Report on HiCS: High Constrast Subspaces for Density-Based Outlier Ranking

STUDENT : AKASH CHANDRA BAIDYA, 230378, Master of Data Science, TU Dortmund, Germany
SUPERVISOR : DANIEL WILMES, Computer Science Department, TU Dortmund, Germany

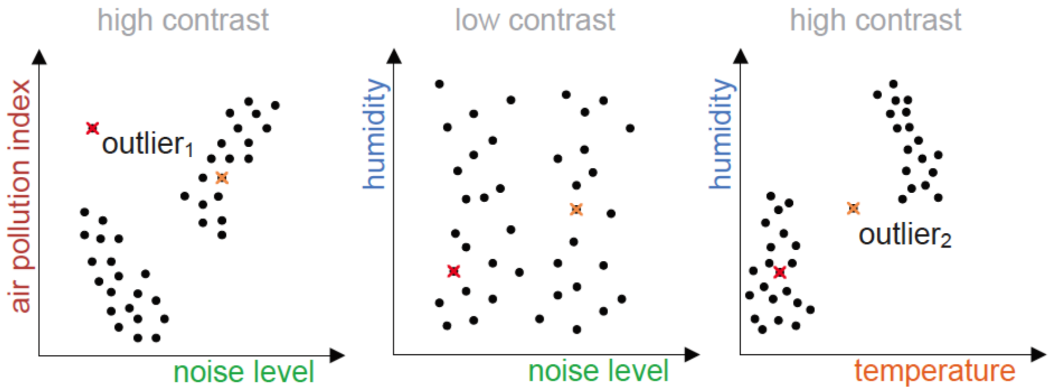

Fig. 1. Outlier Detection in high and low contrast subspaces[9]

Most of the real-world data sets are high-dimensional. Sometimes the data sets may contain hundreds or thousands of dimensions. Traditional outlier detection methods do not perform well in those scenarios. With the increasing dimensionality of datasets, it becomes hard to find outliers. High Contrast Subspaces for Density-Based Outlier Ranking paves the way to introduce the statistical independence for finding the high contrast subspaces and applying the scoring method to find outliers.

CCS Concepts: • **Computing methodologies** → **Anomaly detection**.

Additional Key Words and Phrases: Subspace Search, Curse of Dimensionality, High contrast subspace

ACM Reference Format:
Student : Akash Chandra Baidya, 230378 and Supervisor : Daniel Wilmes. 2022. Report on HiCS: High Constrast Subspaces for Density-Based Outlier Ranking. 1, 1 (February 2022), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

An outlier is a data point that diverges significantly from the rest of the data points in the data set. Outlier often contains abnormal but useful information of systems. Outlier analysis is often applied in recognition of intrusion to computer

Authors' addresses: Student : Akash Chandra Baidya, 230378, akash.baidya@tu-dortmund.de, Master of Data Science, TU Dortmund, Germany; Supervisor : Daniel Wilmes, , daniel.wilmes@cs.tu-dortmund.de, Computer Science Department, TU Dortmund, Germany.

Manuscript submitted to ACM

systems, fraudulent activities of credit cars, interesting sensor events, medical diagnosis, law enforcement, and earth science[2]. In high dimensional data set, there are many attributes. In Fig 1, four dimensions in a dataset are available and example of some subspace projections are depicted. $outlier_1$ appears in highly correlated attribute subspace and $outlier_2$ appears in another highly correlated attribute subspace. These outliers are hidden in the noise vs humidity projection. An outlier may only appear in some distinct combinations of attributes and normal objects can be found in full space or all other attributes. Most of the outlier mining methods search in full space which makes it hard to find the outlier since distances become similar in high dimensions. HiCS separates subspace outlier ranking in two steps- 1. **Subspace search** for finding high contrast of subspaces and 2. **Outlier ranking** for scoring in high contrast subspaces. So, this method increases the quality of traditional outlier rankings by calculating outlier scores in high contrast subspaces only[9].

2 PREVIOUS WORKS

There have been many methods proposed on outlier analysis but all these fails in certain situations creating necessity of a new outlier ranking method.

