A Genetic Algorithm and Discriminant Analysis Based Outlier Detector

Dr. Eren Unlu

Paris, France
datascientist.unlu@gmail.com

Abstract—**Fisher Discriminant Analysis (FDA), also known as Linear Discriminant Analysis (LDA) is a simple in nature yet highly effective tool for classification for vast types of datasets and settings. In this paper, we propose to leverage the discriminative potency of FDA for an unsupervised outlier detection algorithm. Unsupervised anomaly detection has been a topic of high interest in literature due to its numerous practical applications and fuzzy nature of subjective interpretation of success, therefore it is important to have different types of algorithms which can deliver distinct perspectives. Proposed method selects the subset of outlier points based on the maximization of LDA distance between the class of non-outliers via genetic algorithm.**

I. Introduction

Unsupervised anomaly detection (also known as *novelty detection* or *outlier detection*) is probably one of the most interesting and demanded disciplines of machine learning and statistics mainly due to two factors : (1) Its various types are used for countless vital applications in today's automated and data driven world, from server overload forecasting to fraud detection; (2) Its definition of success or accuracy is almost completely subjective, changing based on the context and user interpretation [1] [2]. Some of these numerous fields of practical applications are fraud detection, automated identification of malfunctioning computer servers, medical diagnosis, intrusion detection in cybersecurity, computer vision [3].

The very question *What is an anomaly?* still lies at the heart of this discipline. From a statistical perspective first formal discussion on this issue is by Grubbs in 1969 [4] [5]. It is obvious that the definition of an anomalous point changes from setting to setting, changing demands but also to the human interpreter. Even though there can be certain common understanding to draw boundaries for the limits of normality for a specific application, discriminating small nuances close to these boundaries rests highly dubious. Especially, as the number of dimensions increases, where users can not interpret them visually. Hence, it is paramount of interest to use different types of algorithms to gain valuable insights.

Unsupervised outlier detection algorithms in the literature can be grouped in to three broad categories; *proximity based*, *clustering based* and *statistical modeling based* methods [6]. The common a priori parameter for these algorithms is the *contamination*, the estimated ratio of the outliers in the given dataset. Albeit most of the well known methods in the literature fall either in one of these three groups or their intersections, there also exists certain types of algorithms which can not be explained fully with this taxonomy [6]. Proximity based algorithms characterize each point with their position in the feature space with regard to their closest neighbors. By defining a proper distance metric, either the density of the data points in the vicinity or a direct distance based measure is used the score the anomaly of the point of interest in this neighborhood [7] [8]. On the other hand, clustering based algorithms aim to group data points in the feature space either directly based on the values or transformed metrics such as explaining the local connectivity of an instance [5]. In an iterative or single step fashion, the clustering algorithms encapsulates the most anomalous points in one minority class whose size is determined by the contamination ratio given by the user. Finally as the name suggests, statistical models tries to fit distributions or statistical systems to assign highest anomaly scores to a subgroup of pre-defined contamination size based on the inherent attributes of the data.

One-Class Support Vector Machine (SVM) algorithm is particularly interesting [9]. The central idea of this method is to train a SVM boundary with distances respect to the multidimensional origin by considering all points as the member of a single class. Inspired this technique, [10] proposes to use One-Class Linear Discriminant Analysis with kernels, where this time LDA discrimination is computed by generating *mirror points* in the feature space, a reflection with respect to the origin.

Due to proven ability of discriminant analysis for classification under various circumstances, we have experimented with the possibility to use it for unsupervised anomaly detection. Unlike [10] our model aims to find a subset of anomalies in the dataset which aims to maximize the LDA based separation between normal points and a set of anomalous points with predefined contamination ratio.

II. Proposed Method

Linear Discriminant Analysis (LDA) is based on a simple but very straight-forward rationale [11]. It is a supervised classification and dimensionality reduction technique where the optimal weights of the linear components, w are determined based on the maximization of the ratio of overall inter-class variations to the intra-class variations :

$$S = \frac{\overrightarrow{w}^T \Sigma_b w^T}{\overrightarrow{w}^T \Sigma w^T} \tag{1}$$

where, Σ is the covariance matrix of the whole dataset, C is the number of classes and Σ_b is the inter-class covariance :

$$\Sigma_b = \frac{1}{C} \sum_{i=1}^{C} (\mu_i - \mu)(\mu_i - \mu)^T \tag{2}$$

Conveniently, maximum number of components in this linear equation system can be $C - 1$. In our very specific case of anomaly detection, we have basically two classes, normal versus anomalous points, thus we seek to find a single dimensional vector of optimal weights, and finally a single dimensional projection of data on this component.

