Beyond Defaults: Parameter Free Outlier Detection using GLOSH

by

Kushankur Ghosh

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

In machine learning and data mining, outliers—data points significantly differing from the majority—often pose challenges by introducing irrelevant information. Unsupervised methods are often used for detecting them as the information about outliers is unknown. *Global-Local Outlier Scores based on Hierarchies* (GLOSH) is an unsupervised outlier detection method within HDBSCAN*, a hierarchical clustering approach. GLOSH estimates outlier scores (GLOSH scores) for each data point by comparing its density to the highest density point in its closest cluster within the HDBSCAN* hierarchy. However, GLOSH is sensitive to the min_{pts} parameter, that estimates the density in the hierarchy. Different min_{pts} values may result in different hierarchies, with some representing the underlying cluster structure better than others. Given the lack of prior knowledge about the data in practice, it is unlikely to know an appropriate min_{pts} value beforehand, i.e., that assigns higher GLOSH scores to "true outliers" than to "true inliers". Moreover, to select outliers using GLOSH, one has to pre-define a value n that is used to determine the n datapoints with the highest GLOSH scores. These n datapoints are treated as "potential outliers". However, in practice, one may not know how many outliers are present in a dataset, making it unlikely to know a suitable value for n.

The first contribution of this thesis is an automated approach that aims at determining the value of min_{pts} from a large range of min_{pts} values that results in the best GLOSH performance. Given a range of min_{pts} values, we can obtain a corresponding range of GLOSH scores for each datapoint, which we call the datapoint's *GLOSH–Profile*. We study the behavior of GLOSH–Profiles for distinct outlier types, establishing their ability to distinguish between different kinds of outliers.

Our first major observation is that the min_{pts} value that results in the best overall GLOSH performance corresponds to min_{pts} value where the GLOSH scores in the GLOSH–Profiles start changing at a similar rate. Based on this observation, we develop an automated, unsupervised method to find the min_{pts} value at which the GLOSH scores in the profiles start changing at a similar rate, thus potentially yielding the best or nearly the best results for GLOSH. We apply our method on a range of different synthetic and real datasets with added synthetic outliers, and show that our approach of selecting the min_{pts} value is able to match the best possible performance of GLOSH in a given range of min_{pts} values.

Our second major observation is regarding the min_{pts} value that yields the best GLOSH performance: the GLOSH scores of outlier datapoints notably deviate from the inlier GLOSH scores (for that min_{pts} value) when arranged in increasing order. This observation serves as the key for the second contribution of this thesis—a strategy to estimate a threshold for classifying points into inliers and (potential) outliers, without relying on a pre-defined value of n. The proposed approach is evaluated across synthetic, semi-synthetic, and real datasets. The results show that our approach yields a consistently effective threshold.

Acknowledgements

Over the past few years, I have had the privilege to interact and work with many great scientists in my field. Firstly, I would like to thank my supervisors, Jörg Sander and Murilo Naldi, for their constant feedback, support, and patience throughout the journey. I would like to express my deepest appreciation to them for teaching me how to approach research problems and most importantly, how to think and present my ideas. Our insightful discussions during weekly meetings has helped me to be a more adept researcher. I am equally grateful to my collaborator, Euijin Choo, for her invaluable feedback on my ideas and her guidance on presenting them. I also extend my thanks to Osmar Zaïane for agreeing to be on my examining committee and providing valuable feedback on my thesis.

I owe my thanks to Sankhadeep Chatterjee, who introduced me to machine learning research. Without his encouragement, I might never have pursued research.

The journey would not have been as smooth without the support of my friends. Arghasree, your constant motivation has been a driving force, and I am grateful for your enduring encouragement. I could not have asked for anything more. I am grateful to Samridhi, Subhojeet, Saqib, Aakash, Dhruv, Sonali, Anindita, Shashank, Mohan, and Kushagra for being the reason of some of the most memorable moments in my life. To my friends from India, Swarnava De and Anushka Ghosh, even a five-minute phone conversation with you can reduce my stress-levels significantly.

Lastly, I would like to thank my family for always having my back. To my parents, Kaushik Ghosh and Kakoli Ghosh, thank you for instilling in me the value of prioritizing knowledge above all. I am equally grateful to my aunt and uncle, Kajari

Ghosh and Subhotosh Banerjee, for being the source inspiration since my childhood. Look where it has brought me.

Table of Contents

1 Introduction 1

1.1 Aim and Scope . . . 3

1.2 Thesis Outline . . . 4

2 Background and Related Work 6

2.1 Unsupervised Outlier Detection . . . 6

2.2 HDBSCAN* . . . 8

2.3 GLOSH . . . 10

2.4 Unsupervised Selection of a min_{pts} value . . . 11

3 Using GLOSH–Profiles for min_{pts} Selection 13

3.1 GLOSH–Profile: Outlierness across min_{pts} values . . . 13

3.1.1 GLOSH–Profiles of different kinds of points . . . 21

3.1.2 Limiting the size of the GLOSH–Profiles . . . 27

3.1.3 GLOSH–Profiles on Different Datasets . . . 31

3.2 GLOSH Performance across min_{pts} values . . . 39

3.3 Dissimilarity between GLOSH score rankings for consecutive min_{pts} values . . . 45

3.4 Finding the “Best” min_{pts} value . . . 53

3.5 Experimental Analysis . . . 55

3.5.1 Setup . . . 55

3.5.2 Auto-GLOSH versus GLOSH . . . 56

3.5.3 Auto-GLOSH versus KNN and LOF . . . 62

3.6 Conclusion . . . 66

4 Unsupervised Labelling of Potential Outliers using GLOSH scores 68

4.1 GLOSH scores at the “Best” min_{pts} . . . 68

4.2 Automatically Labelling Potential Outliers . . . 70

4.2.1 Finding the “knee” GLOSH score . . . 71

4.2.2 Adjusting the Inlier Threshold 71
4.3 Experimental Analysis 73
4.3.1 Setup 73
4.3.2 Results 74
4.4 Conclusion 81

5 Conclusion & Future Work 82
5.1 Conclusions 82
5.2 Future Work 83

Bibliography 84

List of Tables

3.1 Dataset Description 28

3.2 Comparing the Precision@n obtained using shorter GLOSH–Profiles and full GLOSH–Profile 30

3.3 Pearson dissimilarity in a scenario where the GLOSH scores change at a similar rate: Each score in the row A decreases by a factor of 5 in the row B. The *Sorted-A* is the sorted sequence of the scores before they change (A), while the *Sorted-B* is the sorted sequence of the scores after they decrease (B). The Δ is computed between the sorted sequences. 46

3.4 Pearson dissimilarity in a scenario where only one score jumps from low to high, while the rest do not change: All the scores from row A are identical in the row B apart from the score obtained by an outlier x_{20}. The Δ is computed between the sorted sequences of A (*Sorted-A*) and B (*Sorted-B*). 47

3.5 Auto-GLOSH vs GLOSH: Precision@n obtained on Type 3 Datasets . 57

3.6 Auto-GLOSH vs GLOSH: Precision@n obtained on Type 2 Datasets with Outlier Clumps 58

3.7 Auto-GLOSH vs GLOSH: Precision@n obtained on Type 2 Datasets with Global Outliers 59

3.8 Auto-GLOSH vs GLOSH: Precision@n obtained on Type 2 Datasets with Local Outliers 60

3.9 Auto-GLOSH vs GLOSH: precision@n obtained on Type 2 Datasets with mixed outliers 61

3.10 Auto-GLOSH vs KNN and LOF: Precision@n obtained on Type 3 Datasets 62

3.11 Auto-GLOSH vs KNN and LOF: Precision@n obtained on Type 2 Datasets with Outlier clumps 63

3.12 Auto-GLOSH vs KNN and LOF: Precision@n obtained on Type 2 Datasets with Global outliers 64

3.13 Auto-GLOSH vs KNN and LOF: Precision@n obtained on Type 2 Datasets with local outliers 65
3.14 Auto-GLOSH vs KNN and LOF: Precision@n obtained on Type 2 Datasets with mixed outliers 66

4.1 Evaluating POLAR on Type 2 datasets with outlier clumps at $min_{pts} = m^*$: We compare the performance achieved by POLAR with the best performance that can be achieved across all possible thresholds. . . . 76
4.2 Evaluating POLAR on Type 2 datasets with global outliers at $min_{pts} = m^*$: We compare the performance achieved by POLAR with the best performance that can be achieved across all possible thresholds. . . . 77
4.3 Evaluating POLAR on Type 2 datasets with local outliers at $min_{pts} = m^*$: We compare the performance achieved by POLAR with the best performance that can be achieved across all possible thresholds. . . . 78
4.4 Evaluating POLAR on Type 2 datasets with mixed outliers at $min_{pts} = m^*$: We compare the performance achieved by POLAR with the best performance that can be achieved across all possible thresholds. . . . 79
4.5 Evaluating POLAR on Type 1 datasets at $min_{pts} = m^*$: We compare the performance achieved by POLAR with the best performance that can be achieved across all possible thresholds. 80

List of Figures

2.1 A Dendrogram representing the HDBSCAN* hierarchy obtained on a small dataset with eleven datapoints, with $min_{pts} = 3$. The pink arrows here represents datapoints as noise until they become *core points* and forms a cluster. 9

3.1 Inliers: GLOSH scores and Core distances at different min_{pts} values . 16

3.2 Global Outliers: GLOSH scores and Core distances at different min_{pts} values 17

3.3 Outlier Clumps: GLOSH scores and Core distances at different min_{pts} values 18

3.4 Multiple Clusters: GLOSH scores and Core distances at different min_{pts} values 19

3.5 Local Outliers: GLOSH scores and Core distances at different min_{pts} values 20

3.6 Synthetic Datasets with different kinds of points: (a) two *Global* outliers, (b) $P\Gamma_{m_{max}}$ of the *Global* outliers, (c) two *Local* outliers, (d) $P\Gamma_{m_{max}}$ of the *Local* outliers, (e) two Outlier Clumps, and (f) $P\Gamma_{m_{max}}$ of all the points in each of the clumps. 24

3.7 $P\Gamma_{m_{max}}$ of different clusters: (a) A synthetic dataset with three clusters (different sizes) C_0 (91), C_2 (344), and $C3$ (425), (b) $P\Gamma_{m_{max}}$ obtained for a data in cluster C_0, (c) $P\Gamma_{m_{max}}$ obtained for a data in cluster C_2, and (d) $P\Gamma_{m_{max}}$ obtained for a data in cluster C_3. 25

3.8 GLOSH–Profiles of Type 2 Datasets with Global Outliers 33

3.9 GLOSH–Profiles of Type 2 Datasets with Outlier Clumps 34

3.10 GLOSH–Profiles of Type 2 datasets with local outliers 36

3.11 GLOSH–Profiles of Type 1 datasets 37

3.12 Type 3 Datasets: Synthetic 2D Datasets with different kinds of outliers 40

3.13 Precision@n obtained by GLOSH at different min_{pts} on Type 3 Datasets 41

3.14 Highlighting the min_{pts} value at which GLOSH achieves the highest Precision@n with Global Outliers: The red GLOSH–Profiles represent the outlier profiles and the green line denotes the specific min_{pts} value where GLOSH first achieves the best Precision@n (P@n) within the range $[2, 100]$ of min_{pts} values. 42

3.15 Highlighting the min_{pts} value at which GLOSH achieves the highest Precision@n with Outlier Clumps: The red GLOSH–Profiles represent the outlier profiles and the green line denotes the specific min_{pts} value where GLOSH first achieves the best Precision@n (P@n) within the range $[2, 100]$ of min_{pts} values. 43

3.16 Highlighting the min_{pts} value at which GLOSH achieves the best Precision@n with Local Outliers: The red GLOSH–Profiles represent the outlier profiles and the green line denotes the specific min_{pts} value where GLOSH first achieves the best Precision@n (P@n) within the range $[2, 100]$ of min_{pts} values. 44

3.17 Type 3 with Global Outliers: Comparing the ORD–Profiles ($R_{m_{max}}$) with precision@n (P@n) achieved by GLOSH across different min_{pts} values. Highlighting the specific min_{pts} value where GLOSH first records the best precision@n within a range $[2, 100]$ of min_{pts} values. . 49

3.18 Type 3 Datasets with Outlier Clumps: Comparing the ORD–Profiles ($R_{m_{max}}$) with precision@n (P@n) achieved by GLOSH across different min_{pts} values. Highlighting the specific min_{pts} value where GLOSH first records the best precision@n within a range $[2, 100]$ of min_{pts} values. 50

3.19 Type 3 Datasets with Local Outliers: Comparing the ORD–Profiles ($R_{m_{max}}$) with precision@n (P@n) achieved by GLOSH across different min_{pts} values. Highlighting the specific min_{pts} value where GLOSH first records the best precision@n within a range $[2, 100]$ of min_{pts} values. 51

3.20 LETTER and 20news_3 datasets with synthetic Outlier Clumps: Comparing the ORD–Profiles ($R_{m_{max}}$) with Precision@n achieved by GLOSH across different min_{pts} values. Highlighting the specific min_{pts} value where GLOSH first records the best Precision@n within a range $[2, 100]$ of min_{pts} values. 52

3.21 Illustrating the process of finding the Elbow of the Outlier Rank Dissimilarity–Profile 54

4.1 Sorted Sequence of GLOSH scores at $min_{pts} = m^*$ 69

71figure.caption.46

4.3 Illustration of the adjusted threshold estimation on Anisotropic dataset with Outlier Clumps at $min_{pts} = m^*$: Using a linear regression on the sorted sequence of GLOSH scores to estimate the highest GLOSH score R. Then, we find the GLOSH score I that is closest to R. 72

4.4 F-Measure obtained on Type 3 Datasets across different thresholds at $min_{pts} = m^*$: The red dashed line signifies the F-measure obtained when using the knee GLOSH score as the threshold, while the green dotted line signifies the F-measure obtained when using the adjusted threshold. 74

4.5 G-Mean obtained on Type 3 Datasets across different thresholds at $min_{pts} = m^*$: The red dashed line signifies the G-Mean obtained when using the knee GLOSH score as the threshold, while the green dotted line signifies the G-Mean obtained when using the adjusted threshold. 75

List of Symbols

D	An unlabelled dataset
D^m	m-th feature of a dataset D
$P\Gamma$	General representation of GLOSH–Profile
$P\Gamma_{m_{max}}$	GLOSH–Profile in a range of $[2, m_{max}]$ of min_{pts} values
$R_{m_{max}}$	Outlier Rank Dissimilarity-Profile
$S_{min_{pts}}$	Sorted Sequence of GLOSH scores at a particular min_{pts} value
$\Delta(.,.)$	Pearson Dissimilarity between two sequences
$\Gamma_{min_{pts}}$	GLOSH score at a particular min_{pts} value
ϵ_c	Core Distance
$\lambda(x_i)$	Density of a datapoint x_i
λ_{max}	Maximum density of a cluster
d_{mrd}	Mutual Reachability Distance
m^*	"Best" min_{pts} value estimated using Auto-GLOSH
$r^{(k)}$	Dissimilarity at index k of Outlier Rank Dissimilarity-Profile $R_{NP\Gamma}$

Abbreviations

Auto-GLOSH Automatic GLOSH parameter selection based on outlier profiles.

GLOSH Global-Local Outlier Scores based on Hierarchies.

GMM Gaussian Mixture Model.

KNN K-Nearest Neighbor.

LOF Local Outlier Factor.

LRD Local Reachability Density.

MST Minimum Spanning Tree.

OCC One-Class Classification.

ORD-Profile Outlier Rank Dissimilarity-Profile.

P@n Precision@n.

POLAR Potential Outlier Labelling Approach.

TNR True Negative Rate.

Chapter 1

Introduction

Outliers are rare examples in a data-space that significantly deviate from the rest of the data (the inliers). The most cited definition of an outlier as proposed by Hawkins [1] is stated as "an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". The presence of outliers is a common occurrence in real-world datasets. In many cases, these instances do not contribute to new knowledge and can result in performance degradation of machine learning models. In some other cases, they can exhibit unique behaviors that can eventually give new insights in areas such as intrusion detection [2], machine failure detection [3], twitter bot detection [4], and forest fire detection [5]. Hence, detecting outliers is an important task in machine learning and data mining.

The task of outlier detection is to capture these deviating instances using supervised, unsupervised, or one-class classification (OCC) strategies. In a supervised setting, there are enough labeled examples of outliers and inliers (normal) examples to train a binary classifier (typically with an imbalance between the number of outliers and the number of normal examples) [6]. In an OCC setting, we have the information about inliers and little or no information about outliers [7], and a model is learned solely based on information about inliers. In an unsupervised setting, there is no prior knowledge about the data, and unsupervised methods identify outliers as instances that deviate from the rest of the data.

Global-Local Outlier Scores based on Hierarchies (GLOSH) is an unsupervised outlier detection method which is a part of the HDBSCAN* clustering framework [8]. It can detect the datapoints that deviate from their local neighborhood (*local* outliers) and also the datapoints that differ more globally from the rest of the data (*global outliers*). The GLOSH score of a datapoint p is computed as the normalized difference between the density estimated around p and the highest density estimated in the cluster closest to p in the HDBSCAN* hierarchy. The hierarchy represents clusters formed at different density levels and it is constructed w.r.t. a parameter value min_{pts}, a smoothing factor that is set by a user and determines the density estimates and thereby, the GLOSH score. However, different min_{pts} values can yield different HDBSCAN* hierarchies, and one of them may better represent the intrinsic cluster structure of the data than others. In practice, the underlying structure of the data is unknown and a unsuitable value of min_{pts} can result in an outlier getting a low GLOSH score. It has also been shown that one might need to use multiple values of min_{pts} to reveal all the clusters in a dataset [9]. This means that the closest clusters of a datapoint could be different for different values of min_{pts}, and hence different outlier detection results are possible for different values of min_{pts}. In practice, as the underlying data distribution is unknown, it is improbable to choose an appropriate min_{pts} value that correctly assigns most, if not all, "true outliers" higher GLOSH scores than the "true inliers".

Moreover, to find the outliers using GLOSH, it is essential to pre-define a value n, which is used to find the n datapoints with the highest GLOSH scores. These n datapoints are treated as "potential outliers" and are subsequently presented to the domain experts for labelling the "true outliers" [10]. In practice, one may not have any knowledge about how many outliers are present in a dataset. This makes it difficult to pick a suitable value for n beforehand.

1.1 Aim and Scope

In this thesis, we propose an approach that does not require to pick a single min_{pts} value in GLOSH, but uses GLOSH scores obtained at a range of different min_{pts} values to find the min_{pts} value that can yield often the best results for GLOSH in the given range of min_{pts} values. We also propose an automated technique to find a threshold for GLOSH scores that distinguishes inliers from potential outliers, using the distribution of GLOSH scores. Our approaches can be applied in a fully unsupervised way and can be computed efficiently by running the HDBSCAN* for a range of min_{pts} values as shown in [11]. This thesis primarily addresses two questions:

- Given a dataset, how to select a min_{pts} value that can yield best or nearly the best results with GLOSH, by using the GLOSH scores obtained for a range of different min_{pts} values?
- Given a min_{pts} value that assigns most, if not all, "true outliers" higher GLOSH scores than the "true inliers", how to select a threshold to classify inliers and potential outliers?

Our key contributions can be summarized as follows:

1. We introduce the notion of a GLOSH–Profile of a datapoint p as a sequence of GLOSH scores for p with respect to a range of min_{pts} values, and discuss its properties.
2. We show empirically that the GLOSH–Profiles behave differently for different kinds of outliers.
3. We show empirically that the GLOSH scores in the GLOSH–Profiles of each datapoint show a pattern that allows us to identify the min_{pts} value that results in the GLOSH scores best distinguishes between inliers and outliers in a dataset.

4. We introduce Automatic GLOSH parameter selection using Outlier Profiles (Auto-GLOSH), an unsupervised strategy that uses the GLOSH–Profiles to find a min_{pts} value that can often result in best results with GLOSH on a dataset.

5. We introduce Potential Outlier Labelling AppRoach (POLAR), an unsupervised strategy that uses the distribution of the GLOSH scores obtained using the min_{pts} value estimated by Auto-GLOSH, to find a GLOSH score that can serve as a threshold for labelling inliers and potential outliers in a dataset.

1.2 Thesis Outline

The thesis is organized as follows:

In Chapter 2 we provide necessary background on Unsupervised Outlier Detection, Hierarchical DBSCAN* (HDBSCAN*), Global-Local Outlier Scores based on Hierarchies (GLOSH), and Unsupervised Selection of a min_{pts} value.

Chapter 3 addresses the problem of selecting a min_{pts} value for GLOSH and has three parts. Firstly, we propose GLOSH–Profiles as a tool to study the behavior of GLOSH scores across different min_{pts} values. Secondly, we systematically establish the properties of GLOSH–Profiles for different kinds of datapoints, show the effectiveness of using shorter profiles over full profiles, and establish the link between the uniform rate of change in GLOSH scores within the GLOSH–Profiles and the best min_{pts} value for GLOSH results. Thirdly, we design an unsupervised method, Auto-GLOSH, to find the min_{pts} value that can potentially yield the best results for GLOSH. We evaluate our method on a series of datasets and compare it with GLOSH and other state-of-the-art outlier detection methods such as KNN and LOF.

In Chapter 4 we address the problem of choosing a threshold for labelling inliers and potential outliers. Firstly, we investigate the distribution of the GLOSH scores at the best min_{pts} value. Secondly, we design an unsupervised strategy to find a threshold to label inliers and potential outliers in a dataset. We evaluate our approach on fully synthetic datasets, real datasets with added synthetically generated outliers, and real one-class classification datasets, and compare it with the best achievable result.

Finally, in Chapter 5, we conclude this thesis by summarizing the entire study and providing possible future directions of research.

Chapter 2

Background and Related Work

2.1 Unsupervised Outlier Detection

In Unsupervised Outlier Detection, we are provided with data without access to ground truth labels and the goal is to segregate outliers from inliers. The primary idea behind unsupervised outlier detection is based on the notion that outlier instances will deviate significantly in terms of their *local* or *global neighbourhood* from that of inliers. The neighbourhood is determined based on a reference set of objects. A *global neighbourhood* is used when the reference set contains all or most of the data in the dataset and a *local neighbourhood* of an instance is a reference set constructed using its neighbours based on a distance measure [12]. In a density based setting the primary assumption is that outlier instances will mostly belong to regions in a data-space with comparatively lower density. Whereas, inliers are assumed too be lying in the higher density regions. In an unsupervised scenario, the density of each instance from a given set of unlabelled examples are computed to measure its outlier score.

Definition 2.1 (Outlier score). An *outlier score* is a numeric measure that quantifies the degree to which a datapoint deviates from the majority of datapoints in a collection mostly composed of so-called inliers, based on certain assumptions about the nature of the distribution of the inliers.

One of the earliest works on unsupervised outlier detection in the field of data mining is the work by Knorr and Ng [13] where the authors define outliers as instances that lie beyond a pre-determined distance from a large proportion of data. In the following years, unsupervised algorithms were developed that assigned an outlier score to each instance based on their *k-Nearest Neighbor* distance (kNN distance) [14]. The kNN distance kNN$-$dist(.) of a datapoint is treated as the outlier score, which is computed as the distance between the datapoint and its k^{th} nearest neighbor. The datapoints that have significantly larger k^{th} nearest neighbor distance are treated as potential outliers. Outliers determined in this way can be considered ”global” outliers since only their distance from ”the rest” of the data, as measure by the distance to their k-nearest neighbors, is used in the computation of the score. Another family of algorithms originated from the work that presented the *Local Outlier Factor* (LOF) [15], where outliers are segregated based on their local neighborhood. To compute LOF for any point x_i, one has to compare the local reachability density $LRD_k(.)$ of x_i with the $LRD_k(.)$ of each of x_i's k nearest neighbours. The local reachability density of x_i if defined as:

$$LRD_k(x_i) = \frac{1}{\frac{\sum_{o \in k\mathrm{NN-set}(x_i)} R{-}dist_k(x_i,o)}{|k\mathrm{NN-set}(x_i)|}} \tag{2.1}$$

where, $R{-}dist_k(x_i, o)$ measures the reachability distance between x_i and o as $max(k\mathrm{NN{-}dist}(o), d(x_i,$ with $d(x_i, o)$ being the distance between x_i and o. The LOF of x_i is then defined as:

$$\mathrm{LOF}(x_i) = \frac{\sum_{o \in k\mathrm{NN-set}(x_i)} \frac{LRD_k(o)}{LRD_k(x_i)}}{k\mathrm{NN{-}set}(x_i)} \tag{2.2}$$

Outlier detection methods are often characterized into *local* and *global* methods in the literature [16].