2.1 Distance, Density and Deviation Based Outlier Ranking

2.1.1 Distance Based Outlier Ranking: The idea behind the distance-based method is to determine if a point is an outlier or not based on the distance(s) to its neighbors. Distance-based outlier-detection (DB(r,π)), detects an outlier by calculating its distance relative to other objects. The assumption is that normal data points have a dense neighborhood and outliers are far apart from their neighbors having less density[10]. This approach can detect a significant global outlier among all data based on the parameter distance threshold r and the outlier fraction threshold π, but it cannot detect a local outlier from a cluster.

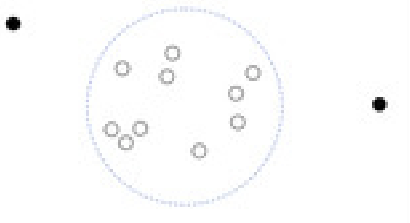

Fig. 2. Distance Based Outlier Ranking[13]

2.1.2 Density Based Outlier Ranking: The idea behind the density-based method is to compare the density around a point with the density around its local neighbors and The relative density of a point compared to its neighbors is computed as an outlier score. Density-based outlier-detection (LOF(n)), detects a local outlier by detecting a significant object that is far from the others among a set of closely related objects based on the parameter number of the n closest-neighbor[6]. The basic assumption is that the density around a normal data object is similar to the density around its neighbors and The density around an outlier is considerably different from the density around its neighbors.

This model is based on three key factors -

- `Reachability distance`: introduces a smoothing factor.

- `Local reachability distance (lrd) of a point`: inverse of the average reachability distances of the kNNs of a point.
- `Local outlier factor (LOF) of a point`: average ratio of local reachability distances of neighbors of a point and local reachability distance of it.

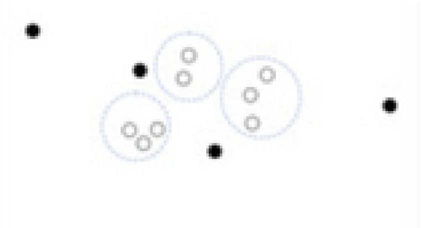

Fig. 3. Density Based Outlier Ranking[13]

2.1.3 Deviation Based Outlier Ranking: The general idea behind the deviation-based outlier mining is that Outliers are points which do not fit the general characteristics of that set, i.e., the variance of the set is minimized when removing the outliers. The basic assumption is that outliers are the outermost points of the data set.

2.1.4 Limitations of traditional approaches: These outlier mining ranking have one problem in common. These approaches calculates outliers in full data space so these outlier mining ranking are unable to detect outliers in subspaces [9].

2.2 Feature Bagging

In the Feature Bagging method, the outlier detection method (LOF) is run in several random feature subsets and the results are combined to an ensemble[11]. Each base component of the ensemble follows the random selection and applies an outlier detection algorithm. Different outlier detection algorithms can be applied in each iteration after the normalization of the scores though in the paper only LOF is used. After that, the outlier scores can be combined with either breadth-first search or cumulative-sum approach.

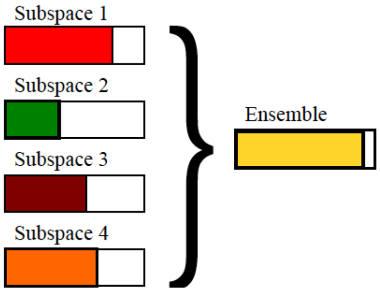

Fig. 4. Feature Bagging Method

In this approach, it takes the benefit of finding the true outliers from normal data which was not possible in traditional full-dimensional outlier algorithms. It is not a specific approach for high-dimensional data but provides efficiency gains by computations on subspaces and effectiveness gains by the ensemble. Feature bagging uses random subspaces to find a full dimensional result.

2.2.1 Limitations of Feature Bagging: Since the selection of subspaces is done through random selection, this does not ensure high-quality results. The random selection of subspaces will result in random ranking which is similar to full dimensional data space.

If more dimensions are found irrelevant, at least some of them are likely to be added in every sub-space sample. Again, There is a chance that information will be lost because many of the dimensions are dropped randomly. This imposes obstacles to the accuracy of the approach[2].

