Project 1 : develop scikit-learn machine learning models to predict the function of the proteins on InterPROscan dataset

For this project, Spark framework is used that is an open-source distributed processing system for big data. Spark properties is configured to run with 16 worker threads and both spark.executor.memory and spark.driver.memory are set to 128g due to facing with “OutOfMemoryError” error.

Generally, three main steps are performed:

1. Data preparation
2. Machin Learning (ML) Implementation:
3. Evaluation

1. Data preparation:

This step is the most important step and exploring and identifying the data should be done very strictly. There are 1921817 records in the main dataframe consisting of information of different annotations for every protein and 332775 distinct proteins. Following measures are taken to prepare a subset ready for ML implementation:

- drop out the columns that are not required in the next steps.
- filter out the rows consisting of no protein annotation.(i.e InterPro_annotations is “-’)
- check whether there are any null records in dataframe
- I added a new column "size" that is the percentage of protein annotation size to protein size:
- df = df.withColumn('size', (((df['Stop'] - df['Start']) / df.Seq_len) * 100))
- create a separate dataframe “large_df” for large protein annotations (size is greater than 90%). Since there is more than one large annotation for one protein, I picked only the largest one for classification.
- make a list of distinct Proteins that have large features for finding small annotations named all_protein (208741 Protein).
- create a new dataframe “small_df” from the main dataframe for small features including only proteins that have the large feature (only the protein from the previous step). I also restricted the size of the feature to be less than 90%. The number of distinct proteins in this dataframe is less than all_protein. so some proteins have only a large annotation. (5202 small feature)
- transpose small_df to pivotDF. I need a dataframe having a row for every protein, and columns are the small annotations and filled with the count of small annotations for every protein. I made it using Pivot() on small_df. Pivot() It is an aggregation that one of the grouping column values is transposed into individual columns with distinct data.

- fill null values with zero. dataframe filled with null value If there were no small features in the previous step.
- change the type of all numerical columns from long to integer to dedicate less memory.
- join pivotDF and large_df to add the corresponding large annotation to protein.
- Combine multiple columns into a single feature vector in order to train ML models. all numeric columns are merged into a vector column using VectorAssembler.
- encode label column (InterPro_annotations) which is a string to a column of label indices using StringIndexer.
- split the prepared data to train and test dataset.

2. Machin Learning (ML) Implementation:

In this stage, several ML methods are applied to identify a pattern. LogisticRegression, Naive Bayes and RandomForestClassifier are implemented.

LogisticRegression is a popular method to predict categorical labels.

Naive Bayes is a family of multiclass classifiers based on applying Bayes' theorem.

RandomForestClassifier is an ensemble machine learning method and high predictive performance is expected.

3. Evaluation

In this step, the accuracy of ML model is assessed. MulticlassClassificationEvaluator is used for evaluation of the Multi-class(large annotations). Accuracy of different classifiers are listed below:

- NaiveBayes; 73%
- LogisticRegression : 78%
- RandomForestClassifier: 15%

Actually, I did not expect high accuracy because:

- only the largest features are picked out.
- the total size of protein did not take into account, probably two protein have the same small features, still they have different size and there may different nucleotides between small annotations.
- other attributes, such as protein family are not considered.

In my opinion, if more specific attributes are examined, more reliable model can be applied.