obtained now assumes a value of 0.31. We can therefore deduce that the solution is not only valid (at an alpha level of 0.01), but is also significantly better than the 11 clusters case, according to the index 'Silhouette'.

3.7 Final clustering solution

Given the evidence obtained in the course of our research, the results of hierarchical clustering are as follows. Applying a Hierarchical cluster, with average linkage, the optimal number of clusters is equal to four, where the number of elements is very different within the clusters. In fact, looking at the records and their labels, we can classify the records in this way:

- a very small group of Pokémon with very high values in Defense;
- a fairly large group made up of the most scarce, non-evolved or rarely considered Pokémon;
- a very large group that takes into account all medium - strong Pokémon;
- a quite small group that contains most of the Legendary Pokémon, therefore characterized by the highest values possible.

Indeed, the solution seems to be lawful and intuitive. But, for greater clarity, we report the radar graph related to the averages (with relative standard deviation) within each group: the clear difference between the different clusters clearly emerges. The solution therefore seems to be good not only at an algorithmic level, as verified in the workflow, but also at an intuitive level.

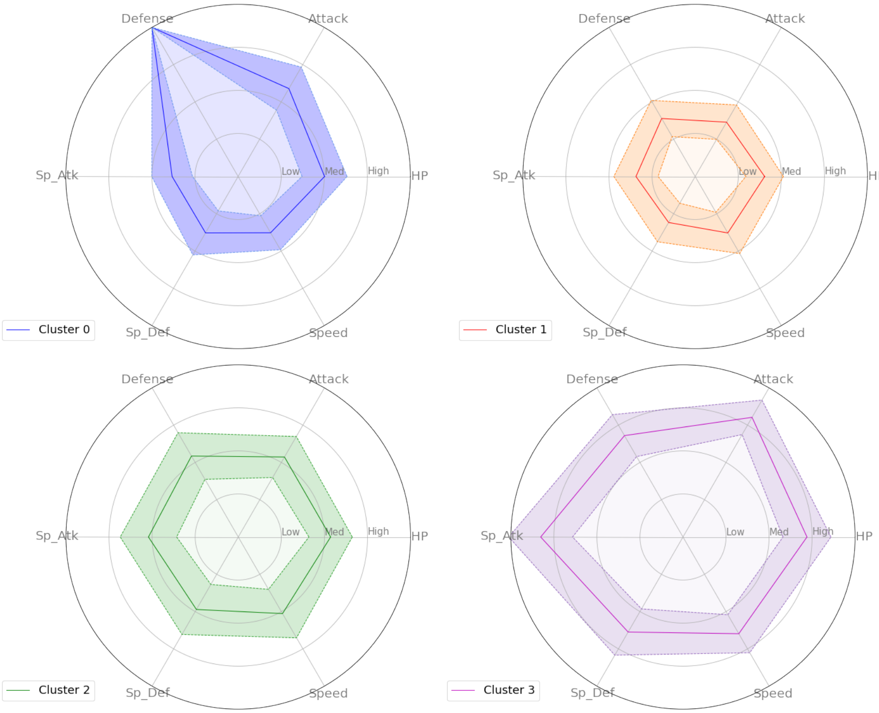

Figure 3.1: Mean values and standard deviation of different clusters

While trying to visualize the dataset with a radar graph, it has been noticed that the parameters (strength, speed and so on) had many outliers. Applying a min-max normalization from 0 to 1, even Pokémon with the highest stats like Mewtwo (154 Sp.Atk) were being visualized as not very strong in the graph, because of "outliers" such as Steelix (200 Def).

Since most of the Pokémon had parameters under 150, in order to make a better visualization we had to normalize data in a different way. To fix this issue, all the Pokémon having a parameter over 150 had been removed and the remaining data had been normalized from 0.1 to 1. The 0.1 minimum is to make sure that even Pokémon with the lowest stats had a visualization on the radar graph. With this normalization applied, all the parameters have a very well-built Gaussian distribution.

Then the normalization model has been applied to the whole dataset. The Pokémon that have a parameter over 150 become then over 1, but since it can be assumed that, only for visualization purposes, there is not a huge difference in very high values, the-out-of-bounds parameters have been modified to be equal to 1. In this way, Pokémon with parameters over or equal to 150 are visualized as having the maximum strength in that particular area.

The extra material to obtain the scraped data, the radar graph visualization and the dataset is available at this link, as a Google Drive folder freely accessible.