Of course the ultimate question in this context is how to decide which point belongs to which class, the precursor of the supervised classification algorithm of LDA. However, rather than using LDA as a classification tool, we would like to harness its separative potency. For this purpose, we propose to define the anomaly detection problem as finding the optimal subset of anomalous points where the difference of means of normal and outlier classes' on LDA projection is maximized.

The problem of finding a subset of points is NP hard by definition, and considering the computational demand for calculating LDA projection at each iteration, it is necessary to come up with an optimization algorithm. We have used Genetic Algorithm for this purpose [12] [13].

As most of the other unsupervised anomaly detection algorithms presented in the literature, the proposed method also assumes that the contamination ratio is given. At the first iteration (generation) of the algorithm, we create 500 random solutions (initial population). Each solution represents an arbitrary subset of points designated as outliers. For each solution, the LDA projection is calculated for the binary classification case of normal points versus anomalies. Next, the *fitness score* for each solution is calculated, which is the absolute difference of mean LDA scores of normal and outlier points (projected values on the found LDA component).

We have defined a simple mutation function, where a mutation on an arbitrary solution corresponds to swapping a single anomalous point with a normal one. We have also defined a basic two parent *mating* (cross-over) function, where a new solution is bred by selecting random anomaly points from the anomalies of each parent, of course by discarding the possibility of duplicates. We generate 1.2 mutations on average per solution at each generation. (Meaning, on average 1.2 points are swapped between normals and outliers). At each generation, we keep the 40% of the best solutions. Only top 10% of solutions are chosen eligible for mating, and we create new offsprings from these parents, where the number of new solutions are equal to 45% of the population of current generation. As it can be seen, at each generation the population grows geometrically. Based on our extensive experiments on various data sets, we have found this scheme as most plausible. Of course, if the number of generations is kept high for optimization; computation burden exceeds the limits. However, we have observed that keeping a relatively low number of generations and using a geometrically increasing population gave better empirical results. We have set the number of generations in our experiments to 50.

III. Experimental Evaluation

Proposed method is tested with 3 different well known datasets. In order to compare the performance with a baseline algorithm we have chosen *Isolation Forest*, one of the widely used unsupervised outlier detectors in the literature [14]. In order to observe the performance of the algorithms with increasing dimensionality gradually, the results for datasets in this section are ordered with increasing number of features. Without loss of generality, for the isolation forest we have set number of trees to 100, where all features are used for training each of them. For all 3 datasets we have used a contamination ratio of 5%.

A. Geyser Dataset

This dataset contains the 272 observations of the *Old Faithful Geyser* in Yellowstone National Park, Wyoming. It has only 3 features : Waiting time between eruptions, the duration of the eruption in seconds and a binary feature indicating whether it is considered a short or long eruption.

[15].

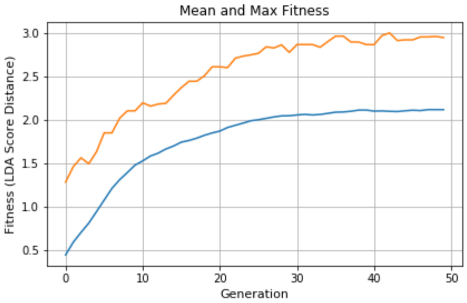

Fig. 1. The maximum and mean fitness score of each generation for *Geyser* dataset.

Fig.1 shows the maximum fitness score (the score of the best performing individual solution) and the mean fitness score of each generation for the Geyser dataset. We see that both mean and maximum fitness scores converge to a global maximum consistently over generations, which indicates the stability of the approach. Even though the best solution may not be at the last generation, we always keep the best solution in the history of optimization, thus use it as the final solution at the end.

What we observe from Fig. 2 is that the data has a bimodal nature, where two main clusters of waiting time-eruption duration patters exist. The upper graph shows the detected outliers in red with our proposed method and the bottom graph presents the results of the baseline isolation forest algorithm. As it can be seen, there are many common outliers identified by both of the algorithms, which indicates that there is an overall consensus on the universal accuracy given

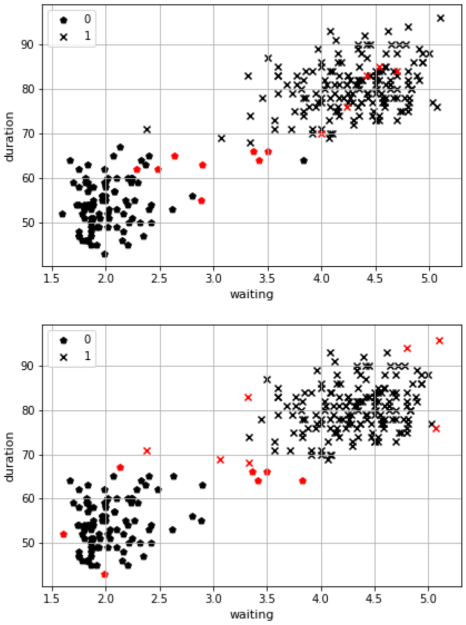

Fig. 2. The detected outlier points with proposed LDA algorithm (top graph) and isolation forest (bottom graph) for *waiting time between eruptions* (x-axis) versus *duration of eruption* (y-axis) in *Geyser* dataset. The type of the eruption 1 or 0 (long or short) is encoded by the shape.