2.3 PCA

Principal component analysis (PCA) is an unsupervised algorithm. It is mainly used for dimensionality reduction but it is used to analyze the internal structure of the data. When PCA receives N samples and D attributes, each attribute represents one coordinate axis. PCA reduces these dimensions by finding different attributes(coordinate axes). Principle Components are the axes to which the attributes D is reduced. First Principal Component (PC1) is the line in the d-dimensional space that goes through the mean of the data points and predicts the data in the least square method. It denotes the highest of the total variance in the seen variables. The second Principal Component (PC2) is orthogonal to PC1 in k-dimensional space. This line goes through the mean point and enhances the approximation of the X-data. This assists in finding out the points which are far from other data points and are known as outliers. In PCA, the original variables are lost and so the newly formed Principle components are not as readable and interpretable as the earlier ones. Data are required to be normalized and scaled before the application of PCA [1]. PCA in itself is computationally expensive but helps in removing outliers and improving visualization.

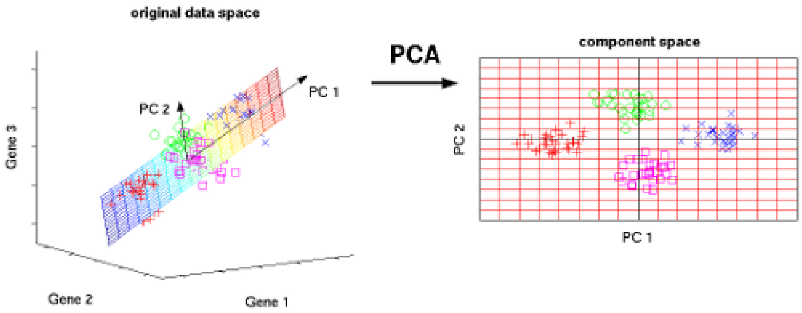

Fig. 5. PCA[5]

2.3.1 Limitations of PCA:. PCA finds the lower-dimensional projections but does not work well as preprocessing step for outlier ranking. The variance of the data which is used as a general measure in PCA can not be applied for objective functions for outlier ranking.

2.4 AutoEncoder

Autoencoders can be used for the purpose of anomaly detection. The main idea is to train the autoencoder with the normal sequence without any outliers. After feeding the autoencoders with the data containing outliers, high reconstruction errors for outliers are obtained and low reconstruction error is obtained for normal data. The reconstruction error is defined as an anomaly score. At the time of encoding and decoding outliers, a portion of the information is lost because of which a high outlier score for outliers is obtained[2].

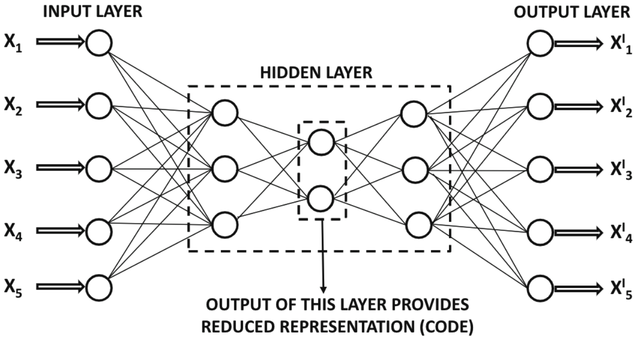

Fig. 6. AutoEncoder[2]

2.4.1 Limitations of AutoEncoder: AutoEncoder is slow to train and also faces a problem of missing significant dimension for keeping the lower number of units in layers of AutoEncoder. AutoEncoder is algorithm specific and faces problem for new type of input data. AutoEncoder also faces the problem of overfitting.

2.5 ENCLUS and RIS

Enclus[4] proposes a selection of subspaces based on the entropy measure for databases of numerical data. The quality measurement of subspaces relies mostly on the subspace clustering algorithm CLIQUE. Data space is divided into equal-sized grid cells. A subspace of low entropy is selected[9].

RIS[7] proposes a selection of subspaces based on density. It relies specifically on DBSCAN algorithm. The core objects are counted in a specific subspace selection and this count of core objects is used as the measure for subspace selection[9].

2.5.1 Limitations of ENCLUS and RIS:. These two methods rely on the specific clustering method. So, the selection of subspace depends on the clustering model. These two search methods do not promise to work well in the general scenario[9].