2.2 HDBSCAN*

HDBSCAN* or *Hierarchical DBSCAN** [8] is a hierarchical way of clustering data and is based on DBSCAN*, which is an improvement of one of the most popular density-based clustering algorithms, DBSCAN [17]. The density hierarchy is formed following Hartigan's model of density trees providing a complete hierarchy of all possible clustering solutions that can be obtained at an infinite range of different density thresholds or radii. In DBSCAN*, density-based clusters are constructed using two parameters: (I) a radius ϵ, and (II) a minimum number of points min_{pts}. Any point x_i in a dataset D is a *core point* if it has min_{pts} many datapoints in its ϵ*–neighborhood* $N_\epsilon(x_i)$ (i.e. in the set of points that are located less than or equal to distance ϵ away from x_i); otherwise, a point is called a noise point. Two *core points* are *density connected* w.r.t. ϵ and min_{pts} if they directly or transitively belong to each other's N_ϵ. DBSCAN* defines a cluster as a maximal set of points where each pair of points in the set are *density connected.*

HDBSCAN* improves on DBSCAN* and does not require the user to pre-define an ϵ radius. For a given min_{pts} value, it can provide DBSCAN*'s solutions at all possible ϵ values. For each $x_i \in D$, HDBSCAN* computes the *core distance* $\epsilon_c(x_i)$, which is the minimum ϵ distance required for x_i to be a *core point*, and the *mutual reachability distance* $d_{mrd}(.,.)$, which is the minimum distance required for x_i and x_j to be in each other's ϵ-neighborhood while both are also *core points.* The $d_{mrd}(.,.)$ between two points x_i and x_j is defined as:

$$d_{mrd}(x_i, x_j) = max\{\epsilon_c(x_i), \epsilon_c(x_j), d(x_i, x_j)\} \tag{2.3}$$

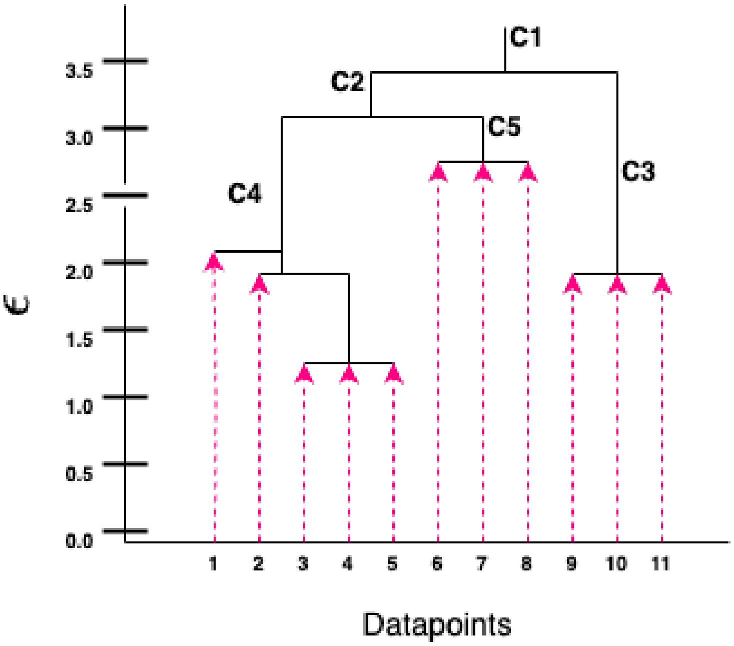


Figure 2.1: A Dendrogram representing the HDBSCAN* hierarchy obtained on a small dataset with eleven datapoints, with $min_{pts} = 3$. The pink arrows here represents datapoints as noise until they become *core points* and forms a cluster.

$d(.,.)$ represents the distance between two points in the dataset. The mutual reachability distance d_{mrd} is dependent on the value of min_{pts}. The HDBSCAN* hierarchy w.r.t. a single min_{pts} value is obtained by computing the *minimum spanning tree* $MST_{min_{pts}}$ of a complete, edge-weighted, virtual graph $G_{min_{pts}}$ [11], where the vertices represent the datapoints in the dataset, and the edge weights are the values of the *mutual reachability distance* d_{mrd} between them. The edges from the MST are then removed in decreasing order of edge weights and at every step of the removal, we obtain a set of connected core objects (which are the clusters) and the remaining noise. This creates a hierarchy of all possible DBSCAN* clustering solutions obtained at $\epsilon \in [0, \infty]$ for a particular min_{pts} value. One can represent the HDBSCAN* hierarchy using a dendrogram, as presented in Figure 2.1. A clustering dendrogram is a diagram of a tree in which the root represents the whole dataset as a single cluster, and internal nodes represent clusters (C1 to C5) that are the result of splitting their parent cluster.

2.3 GLOSH

The HDBSCAN* clustering framework also includes the *Global-Local Outlier Scores based on Hierarchies* (GLOSH) algorithm to detect outliers using the hierarchical density estimates. GLOSH uses the cluster structure present in the hierarchy of HDBSCAN*. The primary idea of GLOSH is to compute a score (called GLOSH score) for each datapoint in a dataset by comparing the density estimate for the datapoint with the density estimate of the most dense point of its nearest cluster in the hierarchy. The reference set of points for any datapoint (found in the closest cluster) is dynamically selected, therefore, allowing GLOSH to find outliers across both local and global scales, depending on the specific hierarchy structure and the datapoint's placement within the clusters in the hierarchy.

To compute the GLOSH score for a point $x_i \in D$, the information of the cluster C_{x_i} closest to x_i, in the hierarchy is used. In the hierarchy, C_{x_i} is the cluster where the datapoint x_i gets assigned to (when ϵ is large enough to make x_i a core point) following a bottom-up approach. The density of x_i, $\lambda(x_i)$, is compared to that of the densest point in C_{x_i}, $\lambda_{max}(C_{x_i})$, which is the point that is assigned to the C_{x_i} when it is first formed in the hierarchy (making it the longest surviving point in C_{x_i}). The GLOSH score $\Gamma_{min_{pts}}(x_i)$, of a point x_i for a particular min_{pts} value is defined as:

$$\Gamma_{min_{pts}}(x_i) = \frac{\lambda_{max}(C_{x_i}) - \lambda(x_i)}{\lambda_{max}(C_{x_i})} \tag{2.4}$$

where, λ for any point x_i is defined as $\frac{1}{\epsilon_c(x_i)}$. The density of the densest point in C_{x_i}, represented as $\lambda_{max}(C_{x_i})$, is used here as the referential density to compare the density of any point x_i that has C_{x_i} as its closest cluster in the hierarchy. The densest point in cluster C_{x_i}, is considered here as the most inlier point in C_{x_i} and its density serves as the standard for all the points that are closest to C_{x_i}. The GLOSH score falls into the range $[0, 1)$. If x_i resides in a dense region of a cluster, then $\lambda_{max}(C_{x_i}) - \lambda(x_i)$ tends to be 0 whereas, for points that are further away from the cluster, $\lambda_{max}(C_{x_i}) - \lambda(x_i)$ tends to become larger, resulting in higher GLOSH scores.

HDBSCAN* and GLOSH can be used as an one-class learner (OCL) by first obtaining the hierarchy using the available training data w.r.t. a single min_{pts} value. Then, the GLOSH scores of unknown instances are computed using the hierarchy as a fixed model [18]. An unknown instance u_i is first added to the pre-computed MST underlying the hierarchy by connecting it to the training datapoint x_i that has the smallest d_{mrd} to u_i. Then the closest cluster of x_i becomes the closest cluster of u_i and $\Gamma_{min_{pts}}(u_i)$ is computed as in equation 2.4. To classify the instance, i.e., assign a label, it is possible to apply a predetermined threshold to the GLOSH outputs. Points that receive a GLOSH score falling below the threshold are classified as 'inliers', while the points receiving a score higher than the threshold are classified as 'outliers' [19].

2.4 Unsupervised Selection of a min_{pts} value

To our knowledge, there is no existing work on automated selection of a best value for min_{pts}, in GLOSH. Finding the best min_{pts} value for GLOSH is challenging because in an unsupervised setting, we do not have prior knowledge about the underlying distribution of the data. In the absence of prior knowledge, it is impossible to assess the outlier detection results at different min_{pts} values and determine which value is best. The closest related work focuses on outlier detection model selection through meta learning [20]. Here, the authors proposed a method to select an outlier detection model for a new dataset based on its past performance on similar datasets. This is different from our setting. Firstly, we focus on a single outlier detection model, GLOSH. Secondly, we operate in a scenario where there is no prior information about the model or the dataset, precluding any history regarding the performance of GLOSH on similar datasets.

Chapter 3

Using GLOSH–Profiles for min_{pts} Selection

3.1 GLOSH–Profile: Outlierness across min_{pts} values

GLOSH scores w.r.t. a single min_{pts} value ($\Gamma_{min_{pts}}$) for any datapoint x_i in dataset D compare the density of x_i with that of the densest point in the cluster structure it density-connects to within the HDBSCAN* hierarchy after it becomes a *core point*. The density of x_i is computed as the inverse of its core distance $\frac{1}{\epsilon_c(x_i)}$. The impact of changing the min_{pts} values on cluster formation has been estimated in prior studies [11, 21]. An increase in min_{pts} values requires each point to have more points in its ϵ*–neighborhood* to become a *core point*, resulting in larger core distance $\epsilon_c(.)$ values for each point (forming its ϵ_c*–neighborhood* with points that are further away). This, in turn, leads to the formation of larger clusters in the hierarchy. Conversely, decreasing min_{pts} values allows points to become *core points* with smaller $\epsilon_c(.)$ values, potentially forming smaller clusters within the hierarchy. Therefore, $\Gamma_{min_{pts}}(x_i)$ at different min_{pts} values may compare its density with that of clusters of different sizes (i.e. neighborhoods of different sizes) formed at those min_{pts} values, influencing the value of $\Gamma_{min_{pts}}(x_i)$.

For a given value of min_{pts}, an outlier ranking can be constructed using the GLOSH scores $\Gamma_{min_{pts}}$ of each datapoint x_i in a dataset D by arranging the GLOSH scores

$\Gamma_{min_{pts}}$ of each datapoint x_i, either in increasing or decreasing order. A popular approach is giving higher ranks to the datapoints getting higher GLOSH scores (i.e. arranging the GLOSH scores in decreasing order). For example, the datapoint in D with the highest GLOSH score gets rank 1, and the point with the lowest score gets rank $|D|$. To apply GLOSH effectively, setting an appropriate value for min_{pts} to construct the density hierarchy using HDBSCAN* is crucial. An appropriate min_{pts} value is expected to assign higher GLOSH scores to the outliers than the inliers. However, in practice, choosing a single min_{pts} value where most, if not all, "true outliers" receive higher GLOSH scores than the "true inliers" is improbable as we do not have prior knowledge about the underlying data distribution.

To address the problem of finding a single value for min_{pts} for GLOSH we propose GLOSH–Profiles $P\Gamma_{m_{max}}$ to capture the behavior of GLOSH scores for each datapoint across different HDBSCAN* hierarchies obtained over a range of min_{pts} values.

Definition 3.1 (GLOSH–Profile). A *GLOSH–Profile* $P\Gamma_{m_{max}}(x_i)$, is an array of $\Gamma_{min_{pts}}(x_i)$ values for a given point x_i for a range $[2, m_{max}]$ of min_{pts} values between 2 and a maximum value m_{max}:

$$P\Gamma_{m_{max}}(x_i) = \begin{bmatrix} \Gamma_2(x_i) \\ \Gamma_3(x_i) \\ \vdots \\ \Gamma_{m_{max}}(x_i) \end{bmatrix} \tag{3.1}$$

When m_{max} is clear from the context or not important for the discussion, we also abbreviate $P\Gamma_{m_{max}}(.)$ as $P\Gamma(.)$. We use GLOSH–Profiles as a tool to investigate the behavior of the GLOSH scores across different min_{pts} values. The investigation will eventually guide us in selecting a best min_{pts} value using the GLOSH–Profiles.

Three different kinds of outliers have been popularly explored in the literature [22, 23]: local, global, and small groups of points (sometimes called "outlier clumps" [22]) having similar behavior but being dissimilar from the rest of data. These types of outliers differ from each other in terms of their spatial characteristics. *Global outliers* are datapoints that are far away from the rest of the points in the data-space, whereas *local outliers* are datapoints deviating from their local neighborhood but are not necessarily far away from the rest of the data. Outliers in a clump, on the other hand, are close to one another, while the whole group deviates from the majority of the data. As each of the different kinds of outliers has different spatial characteristics, we hypothesize that the behavior of their GLOSH–Profiles will differ.

The GLOSH scores forming the profile are derived from different HDBSCAN* hierarchies w.r.t. different min_{pts} values. In a single run of HDBSCAN* at $min_{pts} = m_{max}$, one can access the $MST_{min_{pts}}$ and the m_{max}-*Nearest Neighbor Graph* to construct the *CORE-distance based Spanning Graph*, *CORE-SG*$_{m_{max}}$ [11]. Replacing the complete graph in HDBSCAN* with *CORE-SG*$_{m_{max}}$ allows the extraction of all $MST_{min_{pts}}$ independently, for $min_{pts} \in [1, m_{max} - 1]$. This allows to extract all possible HDBSCAN* hierarchies and, consequently, the GLOSH scores with the same asymptotic computational complexity as running HDBSCAN* once w.r.t. a single min_{pts} value. In practice, it has been shown in [11] that efficient extraction of the HDBSCAN* hierarchies in the range $[1, 100]$ is possible in a run-time that is close to running HDBSCAN* twice.

As the *core distances* ϵ_c for each point increases as we increase the min_{pts} value, we are able to obtain different density estimates (computed as $\frac{1}{\epsilon_c(.)}$) of the points and thereby, different GLOSH scores $\Gamma_{min_{pts}}$. The intuition behind the GLOSH–Profiles $P\Gamma$ is to capture the outlierness (based on GLOSH scores) of a point w.r.t. different ϵ_c–*neighborhoods*. If for any $x_i \in D$, its GLOSH–Profile contains consistently low GLOSH scores across most min_{pts} values, it indicates that x_i is a relatively dense

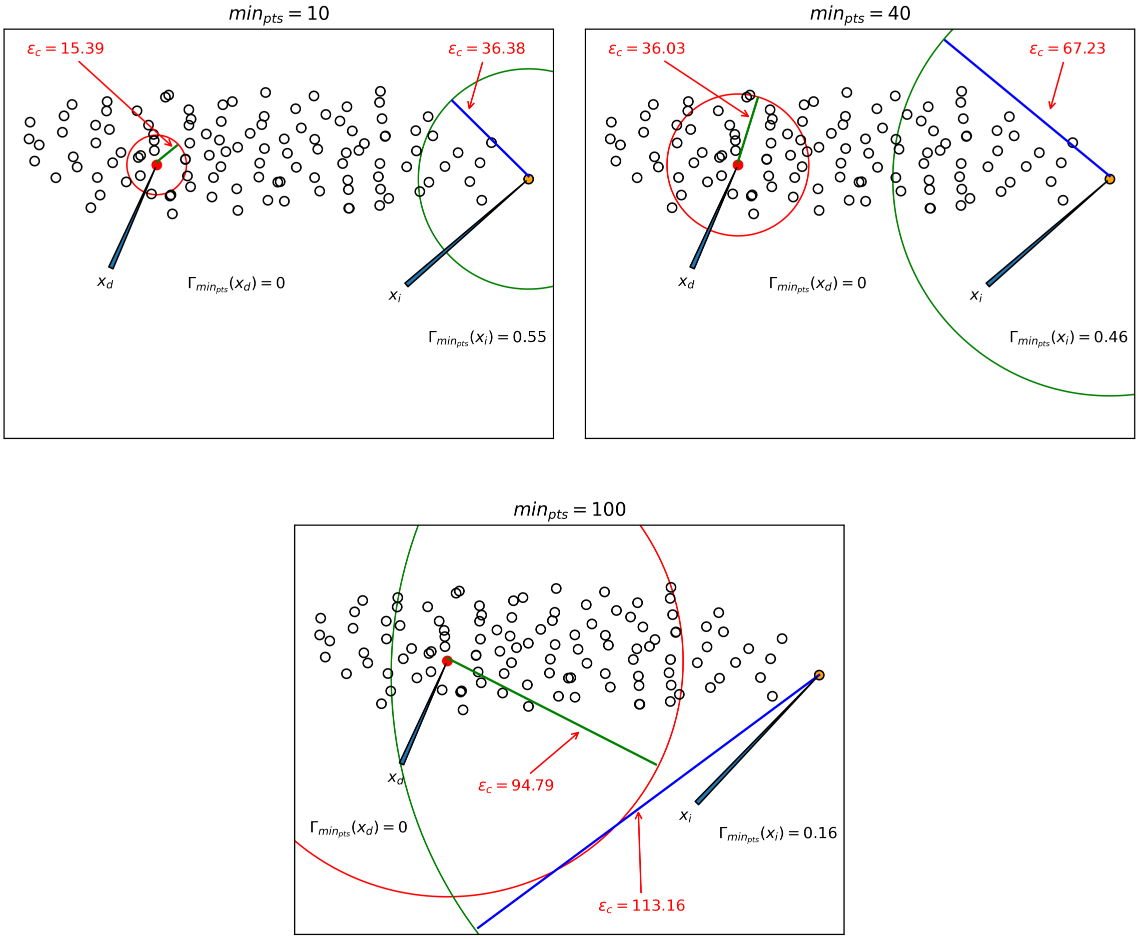


Figure 3.1: Inliers: GLOSH scores and Core distances at different min_{pts} values

point when compared to neighborhoods of points with different sizes. Therefore, we call these points *global inliers*. We use Figure 3.2 to illustrate *global inliers*. In Figure 3.2 the min_{pts} value increases from 10 to 100. The *core distance* of the inlier point x_i (a point residing in a low density region the cluster), represented as $\epsilon_c(x_i)$ and the *core distance* of x_d (the densest point inside the cluster), represented as ϵ_c, both increases as the min_{pts} value increases. One can see that for lower min_{pts} value, x_i records a moderately high GLOSH score $\Gamma_{min_{pts}}$ of 0.55. In the figure, x_d can be seen as a *global inlier* as it records a low GLOSH score for all the values of min_{pts}. This shows that x_d stays a dense point when compared both large and smaller neighborhoods of points. As the min_{pts} value increases, the *core distances* of x_i becomes close to that

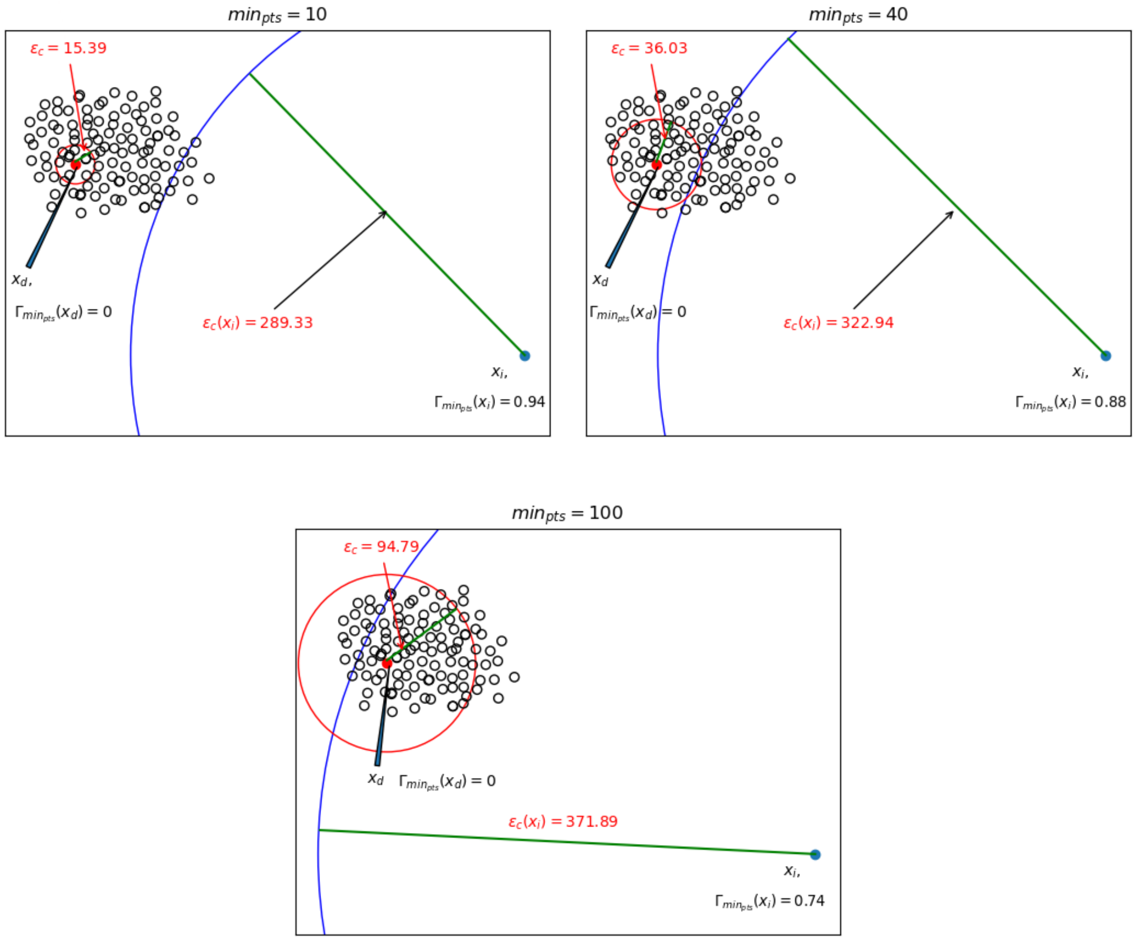


Figure 3.2: Global Outliers: GLOSH scores and Core distances at different min_{pts} values

of x_d. The figure also shows us that it is possible for certain inlier points to get a high GLOSH score for certain min_{pts} values.

