Noa Ecker and Ronit Feldman

[1]Unsupervised learning project

Selected paper

The paper we selected is : Zhiying Jiang, Bo Gao, Yanlin He, Yongming Han, Paul Doyle, Qunxiong Zhu, "**Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports**", *Mathematical Problems in Engineering*, vol. 2021, Article ID 6619088, 30 pages, 2021. https://doi.org/10.1155/2021/6619088. The paper aims to improve the performance of TF-IDF term weighting scheme for text classification by making it more robust to unbalanced categories. The original TF-IDF score is calculated as follows: $TF - IDF = TF(t,d) * IDF(t,d,D)$. The IDF value is evaluated as: $IDF(t,d,D) = log(\frac{|D|+1}{DF(t,d)+1})$ where $DF(t,d)$ is the number of documents in which the term *t* occurs. The TF value $TF(t,d)$ is the frequency of term *t* in document *d.* The authors noticed that when the corpus is not balanced with respect to the true text categories, terms belonging to more dominant categories will be assigned relatively low weights using the TF-IDF formulation. To this end, the authors suggested to adjust the $IDF(t,d,D)$ metric by using the deviation of document frequency from the average document frequency instead of using the document frequency itself. They defined the following metrics: $A_{DF}(t,D) = \frac{DF(t,D)-mean(DF)}{n}$ where *n* is the number of different terms. Then, they define $IADF(t,D) = log(\frac{|D|+1}{ADF(t,D)+1})$ to replace the original IDF formulation. To further reduce the weights of extremely high or low DF values, the authors suggested using the following metric: $IADF^{+}(t,D) = log(\frac{|D|+1}{DF(t,D)+1}) * \frac{1}{log(ADF(t,D)+1)+1}$. The authors provided additional formulations for collection frequencies by using normalization terms. Finally, the authors showed that the proposed $TF - AIDF$ scores are superior to the original $TF - IDF$ score on text classification tasks based on several unbalanced corpuses, using different classification algorithms.

- While this approach is relatively simple and easy to implement, it lacks a theoretical justification. In addition, it only considers the frequency of each term and thus ignores the semantic environment of each word.
- The authors mentioned previous approaches to handle class imbalance. (1) Data driven approaches- which are based on oversampling and/or undersampling. (2) Algorithm driven approaches- which are based on building more robust final classification algorithms. According to the authors, these methods become less effective as the imbalance becomes more prominent. This is due to the naive TF-IDF calculation, which could

favor less common terms in large-scale categories over more common terms in small-scale categories.

- We did not find other papers citing this paper which suggest further improvement to the suggested TF-IDF score.

Application on restriction enzymes detection

What are restriction enzymes?

Restriction enzymes, also known as restriction endonucleases, are a group of enzymes that are commonly found in bacteria and archaea. These enzymes function to cleave foreign DNA molecules at specific locations known as restriction sites. This mechanism serves as a defense against viral invasions while protecting the host DNA through the action of modification enzymes. In laboratory settings, restriction enzymes have been utilized for the purpose of DNA modification, making them a critical tool in contemporary biological research. To date, over 3600 distinct restriction enzymes have been characterized and classified. Despite similarities in structure and mechanism, significant variations exist between different restriction enzymes, which is thought to be a result of horizontal gene transfer.

The classification task

Classification of restriction enzymes presents a challenge due to its large variance, which has yet to be addressed through the utilization of machine-learning models. This study proposes the implementation of a variant of the TF-IDF algorithm to generate a numerical representation of DNA sequences, thereby facilitating the differentiation between restriction enzymes and non-restriction enzymes. The approach involves the extraction of K-mers with length 4 from each sequence, treating them as "words" and each sequence as a "document." The application of the TF-IDF approach aims to generate a numerical representation based on the discriminative power of the K-mers. TF-IDF assumes that K-mers(words) appearing in many sequences(documents) are less important. The dataset used for this study consisted of 243 known restriction enzyme sequences from 48 different bacterial organism families, along with on average around 50 additional non-restriction enzyme sequences per organism, resulting in an unbalanced dataset with a higher number of negative examples. In light of this, the methodology described in the paper mentioned above should be beneficial.

Results

The relevant code and detailed results can be found in the ***github repository*** *https://github.com/noaeker/re_project.*

As described above, we extracted Kmers of size 4 from each sequence. Then, we applied the standard TF-IDF method, the $IADF$ method and the $IADF^{+}$ methods on the obtained dataset. Next, we applied the PCA dimensionality reduction algorithm using 2 dimensions. Not surprisingly, when applying the PCA algorithm on the entire set of sequences, the data is mainly clustered according to the bacterial taxonomy.

