CSIT5210 Group Project Proposal

The COPOD Algorithm and its Applications under Multiple Scenarios

GROUP 02

20786466, BAI,Qian

20788658, TONG,Jinjian

20803628, YU,Ximeng

20785278, YUAN,Dian

20789341, WANG,Weihong

In data science, outliers refer to the data points which deviate dramatically from other normal observations. It might be bad data from human input error, or it might be important data which is worth studying. In either way, the existence of outliers could have a great impact on the model training. For example, outliers can importantly affect the choice of cluster center. To avoid its influence on our insight to the observation, it is important to detect the outliers, and study it when it is natural, or remove it when we need to preserve the integrity of the model. Over the last decades, different kinds of outlier detection algorithm are proposed, including covariance based model (i.e. MCD), proximity based model (i.e. KNN), linear model (i.e, SVM), ensemble base model (i.e. LSCP), and neural network based models (i.e. SO-GAAL and MO-GALL). All these models have different drawbacks. First, they all suffer from the curse of dimensionality, and their performance drops and the response time climbs as the dimensionality of data increases. Second, many of those models need extra amounts of work, including hyperparameter selection, tuning, or choosing layers of classification. Not to mention that different choices of hyperparameters can lead to different outcomes and have different performance. Therefore, it is important to develop a new algorithm that both has great performance, even with high dimension data, and does not need a heavy amount of work.

The author of the article proposes a method called COPOD which stands for **Cop**ula-Based **O**utlier **D**etection. The first question we need to solve is what copula is before we understand COPOD. Copula is a Latin word which means Link and Copula does play an important role in connecting. Basically, the Copula theory says that suppose we have N random variables which are X_1 to X_n, and their marginal distribution and joint distribution will be $F(X_1)$ to $F(X_n)$ and $H(X_1)$ to $H(X_n)$. The Copula theory believes there must exist a function C which satisfied:

$$H(x_1, x_2, ..., x_N) = C(F_1(x_1), F_2(x_2), ..., F_N(x_N))$$

according to the inverse transformation of CDF. We will get:

$$C(u_1, u_2, ..., u_N) = H[F^{-1}(u_1), F^{-1}(u_2), ..., F_N(x_N)]$$

Then we apply the inverse transform of cumulative distribution Function, and we will get the Copula formula. The COPOD uses the Copula formula to build a model with multi-dimensional cumulative distribution, and it can be used to effectively build a model based on the dependency between multiple random variables.

According to the article, COPOD uses a non-parametric method to obtain the empirical copula through Empirical CDF. After that, we can simply use the empirical copula to estimate the tail probability of the joint distribution in all dimensions. we should care about the possibility that a sample falls on the tail on the left and the tail on the right at the same time, but the actual situation may be more complicated. Anomalies may appear on the left side of the distribution, or on the right side of the distribution, or on both sides. In different situations, using different tail probabilities will get different results. In this case, we calculate the skewness of the distribution, in other words, the distribution is skewed to the left or to the right. If it is to the left, then we care more about the end on the right, and vice versa.

The proposed method has the following key advantages. First of all, it is different from most anomaly detection algorithms. COPOD does not need to calculate the distance between samples, so the running overhead is small and the speed is faster, and it can be easier to expand large data sets. COPOD can easily process data sets with 10,000 features and 1,000,000 observations on a standard personal laptop. Secondly, because COPOD is based on the empirical cumulative distribution function (ECDF), it does not need to be adjusted and can be called directly, which avoids the challenges of hyperparameter selection and potential deviations. The third is the effect of COPOD, which ranks the highest in comparison with other 9 mainstream algorithms (such as isolated forest, LOF) on more than 30 data sets. Its ROC-AUC score is 1.5% higher, and the average accuracy is 2.7% higher than the second-best detector. In addition, COPOD can provide some interpretability for which dimensions caused the anomaly, and quantify the anomaly contribution of each dimension through the dimension outline map. In this way, we can directly find the dimensions that cause the most anomalies for in-depth analysis. This allows practitioners to focus on certain subspaces to better improve data quality.

To verify the advantages of COPOD mentioned above, researchers designed several experiments on 30 public benchmark datasets. They implemented the COPOD-N algorithm as an API in the popular open source outlier detection Python toolbox-PyOD, which contains a wide range of outlier detection algorithms and models. Performance of each algorithm was evaluated by average ROC(receiver operating characteristic) and AP(average precision) scores of 10 independent trials. Firstly, researchers compared performances of 4 variants of COPOD algorithm, which have different calculation methods for tail probabilities andw corrections of statistics. Secondly, researchers compared COPOD with 9 other leading outlier detectors. COPOD shows better performance than leading outlier detection algorithms in most cases. Thirdly, researchers scaled COPOD to large dimensions to indicate its efficiency in computation by conducting experiments in a synthetic high dimension dataset. And researchers also visualized outlier detection results in the Breast Cancer Diagnostic Dataset to provide an explanation of the outlier detection process of COPOD.

In our implementation, we plan to build a model based on COPOD algorithm by coding. And we will follow the settings of the experiments in this paper and evaluate the model by comparing the performance of our implementation with researchers' implementation and other baseline models. We will also conduct some analysis if there is a discrepancy in the experimental results.

With the details and advantages of the COPOD algorithm investigated, we believe some hands-on implementations on real-life problems can further improve our understanding of this algorithm. Although many outlier detection algorithms have been focusing on industrial requirements such as anti-fraud and big-data analysis, we noticed that the efficiency and the minimal size of statistical algorithms like COPOD also reveals their application potential in small everyday problems. Therefore, we propose multiple outlier detection scenarios where this algorithm is worth trying, which can be roughly categorized into three categories: scenarios based on text, image, and sound.

- Text Scenarios
  - In some categorized folder of entries (in file system, email inbox, etc.), find out entries that are wrongly classified.
  - In the context of a certain written language, find out texts that are from a different language but using the same alphabet. (For example, a french comment in an English community.)
  - In research papers in a certain academic field, find papers whose titles are distinct from the rest papers in that field.
- Image Scenarios

  - In the collected ID photos, find photos that do not meet the standard.

- Sound Scenarios
  - In the context of a certain oral language, find out the voice clips that are spoken in a different language.
  - In music pieces categorized by their metadata, find out the wrongly classified pieces.

These scenarios cover a wide range of outlier detection motivations that emerge in our daily life, which are great for the evaluation of whether the COPOD algorithm can be applied to similar problems. If time permits, we will deal with two or three of the above problems using COPOD, and explain the results and the implications for the algorithms characteristics in our project report.

References

[1] Li, Zheng, et al. *COPOD: copula-based outlier detection.*
2020 IEEE International Conference on Data Mining (ICDM). IEEE, 2020.

This project is done solely within the course but not other scopes.