As the min_{pts} value increases, smaller clusters tend to be smoothed and vanish in the hierarchy. Consequently, points tend to become *core* points at higher levels of the HDBSCAN* hierarchy, and become part of larger clusters that resides at these levels. If the GLOSH–Profile of any point contains high GLOSH scores across most min_{pts} values, this indicates that the density of x_i is low (high *core distance*) compared to the points in both large and small clusters. For instance, consider Figure 3.2 where the min_{pts} value increases from 10 to 100. The *core distance* of x_i (a point that is

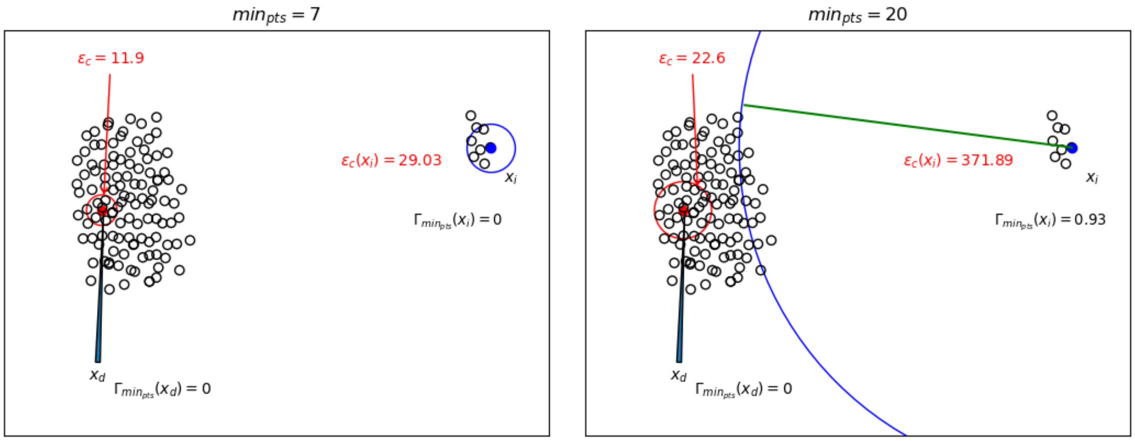


Figure 3.3: Outlier Clumps: GLOSH scores and Core distances at different min_{pts} values

very far away from the cluster), represented as $\epsilon_c(x_i)$ and the *core distance* of x_d (the densest point inside the cluster), represented as ϵ_c, both increases as the min_{pts} value increases. However, even if the min_{pts} value increases, the core distance of x_i still stays significantly larger than that of x_d, as x_d is closely surrounded by more points than x_i. Therefore, x_i keeps getting high GLOSH scores as the min_{pts} value increases. In Figure 3.2, x_i is not a dense point when compared both large and smaller neighborhoods of points, resulting in high GLOSH scores ($\Gamma_{min_{pts}}(x_i)$) for each of the min_{pts} values. We will refer to such points as *global outliers.*

In cases where a GLOSH–Profile starts with initially low GLOSH scores for lower min_{pts} values, but then has larger GLOSH scores for higher min_{pts} values, this means that the point is a dense point when compared to its close neighborhood points, but not a dense point when compared to larger neighborhoods in the dataset (i.e., formed by other points in the dataset) that are away from the point. This indicates that x_i is a part of a small group of points that is far away from a large section of points in the dataset. In Figure 3.3, one can see that for a lower min_{pts} value ($min_{pts} = 7$) x_i (a point in a smaller clump of points) has a low GLOSH score ($\Gamma_{min_{pts}}(x_i)$) since the number of points closely surrounding it is larger than 7 (resulting in a low core

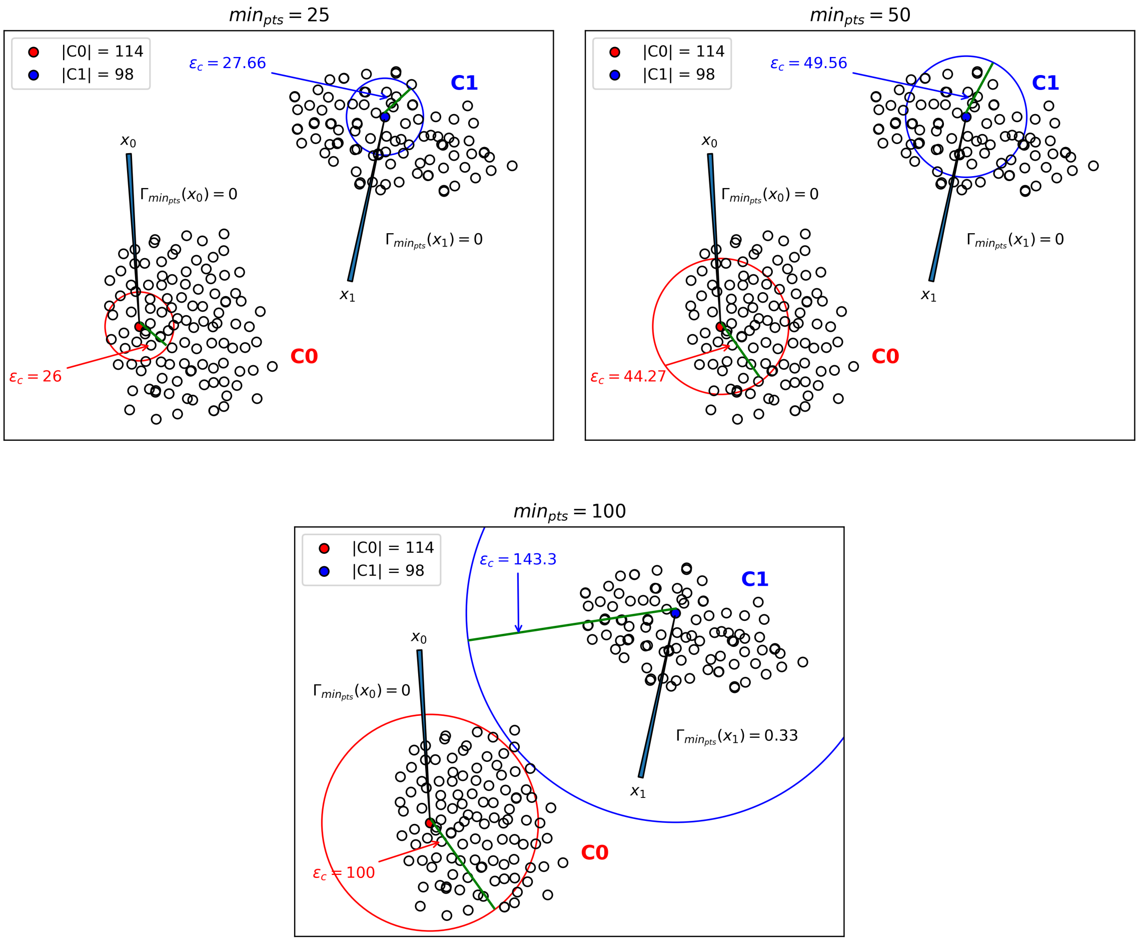


Figure 3.4: Multiple Clusters: GLOSH scores and Core distances at different min_{pts} values

distance, i.e., high density). However, for a larger value such as $min_{pts} = 20$, to become a *core point*, the radius around x_i has to include points from the larger cluster that is far away, resulting in a much larger *core distance* $\epsilon_c(x_i)$ (lower density) than that of the densest point x_d in the larger cluster. Therefore, $\Gamma_{min_{pts}}(x_i)$ (the GLOSH score of x_i) increases to 0.94. We call such points *outlier clumps*. Similar behavior in GLOSH scores may be observed for points in different clusters. If there are multiple clusters in the dataset, for the points in the smaller cluster, one can expect that the GLOSH scores will have a sharp increase, as we keep increasing the min_{pts} values. These sharp increase in the GLOSH scores can occur as the points in

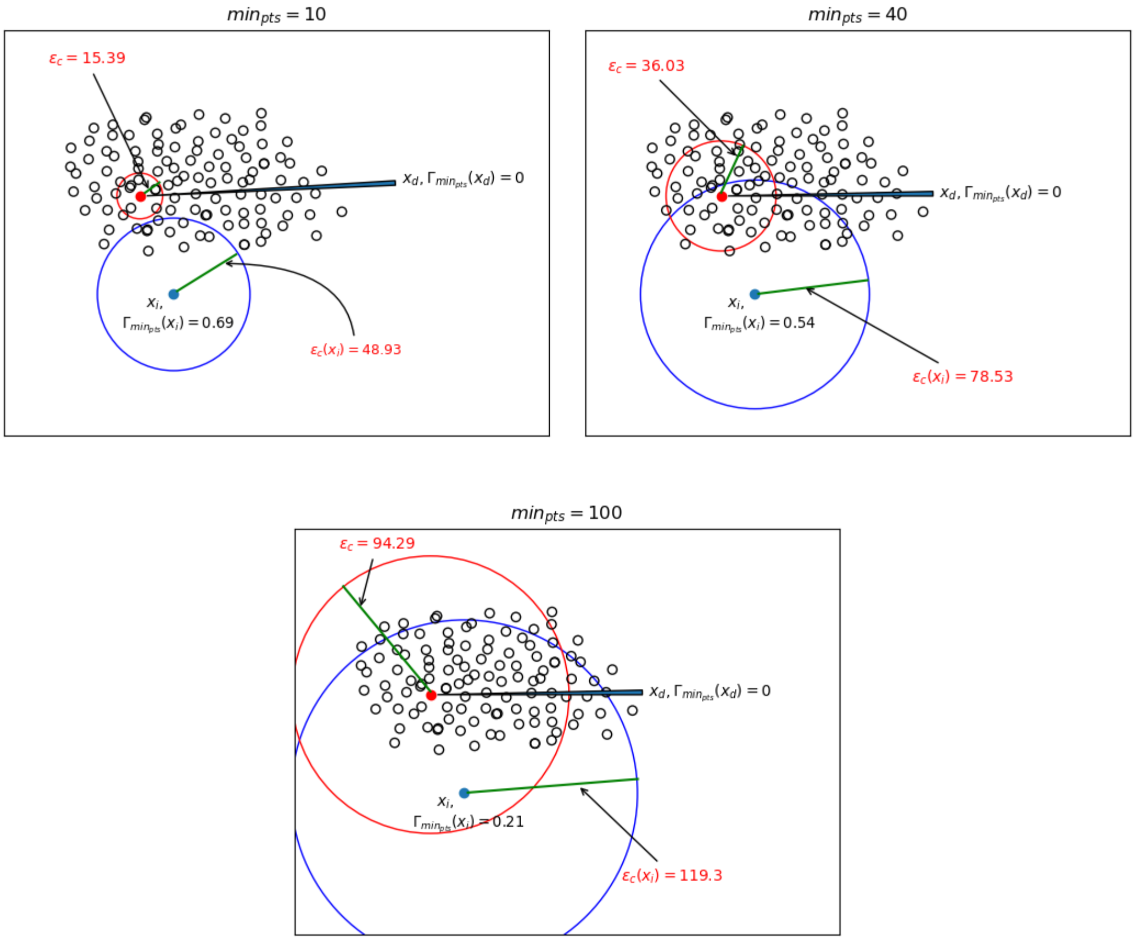


Figure 3.5: Local Outliers: GLOSH scores and Core distances at different min_{pts} values

the smaller cluster will have to include more points from other larger clusters that are further away, in order to become *core points*. For example, in Figure 3.4, one can see two different clusters C0 and C1, where C1 is the smaller cluster with 98 points. One can see that as the min_{pts} becomes greater than 98, the ϵ radius of the densest point in C1, i.e. x_1, has to include points from the larger cluster C0. This results in a increase in the GLOSH score of x_1.

In other cases, the GLOSH–Profile of a point may initially capture high GLOSH scores at lower min_{pts} values, followed by decreasing GLOSH scores as the min_{pts} value increases. This shows that when the point is compared to smaller neighborhoods, it looks like an outlier. However, when compared to larger neighborhoods (at larger min_{pts} values), the point looks more and more like an inlier. As the value for min_{pts} increases, the *core distances* of all the points lying inside or on the border of a cluster becomes more and more similar, and therefore, the differences between their densities tends to be reduced. For instance, in Figure 3.5, as we increase min_{pts} from a smaller value (10) to a larger value (100), the *core distances* of x_i and x_d (the densest point in the cluster) become more and more close to each other and the difference between the densities of x_d $(1/\epsilon_c)$ and x_i $(1/\epsilon_c(x_i))$ decreases. Therefore, as presented in the Figure 3.5, x_i records a high GLOSH score ($\Gamma_{min_{pts}(x_i)} = 0.69$) for a smaller min_{pts} value and $\Gamma_{min_{pts}}(x_i)$ keeps decreasing rapidly after that (shown for $min_{pts} = 40$ and $min_{pts} = 100$). We refer to these points as *local outliers*.

3.1.1 GLOSH–Profiles of different kinds of points

In Chapter 2 and section 3.1, we described the mechanism of GLOSH and the characteristics of GLOSH–Profiles ($P\Gamma$) on an intuitive basis. In this section we investigate the behavior of inlier and outlier GLOSH–Profiles in a simple experimental setting.

In summary, we ask two questions:

- **RQ1:** Can GLOSH–Profiles behave differently for different kinds of outliers?
- **RQ2:** Can GLOSH–Profiles behave differently for different inliers?

Following [23, 24], we concentrate on three kinds of outliers: *global*, *local*, and small groups of points with similar behaviour but dissimilar from the rest of the data (*outlier clumps*). As discussed in the previous section, we hypothesize that GLOSH–Profiles will exhibit different behaviour for different kinds of outliers. Although inlier points lie in the dense region of clusters, they may also exhibit high GLOSH scores in the profiles at certain min_{pts} values. For a cluster C, the profiles of each datapoint in C will show an elevation in GLOSH scores once the min_{pts} value becomes greater than $|C|$. When the min_{pts} value surpasses the size of the cluster, the core distances of each datapoint in the cluster, must include other points in the dataset that are further away from the cluster. Therefore, datapoints from the same cluster tend to show an elevation in GLOSH scores more or less at "the same time" i.e. in more or less the same range of min_{pts} values, and we hypothesize, that the GLOSH–Profiles of inliers from the same cluster will show a similar behavior.

We study the behaviour of GLOSH–Profiles $P\Gamma_{m_{max}}$ using simple 2-dimensional synthetic datasets. Synthetic datasets are preferred here to provide more control over the systematic evaluation of the GLOSH–Profiles. We use synthetic datasets containing clusters following a Gaussian distribution with the addition of three kinds of outliers points: (I) global outliers, (II) local outliers, and (III) outlier clumps [23]. The Gaussian clusters are generated in a randomized way (random sizes, and positions) as proposed by Prati et al. [25]. Then, we generate and add different kinds of outliers to the datasets.

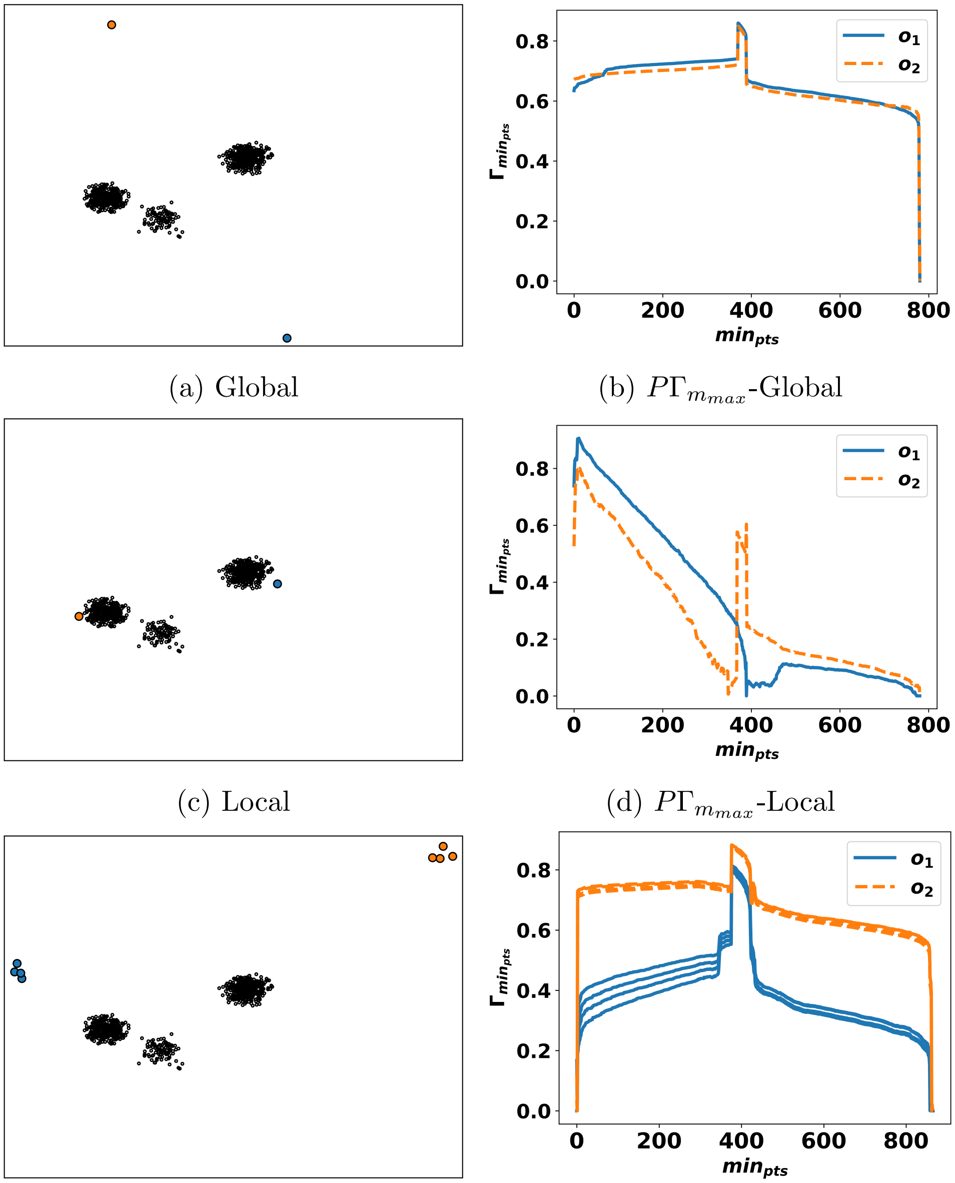
To generate *global outliers*, we follow [23] to generate instances that are far from the Gaussian clusters. The instances are generated from a uniform distribution with boundaries defined as $(\alpha \cdot max(D^m), \alpha \cdot min(D^m))$ where D^m represents the m-th feature of the generated Gaussian data D. However, this process of generating outliers has a drawback as there is a chance that a generated outlier could fall inside on of the clusters. To address this problem, we follow [26] and apply the *tomek links* technique as a data-cleaning method to filter out such outliers. The *tomek links* technique is based on looking at pairs of datapoints that are closest neighbors w.r.t. some distance metric, but are from opposite classes. In our setting we have two classes, the inlier class and the outlier class. Considering an instance x_i from the generated Gaussian clusters , i.e., an instance of the inlier class, and a generated outlier o, x_i and o can form a *tomek link*, w.r.t. a distance measure $d(.,.)$, if there is no other outlier instance x_k in the dataset that satisfies $d(x_i, x_k) < d(x_i, o)$ or $d(o, x_k) < d(x_i, o)$. The intuition behind a *tomek link* in this context is that if two instances form a *tomek link* then the generated outlier instance falls inside a cluster or is located close to the border of a cluster, i.e., the generated outlier is not really a global outlier. We thus remove generated outliers forming a *tomek link* from the set of global outliers.

To create *local outliers*, we randomly choose two different clusters from our generated Gaussian clusters. Then, we follow [23] and scale the covariance matrix Σ by a factor of 5 ($\hat{\Sigma} = \alpha\Sigma$ where, $\alpha = 5$) to generate the outliers using the generator that we initially used to generate the clusters. The covariance matrix controls the spread of the datapoints that are generated. We use the parameter $\alpha = 5$, as proposed in [23] to generate points that are nearby to the clusters. Opting for two different clusters ensures that the outliers are generated in the vicinity of different clusters, avoiding concentration in the same region.

To generate "outlier clumps" [23], we again choose two clusters from our generated Gaussian clusters. Then we scale the mean feature vector μ of each of the clusters by a factor of 5 ($\hat{\mu} = \alpha\mu$ where $\alpha = 5$) to generate the outlier samples using the generator. As the mean feature vector is shifted by α, that ensures that the outlier samples are generated in the same region but away from the cluster.

Fig. 3.6 shows three 2–D datasets (containing 3 Gaussian clusters and different kinds of outliers; it also shows the $P\Gamma_{m_{max}}$ plots for the outliers.

The first major observation suggests an answer to our first research question of this section. In case of *global outliers*, one can see that the GLOSH scores in profile $P\Gamma_{m_{max}}$ of a *global outlier* is overall higher across most min_{pts} values than that of a *local* outlier. This finding supports our characterization of *global outliers* in section 3.1. In case of *local outliers*, one can see that their $P\Gamma_{m_{max}}$ show high GLOSH scores for lower min_{pts} values, and the scores decrease as the min_{pts} value increases. Unlike the global outlier profiles, the GLOSH–Profiles of *local outliers* show a GLOSH score close to 0 for a much smaller value of min_{pts}. This finding supports our characterization of *local outliers* in section 3.1. In the case of small groups of outliers or "oulier clumps", the GLOSH scores in the profiles of the points start from a low value and



(a) Global

(b) $P\Gamma_{m_{max}}$-Global

(c) Local

(d) $P\Gamma_{m_{max}}$-Local

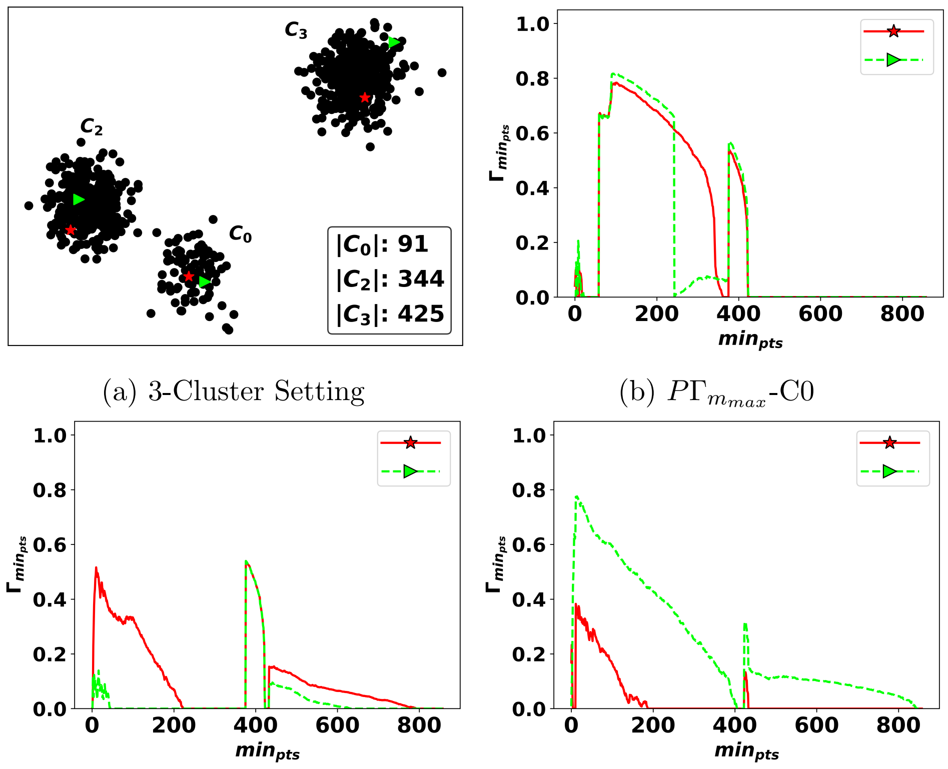
(e) Outlier Clumps

(f) $P\Gamma_{m_{max}}$-Outlier Clumps

Figure 3.6: Synthetic Datasets with different kinds of points: (a) two *Global* outliers, (b) $P\Gamma_{m_{max}}$ of the *Global* outliers, (c) two *Local* outliers, (d) $P\Gamma_{m_{max}}$ of the *Local* outliers, (e) two Outlier Clumps, and (f) $P\Gamma_{m_{max}}$ of all the points in each of the clumps.

increase significantly as we increase the min_{pts} value beyond the size of the clump, and the values stay overall high for most min_{pts} values. In a sense, the whole clump behaves similarly to global outliers for min_{pts} values greater than the size of the outlier clump. This finding supports our characterization of *outlier clumps* in section

3.1. The $P\Gamma_{m_{max}}$ profiles shown here for different kinds of outliers have distinct characteristics for different types of outliers, which suggests that the GLOSH–Profiles $P\Gamma_{m_{max}}$ may be used to characterize and detect diverse outlier patterns in a dataset.