3 METHOD

There are many challenges with high dimensional data such as-

- `Concentration of Scores`: distances of attribute wise objects converge to normal distribution with low variance
- `Noise attributes`: large part of irrelevent attributes can influence the distances

- `Definition of Reference-Sets`: right subspace is required to be known to find neighbors
- `Bias of Scores`: without normalization, L_p norms are biased and distances are not comparable
- `Interpretation & Contrast of Scores`:due to density, scores become similar
- `Exponential Search Space`: with increase of dimensionality, the number of subspaces grows exponentially

3.1 Main focus

The main idea of HiCS[9] is finding subspaces with high contrast. For finding the contrast, correlation among the attributes of a subspace is calculated. It searches for violation of statistical independence assuming statistical dependence in high dimensions. For statistical independence, marginal and conditional distributions are compared. In one dimensional subspace, trivial outliers are identifiable and in high contrast subspaces, outliers are not trivial but deviate from the correlation trend shown by the majority of data in this subspace.

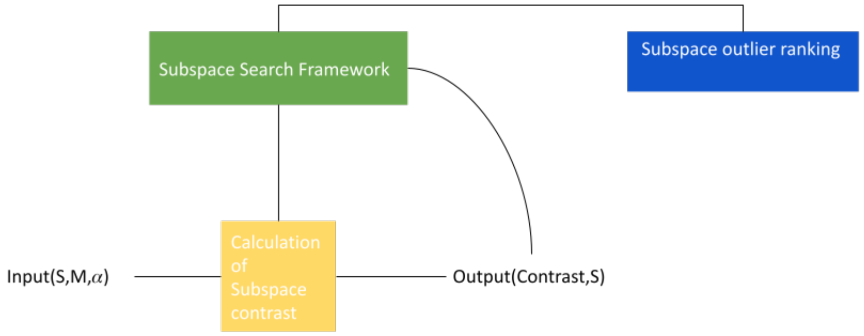

Fig. 7. Flow of HiCS

3.2 Subspace Contrast

Consider a subspace of dimensionality d, in which the dimensions are indexed as1...d. The conditional probability $P(x_1|x_2...x_p)$ for an attribute value x_1 is the same as its unconditional probability $P(x_1)$ for the case of uncorrelated data. High-contrast subspaces are likely to violate this assumption because of non-uniformity in data distribution. The idea is that subspaces with non-uniformity have more chances to contain outliers[2].

For a low contrast subspace and uncorrelated data,

$$\underbrace{p_{s_1|S_2,....S_d}(X_{S_1}|X_{S_2},....X_{S_d})}_{p_{S_i|C_i}} = \frac{p_{s_1,....S_d}(X_{S_1},....X_{S_d})}{p_{S_2,....S_d}(X_{S_2},....X_{S_d})} = \underbrace{p_{s_1}(X_{S_1})}_{p_{S_i}} \tag{1}$$

For evaluation of a subspace by a Monte Carlo Algorithm with M iterations are conducted. In each iteration-

- A random marginal attribute S_i is selected
- A random condition set C_i is imposed
- Violation of equation 1 is determined

After, all deviation results will be combined to obtain a single contrast value for the subspace

$$contrast(s) = \frac{1}{M}\sum_{i=1}^{M} deviation(p_{S_i}^{(m)}, p_{S_i|C_i}^{(c)}) \tag{2}$$

3.3 Deviation function

To check whether the independence assumption is violated or not, sample deviation is compared with hypothesis testing. The null hypothesis is assumed that both samples are from the same distribution. Welch's t-Test and Kolmogorov Smirnov test can be used to measure the deviation of a subspace from the basic hypothesis of independence. This provides a measure of the non-uniformity of the subspace and a way to measure the quality of the subspaces in terms of likeliness to contain outliers.

3.4 Adaptive Condition Sets

For a subspace with dimensions, a number of conditions are $|C| = d - 1$ which creates the problem of the condition set scaling to subspaces of dimension d. For making sure of a fixed size of selected sample, a random object is chosen as centroid for the candidate slice and the candidate slice is scaled according to $N. \sqrt[|C|]{\alpha}$)

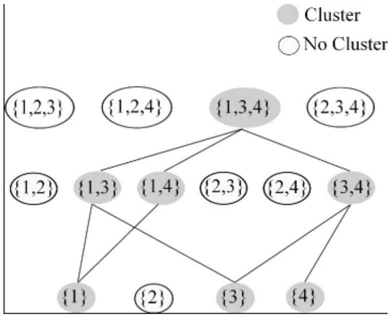

Fig. 8. Bottom up Subspace Search(Apriori like)[8]

3.5 Subspace Search Framework

A bottom-up Apriori-like approach is used to identify the relevant projections. In this bottom-up approach, the subspaces are continuously extended to higher dimensions for non-uniformity testing. Similar to the Apriori method, only subspaces having sufficient contrast are provided for non-uniformity testing as potential candidates. After the end of this stage, an additional pruning step is required which can remove redundant subspaces. A subspace of dimensionality d is removed, if another subspace of dimensionality(d+ 1) exists (among the reported subspaces) with higher contrast.