the ambiguous nature of unsupervised anomaly detection. In addition to these, we see that the proposed algorithm is more capable of selecting short eruptions with longer waiting times. On the other hand, in this relatively low dimensional feature space, isolation forest seems to suggest semantically more valid anomalies. However, in order to understand the novelty introduced by the new algorithm we should test it with much more features, as in next 2 subsections.

B. *Cities Living Costs Indices Dataset*

The second dataset we have evaluated is called *Cities Living Costs Indices*, composed of 6 different living cost related indices of 536 cities around globe. Each index is a relative measure of cost compared to New York City. These features are *Cost of Living Index*, *Rent Index*, *Cost of Living Index plus Rent Index*, *Groceries Index*, *Restaurant Price Index* and *Local Purchasing Power Index*.

As it can be seen from Fig. 4, in addition to several commonly identified outliers, we see that isolation forest can easily detect the relatively extreme values (points on the right-hand side of the graph) due to its algorithmic nature, which depends on the overall depth of the trees to isolate a point. However, note that proposed approach can identify numerous points in a cluster which deviates from the linear pattern. This obviously suggests that our method can provide novel insights on unsupervised outlier detection.

C. *Wine Dataset*

The final dataset we experiment with, *Wine Dataset* has much higher dimensionality. It is composed of 178 samples of 3 classes of different wines (from 3 different regions) and their 13 numerical features. Without loss of generality, we have one hot encoded the wine type and included in the features. Thus, at the end there are 16 numerical features.

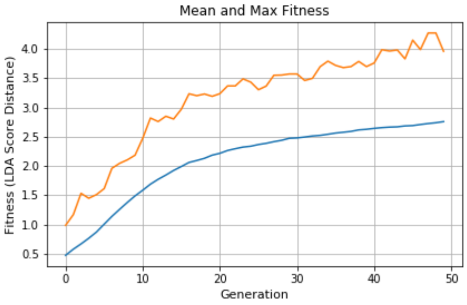

Fig. 3. The maximum and mean fitness score of each generation for *Cities Living Costs Indices Dataset* dataset.

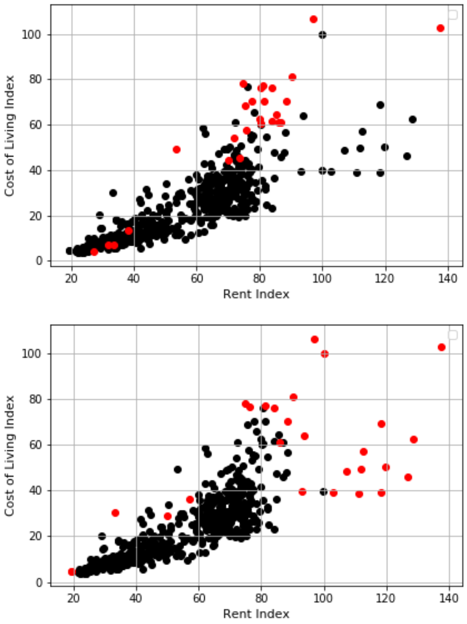

Fig. 4. The detected outlier points with proposed LDA algorithm (top graph) and isolation forest (bottom graph) for *Rent Index* (x-axis) versus *Cost of Living Index* (y-axis) in *Cities Living Costs Indices* dataset.

For instance in Fig. 5 where the *magnesium* versus *alcohol* is plotted, in addition to commonly identified outliers with isolation forest, it can be observed that our algorithm suggests visually interesting novel outliers, especially of wine type 1. A similar conclusion can be drawn from Fig. 6 for the visualization of *Nonflavonaid Phenols* versus *Proanthocyanin*. Our algorithm can identify two more wine samples of type 1 with significantly high nonflavonaid phenol content.