(a) 3-Cluster Setting

(b) $P\Gamma_{m_{max}}$-C0

(c) $P\Gamma_{m_{max}}$-C2

(d) $P\Gamma_{m_{max}}$-C3

Figure 3.7: $P\Gamma_{m_{max}}$ of different clusters: (a) A synthetic dataset with three clusters (different sizes) C_0 (91), C_2 (344), and $C3$ (425), (b) $P\Gamma_{m_{max}}$ obtained for a data in cluster C_0, (c) $P\Gamma_{m_{max}}$ obtained for a data in cluster C_2, and (d) $P\Gamma_{m_{max}}$ obtained for a data in cluster C_3.

We also study the behavior of GLOSH profiles for inliers. In Figure 3.7, we show the GLOSH–Profile $P\Gamma_{m_{max}}$ plots of two representative inlier points (represented by a red star and a green triangle) from each of the clusters in the 2–D dataset.

Starting with cluster C_0, one can see that GLOSH–Profile plots of points in the cluster C_0 have initially a GLOSH score $\Gamma_{min_{pts}}$ close to 0. The GLOSH scores have a first spike at $min_{pts} > 91$ and a second spike at $min_{pts} > 344$. The first spike in the GLOSH scores happens when min_{pts} exceeds the size of the cluster C_0. The core distances of each point in C_0 increases at that point significantly to include points from cluster C_2 into their neighborhood to become a *core point*. The second spike in the GLOSH scores happens when the ϵ-neighborhoods of datapoints in C_0 need to include points from the cluster C_3 to become a *core point* (at min_{pts} value > 344, which is the combined size of clusters C_0 and C_2). After the first spike, one can see that the profile of the green point in cluster C_0 comes down to a GLOSH score close to 0 more quickly (for a smaller min_{pts} value) compared to the red point. This happens because the green point is located in a denser region within the cluster compared to the red point. From these observations, we can conclude that the GLOSH–Profiles of inlier points in cluster C_0 behave similarly.

In the cluster C_2 of Figure 3.7c, the GLOSH–Profiles $P\Gamma_{m_{max}}$ of both points are found to exhibit three spikes. One can see that GLOSH–Profiles of the points starts from a GLOSH score close to 0 and the first spike in the score occurs for a low min_{pts} value. The green point is situated in a denser region of the cluster compared to the red point and therefore, gets a relatively lower GLOSH score in the first spike. The second spike in the GLOSH scores occurs when the min_{pts} exceeds $|C_2|$, at $min_{pts} = 344$. This spike in the scores happens as the core distances of the points in the cluster C_2 increase together as they need more points (more than $|C_2|$) in their neighborhood to become a *core point*. Therefore, the core distance radius of points in cluster C_2 start including points from cluster C_0, which is closer to C_2 than C_3.

For the cluster C_3 in Figure 3.7, we again see that the GLOSH–Profiles $P\Gamma_{m_{max}}$ start from a low GLOSH score. The points in C_3 show a spike in their GLOSH scores for a lower min_{pts} value and then a second spike when the min_{pts} value exceeds the size of C_3. The second spike in the GLOSH scores occurs as the core distance radius of points in the cluster C_3 needs to include points from the clusters C_0 and C_2 to become a core point. Overall, one can see that the GLOSH–Profiles of inlier points from the same cluster are not identical but they share similarities in terms of where (i.e. at what min_{pts} value) certain spikes in the GLOSH scores occur. This answers our second research question. When comparing the GLOSH–Profiles for different clusters, one can see that the position (i.e. the min_{pts} value) of the spikes in the GLOSH–Profiles differ for different clusters w.r.t their sizes and positions in the data-space.

We summarize the findings of this section as follows:

- Local outliers tends to have a higher GLOSH score initially for low min_{pts} values but decrease rapidly as we increase the min_{pts} values. Global outliers seem to exhibit consistently higher GLOSH scores across most min_{pts} values compared to local outliers. Points in outlier clumps, seem to behave similarly to global outliers for min_{pts} values beyond the size of the outlier clump.
- The GLOSH–Profiles of inlier points from the same cluster tends to exhibit similar behavior.

3.1.2 Limiting the size of the GLOSH–Profiles

As we have shown in section 3.1 and 3.1.1, the behavior of the GLOSH scores in the profiles of outliers either become similar to that of inliers or do not change after a certain value of min_{pts}. Extracting the full GLOSH–Profiles $P\Gamma_{m_{max}}$ is straightforward as it combines all possible GLOSH scores at all possible min_{pts} values, and one does not have to set a limit to it. However, if a shorter GLOSH–Profile can perform as well or better than a full profile, then one can avoid the additional computational effort of extracting full GLOSH–Profiles by setting a limit on the m_{max} value. Note that *CORE-SG* [11], which introduces an efficient way to compute multiple HDBSCAN* hierarchies, has been proposed with values of m_{max} typically up to 100 in mind, since effective values of min_{pts} for hierarchical clustering are typically much smaller than 100. Motivated by this, we compare the performance of GLOSH–Profiles of different sizes with a full GLOSH–Profile in a simple setting that uses the profiles to define a simple aggregate outlier score for each point based on its profile. We then use the aggregate score to rank the points and compare the performance of these rankings

Dataset	**Number of Samples**	**Dimensions**	**Outlier %**	**Dataset Type**
MVTec-AD_zipper	391	512	5	Image
HEPATITIS	80	19	16.25	Healthcare
LETTER	1600	32	6.25	Image
PIMA	768	8	34.9	Healthcare
STAMPS	340	9	9.12	Document
VERTEBRAL	240	6	12.5	Biology
VOWELS	1456	12	3.43	Linguistics
WDBC	367	30	2.72	Healthcare
WINE	129	13	7.75	Chemistry
WPBC	198	33	23.74	Healthcare
YEAST	1484	8	34.16	Biology
BREASTW	683	9	34.99	Healthcare
CARDIO	1831	21	9.61	Healthcare
CARDIOTOCOGRAPHY	2114	21	22.04	Healthcare
20news_3	615	768	5	NLP

Table 3.1: Dataset Description

for outlier detection on some labelled benchmark datasets.

We use fifteen different real-world datasets that are popularly used in outlier detection and one-class classification studies [23]—we refer to them as **Type 1** datasets. The datasets used are: MVTec-AD_zipper, HEPATITIS, LETTER, PIMA, STAMPS, VERTEBRAL, VOWELS, WDBC, WINE, WPBC, YEAST, BREASTW, CARDIO, CARDIOTOCOGRAPHY, and 20news_3. We obtain the one-class datasets (**Type 1**) from the AdBench repository [23]. These datasets are originally classification datasets that have been converted into one-class classification datasets in two ways: (I) For binary classification datasets, one class is labelled as the inlier class and the other as the outlier class which is downsampled, and (II) for multi-class datasets, all but one of class labels are relabelled as inlier class, leaving one class that is labelled as the outlier class which is downsampled. The datasets are described in Table 3.1.

We extract the following GLOSH–Profiles $P\Gamma_{m_{max}}$ for each datapoint:

- $P\Gamma_5$: $m_{max} = 5$

- $P\Gamma_{10}$: $m_{max} = 10$

- $P\Gamma_{25}$: $m_{max} = 25$

- $P\Gamma_{50}$: $m_{max} = 50$

- $P\Gamma_{100}$: $m_{max} = 100$

- Full Profile: m_{max} = size of the dataset

As aggregate outlier score for a datapoint, we use the *Area Under the Curve* (AUC) [27] that has been widely used in the literature to aggregate values of a function.

As a measure of performance of the resulting rankings, we use precision@n, which is a popular evaluation metric in outlier detection [12]. Precision@n in this context measures the fraction of "true" outliers among the top n points in a ranking of the points from largest to smallest aggregated GLOSH-Profile value.

Table 3.2 compares the precision@n obtained by the aggregated GLOSH–Profiles of different sizes across the different test datasets. When comparing the results, one can see that aggregating a full GLOSH–Profile has no added benefits for most of the datasets. One can see that in datasets such as MVTec-AD_zipper, HEPATITIS, LETTER, PIMA, VERTEBRAL, and 20news_3, aggregating the full GLOSH–Profiles does not improve the precision@n over shorter profiles ($P\Gamma_5$ to $P\Gamma_{100}$). In datasets such as LETTER, VOWELS, and YEAST one can see that aggregating the full profile even reduces the precision@n compared to that what is obtained with any of the shorter profiles.

In datasets such as STAMPS, BREASTW, CARDIO, and CARDIOTOCOGRAPHY, using a full GLOSH–Profile results in minor improvements over shorter profiles. However, in the cases of STAMPS and BREASTW datasets, one can only see minor improvements of 0.01, and 0.04 over the maximum precision@n obtained using the

Dataset	$\boldsymbol{P\Gamma_5}$	$\boldsymbol{P\Gamma_{10}}$	$\boldsymbol{P\Gamma_{25}}$	$\boldsymbol{P\Gamma_{50}}$	$\boldsymbol{P\Gamma_{100}}$	**Full Profile**
MVTec-AD_zipper	0.60	0.60	0.60	0.60	0.61	0.60
HEPATITIS	0.15	0.23	0.23	0.23	-	0.23
LETTER	0.12	0.28	0.29	0.29	0.27	0.08
PIMA	0.53	0.52	0.54	0.54	0.54	0.53
STAMPS	0.22	0.19	0.19	0.19	0.19	0.26
VERTEBRAL	0.03	0.03	0.03	0.03	0.00	0.00
VOWELS	0.40	0.54	0.56	0.48	0.44	0.12
WDBC	0.20	0.40	0.40	0.40	0.50	0.60
WINE	0.00	0.00	0.10	0.30	0.30	0.2
WPBC	0.19	0.21	0.19	0.17	0.19	0.19
YEAST	0.29	0.28	0.27	0.27	0.27	0.24
BREASTW	0.63	0.54	0.77	0.84	0.91	0.95
CARDIO	0.19	0.25	0.31	0.37	0.42	0.57
CARDIOTOCOGRAPHY	0.30	0.29	0.21	0.24	0.34	0.50
20news_3	0.17	0.17	0.13	0.13	0.13	0.13

Table 3.2: Comparing the Precision@n obtained using shorter GLOSH–Profiles and full GLOSH–Profile

shorter profiles ($P\Gamma_5$ to $P\Gamma_{100}$). In the cases of CARDIO and CARDIOTOCOGRAPHY datasets using the full profiles results in a slightly larger gains. Overall, the findings of this section suggest that using Full GLOSH–Profiles has little or no benefits over using GLOSH–Profiles with lower m_{max} values (typically, $m_{max} \leq 100$) and does not justify the computational cost required. Therefore, in subsequent sections of this thesis, we adopt shorter GLOSH–Profiles (with $m_{max} = 100$) for our investigations.

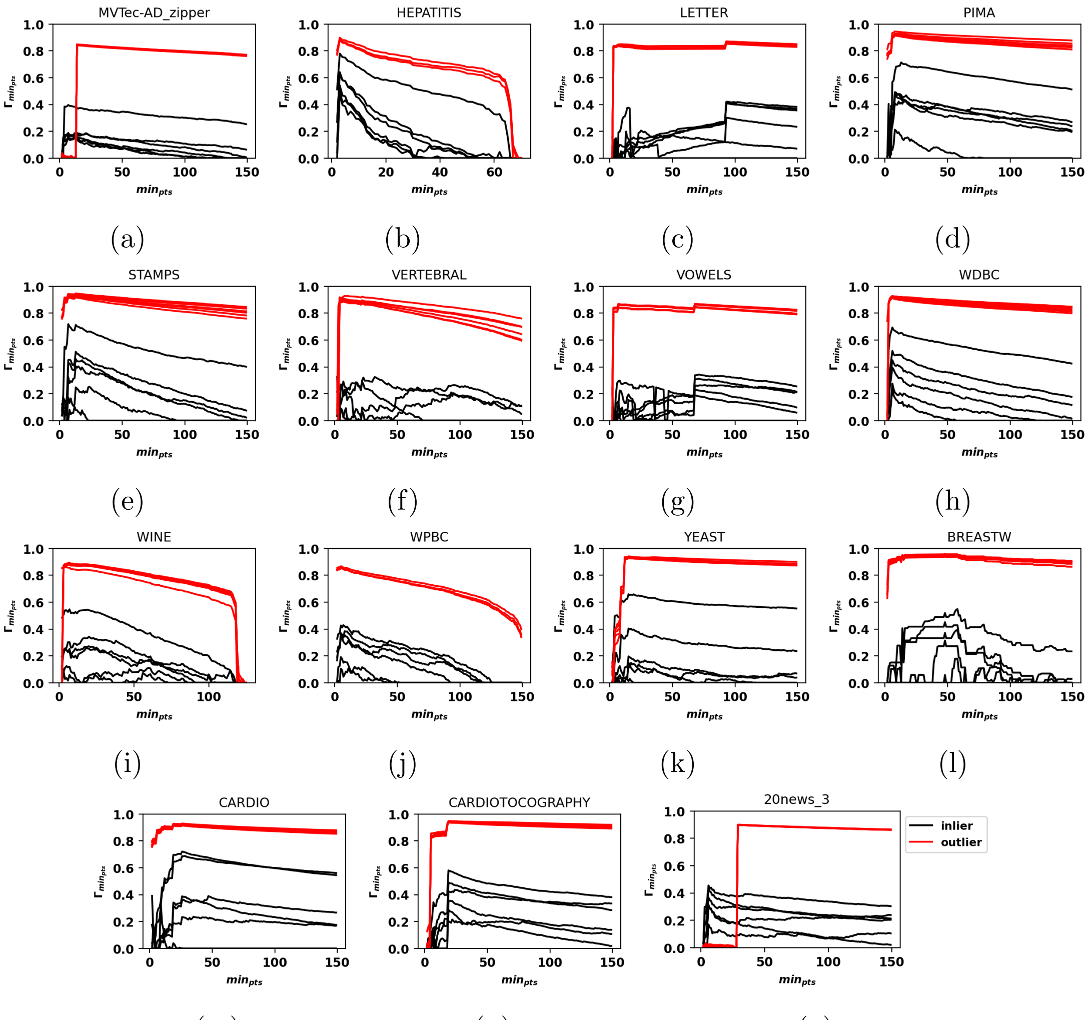
3.1.3 GLOSH–Profiles on Different Datasets

In section 3.1.1, we analyzed the behavior of GLOSH–Profiles of different kinds of outlier and inlier points on simple 2-dimensional synthetic datasets. In this section we analyse the behavior of the GLOSH–Profiles of different outliers and inliers on more complex and more realistic datasets. The goal of this section is to confirm our findings of section 3.1.1 on a wider range of (I) **Type 1**: Real One-Class Classification datasets, and (II) **Type 2**: Datasets with real Inliers and added Synthetic Outliers of controlled types ("semi-real" datasets).

As the **Type 1** datasets are originally classification datasets, we hypothesize that the datapoints "labelled as outliers" may not necessarily follow the characteristics of outliers defined in the outlier detection literature. Consequently, the GLOSH–Profiles of such points may in some cases behave more like that of inlier profiles and not exhibit the characteristics discussed in sections 3.1 and 3.1.1.

The **Type 1** datasets are the same that we used in Section 3.1.2: MVTec-AD_zipper, HEPATITIS, LETTER, PIMA, STAMPS, VERTEBRAL, VOWELS, WDBC, WINE, WPBC, YEAST, BREASTW, CARDIO, CARDIOTOCOGRAPHY, and 20news_3. Although, these datasets are popular in outlier detection studies, they provide very limited scope to analyze the properties of different kinds of outliers. This happens because one does not know what kind of outliers are in the so-called "outlier class", and datapoints in the "outlier class" often deviate from the characteristics typically associated with outliers in the outlier detection literature (global, local, clumps)—which is also reflected in the overall very low precision@n values shown in Table 3.2, for all but one dataset. To address these limitations we derive **Type 2** datasets from **Type 1** datasets. The goal here is to use real inlier datapoints from the one-class datasets and add different kinds of synthetic outliers in a controlled manner. As we did in section 3.1.1, we follow [23] to generate different kinds of synthetic outliers (local, global, and clumps) on real inlier data. For generating local outliers, we learn a Gaussian Mixture Model (GMM) on the inlier samples of the one-class datasets. The covariance matrix Σ of the inlier samples is estimated using the GMM and scaled to $\hat{\Sigma} = \alpha\Sigma$ (where $\alpha = 5$ as in [23]). We use the GMM model to generate the local outlier samples using $\hat{\Sigma}$. To generate outlier clumps, we use the GMM to estimate the mean feature vector μ of the inlier class samples. Then, we scale the mean feature vector by a factor of 5 ($\hat{\mu} = \alpha\mu$ where $\alpha = 5$ as in [23]) to generate the clumps using the GMM model. We generate the global outliers from a uniform distribution with boundaries defined by $(\alpha \cdot max(D^m), \alpha \cdot min(D^m))$ where D^m represents the m-th feature of the inlier class data. Then, as we did in section 3.1.1, we apply the *tomek links* technique [26] to filter out generated points that fall "too close" to the existing inlier points, i.e., are not really global outliers.

In Figure 3.8, we illustrate the GLOSH–Profiles of inliers and global outliers from the **Type 2** datasets (datasets with real inliers and synthetic outliers). To have a clear view of the individual GLOSH–Profiles, we randomly choose six inlier and outlier datapoints and plot their GLOSH–Profiles. Firstly, one can see that there exists a certain min_{pts} value for each of the datapoints when their GLOSH–Profile reaches its highest GLOSH score. Following that min_{pts} value, the GLOSH scores $\Gamma_{min_{pts}}$ gradually decreases as we increase the min_{pts} value. In most datasets, it appears that the GLOSH scores in all the profiles change (decrease) at an almost similar rate after a certain min_{pts} value. Moreover, one can see that, beyond that min_{pts} value, outliers show higher GLOSH scores than inliers. For example, in Figures



(a) (b) (c) (d)

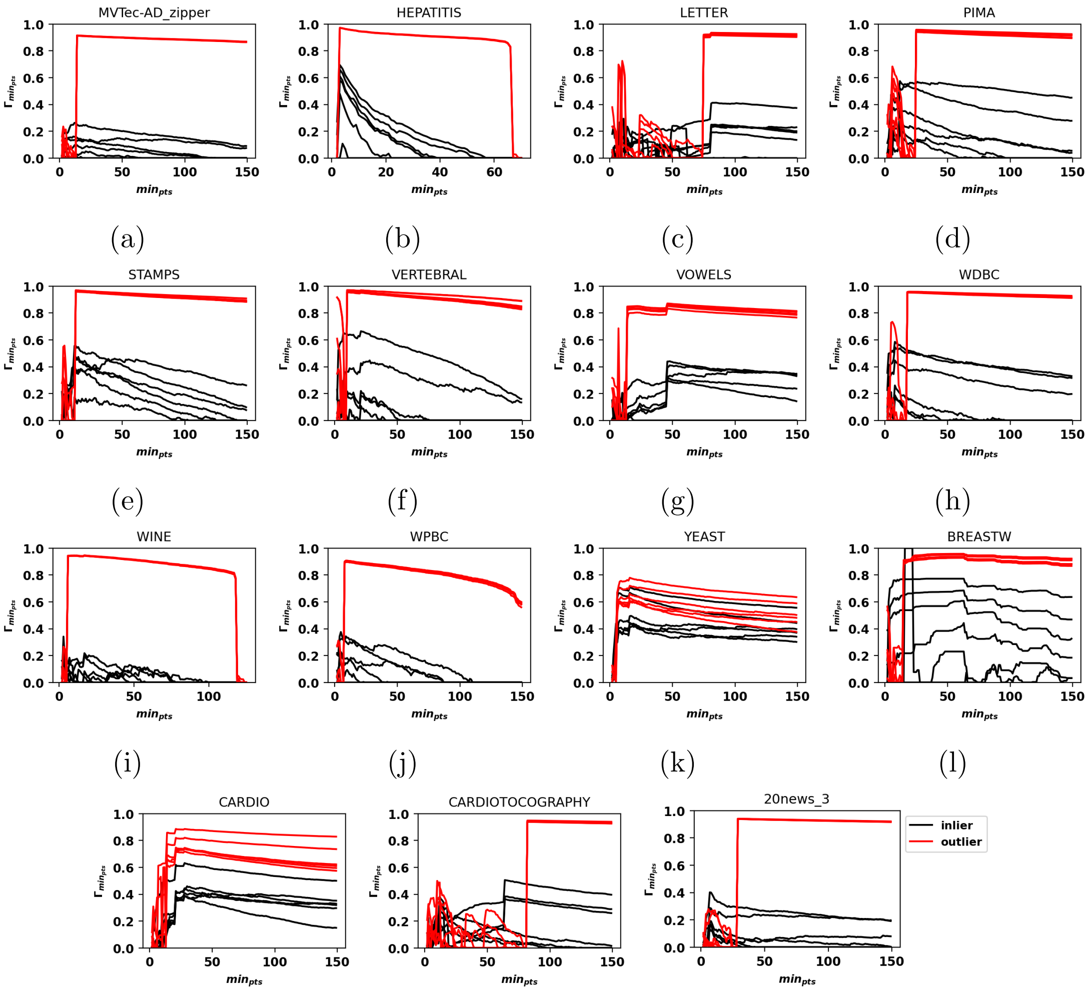
(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o)

Figure 3.8: GLOSH–Profiles of Type 2 Datasets with Global Outliers

3.8a and 3.8h, one can see that the GLOSH scores in both inliers and outlier profiles changes (decreases) at an almost similar rate as the min_{pts} value increase from 22 and 15, respectively. Given the similar rate of change in the GLOSH scores across different min_{pts} values, if one would rank the datapoints based on GLOSH scores (highest to lowest), the relative positions of the datapoints in the rankings would remain nearly the same for min_{pts} values larger than the min_{pts} value at which the GLOSH scores start decreasing. For example, considering the GLOSH–Profiles of the twelve datapoints in Figure 3.8h, if one ranks the datapoints based on their GLOSH scores $\Gamma_{min_{pts}}$ after $min_{pts} = 15$, their positions in the rankings will remain unchanged



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

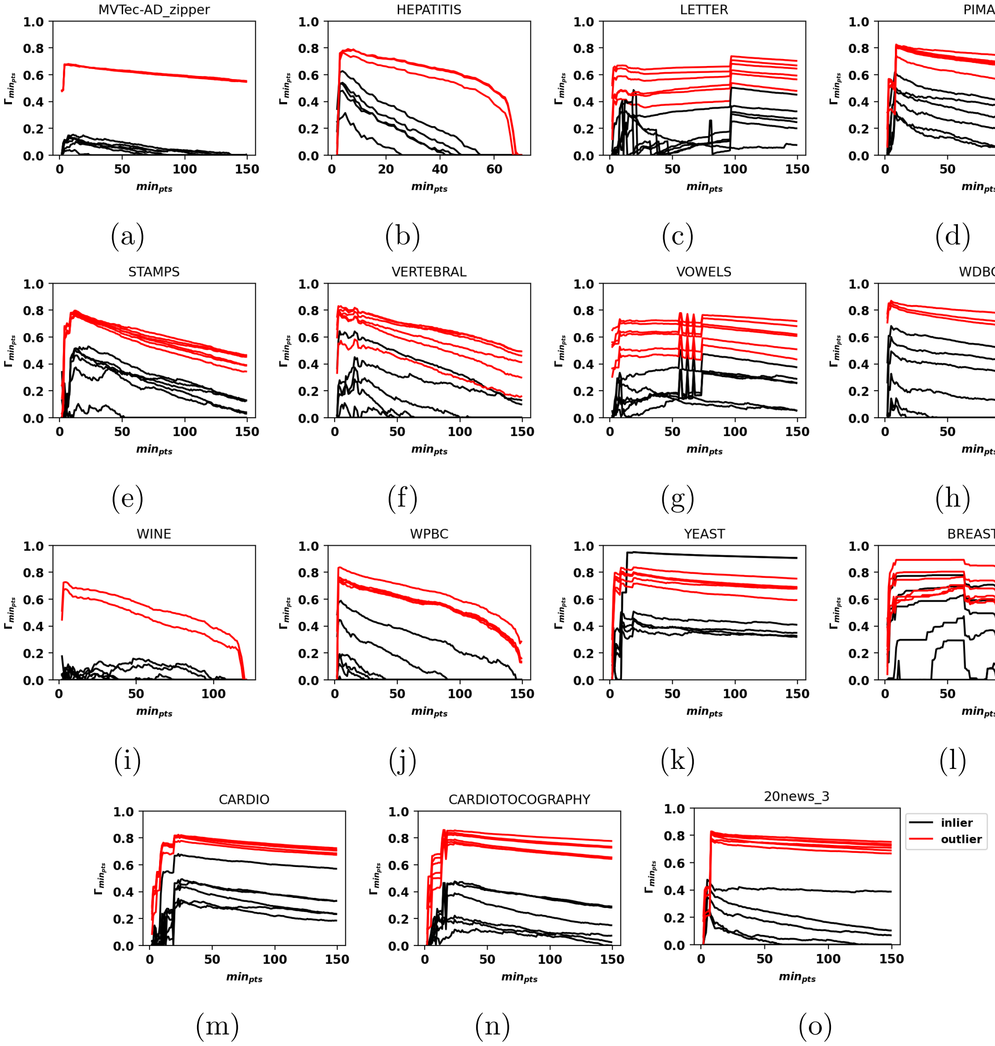
(m) (n) (o)

Figure 3.9: GLOSH–Profiles of Type 2 Datasets with Outlier Clumps

for subsequent min_{pts} values. Consequently, their rankings will stay more or less the same (“stabilize”) for consecutive min_{pts} values.