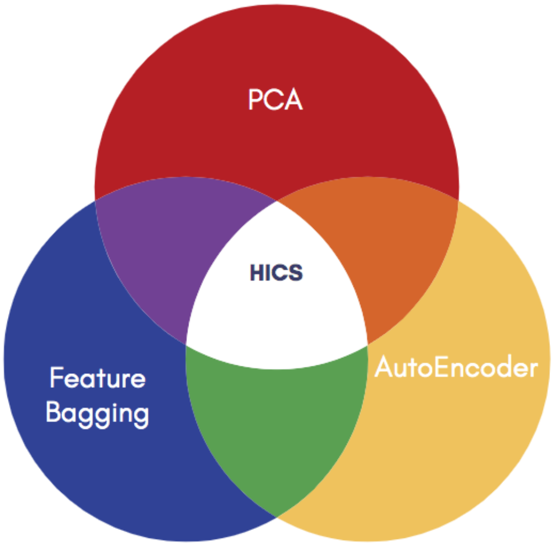

Fig. 9. Venn Diagram of PCA, AutoEncoder, Feature Bagging and HiCS

4 COMPARISON WITH OTHER METHODS

HiCS shows robust parameterization for No of Statistical Tests, M, and Ratio of Test Statistic, α. HiCS outperforms other methods with increasing dimensionality and scalable processing of HiCS.

PCA(Principal Component Analysis) turns the data into lower-dimensional sub-spaces using linear correlations. After that, the distance of each data point from the plane that fits the sub-space is calculated. This distance is applied to find outliers. PCA is an example of linear models for anomaly detection. PCA relies on the variances of individual dimensions which gives little information about the dependencies of dimensions. HiCS which uses biased subspace selection, almost always use the dependencies among dimensions as a key selection criterion.

As discussed earlier, Feature bagging randomly samples subspaces and the results of outlier ranking are joined to an ensemble. It becomes hard to find the nontrivial outliers which are not identifiable in full space and one-dimensional space and also information is lost. On the other side, HiCS selects high contrast subspaces where nontrivial outliers are discovered.

Heterogeneous Detector Ensemble on Random Subspaces (HeDES) extends Feature Bagging by using of different outlier detection technique to estimate outlier scores for each point on random subspaces[12].

Autoencoders are an easier choice for outlier detection because they are commonly used for dimensionality reduction of multidimensional data sets as an alternative to PCA or matrix factorization. Autoencoders are slow to train and sensitive to noise-causing overfitting. AutoEncoder can handle nonlinear complex structures. HiCS can also handle complex nonlinear distributions.

One example of an ensemble-based autoencoder is "Outlier Detection with Autoencoder Ensembles" [3]. The main idea is based on varying randomly the connecting architecture of the autoencoder to achieve better quality performance.

Boosting-based Autoencoder Ensemble approach (BAE) [14] is an unsupervised ensemble method that, similarly to the boosting approach. It builds an adaptive set of autoencoders to gain improved and robust results better than HiCS[9].

5 LIMITATION

HiCS combines LOF scores from subspaces of different dimensionality without score normalization which is the bias of scores, a common problem of curse of dimensionality. This combination of scores is rather naive and ensemble reasoning can be applied. The implicit notion of density is appropriate only for density-based outlier scores.

There are many alternatives available for finding high-contrast subspaces, which might be worth exploring. For example, one can use the multidimensional Kurtosis measure in order to test the relevance of a subspace for high-dimensional outlier detection and Kurtosis measure measures the non-uniformity. Steps of Kurtosis measure-

- Calculation of mean μ and σ
- Standardization to zero mean and unit variance with the formula

$$z_i = \frac{x_i - \mu}{\sigma} \tag{3}$$

- Computation of Kurtosis Measure:

$$K(z_1, ...z_N) = \frac{\sum_{i=1}^{N} z_i^4}{N} \tag{4}$$

This measure is easy to compute and also considers the interactions between the dimensions into account because of the Mahalanobis distance.