IV. Conclusion and Perspectives

In this paper, we have presented an innovative unsupervised anomaly detection algorithm which aims to leverage the discriminative potential of Fisher Discriminant Analysis. The

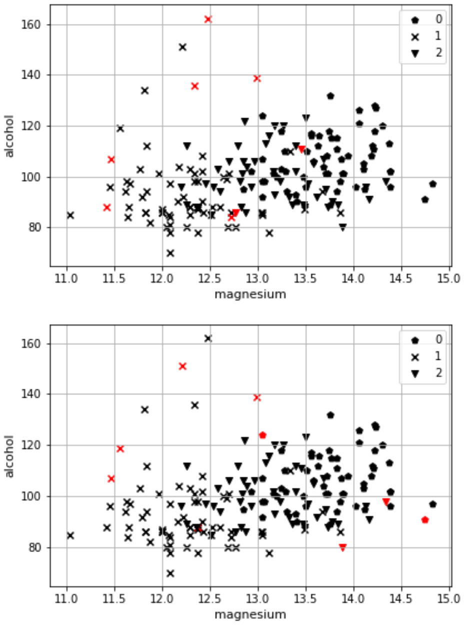

Fig. 5. The detected outlier points with proposed LDA algorithm (top graph) and isolation forest (bottom graph) for *Magnesium* (x-axis) versus *Alcohol* (y-axis) in *Wine* dataset. The three types of wines are encoded with shapes

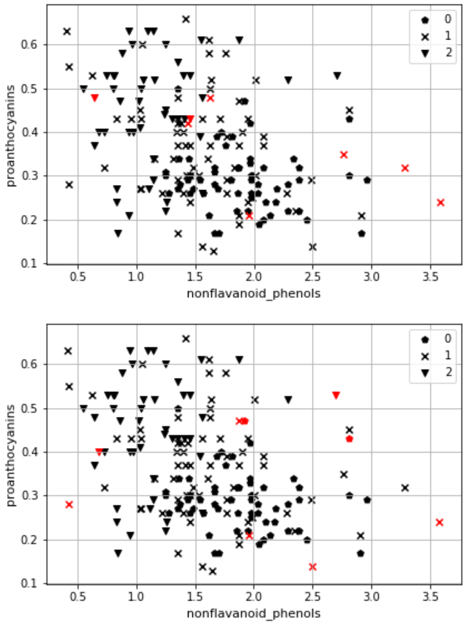

Fig. 6. The detected outlier points with proposed LDA algorithm (top graph) and isolation forest (bottom graph) for *Nonflavonaid Phenols* (x-axis) versus *Proanthocyanin* (y-axis) in *Wine* dataset. The three types of wines are encoded with shapes

central idea is to use genetic algorithm to find a bipartition of the dataset between normal and outlier points, which gives the largest mean difference of projections on the discriminant component. Unsupervised anomaly detection is a highly significant field due to its numerous vital practical applications. However, its performance is loosely defined and varies based on the context and semantic perception. Therefore, it is at paramount of interest to introduce new kind of algorithms such as the method proposed in this paper, which can provide novel insights and perspectives to users.

References

[1] C. Fan, F. Xiao, Z. Li, and J. Wang, “Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review,” *Energy and Buildings*, vol. 159, pp. 296–308, 2018.

[2] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, “A geometric framework for unsupervised anomaly detection,” in *Applications of data mining in computer security*. Springer, 2002, pp. 77–101.

[3] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, “Deep autoencoding gaussian mixture model for unsupervised anomaly detection,” in *International Conference on Learning Representations*, 2018.

[4] F. E. Grubbs, “Procedures for detecting outlying observations in samples,” *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[5] M. Goldstein and S. Uchida, “A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data,” *PloS one*, vol. 11, no. 4, p. e0152173, 2016.

[6] M. Goldstein and A. Dengel, “Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm,” *KI-2012: Poster and Demo Track*, pp. 59–63, 2012.

[7] M. J. Prerau and E. Eskin, “Unsupervised anomaly detection using an optimized k-nearest neighbors algorithm,” *Undergraduate Thesis, Columbia University: December*, 2000.

[8] F. Falcão, T. Zoppi, C. B. V. Silva, A. Santos, B. Fonseca, A. Ceccarelli, and A. Bondavalli, “Quantitative comparison of unsupervised anomaly detection algorithms for intrusion detection,” in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 2019, pp. 318–327.

[9] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, “High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning,” *Pattern Recognition*, vol. 58, pp. 121–134, 2016.

[10] V. Roth, “Outlier detection with one-class kernel fisher discriminants,” in *Advances in Neural Information Processing Systems*, 2005, pp. 1169–1176.

[11] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, “Fisher discriminant analysis with kernels,” in *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*. Ieee, 1999, pp. 41–48.

[12] D. Whitley, “A genetic algorithm tutorial,” *Statistics and computing*, vol. 4, no. 2, pp. 65–85, 1994.

[13] J. Yang and V. Honavar, “Feature subset selection using a genetic algorithm,” in *Feature extraction, construction and selection*. Springer, 1998, pp. 117–136.

[14] A. Zimek, E. Schubert, and H.-P. Kriegel, “A survey on unsupervised outlier detection in high-dimensional numerical data,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 5, no. 5, pp. 363–387, 2012.

[15] W. K. Härdle *et al.*, *Smoothing techniques: with implementation in S*. Springer Science & Business Media, 1991.