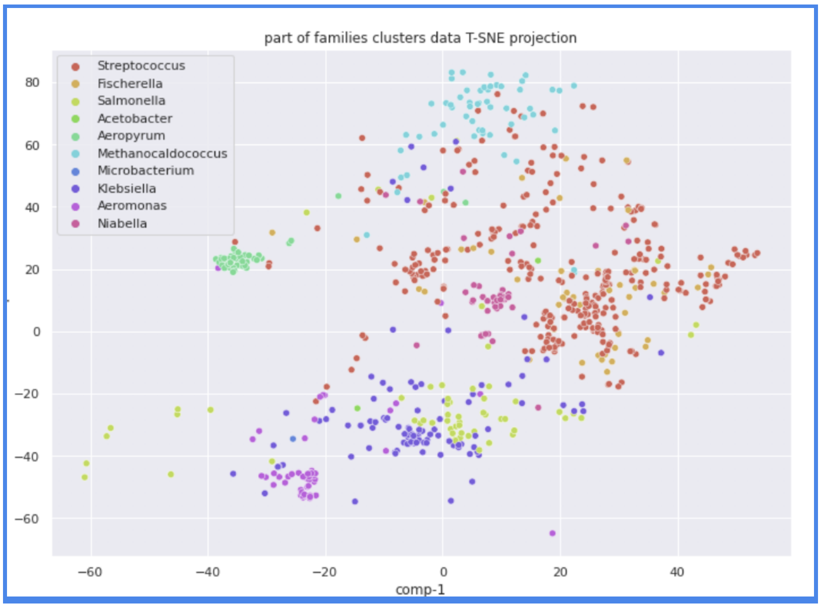

Fig 1. TSE 2d projection on the entire set of bacterial families

Hence, we decided to perform a separate dimensionality reduction for each bacterial family.

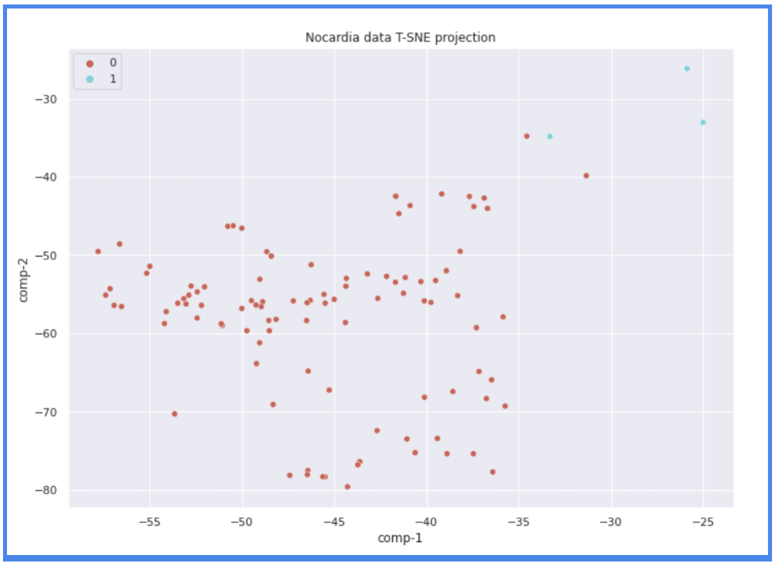

Fig 2. TSE 2d projection on the Nocardia family.

Looking at the obtained data distribution per family, we observed that restriction enzymes tend to be outliers compared to other proteins. Hence, this motivated us to use outlier detection algorithms and test their ability to detect restriction enzymes. Specifically, we used the following outlier detection algorithms: Robust covariance, one class SVM, Isolation forest and DBSCAN. To measure the performance of each algorithm we treated the outliers found by the algorithms as if they are classified as restriction enzymes. We used the following metrics for performance evaluation: true positives (TP), true negatives (TN), false positives (FP), false negatives(FN), Adjusted Rand Index(ARI) and The Fowlkes-Mallows index (FMI) iIn addition to homogeneity, completeness and the v_measure (i.e. the harmonic mean between homogeneity and completeness). In the notebook found in our github repository ('results summary'), we summarized the performance of each combination of TF-IDF version and outlier detection algorithm. The best performing combination was TF-IDF with Elliptic Envelope algorithm reaching an median FMI score of 0.895. Hence, in our case, the original TF-IDF score was better than the two new alternatives scores,although they are expected to perform better based on our initial hypothesis.