HiCS [9] can fall victim to the curse of dimensionality with respect to conditional distributions in high-dimensional spaces. It may miss relevant subspaces because of its random nature.

6 CONCLUSION

HiCS is a subspace search method for finding outliers in high dimensions and it focuses on nontrivial outliers. It is a two-step process, one step is subspace search which looks up for high contrast subspaces and another step is outlier ranking where scoring algorithms are applied in those high contrast subspaces. Different outlier scoring can be applied for further improvement.

REFERENCES

[1] Amulya Agarwal and Nitin Gupta. 2021. *Comparison of Outlier Detection Techniques for Structured Data.*

[2] Charu C. Aggarwal. 2017. *Outlier Analysis* (2nd. ed.). Springer, New York, NY.

[3] Jinghui Chen, Saket Sathe, Charu Aggarwal, and Deepak Turaga. 2017. *Outlier Detection with Autoencoder Ensembles.* 90–98. https://doi.org/10.1137/1.9781611974973.11

[4] C.-H. Cheng, A.W. Fu, and Y. Zhang. [n. d.]. Entropy-based subspace clustering for mining numerical data. In *KDD.* 84–93.

[5] Samadrita Ghosh. 2019. Principal Component Analysis in R. Retrieved February 6, 2022 from https://dimensionless.in/principal-component-analysis-in-r/

[6] Wen Jin, Anthony KH Tung, and Jiawei Han. 2001. Mining top-n local outliers in large databases. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.* 293–298.

[7] Karin Kailing, Hans-Peter Kriegel, Peer Kröger, and Stefanie Wanka. 2003. Ranking Interesting Subspaces for Clustering High Dimensional Data. *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)* 2838, 241–252. https://doi.org/10.1007/978-3-540-39804-2_23

[8] Amardeep Kaur and Amitava Datta. 2015. A novel algorithm for fast and scalable subspace clustering of high-dimensional data. Retrieved February 6, 2022 from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0027-y/figures/1

[9] Fabian Keller, Emmanuel Muller, and Klemens Bohm. 2012. HiCS: High Contrast Subspaces for Density-Based Outlier Ranking. In *2012 IEEE 28th International Conference on Data Engineering*. 1037–1048. https://doi.org/10.1109/ICDE.2012.88

[10] Edwin M. Knorr, Raymond T. Ng, and Vladimir Tucakov. 2000. Distance-based outliers: algorithms and applications. *The VLDB Journal* 8, 3 (01 Feb 2000), 237–253. https://doi.org/10.1007/s007780050006

[11] Aleksandar Lazarevic and Vipin Kumar. 2005. Feature Bagging for Outlier Detection. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (Chicago, Illinois, USA) *(KDD '05)*. Association for Computing Machinery, New York, NY, USA, 157–166. https://doi.org/10.1145/1081870.1081891

[12] Hoang Nguyen, Hock Ang, and Vivekanand Gopalkrishnan. 2010. Mining Outliers with Ensemble of Heterogeneous Detectors on Random Subspaces. 368–383. https://doi.org/10.1007/978-3-642-12026-8_29

[13] Iq Reviessay Pulshashi, Hyerim Bae, Hyunsuk Choi, Seunghwan Mun, and Riska Asriana Sutrisnowati. 2019. Simplification and Detection of Outlying Trajectories from Batch and Streaming Data Recorded in Harsh Environments. *ISPRS International Journal of Geo-Information* 8, 6 (2019). https://doi.org/10.3390/ijgi8060272

[14] Hamed Sarvari, Carlotta Domeniconi, Bardh Prenkaj, and Giovanni Stilo. 2021. Unsupervised Boosting-Based Autoencoder Ensembles for Outlier Detection. In *Advances in Knowledge Discovery and Data Mining*, Kamal Karlapalem, Hong Cheng, Naren Ramakrishnan, R. K. Agrawal, P. Krishna Reddy, Jaideep Srivastava, and Tanmoy Chakraborty (Eds.). Springer International Publishing, Cham, 91–103.