We make similar observations on the **Type 2** datasets with Outlier Clumps. In Figure 3.9, we present the GLOSH–Profiles of six randomly chosen inlier datapoints and an additional six datapoints randomly chosen from outlier clumps within the **Type 2** datasets. As we already observed in section 3.1.1, GLOSH–Profiles of outlier clumps start from a low GLOSH score and increase significantly after a certain small min_{pts} value corresponding to the clump size. As we increase the min_{pts} value beyond the min_{pts} value for which the GLOSH–Profile shows the highest value for the first time, the GLOSH scores $\Gamma_{min_{pts}}$ in each of the profiles (inliers or outliers) decrease gradually. Additionally, one can see that beyond a certain min_{pts} value, the rate of decrease in the GLOSH scores becomes similar for both outliers and inliers and that, beyond that min_{pts} value, the outliers have higher GLOSH scores than inliers. Consequently, if one were to rank the datapoints based on their GLOSH scores (highest to lowest) beyond that specific min_{pts} value, the relative positions of the datapoints in the rankings would remain mostly unchanged. For example, if one ranks the twelve datapoints in Figure 3.9m, beyond $min_{pts} = 30$, their positions in the rankings will remain unchanged for the subsequent min_{pts} values. In Figure 3.9m, beyond $min_{pts} = 30$, the GLOSH–Profiles follow an almost straight line, which means, the GLOSH scores in each of the profiles change at a similar rate.

In Figure 3.10, we present the GLOSH–Profiles of six randomly chosen inliers and outliers from the **Type 2** datasets with local outliers. When comparing Figures 3.8, 3.9, and 3.10, one can observe certain differences. For the global outliers and clumps, one can see a gap between the outlier profiles and inlier profiles in most datasets. Conversely, for local outliers, their profiles are similar to inliers, with a smaller gap compared to global outliers and clumps. This is expected as by definition, local outliers are typically closer to inlier, and hence, their core distances are not significantly higher or different than those of borderline points of the clusters. In datasets like STAMPS and WPBC, it is also apparent that, in contrast to the GLOSH–Profiles of global outliers and clumps, the GLOSH scores within the local outlier profiles exhibit a more rapid decrease with increasing min_{pts} value. Similar to what we observed for global outliers and clumps, even for local outliers, the GLOSH scores in each of the profiles decrease at an almost similar rate. One can see that in most cases, when the GLOSH scores start to decrease, the outlier profiles tend to exhibit higher GLOSH scores than the inlier profiles. However, as there is no significant gap between the outlier and inlier profiles, the outliers might not consistently exhibit higher GLOSH scores over an extended range of min_{pts} values. As the min_{pts} value increases, the



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o)

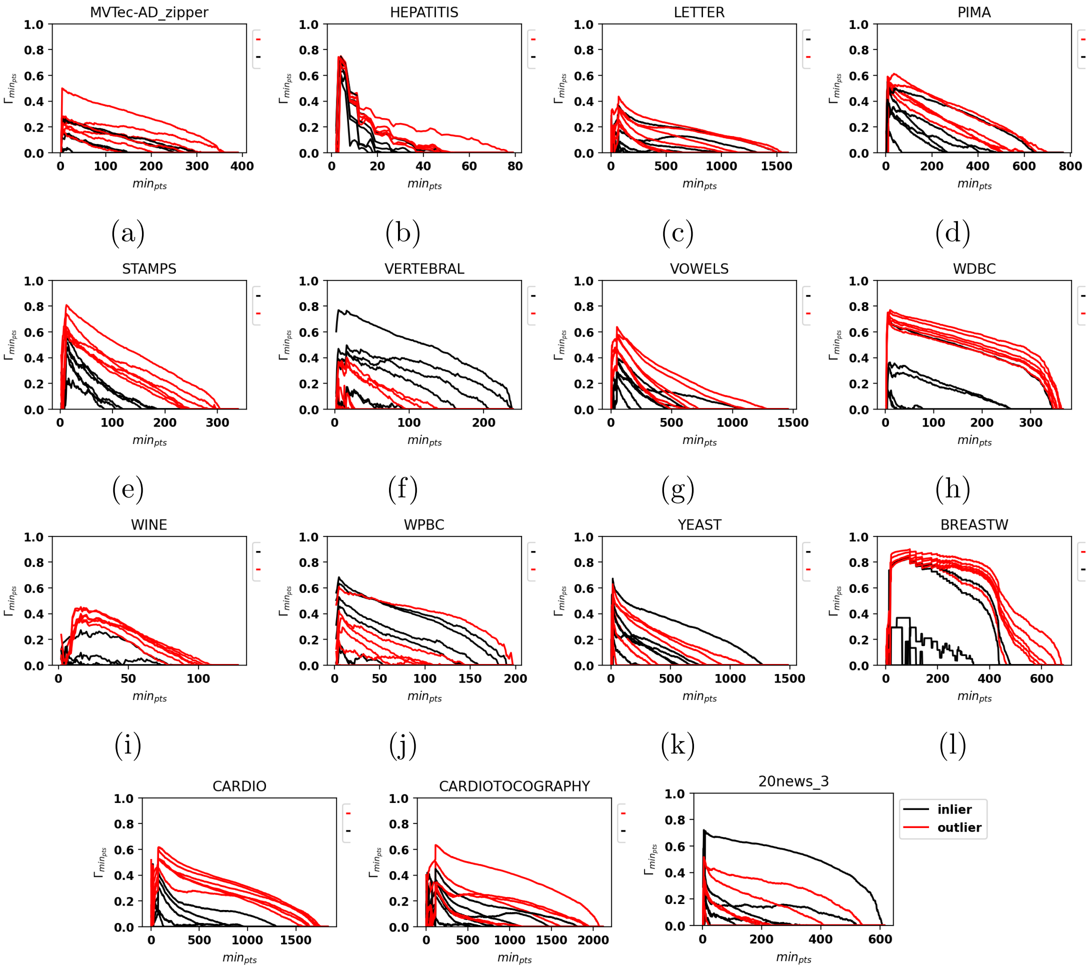
Figure 3.10: GLOSH–Profiles of Type 2 datasets with local outliers

outlier profiles may eventually get lower GLOSH scores than some inliers. Therefore, intuitively, one would expect the outlier rankings to change in a way that affects the precision@n.

Across the three different kinds of outliers we make two common observations:

- There exists a min_{pts} value beyond which the GLOSH scores in all the GLOSH–Profiles (inliers or outliers) change at almost a similar rate.

- Beyond that min_{pts} value, the outlier profiles generally show higher GLOSH scores than inlier profiles.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o)

Figure 3.11: GLOSH–Profiles of Type 1 datasets

Continuing our investigation, we also analyse the GLOSH–Profiles obtained on **Type 1** datasets (real one-class datasets). Firstly, one can see that in many of the datasets, the GLOSH–Profiles of outlier datapoints overlap with the inlier profiles. Secondly, there is also a range of min_{pts} values where the GLOSH scores exhibit a gradual decrease for all the datapoints and the rate of decrease in the GLOSH scores becomes almost similar for all the datapoints. Therefore, one can again expect that the outlier rankings will be similar between consecutive min_{pts} values in that range. However, in most of these datasets, the outlier datapoints either get GLOSH scores lower than the inliers, or the profiles overlap in a way that makes ranking them higher more challenging. have GLOSH–Profiles that are not unlike GLOSH–Profiles of inliers, which explains the relatively low precision@n results reported in Table 3.2, earlier in this chapter. In these datasets, the datapoints labelled as outlier, do often not follow the characteristics of local, global, or clumps outliers.

We summarize the findings of this section as follows:

- In most of the benchmark datasets often used to evaluate outlier detection methods (Type 1 datasets), the GLOSH–Profiles of the datapoints labelled as "outliers" do not show the characteristics of the common outlier types assumed in the literature and discussed in sections 3.1 and 3.1.1. This supports our hypothesis that the labelled outliers in these datasets may not exhibit the typical traits of real outliers.

- In the Type 2 datasets, we observed that there exists a certain min_{pts} value beyond which the GLOSH scores in most profiles tends to change at a similar rate.

- Furthermore, in the Type 2 datasets, we observed that once the GLOSH scores in the profiles start changing at a similar rate, the outlier profiles tends to show higher scores than the inlier profiles.

- Compared to global outliers and outlier clumps, the GLOSH–Profiles of local outliers are similar in behavior to inlier profiles.

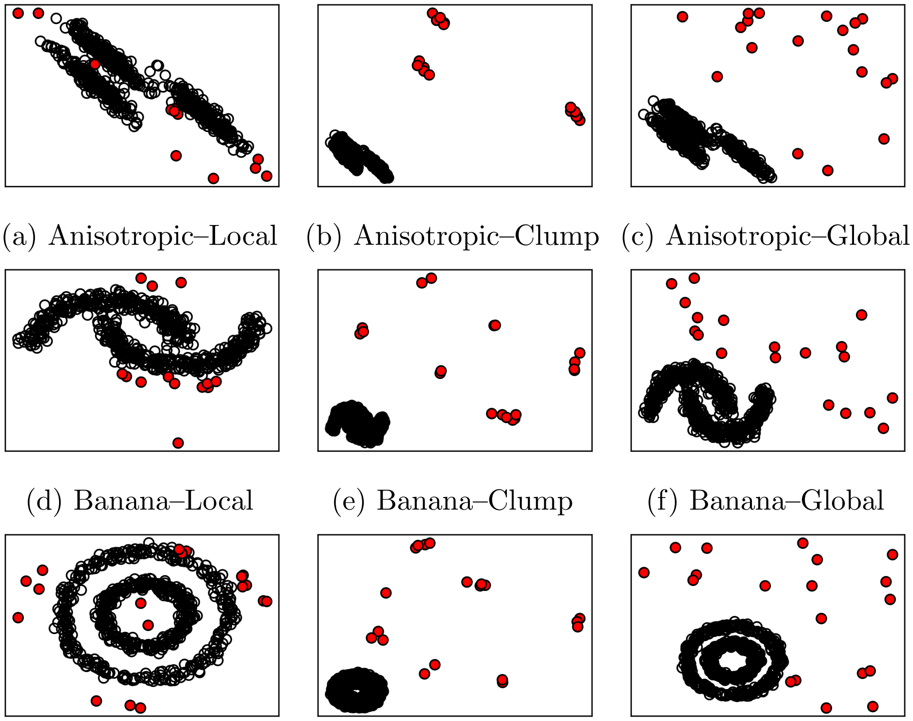
- The GLOSH–Profiles of global outliers and outlier clumps show higher GLOSH scores than inliers for a wide range of min_{pts} values, whereas, in many cases the GLOSH scores in local outlier profiles decrease rapidly with the increase in the min_{pts} value.

3.2 GLOSH Performance across min_{pts} values

In this section we analyze the performance of GLOSH at different min_{pts} values. This analysis is necessary as one of the primary goals of this thesis is to find the min_{pts} value that leads to the best GLOSH performance. To achieve this goal, one has to first understand how GLOSH performs across different min_{pts} values and identify the specific ranges of min_{pts} values where it achieves its best performance

To assess the performance of GLOSH we use again the precision@n metric; it measures the fraction of "true outliers" among the points with the n-th highest GLOSH scores, where n is taken as the total number of labelled outliers in a test dataset. One of the key observations of section 3.1.3 was that GLOSH scores in the profiles of global outliers and outlier clumps stay higher than that of the inliers for a long range of min_{pts} values. Therefore, for global outliers and clumps, we hypothesize that GLOSH will have its highest precision@n score for a large range of min_{pts} values. Therefore, there can be multiple min_{pts} values that can yield the best result. In section 3.1.3, we also saw that in many datasets, the GLOSH–Profiles of local outliers are more similar to inlier profiles and they may have high GLOSH scores for smaller values of min_{pts}, but the GLOSH scores decrease rapidly as the min_{pts} value increases. Therefore, we hypothesize that for local outliers, GLOSH will show high precision@n scores for a shorter range of min_{pts} compared to global outliers and clumps.

To gain a better intuitive understanding, we use three different kinds of synthetic two dimensional datasets in this investigations: Anisotropic, Banana, and Circular, obtained from [28]. To conduct a systematic investigation of different kinds of outliers, we follow the same approach as discussed in section 3.1.3 to generate different kinds of outliers using the inlier samples provided in the original datasets. The datasets including the generated outliers are shown in Figure 3.12. Each of these datasets has been enriched with different types of outliers, namely Local Outliers, Global Outliers, and Outlier Clumps, hereafter referred to simply as 'Clumps'. We refer these datasets



(a) Anisotropic–Local (b) Anisotropic–Clump (c) Anisotropic–Global

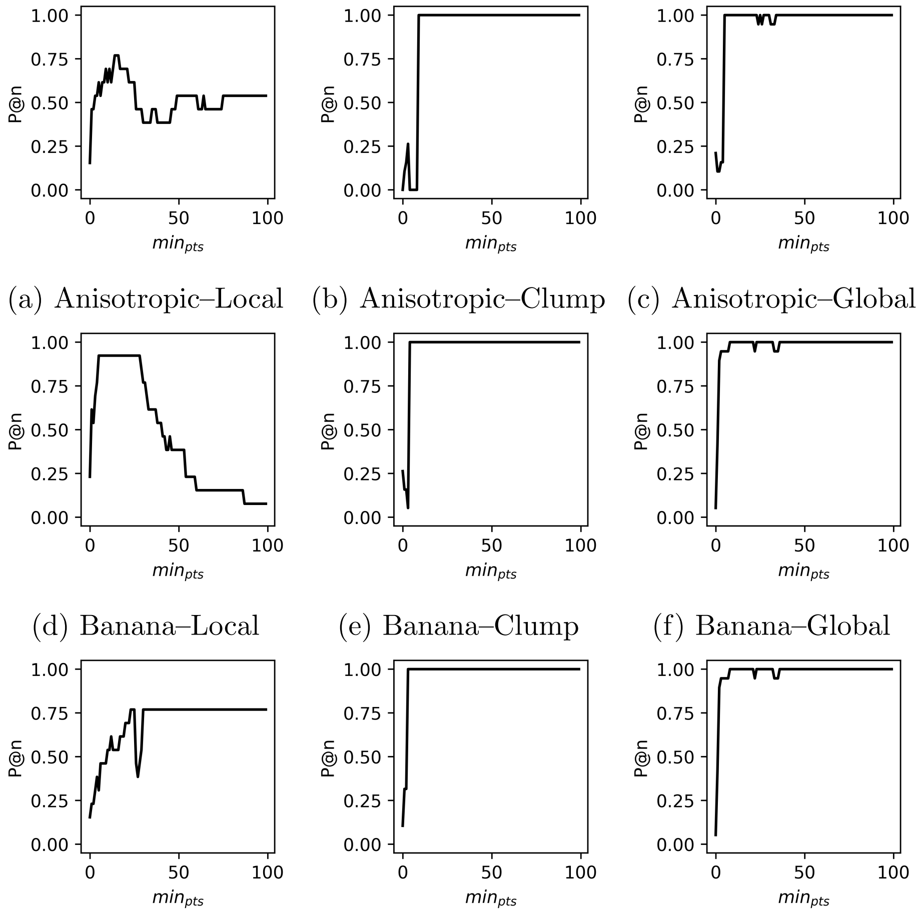
(d) Banana–Local (e) Banana–Clump (f) Banana–Global

(g) Circular–Local (h) Circular–Clump (i) Circular–Global

Figure 3.12: Type 3 Datasets: Synthetic 2D Datasets with different kinds of outliers

as **Type 3** datasets.

Figure 3.13 illustrates the Precision@n (P@n) scores obtained by GLOSH on each of the datasets, for different min_{pts} values. Firstly, as we expected, one can see for global outliers and clumps, GLOSH records its highest P@n scores for a large range of min_{pts} values. However, one can also see initially for lower min_{pts} values, GLOSH records low precision@n. Secondly, as per our expectations, for most of the datasets with local outliers, GLOSH records a high P@n for a small range of min_{pts} values and then the P@n starts to decrease as the min_{pts} value increases. In practice, we do not know what kind of outliers are present in the dataset, therefore, one cannot easily guess the min_{pts} values where GLOSH achieves its best performance. However, since it is easy for us to obtain the GLOSH scores for a large range of min_{pts} values (i.e., obtain GLOSH–Profiles of a significant length) efficiently, a natural question to ask is: Can we use the GLOSH–Profiles to estimate the range of min_{pts} values where one



(a) Anisotropic–Local (b) Anisotropic–Clump (c) Anisotropic–Global

(d) Banana–Local (e) Banana–Clump (f) Banana–Global

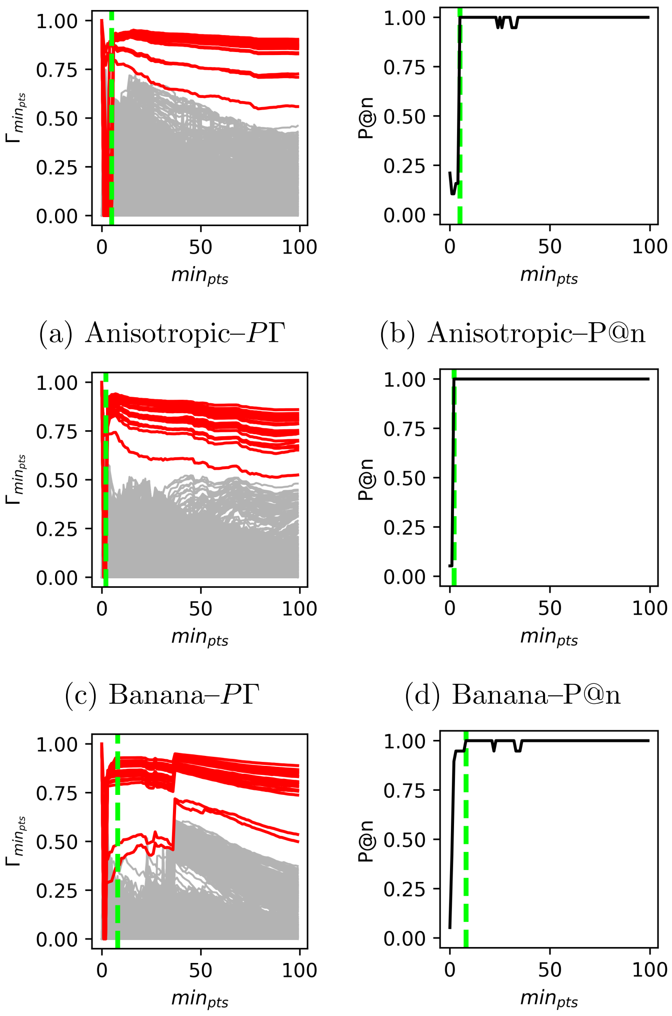
(g) Circular–Local (h) Circular–Clump (i) Circular–Global

Figure 3.13: Precision@n obtained by GLOSH at different min_{pts} on Type 3 Datasets

can achieve the best or close to best P@n? To answer this question, one first has to systematically investigate how the GLOSH–Profiles look like (i.e. how the GLOSH scores behave) in the range of min_{pts} values where GLOSH achieves the highest P@n.

In Figures 3.14 to 3.16, we compare GLOSH–Profiles with the Precision@n (P@n) values obtained by GLOSH for different min_{pts} values.

For the datasets with global outliers, in Figure 3.14, one can observe that initially for low min_{pts} values, the GLOSH scores in the profiles fluctuates, i.e., the GLOSH scores change significantly between consecutive min_{pts} values and at different rates for different profiles. The variation in the scores results in several outliers getting lower GLOSH scores than inliers. Consequently, it is not a surprise that GLOSH records



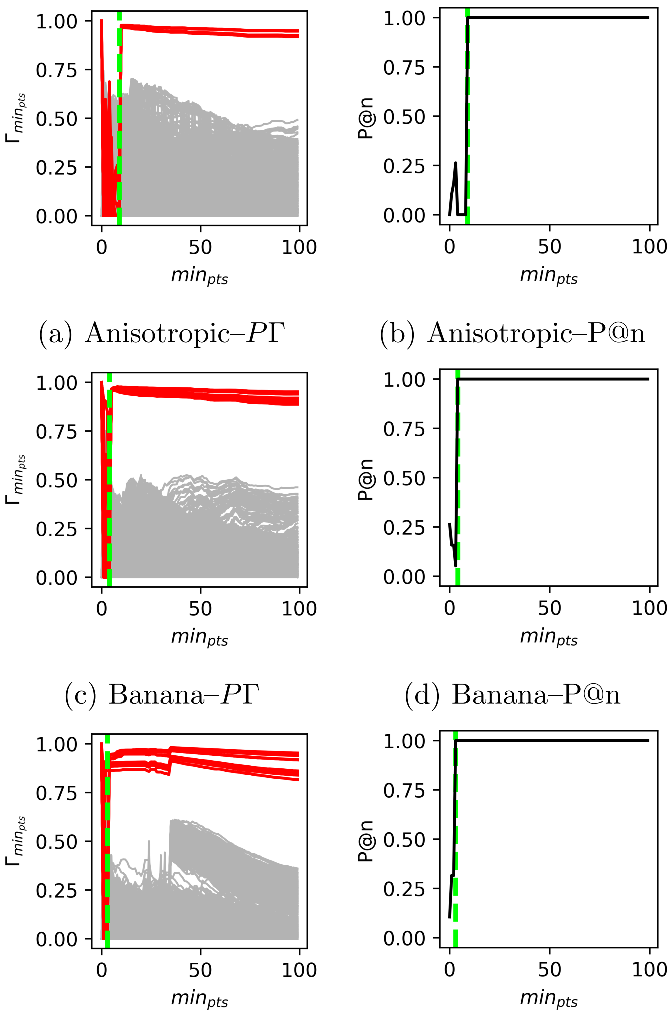
(a) Anisotropic–$P\Gamma$ (b) Anisotropic–P@n

(c) Banana–$P\Gamma$ (d) Banana–P@n

(e) Circular–$P\Gamma$ (f) Circular–P@n

Figure 3.14: Highlighting the min_{pts} value at which GLOSH achieves the highest Precision@n with Global Outliers: The red GLOSH–Profiles represent the outlier profiles and the green line denotes the specific min_{pts} value where GLOSH first achieves the best Precision@n (P@n) within the range $[2, 100]$ of min_{pts} values.

low precision@n intially for low min_{pts} values. Beyond a certain min_{pts} value, the GLOSH scores in the profiles start undergoing minimal changes between consecutive min_{pts} values, and when they do change, they change at similar rates so that the resulting rankings are likely not changing in a way that the Precision@n is affected. A major observation is that the min_{pts} value where the GLOSH scores start to change at a similar rate, corresponds to the min_{pts} value at which GLOSH starts to yield the best precision@n (P@n).