Although the overall performance is good, for some families the algorithm performed very well (for example, for the Moraxella family and Enterobacter families). While for other families, the algorithm performed poorly. We tried to test the correlation between the performance of our approach and the size of the family and the percentage of restriction enzymes within the family. However, these properties were not correlated to the overall performance.

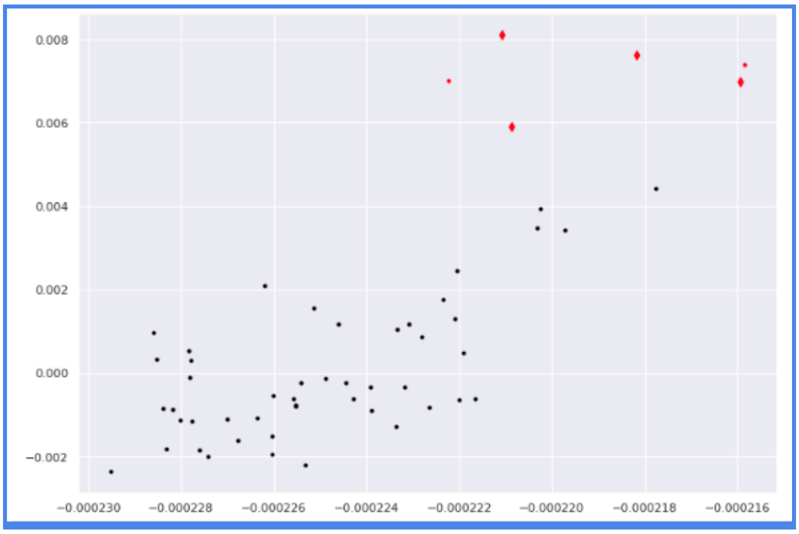

Fig 3. PCA 2d projection on the Moraxella family. Red dots represents restriction enzymes while black dots represent non restriction enzymes. The elliptical dots represent points which are considered outliers based on the outlier detection algorithm, while the classic dots represent non-outliers.

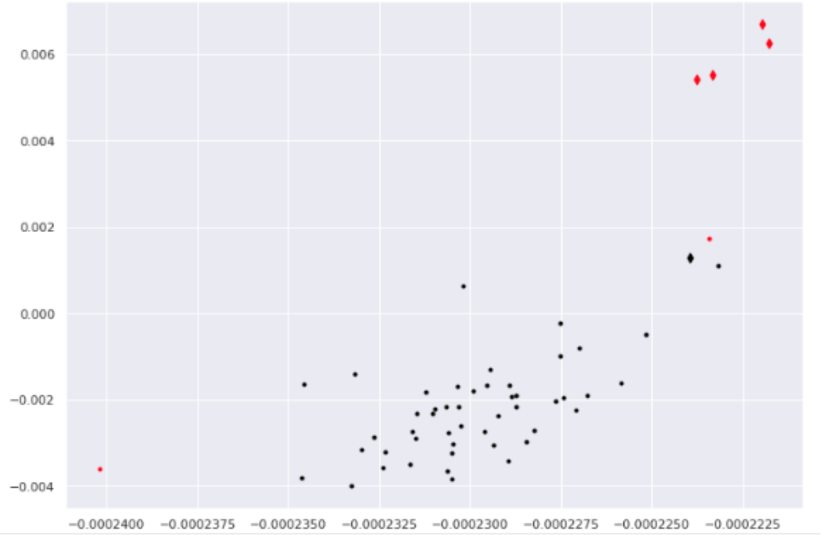

Fig 4. PCA 2d projection on the Enterobacter family. Red dots represents restriction enzymes while black dots represent non restriction enzymes. The elliptical dots represent points which are considered outliers based on the outlier detection algorithm, while the classic dots represent non-outliers.

To conclude, our new approach was shown to have a predictive power for detecting restriction enzymes. We observed that extracting K-mers from sequences and treating K-mers as tokens within documents has beneficial properties with respect to clustering and classification of sequences. In our case, restriction enzymes were shown to be outliers with respect to the rest of the proteins, and we have shown that using outlier detection algorithms is an appropriate methodology for this problem.