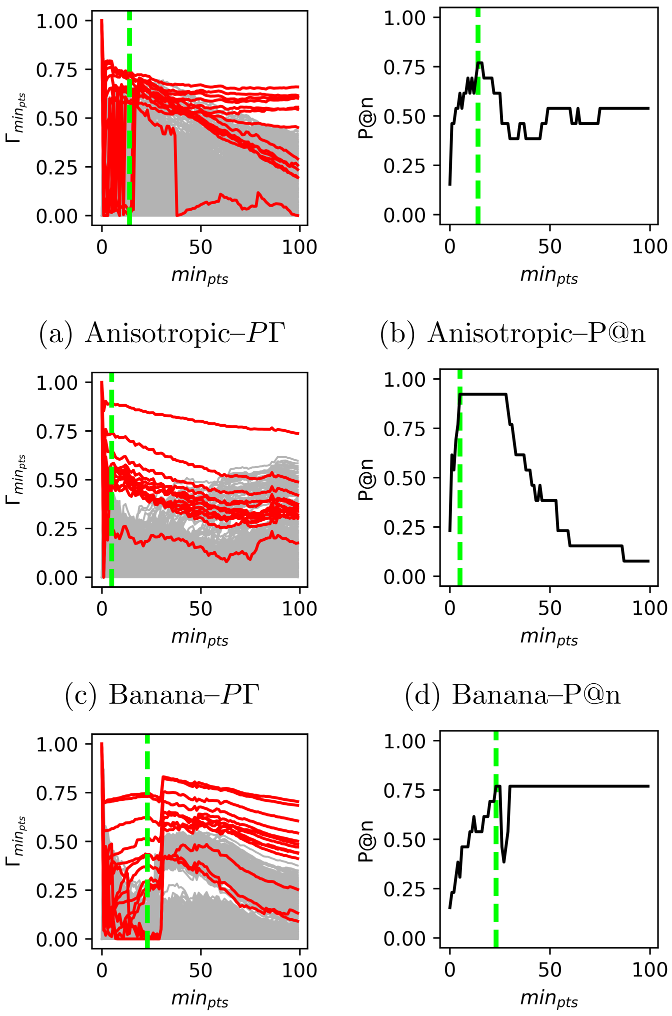
(a) Anisotropic–$P\Gamma$ (b) Anisotropic–P@n

(c) Banana–$P\Gamma$ (d) Banana–P@n

(e) Circular–$P\Gamma$ (f) Circular–P@n

Figure 3.15: Highlighting the min_{pts} value at which GLOSH achieves the highest Precision@n with Outlier Clumps: The red GLOSH–Profiles represent the outlier profiles and the green line denotes the specific min_{pts} value where GLOSH first achieves the best Precision@n (P@n) within the range $[2, 100]$ of min_{pts} values.

For datasets with outlier clumps, in Figure 3.15, we can make similar observations. The GLOSH scores in the profiles show significant fluctuations initially for lower min_{pts} values, however, beyond a certain min_{pts} value, the GLOSH scores start to change at a similar rate in the profiles, with the outlier profiles showing higher GLOSH scores than the inlier profiles. Notably, the min_{pts} value at which the GLOSH scores in the profiles start to change at a similar rate, corresponds to the min_{pts} value at which GLOSH starts yielding the highest precision@n (P@n).



(a) Anisotropic–$P\Gamma$ (b) Anisotropic–P@n

(c) Banana–$P\Gamma$ (d) Banana–P@n

(e) Circular–$P\Gamma$ (f) Circular–P@n

Figure 3.16: Highlighting the min_{pts} value at which GLOSH achieves the best Precision@n with Local Outliers: The red GLOSH–Profiles represent the outlier profiles and the green line denotes the specific min_{pts} value where GLOSH first achieves the best Precision@n (P@n) within the range $[2, 100]$ of min_{pts} values.

In Figure 3.16, we observe that even for local outliers, beyond a specific min_{pts} value the GLOSH scores, in general, start to decrease at a similar rate but not as consistently as for other types of outliers. Nonetheless, this specific min_{pts} value corresponds to a min_{pts} value that results in the best or nearly the best precision@n for GLOSH. However, one cannot randomly choose a value beyond this specific min_{pts} value and expect to achieve the best precision@n. Unlike global outliers and clumps, local outlier profiles are more similar to inlier profiles and there is no significant gap between them. In addition, as the min_{pts} value increases significantly beyond the value where the GLOSH scores start changing at a similar rate, the rate of change may become more dissimilar again and some local outliers end up receiving scores lower than several inlier points. This effect is particularly strong in the Banana dataset and reduces the precision@n. When the scores change at a similar rate, ordered sequence of the GLOSH scores may be similar between consecutive min_{pts} values, even though the relative ordering of the datapoints may differ.

Overall, we summarize the key observations of this section as follows:

- As the min_{pts} value increases from 2 to 100, GLOSH initially tends to result in low performance w.r.t. precision@n. However, this performance tends to improve as the min_{pts} value increases.

- For lower min_{pts} values, the GLOSH scores in outlier profiles tends to exhibit significant variations—resulting in outliers getting lower GLOSH scores than inliers

- The min_{pts} value at which the GLOSH scores starts to change at a similar rate for most of the GLOSH–Profiles, corresponds to the min_{pts} value that yields the best Precision@n with GLOSH, however, choosing a min_{pts} value beyond that value does not always guarantee the best Precision@n.

3.3 Dissimilarity between GLOSH score rankings for consecutive min_{pts} values

In this section we study how the GLOSH scores change between consecutive min_{pts} values. One way to do that is by measuring the differences in the outlier rankings at consecutive min_{pts} values. However, it may happen that two datapoints get identical GLOSH scores at consecutive min_{pts} values. If one ranks the datapoints based on their GLOSH scores, any specific ordering of those two datapoints may differ between

Row	Datapoints																				Δ
	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$	$\boldsymbol{x_6}$	$\boldsymbol{x_7}$	$\boldsymbol{x_8}$	$\boldsymbol{x_9}$	$\boldsymbol{x_{10}}$	$\boldsymbol{x_{11}}$	$\boldsymbol{x_{12}}$	$\boldsymbol{x_{13}}$	$\boldsymbol{x_{14}}$	$\boldsymbol{x_{15}}$	$\boldsymbol{x_{16}}$	$\boldsymbol{x_{17}}$	$\boldsymbol{x_{18}}$	$\boldsymbol{x_{19}}$	$\boldsymbol{x_{20}}$	
A	0.001	0.046	0.107	0.005	0.207	0.198	0.08	0.167	0.017	0.015	0.101	0.048	0.221	0.022	0.065	0.233	0.001	0.13	0.044	0.893	0.0
Sorted-A	0.001	0.001	0.005	0.015	0.017	0.022	0.044	0.046	0.048	0.065	0.08	0.101	0.107	0.13	0.167	0.198	0.207	0.221	0.233	**0.893**	
B	0.0002	0.0092	0.021	0.001	0.041	0.039	0.016	0.033	0.0034	0.003	0.02	0.0096	0.044	0.004	0.013	0.046	0.0002	0.026	0.008	0.178	
Sorted-B	0.0002	0.0002	0.001	0.003	0.0034	0.004	0.008	0.0092	0.0096	0.013	0.016	0.02	0.021	0.026	0.033	0.039	0.041	0.044	0.046	**0.178**	

Table 3.3: Pearson dissimilarity in a scenario where the GLOSH scores change at a similar rate: Each score in the row A decreases by a factor of 5 in the row B. The *Sorted-A* is the sorted sequence of the scores before they change (A), while the *Sorted-B* is the sorted sequence of the scores after they decrease (B). The Δ is computed between the sorted sequences.

the consecutive min_{pts} values but the scores of adjacent points in large stretches of a ranking may be very similar. In these cases any measure of point ordering similarity will be dominated by how the large number of inliers are ordered, which is not very informative for outlier detection. For outlier detection, we are mostly interested in the large values or a small portion at the large end of the sorted order. In such scenarios, it is not clear how one can measure the difference between the two rankings. To bypass this problem, we measure the dissimilarity not between the ordered point sequences but between the sorted GLOSH scores, between consecutive min_{pts} values in a way that is sensitive to GLOSH scores that can indicate outliers.

To measure the dissimilarity between the sorted sequences we use Pearson correlation. Pearson correlation is a popular symmetric metric that can measure the relationship between two outlier score sequences [29]. For two sorted GLOSH score sequences obtained at $min_{pts} = k$ (S_k) and $min_{pts} = l$ (S_l), the Pearson dissimilarity measure Δ is defined as:

$$\Delta(S_k, S_l) = 1 - \left| \frac{Cov(S_k, S_l)}{\sqrt{Var(S_k)Var(S_l)}} \right| \tag{3.2}$$

where, Cov measures the covariance between the two outlier score sequences obtained at $min_{pts} = k$ and $min_{pts} = l$, and Var measures the variance of the values in each of the sequences. When GLOSH–Profiles for most points start changing at a similar rate (as observed in the previous section), the resulting relative order and relative

Row	Datapoints																				Δ
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	
A	0.001	0.046	0.107	0.005	0.207	0.198	0.08	0.167	0.017	0.015	0.101	0.048	0.221	0.022	0.065	0.233	0.001	0.13	0.044	0.152	0.26
Sorted-A	0.001	0.001	0.005	0.015	0.017	0.022	0.044	0.046	0.048	0.065	0.08	0.101	0.107	0.13	0.152	0.167	0.198	0.207	0.221	**0.233**	
B	0.001	0.046	0.107	0.005	0.207	0.198	0.08	0.167	0.017	0.015	0.101	0.048	0.221	0.022	0.065	0.233	0.001	0.13	0.044	0.893	
Sorted-B	0.001	0.001	0.005	0.015	0.017	0.022	0.044	0.046	0.048	0.065	0.08	0.101	0.107	0.13	0.167	0.198	0.207	0.221	0.233	**0.893**	

Table 3.4: Pearson dissimilarity in a scenario where only one score jumps from low to high, while the rest do not change: All the scores from row A are identical in the row B apart from the score obtained by an outlier x_{20}. The Δ is computed between the sorted sequences of A (*Sorted-A*) and B (*Sorted-B*).

magnitude of GLOSH scores for the points is likely to change only very little as a consequence, between consecutive values of min_{pts}. As the Pearson correlation measures the covariance between the sequences, it remains unaffected to changes in the absolute values as it measures the variations from the mean. When scores change at a similar rate, that results in positive covariance, yielding a Pearson correlation close to 1 and, consequently, a Δ score close to 0. In Table 3.3, we show the scores obtained on 20 points from the Anisotropic dataset with global outliers. One can see that when the scores change at a similar rate (decreases by a factor of 5), we obtain a Pearson dissimilarity Δ of zero between the ordered sequences of the scores. In practice, scores in the profiles may not change at exactly the same rate but if they change at "almost" the similar rate, we expect the dissimilarity to be close to 0.

As we saw earlier, initially for lower min_{pts} values, the GLOSH scores in outlier profiles fluctuate significantly. In other words, there are significant jumps in the GLOSH scores (either decreasing or increasing) between consecutive min_{pts} values. When this happens for a few points and the scores of most points (inliers) are not changing much relative to each other, we expect the Pearson dissimilarity to be higher. In Table 3.4, we show the simplest case when the score of only a single outlier datapoint (x_{20}) jumps from 0.152 to 0.893 while the rest scores are identical, we get a high dissimilarity of 0.26 between the sorted sequences. When there are more outlier points present in the sequence we expect the Pearson dissimilarity to be even higher between sorted GLOSH scores sequences for certain consecutive min_{pts} values. Therefore, initially for lower min_{pts} values, we expect a high Pearson dissimilarity between the sorted sequences of GLOSH scores at consecutive min_{pts} values until the point when the scores start changing at a similar rate where we expect the Pearson dissimilarity to be converge to 0.

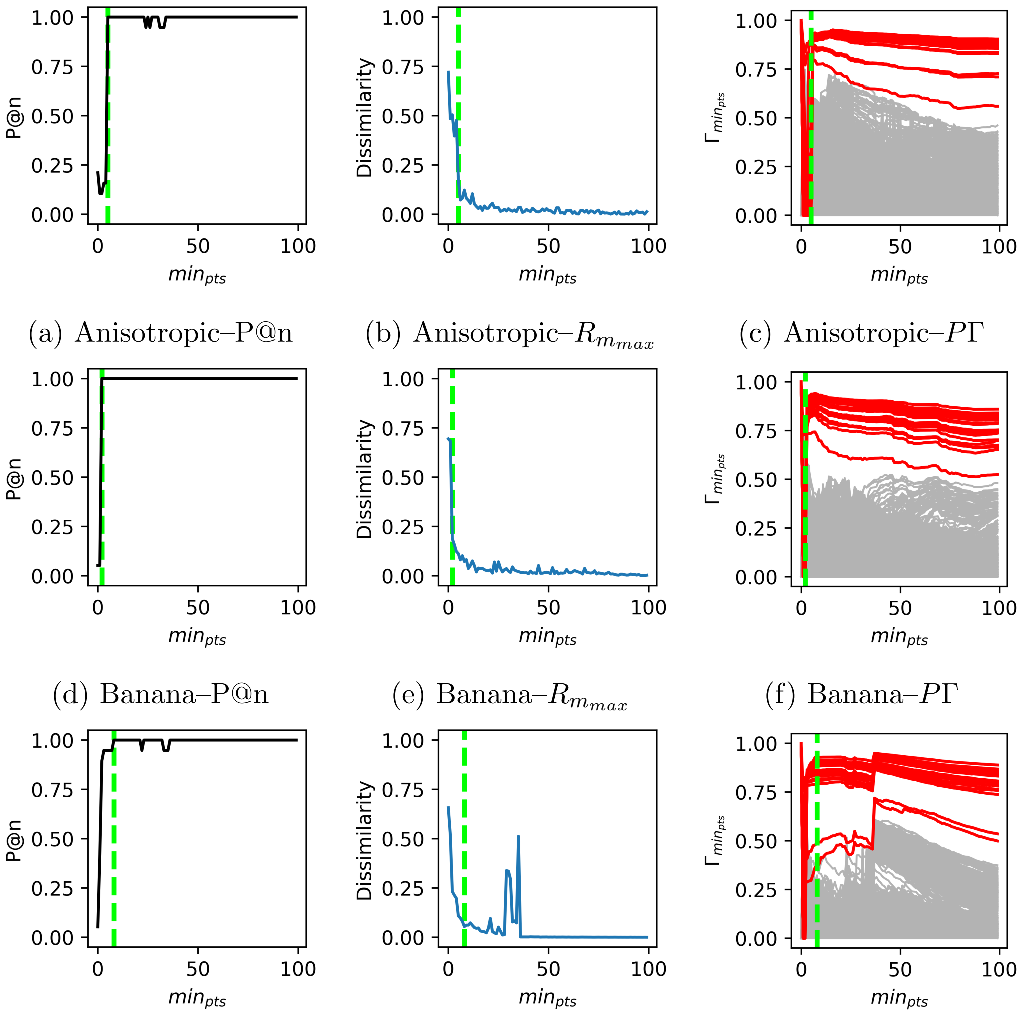
To study the behavior of GLOSH score rankings in more detail, we first define the GLOSH Outlier Rank Dissimilarity–Profile (ORD–Profile) $R_{m_{max}}$ more formally. The ORD–Profile basically quantifies the dissimilarity between the sorted sequences of GLOSH scores at consecutive min_{pts} values, as we increase the min_{pts} values.

Definition 5.2. GLOSH Outlier Rank Dissimilarity–Profile: The GLOSH Outlier Rank Dissimilarity–Profile $R_{m_{max}}$ is an array of dissimilarity values for pairs of sorted sequences of GLOSH scores $S_{min_{pts}}$ obtained at consecutive min_{pts} values in a range $[2, m_{max}]$:

$$R_{m_{max}} = \begin{bmatrix} \Delta(S_2, S_3) \\ \Delta(S_3, S_4) \\ \Delta(S_4, S_5) \\ \vdots \\ \Delta(S_{m_{max}-1}, S_{m_{max}}) \end{bmatrix} \tag{3.3}$$

Assuming the indices i of the dissimilarity scores in $R_{m_{max}}$ starts from 0. As we consider a range $[2, 100]$ of min_{pts} values, each dissimilarity score $r_i \in R_{m_{max}}$ is obtained between $min_{pts} = i + 2$ and $min_{pts} = i + 3$.

For our Type 3 datasets with global outliers, one can see in Figure 3.17 that initially when the GLOSH scores of outlier profiles fluctuate in the profiles, the dissimilarity between the sorted sequences in the ORD–Profile is high. Then, when the GLOSH



(a) Anisotropic–P@n (b) Anisotropic–$R_{m_{max}}$ (c) Anisotropic–$P\Gamma$

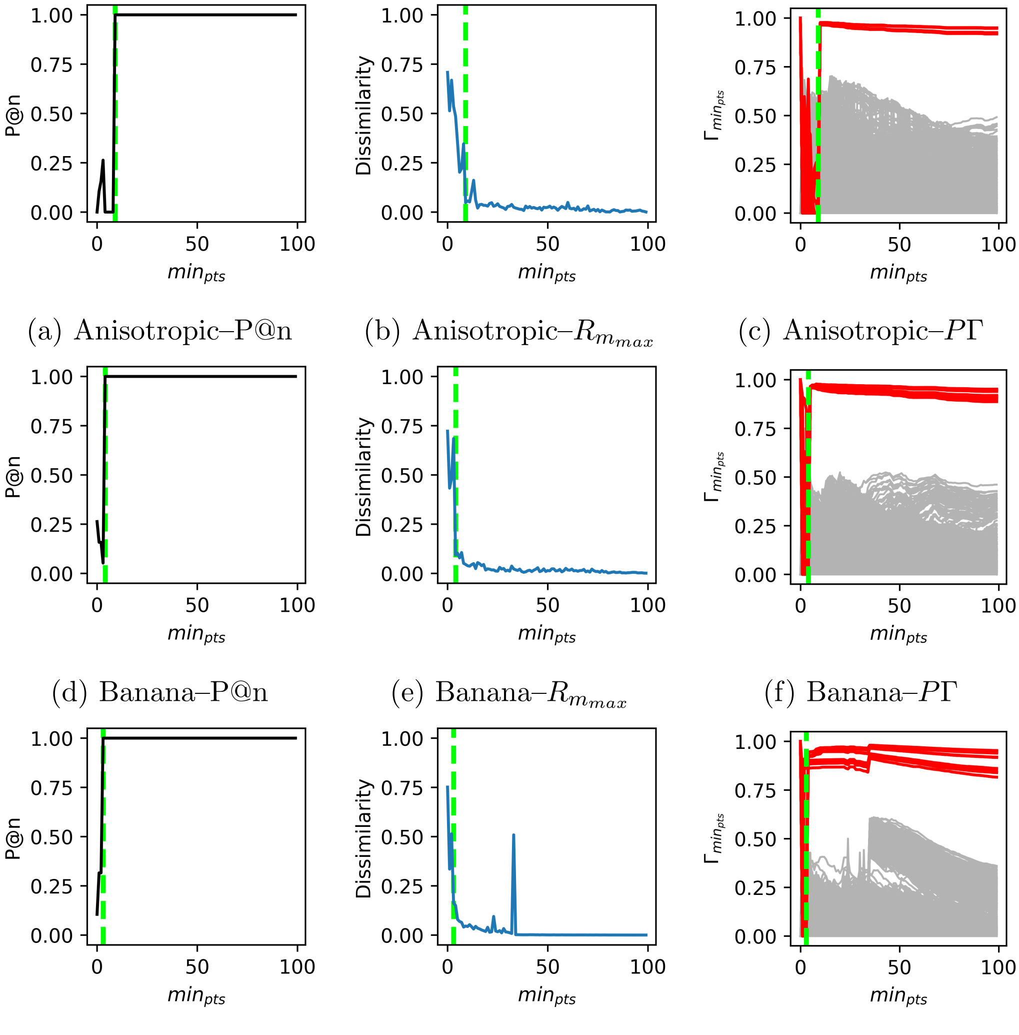
(d) Banana–P@n (e) Banana–$R_{m_{max}}$ (f) Banana–$P\Gamma$

(g) Circular–P@n (h) Circular–$R_{m_{max}}$ (i) Circular–$P\Gamma$

Figure 3.17: Type 3 with Global Outliers: Comparing the ORD–Profiles ($R_{m_{max}}$) with precision@n (P@n) achieved by GLOSH across different min_{pts} values. Highlighting the specific min_{pts} value where GLOSH first records the best precision@n within a range $[2, 100]$ of min_{pts} values.

scores start changing at a similar rate in the profiles, the ORD–Profile plots show an "elbow like" structure for each of the datasets. The "elbow" can be seen as the point from where GLOSH scores in most of the profiles starts changing at almost a similar rate. Moreover, one can see that the min_{pts} value where the "elbow" is formed corresponds to the min_{pts} value where GLOSH achieves its best precision@n. Therefore, choosing the min_{pts} value at the "elbow" may potentially yield the best precision@n. In the last row of Figure 3.17, the dissimilarity score increases again for a min_{pts}

value around 28 when the GLOSH scores in most of the profiles increases. However, this increased dissimilarity is not as high as what is observed at the beginning of the ORD–Profile.



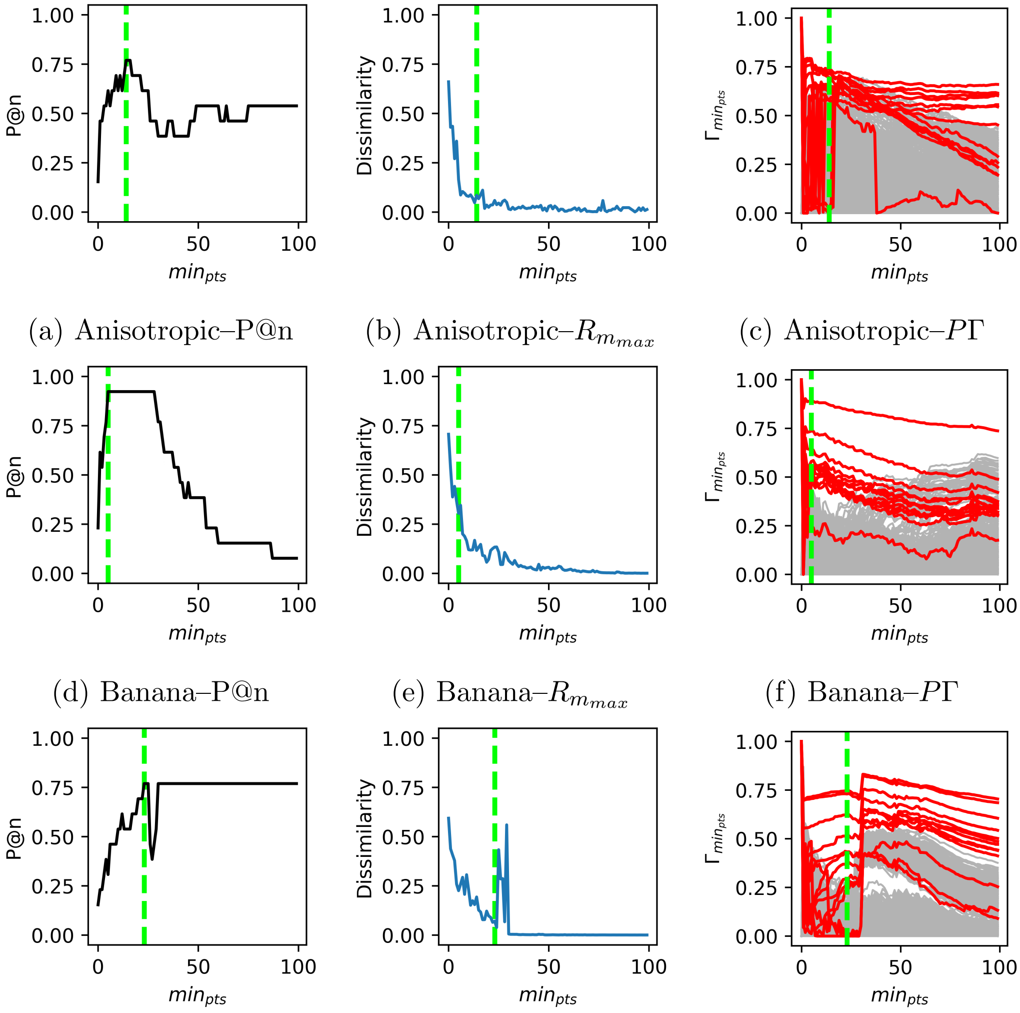
(a) Anisotropic–P@n (b) Anisotropic–$R_{m_{max}}$ (c) Anisotropic–$P\Gamma$

(d) Banana–P@n (e) Banana–$R_{m_{max}}$ (f) Banana–$P\Gamma$

(g) Circular–P@n (h) Circular–$R_{m_{max}}$ (i) Circular–$P\Gamma$

Figure 3.18: Type 3 Datasets with Outlier Clumps: Comparing the ORD–Profiles ($R_{m_{max}}$) with precision@n (P@n) achieved by GLOSH across different min_{pts} values. Highlighting the specific min_{pts} value where GLOSH first records the best precision@n within a range $[2, 100]$ of min_{pts} values.

One can make similar observations for the Type 3 datasets containing outlier clumps in Figure 3.18. In each of the cases, one can see that the ORD–Profile forms an “elbow like” structure when the GLOSH scores in the profiles start to change at a similar rate. The min_{pts} value at the “elbow” corresponds again to the min_{pts} value that results in the best precision@n for GLOSH. These observations suggests that choosing the min_{pts} value at the "elbow" may potentially yield the best precision@n.



(a) Anisotropic–P@n (b) Anisotropic–$R_{m_{max}}$ (c) Anisotropic–$P\Gamma$

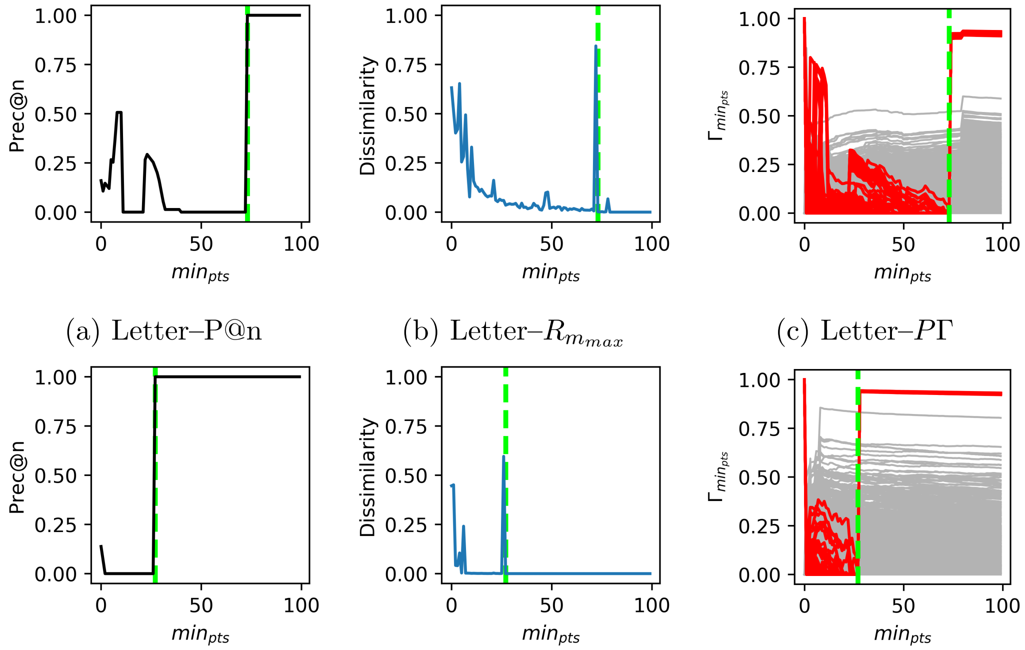
(d) Banana–P@n (e) Banana–$R_{m_{max}}$ (f) Banana–$P\Gamma$

(g) Circular–P@n (h) Circular–$R_{m_{max}}$ (i) Circular–$P\Gamma$

Figure 3.19: Type 3 Datasets with Local Outliers: Comparing the ORD–Profiles ($R_{m_{max}}$) with precision@n (P@n) achieved by GLOSH across different min_{pts} values. Highlighting the specific min_{pts} value where GLOSH first records the best precision@n within a range $[2, 100]$ of min_{pts} values.

Similarly, for the Type 3 datasets containing local outliers, one can see in Figure 3.19 that the plots of the ORD–Profiles show again an "elbow" when the GLOSH scores in most of the profiles starts to change at a similar rate. As before, one can also see that for the min_{pts} value at the "elbow" in the ORD–Profile plot, GLOSH obtains its best or nearly the best precision@n. However, in each of the datasets, if one selects a min_{pts} value that is beyond the "elbow", then GLOSH may potentially end up with a very low precision@n. For example, in the second row of Figure 3.19, one can see that beyond $min_{pts} = 30$, even though the ORD–Profile show low dissimilarity scores, the precision@n reduces for GLOSH.

Overall, the observations suggest that the min_{pts} value at the "elbow" in the ORD–Profile plot may potentially yield the best results for GLOSH. For global outliers and clumps, a min_{pts} value greater than the "elbow" may still lead to good results for GLOSH, however, this is not true for local outliers. In practice, we do not know what kind of outliers are present in a dataset, therefore, selecting the min_{pts} directly at the "elbow" seems to be the best option.



(a) Letter–P@n

(b) Letter–$R_{m_{max}}$

(c) Letter–$P\Gamma$

(d) 20news_3–P@n

(e) 20news_3–$R_{m_{max}}$

(f) 20news_3–$P\Gamma$

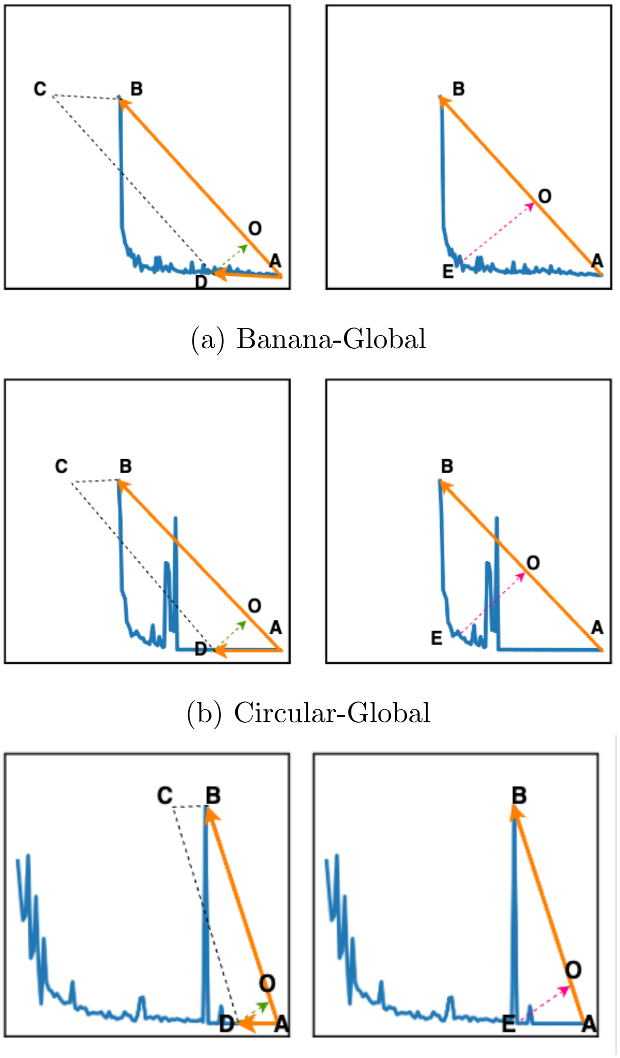
Figure 3.20: LETTER and 20news_3 datasets with synthetic Outlier Clumps: Comparing the ORD–Profiles ($R_{m_{max}}$) with Precision@n achieved by GLOSH across different min_{pts} values. Highlighting the specific min_{pts} value where GLOSH first records the best Precision@n within a range $[2, 100]$ of min_{pts} values.

It may happen that the ORD–Profile plot shows multiple spikes in the dissimilarity score plots. As we see in Figures 3.17h, 3.18h, and 3.19h, one can observe multiple (in this case two) spikes with two elbows in the profile plots. In each of these cases, the second spike in the dissimilarity scores does not correspond to the maximum dissimilarity. In these datasets, taking the min_{pts} value at the first elbow is the choice that leads to the best or nearly the best precision@n for GLOSH. In Figure 3.20 we present two **Type 2** datasets with outlier clumps where we observe multiple spikes in the ORD–Profile and where the last spike leads to the best precision@n. Therefore, we can conclude that it is not the order of the spikes that can guide us in selecting which value of min_{pts} to select. However, what seems to be a common pattern in all these cases is that the best or nearly best precision@n is obtained for the min_{pts} value at the first "elbow" in the ORD–Profile plots after the point where the ORD–Profile shows a maximum dissimilarity.

3.4 Finding the "Best" min_{pts} value

In this section, we propose a strategy, Auto-GLOSH, to find the min_{pts} value at the first "elbow" in the ORD–Profile plot that occurs after the first spike with the maximum dissimilarity. We refer to this automatically selected min_{pts} value as m^*. Based on our findings in the previous sections, we expect this value to yield the best or near-best results for GLOSH. We limit our search to a maximum min_{pts} value (m_{max}) of 100. It is common in research studies to assess the outlier detection results obtained using GLOSH within a m_{max} value of 100 [30]. Additionally, as discussed in section 3.1.2, using GLOSH scores at all possible min_{pts} values up to the dataset's size is impractical and does not yield improved results compared to using shorter profiles.

Given an ORD–Profile, $R_{m_{max}}$, we first find the (first in case of ties) maximum dissimilarity score in $R_{m_{max}}$. Let this value be b at index i in the ORD-Profile $R_{m_{max}}$. This score corresponds to the largest peak in the ORD–Profile plot. Next, we estimate a displacement vector starting from the last value in $R_{m_{max}}$ to the selected maximum value in the $R_{m_{max}}$. In each of the sub-figures of Figure 3.21, $\overrightarrow{AB}$ is the displacement



(a) Banana-Global

(b) Circular-Global

(c) LETTER-Clumps

Figure 3.21: Illustrating the process of finding the Elbow of the Outlier Rank Dissimilarity–Profile

vector drawn from the last value in $R_{m_{max}}$, $A = (|R_{m_{max}}|, R_{m_{max}}[|R_{m_{max}}|])$, and the maximum value, $B = (i, b)$, in the ORD–Profile. By visual inspection, one can see that the first "elbow" point in the ORD–Profile plot after the selected maximum value B, is the farthest point from the vector $\overrightarrow{AB}$. Intuitively, we will use the vector $\overrightarrow{AB}$ as tool to find the dissimilarity score in the ORD–Profile that has the largest "shortest distance" (orthogonal distance) to $\overrightarrow{AB}$.

With the help of vector algebra [31, 32], we compute first $\overrightarrow{AB}$ as $B - A$. Next, we compute the orthogonal distances between each dissimilarity score lying between A and B, and the vector $\overrightarrow{AB}$. To do this, we first draw displacement vectors from A to each of the dissimilarity scores in the plot of the scores. In the first column of Figure 3.21, we illustrate a displacement vector $\overrightarrow{AD}$ drawn from A to a dissimilarity score D. This is computed as $D - A$. As presented in the figures, $\overrightarrow{AB}$) and $\overrightarrow{AD}$ can be extended to form a virtual parallelogram $ABCD$. The shortest distance (orthogonal distance) between D and $\overrightarrow{AB}$, represented as $\overrightarrow{DO}$, is the height of the parallelogram if one assumes $\overrightarrow{AB}$ as the base. In classical geometry, the area of a parallelogram is computed as *base* × *height*. As showed in [31], in our setting, the area of a parallelogram can be computed as the magnitude of $\overrightarrow{AD} \times \overrightarrow{AB}$ represents the:

$$||\overrightarrow{AD} \times \overrightarrow{AB}|| = ||\overrightarrow{AB}|| \times ||\overrightarrow{DO}|| \implies ||\overrightarrow{DO}|| = \frac{||\overrightarrow{AD} \times \overrightarrow{AB}||}{||AB||} \tag{3.4}$$

where $||.||$ performs euclidean norm, computing magnitude. We compute the orthogonal distances of each dissimilarity scores lying between A and B in the ORD–Profile, and find the score that has the maximum distance as the elbow point. If r_i is the dissimilarity score at the elbow, as discussed earlier, it is obtained at $min_{pts} = i + 3$, on measuring the dissimilarity between the sequences at $min_{pts} = i+3$ and $min_{pts} = i+2$. Therefore, we return $i + 3$ as the value of m^*.

3.5 Experimental Analysis

3.5.1 Setup

The evaluation in the current section answers the following questions: (I) does the "best" min_{pts} value estimated using Auto-GLOSH, m^*, match the best performance of GLOSH? (II) can the estimated min^* value outperform or match the results achieved using the commonly used min_{pts} values in GLOSH? (III) can GLOSH with the m^* value outperform existing state-of-the-art outlier detection methods?

We compare Auto-GLOSH with the traditional GLOSH algorithm, wherein the user has to pre-define a min_{pts} value. For GLOSH, we use the min_{pts} values that are commonly used in the literature and practice [8, 33–36], including 5, 10, 25, 50, and 100. Additionally, we compare the performance achieved at the "best" min_{pts} value, m^*, as estimated by Auto-GLOSH, with the performance obtained at the best min_{pts} value in GLOSH in range $[2, 100]$ where GLOSH obtain its highest performance. Thirdly, we compare Auto-GLOSH with state-of-the-art outlier detection techniques such as LOF and KNN. Specifically, we compare the results obtained using Auto-GLOSH to those of LOF and KNN across various neighborhood parameter values k in $[5, 10, 25, 50, 100]$.

We again use precision@n to evaluate the performance of each of the methods, as it is a popular evaluation metric in outlier detection [12]. For the current investigation, we use a total of 69 datasets, including **Type 2** and **Type 3** datasets (as described in section 3.1.3). The **Type 2** datasets are the datasets with real inliers and synthetically generated different kinds of outliers. As described in section 3.1.3, we generate local outliers, global outliers, and outlier clumps using the real inlier samples. In addition to that, we also generate datasets including a mix of the different kinds of outliers. To create these datasets, as described in 3.1, we use the inlier samples from the following datasets: MVTec-AD_zipper, HEPATITIS, LETTER, PIMA, STAMPS, VERTEBRAL, VOWELS, WDBC, WINE, WPBC, YEAST, BREASTW, CARDIO, CARDIOTOCOGRAPHY, and 20news_3. The **Type 3** datasets are synthetic datasets obtained from [28], which are: Anisotropic, Banana, and Circular. Similar to what we did for **Type 2** datasets, we generate different kinds of outliers (local, global, or outlier clumps) using the inlier samples in the datasets.

3.5.2 Auto-GLOSH versus GLOSH

In this section we compare the results obtained using Auto-GLOSH against the traditional GLOSH, where a user has to pre-define a min_{pts} value. In Table 3.5, we

Dataset	Outlier Type	Auto-GLOSH	GLOSH					
			$\mathbf{\Gamma_5}$	$\mathbf{\Gamma_{10}}$	$\mathbf{\Gamma_{25}}$	$\mathbf{\Gamma_{50}}$	$\mathbf{\Gamma_{100}}$	*Best
Anisotropic	Local	0.69 (21)	0.61	0.61	0.61	0.46	0.53	0.76 (16)
Banana		0.92 (22)	0.69	0.92	0.92	0.38	0.07	0.92 (4)
Circular		0.76 (26)	0.30	0.46	0.76	0.76	0.76	0.76 (25)
Anisotropic	Outlier Clumps	1.0 (12)	0.26	0.0	1.0	1.0	1.0	1.0 (11)
Banana		1.0 (11)	0.05	1.0	1.0	1.0	1.0	1.0 (6)
Circular		1.0 (9)	1.0	1.0	1.0	1.0	1.0	1.0 (5)
Anisotropic	Global	1.0 (9)	0.15	1.0	1.0	1.0	1.0	1.0 (7)
Banana		1.0 (13)	1.0	1.0	1.0	1.0	1.0	1.0 (4)
Circular		1.0 (12)	0.94	1.0	1.0	1.0	1.0	1.0 (10)

Table 3.5: Auto-GLOSH vs GLOSH: Precision@n obtained on Type 3 Datasets

report three key metrics: (I) the precision@n obtained using the "best" min_{pts} value, m^*, as estimated using Auto-GLOSH (we report the m^* value in brackets), (II) the precision@n obtained using GLOSH with min_{pts} values that are commonly used in the literature and practice (Γ_5 to Γ_{100}), and (III) the best precision@n achieved by GLOSH for a min_{pts} value in range $[2, 100]$ (we report the min_{pts} value in brackets).

In Table 3.5, overall, one can see that in most cases, the estimated "best" min_{pts} value, m^*, results in the best precision@n that is achievable using GLOSH. One can see that for the Anisotropic dataset with local outliers, the estimated "best" min_{pts} value, m^*, results in a higher P@n than any of the commonly used min_{pts} values. Notably, the precision@n obtained using m^* (0.69) is the closest to the precision@n achievable at the best min_{pts} value (0.76). For the Banana dataset with local outliers, one can see that with the min_{pts} values 10 and 25, one can achieve the best precision@n. However, one still has to know which one to choose among $[5, 10, 25, 50, 100]$. Without prior knowledge about the data, one might randomly choose a min_{pts} value that might result in a low precision@n. In contrast, our approach, Auto-GLOSH, estimates the m^* value that attains the best precision@n of 0.92, achievable using GLOSH.

Dataset	**Auto-GLOSH**	**GLOSH**					
		$\mathbf{\Gamma_{5}}$	$\mathbf{\Gamma_{10}}$	$\mathbf{\Gamma_{25}}$	$\mathbf{\Gamma_{50}}$	$\mathbf{\Gamma_{100}}$	***Best**
MVTec-AD_zipper	1.0 (16)	0.0	0.0	1.0	1.0	1.0	1.0 (14)
HEPATITIS	1.0 (5)	1.0	1.0	1.0	1.0	-	1.0 (3)
LETTER	1.0 (77)	0.13	0.50	0.29	0.0	1.0	1.0 (75)
PIMA	1.0 (27)	0.0	0.04	1.0	1.0	1.0	1.0 (25)
STAMPS	1.0 (15)	0.0	0.0	1.0	1.0	1.0	1.0 (13)
VERTEBRAL	1.0 (12)	0.2	1.0	1.0	1.0	1.0	1.0 (10)
VOWELS	1.0 (16)	0.14	0.02	1.0	1.0	1.0	1.0 (14)
WDBC	1.0 (20)	0.16	0.0	1.0	1.0	1.0	1.0 (18)
WINE	1.0 (8)	0.0	1.0	1.0	1.0	1.0	1.0 (6)
WPBC	1.0 (10)	0.0	1.0	1.0	1.0	1.0	1.0 (8)
YEAST	0.42 (9)	0.04	0.44	0.42	0.40	0.40	0.46 (13)
BREASTW	1.0 (25)	0.0	0.0	1.0	1.0	1.0	1.0 (23)
CARDIO	0.79 (16)	0.14	0.15	0.85	0.84	0.79	0.85 (25)
CARDIOTOCOGRAPHY	1.0 (84)	0.13	0.10	0.01	0.0	1.0	1.0 (82)
20news_3	1.0 (31)	0.0	0.0	0.0	1.0	1.0	1.0 (29)

Table 3.6: Auto-GLOSH vs GLOSH: Precision@n obtained on Type 2 Datasets with Outlier Clumps

Similarly, for the Circular dataset with local outliers, choosing a min_{pts} value of 4 or 10 yields low precision@n scores of 0.30 and 0.46, respectively. In contrast, with Auto-GLOSH, one is able to automatically estimate the "best" min_{pts} value of 33, which results in the best precision@n score of 0.76. For Oulier Clumps, one can see that when choosing min_{pts} values 5 and 10, one obtains very low precision@n scores. Auto-GLOSH, on the other hand, automatically achieves the best precision@n score of 1.0.

In Table 3.6, one can see that for most Type 2 datasets with outlier clumps, Auto-GLOSH is able to estimate the min_{pts} value that leads to the best precision@n. In contrast, for the LETTER and CARDIOTOCOGRAPHY datasets, among the

Dataset	Auto-GLOSH	GLOSH					
		$\mathbf{\Gamma_5}$	$\mathbf{\Gamma_{10}}$	$\mathbf{\Gamma_{25}}$	$\mathbf{\Gamma_{50}}$	$\mathbf{\Gamma_{100}}$	*Best
MVTec-AD_zipper	1.0 (16)	0.0	0.0	1.0	1.0	1.0	1.0 (14)
HEPATITIS	1.0 (6)	1.0	1.0	1.0	1.0	-	1.0 (2)
LETTER	1.0 (5)	1.0	1.0	1.0	1.0	1.0	1.0 (3)
PIMA	1.0 (10)	1.0	1.0	1.0	1.0	1.0	1.0 (2)
STAMPS	1.0 (10)	1.0	1.0	1.0	0.93	0.93	1.0 (2)
VERTEBRAL	0.9 (7)	0.9	0.9	0.9	0.9	0.9	0.9 (4)
VOWELS	1.0 (5)	1.0	1.0	1.0	1.0	1.0	1.0 (3)
WDBC	1.0 (8)	1.0	1.0	1.0	1.0	1.0	1.0 (3)
WINE	1.0 (6)	1.0	1.0	1.0	1.0	1.0	1.0 (3)
WPBC	1.0 (10)	1.0	1.0	1.0	1.0	1.0	1.0 (2)
YEAST	1.0 (12)	0.04	0.73	1.0	1.0	1.0	1.0 (12)
BREASTW	1.0 (13)	0.9	1.0	1.0	1.0	1.0	1.0 (10)
CARDIO	1.0 (21)	1.0	1.0	1.0	1.0	1.0	1.0 (4)
CARDIOTOCOGRAPHY	1.0 (7)	1.0	1.0	1.0	1.0	1.0	1.0 (5)
20news_3	1.0 (31)	0.0	0.0	0.0	1.0	1.0	1.0 (29)

Table 3.7: Auto-GLOSH vs GLOSH: Precision@n obtained on Type 2 Datasets with Global Outliers

commonly used min_{pts} value, only $min_{pts} = 100$ (Γ_{100}) one can achieve the best precision@n. Similarly, for the 20news_3 dataset, choosing any min_{pts} value from $[5, 10, 25]$ yields a P@n of 0, while the value m^* identified by Auto-GLOSH results in the best precision@n. For YEAST and CARDIO, although the estimated m^* did not yield the best precision@n, the obtained precision@n is close to the best precision@n that can be achieved using GLOSH.

Similar to clumps, in Table 3.7, one can see that for all the Type 2 datasets with global outliers, the estimated "best" min_{pts} value, m^*, is able to yield the best precision@n score.

In Table 3.8, one can see that in many of the Type 2 datasets with local outliers,

Dataset	**Auto-GLOSH**	**GLOSH**					
		Γ_{5}	**Γ_{10}**	**Γ_{25}**	**Γ_{50}**	**Γ_{100}**	***Best**
MVTec-AD_zipper	1.0 (7)	1.0	1.0	1.0	1.0	1.0	1.0 (2)
HEPATITIS	1.0 (5)	1.0	1.0	1.0	1.0	-	1.0 (3)
LETTER	0.90 (99)	0.93	0.83	0.81	0.9	0.9	0.93 (5)
PIMA	0.86 (11)	0.72	0.86	0.68	0.63	0.59	0.86 (4)
STAMPS	0.35 (11)	0.5	0.35	0.35	0.35	0.35	0.57 (4)
VERTEBRAL	0.7 (11)	0.7	0.7	0.7	0.7	0.7	0.7 (2)
VOWELS	0.88 (15)	0.92	0.88	0.88	0.84	0.88	0.92 (5)
WDBC	0.66 (7)	0.66	0.66	0.66	0.66	0.66	1.0 (2)
WINE	1.0 (7)	1.0	1.0	1.0	1.0	1.0	1.0 (2)
WPBC	0.8 (6)	0.8	0.8	1.0	0.8	1.0	1.0 (17)
YEAST	0.87 (7)	0.89	0.87	0.46	0.46	0.44	0.89 (4)
BREASTW	0.22 (23)	0.0	0.09	0.22	0.22	0.22	0.22 (20)
CARDIO	0.87 (14)	0.56	0.87	0.80	0.79	0.79	0.89 (9)
CARDIOTOCOGRAPHY	0.89 (17)	0.76	0.91	0.89	0.89	0.86	0.93 (6)
20news_3	0.95 (10)	0.0	0.95	0.95	0.95	0.91	0.95 (8)

Table 3.8: Auto-GLOSH vs GLOSH: Precision@n obtained on Type 2 Datasets with Local Outliers

Dataset	**Auto-GLOSH**	**GLOSH**					
		$\mathbf{\Gamma_5}$	$\mathbf{\Gamma_{10}}$	$\mathbf{\Gamma_{25}}$	$\mathbf{\Gamma_{50}}$	$\mathbf{\Gamma_{100}}$	***Best**
MVTec-AD_zipper	1.0 (30)	0.56	0.0	1.0	1.0	1.0	1.0 (20)
HEPATITIS	1.0 (15)	1.0	1.0	1.0	1.0	-	1.0 (5)
LETTER	1.0 (71)	0.25	0.09	0.15	0.04	1.0	1.0 (61)
PIMA	1.0 (22)	0.42	0.14	1.0	1.0	1.0	1.0 (11)
STAMPS	1.0 (26)	0.0	0.0	1.0	1.0	1.0	1.0 (16)
VERTEBRAL	0.95 (20)	0.04	0.95	0.95	0.95	0.95	0.95 (10)
VOWELS	1.0 (23)	0.38	0.51	1.0	1.0	1.0	1.0 (13)
WDBC	0.96 (27)	0.59	0.07	0.96	0.96	0.96	0.96 (17)
WINE	1.0 (18)	0.0	1.0	1.0	1.0	1.0	1.0 (8)
WPBC	1.0 (18)	0.0	1.0	1.0	1.0	1.0	1.0 (8)
YEAST	0.76 (15)	0.76	0.79	0.82	0.82	0.81	0.83 (38)
BREASTW	1.0 (28)	0.22	0.08	1.0	1.0	1.0	1.0 (18)
CARDIO	0.92 (23)	0.43	0.42	0.94	0.94	0.91	0.94 (21)
CARDIOTOCOGRAPHY	1.0 (68)	0.14	0.01	0.0	0.005	1.0	1.0 (58)
20news_3	0.98 (39)	0.08	0.25	0.08	0.98	0.98	1.0 (20)

Table 3.9: Auto-GLOSH vs GLOSH: precision@n obtained on Type 2 Datasets with mixed outliers

Auto-GLOSH is able to yield a precision@n score that is close to the best precision@n achievable using GLOSH. Local outliers are more difficult to detect because they are closer to the inlier clusters. Moreover, in real datasets, there may be inlier samples already that deviate from the majority and behave like local outliers. For datasets such as MVTec-AD_zipper, HEPATITIS, PIMA, BREASTW, and 20news_3, one can see that Auto-GLOSH is able to yield the best precision@n. For datasets such as LETTER, VOWELS, YEAST, CARDIO, and CARDIOTOCOGRAPHY, one can see that precision@n acheived by Auto-GLOSH is close to the best precision@n that is achievable using GLOSH.

We make similar observations on datasets with mixed outliers. In Table 3.9, one can see that for most of the datasets using the automatically selected min_{pts} value,

Dataset	**Outlier Type**	**Auto-GLOSH**	$k = 5$		$k = 10$		$k = 25$		$k = 50$		$k = 100$		***Best GLOSH**
			KNN	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	
Anisotropic	Local	0.69	0.53	0.46	0.53	0.61	0.53	**0.84**	0.53	0.53	0.53	0.53	0.76
Banana		0.92	**1.0**	0.72	**1.0**	**1.0**	**1.0**	**1.0**	0.54	**1.0**	0.36	0.63	0.92
Circular		0.76	**0.88**	0.61	**0.83**	**0.83**	**0.77**	**0.83**	0.72	0.5	0.72	**0.77**	0.76
Anisotropic	Outlier Clumps	1.0	0.89	0.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Banana		1.0	1.0	0.21	1.0	0.31	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Circular		1.0	1.0	0.26	1.0	0.52	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Anisotropic	Global	1.0	1.0	0.15	1.0	0.26	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Banana		1.0	1.0	0.42	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Circular		1.0	1.0	0.47	1.0	0.94	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 3.10: Auto-GLOSH vs KNN and LOF: Precision@n obtained on Type 3 Datasets

m^*, resulted in the best precision@n that is achievable using GLOSH. Notably, for datasets such as CARDIOTOCOGRAPHY and LETTER, one can see that among the commonly used min_{pts} values, one can only achieve the best precision@n when using $min_{pts} = 100$. Without any prior knowledge about the data, it is unlikely to know which min_{pts} value to choose. With Auto-GLOSH, one is able to estimate a min_{pts} value that results in the best or close to the best precision@n achievable by GLOSH.

3.5.3 Auto-GLOSH versus KNN and LOF

In this section we compare the precision@n obtained using Auto-GLOSH against precision@n obtained using the KNN and LOF algorithms. The results are presented between Tables 3.10 to 3.14.

In Table 3.10, it is clear that overall, for local outliers, both KNN and LOF, for some of the commonly used parameter values, can outperform the best results achievable using GLOSH. Our proposed approach, Auto-GLOSH, is built upon GLOSH and is designed to extract optimal results achievable with GLOSH. Consequently, it is expected that Auto-GLOSH may not surpass the performance of KNN and LOF when GLOSH itself does not outperform these methods. For global outliers and clumps, KNN and LOF record high precision@n across some of the commonly used neighborhood parameter k values, however, one can notice that it is not the case when consid-

Dataset	**Auto-GLOSH**	$k = 5$		$k = 10$		$k = 25$		$k = 50$		$k = 100$		***Best GLOSH**
		KNN	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	
MVTec-AD_zipper	1.0	0.48	0.07	0.48	0.03	0.81	0.0	1.0	1.0	1.0	1.0	1.0
HEPATITIS	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	-	-	1.0
LETTER	**1.0**	0.61	0.1	0.66	0.06	0.93	0.17	**1.0**	0.0	**1.0**	0.02	**1.0**
PIMA	1.0	0.56	0.12	0.6	0.04	0.98	0.0	1.0	0.0	1.0	1.0	1.0
STAMPS	1.0	0.41	0.29	0.64	0.22	0.90	0.09	1.0	0.93	1.0	1.0	1.0
VERTEBRAL	1.0	0.33	0.23	0.66	0.23	1.0	0.95	1.0	0.95	1.0	1.0	1.0
VOWELS	1.0	0.63	0.17	0.68	0.09	0.90	0.43	1.0	1.0	1.0	1.0	1.0
WDBC	1.0	0.36	0.11	0.38	0.25	0.80	0.25	1.0	0.94	1.0	1.0	1.0
WINE	1.0	0.58	0.08	0.91	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WPBC	1.0	0.4	0.0	0.66	0.0	1.0	0.93	1.0	1.0	1.0	1.0	1.0
YEAST	0.42	0.48	0.15	0.47	0.17	0.54	0.26	0.55	0.44	**0.57**	0.51	0.46
BREASTW	**1.0**	0.31	0.0	0.38	0.0	0.7	0.0	**1.0**	0.04	**1.0**	0.68	**1.0**
CARDIO	0.79	0.62	0.06	0.63	0.07	0.75	0.25	**0.86**	0.50	**0.85**	**0.85**	**0.85**
CARDIOTOCOGRAPHY	**1.0**	0.58	0.07	0.58	0.05	0.68	0.18	0.87	0.22	0.98	0.01	**1.0**
20news_3	1.0	0.46	0.01	0.39	0.03	0.53	0.06	0.74	0.01	1.0	1.0	1.0

Table 3.11: Auto-GLOSH vs KNN and LOF: Precision@n obtained on Type 2 Datasets with Outlier clumps

ering local outliers. For local outliers, as the neighborhood parameter k increases, the precision@n achieved by KNN tends to decrease, whereas, the precision@n obtained by LOF increases initially and then begins to decline for $k \geq 50$. These results show the sensitivity of both KNN and LOF w.r.t. the value of k.

The results for Type 2 datasets with outlier clumps are shown in Table 3.11. One can see that LOF does not perform well until the k value is equal to 100. Even with $k = 100$, LOF results in a precision@n smaller than Auto-GLOSH in many cases. Examples include LETTER, BREASTW, and CARDIOTOCOGRAPHY. Notably, for LETTER and CARDIOTOCOGRAPHY, the precision@n decreases as we increase the value of k. Examples like these makes it unclear which k value to choose for LOF when there underlying data distribution is unknown. This underscores the need for Auto-GLOSH, which achieves a good precision@n on most of the datasets without having to choose a parameter value. KNN is found to be the most competitive to Auto-GLOSH across all the datasets. However, it only performs well when the k value is chosen carefully, which is difficult without knowing the underlying data distribution. One might think that taking a high value for k (typically $k \geq 50$) will

Dataset	**Auto-GLOSH**	$\boldsymbol{k = 5}$		$\boldsymbol{k = 10}$		$\boldsymbol{k = 25}$		$\boldsymbol{k = 50}$		$\boldsymbol{k = 100}$		***Best GLOSH**
		KNN	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	
MVTec-AD_zipper	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0
HEPATITIS	1.0	1.0	0.42	1.0	1.0	1.0	1.0	1.0	1.0	-	-	1.0
LETTER	1.0	1.0	0.35	1.0	0.52	1.0	0.82	1.0	0.94	1.0	0.98	1.0
PIMA	1.0	1.0	0.42	1.0	0.58	1.0	0.84	1.0	0.98	1.0	0.98	1.0
STAMPS	1.0	1.0	0.25	1.0	0.51	1.0	0.74	0.96	0.96	0.96	1.0	1.0
VERTEBRAL	0.9	**0.95**	0.23	**0.95**	0.47	**0.95**	0.9	**0.95**	**0.95**	**0.95**	**0.95**	0.9
VOWELS	1.0	1.0	0.82	1.0	0.97	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WDBC	1.0	1.0	0.08	1.0	0.19	1.0	0.52	1.0	0.94	1.0	1.0	1.0
WINE	1.0	1.0	0.08	1.0	0.16	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WPBC	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
YEAST	1.0	1.0	0.27	1.0	0.29	1.0	0.36	1.0	0.6	1.0	0.78	1.0
BREASTW	1.0	1.0	0.0	1.0	0.02	1.0	0.0	1.0	0.63	1.0	0.9	1.0
CARDIO	1.0	1.0	0.83	1.0	0.89	1.0	0.98	1.0	0.99	1.0	1.0	1.0
CARDIOTOCOGRAPHY	1.0	1.0	0.02	1.0	0.07	1.0	0.18	1.0	0.35	1.0	0.67	1.0
20news_3	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.98	1.0

Table 3.12: Auto-GLOSH vs KNN and LOF: Precision@n obtained on Type 2 Datasets with Global outliers

work all time. However, this is not the case. For example, in BREASTW, one can achieve the highest precision@n (equal to 1) with $k = 50$, while for 20news_3, KNN only achieves a precision@n of 1 when $k = 100$.

We make similar observations for global outliers, as presented in Table 3.12. However, global outliers are fairly easy to detect for all methods. For most datasets, KNN obtains a high precision@n for most values of k. For LOF, one can see that a high value of k (typically $k = 100$) works well for detecting global outliers. However, our approach Auto-GLOSH is able to achieve high precision@n without selecting any parameter.

In Table 3.13 one can see that in the datasets where KNN or LOF (with some typical values for k) outperforms Auto-GLOSH, they also outperform the best precision@n achievable using GLOSH. Examples include STAMPS, VERTEBRAL, BREASTW, and CARDIO, where KNN and LOF outperform GLOSH for some parameter values. However, for many datasets, Auto-GLOSH achieves a precision@n close to the best obtained by KNN and LOF. Additionally, precision@n results appear sensitive to the choice of k in many cases for KNN and LOF. For example, in the LETTER dataset,

Dataset	**Auto-GLOSH**	$\boldsymbol{k=5}$		$\boldsymbol{k=10}$		$\boldsymbol{k=25}$		$\boldsymbol{k=50}$		$\boldsymbol{k=100}$		***Best GLOSH**
		KNN	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	**KNN**	**LOF**	
MVTec-AD_zipper	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
HEPATITIS	1.0	1.0	0.0	1.0	0.75	1.0	1.0	1.0	1.0	-	-	1.0
LETTER	0.90	0.95	0.95	0.95	**1.0**	0.90	0.98	0.90	0.98	0.90	0.98	0.93
PIMA	0.86	0.86	0.90	0.86	**0.95**	0.68	0.90	0.63	0.90	0.59	0.68	0.86
STAMPS	0.35	0.5	**0.78**	0.35	0.64	0.35	0.5	0.35	0.5	0.35	0.42	0.57
VERTEBRAL	0.7	0.71	0.64	**0.78**	**0.78**	**0.78**	**0.78**	**0.78**	**0.78**	**0.78**	**0.78**	0.7
VOWELS	0.88	0.92	0.92	0.88	0.92	0.88	**1.0**	0.84	**1.0**	0.88	**1.0**	0.92
WDBC	0.66	0.66	**1.0**	0.66	**1.0**	0.66	**1.0**	0.66	**1.0**	0.66	**1.0**	**1.0**
WINE	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WPBC	0.8	0.8	0.4	0.8	0.8	**1.0**	0.8	0.8	**1.0**	**1.0**	**1.0**	**1.0**
YEAST	0.87	0.80	0.63	0.76	0.78	0.70	0.73	0.70	0.72	0.69	0.71	**0.89**
BREASTW	0.22	0.27	0.09	0.29	0.13	0.38	0.29	0.38	**0.5**	0.34	0.36	0.22
CARDIO	0.87	0.94	0.81	0.89	0.93	0.82	**0.98**	0.81	0.96	0.8	0.93	0.89
CARDIOTOCOGRAPHY	0.89	0.93	0.71	0.91	0.89	0.90	0.95	0.90	**0.96**	0.87	0.95	0.93
20news_3	0.95	0.95	0.73	0.95	0.95	0.95	0.95	0.95	0.95	0.91	0.95	0.95

Table 3.13: Auto-GLOSH vs KNN and LOF: Precision@n obtained on Type 2 Datasets with local outliers

increasing k decreases precision@n for KNN but keeps it stable for LOF, and for the PIMA dataset, precision@n for both KNN and LOF changes as k increases, making it unclear which k value should be chosen beforehand.

For the mixed outliers in Type 2 datasets, as presented in Table 3.14, one can see that our approach, Auto-GLOSH, consistently achieves precision@n scores that are either the best or very close to the best precision@n obtainable compared with KNN and LOF for some of the commonly used k values. However, the uncertainty about which k value to choose for LOF and KNN can be seen here as well. For example, in BREASTW, KNN achieves a precision@n of 1 for $k = 50$. However, for datasets such as CARDIO and YEAST, one can see that the precision@n obtained using KNN decreases as the k value is increased from 50 to 100. LOF performs well on some data sets such as MVTec-AD_zipper, PIMA, STAMPS, VOWELS, WDBC, and 20news_3, when choosing a high value of k, while for other datasets such as CARDIOTOCOGRAPHY and LETTER, the precision@n of LOF decreases as the value of k increases, and for datsets such as PIMA and STAMPS, the precision@n for LOF decreases at first and then increases. Without having any prior knowledge about

Dataset	Auto-GLOSH	$k = 5$		$k = 10$		$k = 25$		$k = 50$		$k = 100$		*Best GLOSH
		KNN	LOF	KNN	LOF	KNN	LOF	KNN	LOF	KNN	LOF	
MVTec-AD_zipper	1.0	0.86	0.26	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
HEPATITIS	1.0	1.0	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
LETTER	**1.0**	0.90	0.43	0.96	0.44	**1.0**	0.34	**1.0**	0.29	**1.0**	0.13	**1.0**
PIMA	1.0	0.94	0.29	1.0	0.25	1.0	0.20	1.0	0.44	1.0	1.0	1.0
STAMPS	1.0	0.93	0.18	0.96	0.12	1.0	0.15	1.0	0.84	1.0	1.0	1.0
VERTEBRAL	0.95	0.95	0.23	0.95	0.19	0.95	0.95	0.95	0.95	0.95	0.95	0.95
VOWELS	1.0	0.95	0.55	0.99	0.81	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WDBC	0.96	0.77	0.67	0.96	0.25	0.96	0.07	0.96	0.96	0.96	1.0	0.96
WINE	1.0	1.0	0.0	1.0	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WPBC	1.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
YEAST	0.76	0.76	0.35	0.79	0.47	0.82	0.61	0.82	0.67	0.81	0.75	**0.83**
BREASTW	**1.0**	0.58	0.0	0.72	0.0	0.97	0.0	**1.0**	0.06	1.0	0.87	**1.0**
CARDIO	0.92	0.86	0.39	0.87	0.43	0.94	0.77	0.94	0.92	0.91	**0.95**	0.94
CARDIOTOCOGRAPHY	**1.0**	0.86	0.33	0.91	0.39	0.98	0.33	0.99	0.13	**1.0**	0.08	**1.0**
20news_3	0.98	0.88	0.41	0.83	0.61	0.98	0.26	0.98	0.1	0.98	**1.0**	**1.0**

Table 3.14: Auto-GLOSH vs KNN and LOF: Precision@n obtained on Type 2 Datasets with mixed outliers

the underlying data distribution, it is difficult to choose a value of k that can yield the best performance. For GLOSH, we have addressed this problem using our approach Auto-GLOSH that can provide the best or nearly the best precision@n achievable by GLOSH without choosing any parameters.

3.6 Conclusion

This chapter successfully answers the first research question of this thesis: "How to select a min_{pts} value to yield the best or nearly the best results for GLOSH by utilizing the GLOSH scores obtained for a range of different min_{pts} values?" Our investigation showed that outlier GLOSH–Profiles behave differently than inlier profiles. Notably, beyond a specific min_{pts} value, the outlier profiles start showing higher GLOSH scores than inliers. We make two major observations at this specific min_{pts} value: (I) GLOSH scores within the profiles start changing at a similar rate, (II) GLOSH achieves its best or nearly the best performance, in terms of precision@n. We use these observations to develop an unsupervised strategy, Auto-GLOSH (Automatic GLOSH parameter selection based on outlier profiles), to find the min_{pts} value where the GLOSH scores in the profiles start changing at a similar rate—we refer to it as m^*. We evaluate GLOSH using the m^* value on various datasets, including Type 2 datasets with real inliers and synthetically generated outliers (local, global, and clumps). Overall, our evaluation showed that Auto-GLOSH is able to match the best or nearly best precision@n that can be achieved using GLOSH. When comparing Auto-GLOSH with the commonly used min_{pts} values in GLOSH, we found that the precision@n differs a lot as we change the values. Similarly, when comparing with KNN and LOF, we found that among the commonly used neighborhood k values, the precision@n can differ a lot. This shows the importance of Auto-GLOSH in achieving an automated method for selecting a parameter that can result in high precision@n values for one of the methods, GLOSH. Studying whether a similar approach could be effective for KNN and LOF is left for future work.

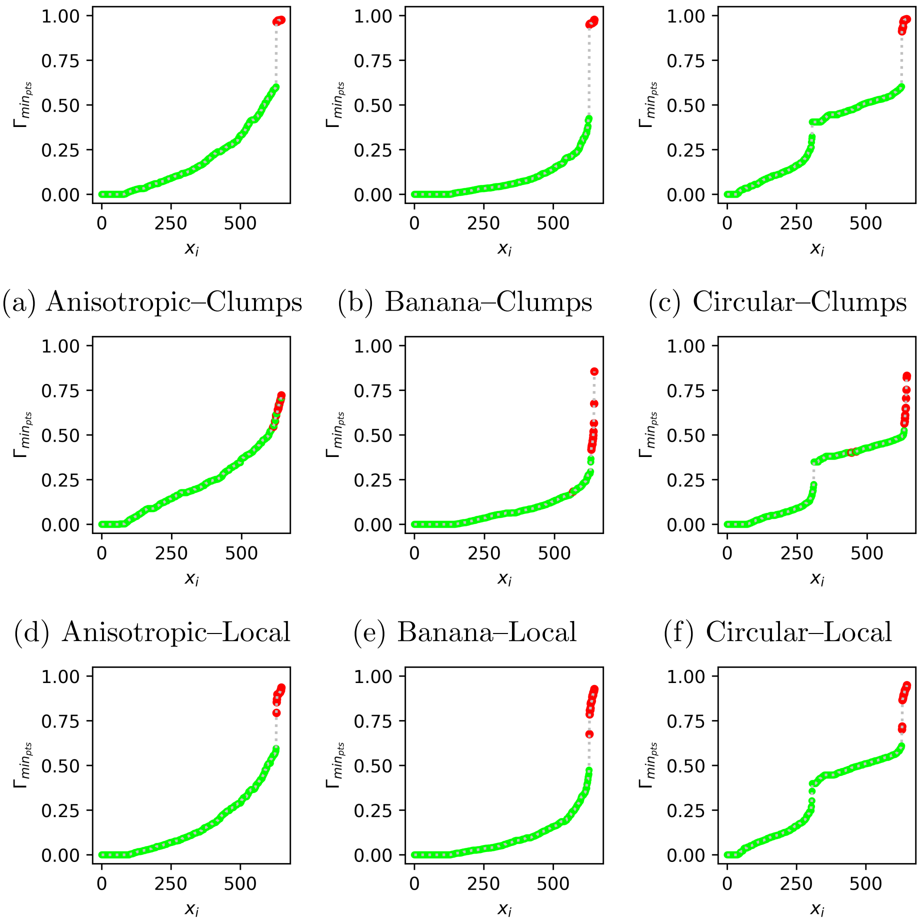
Chapter 4

Unsupervised Labelling of Potential Outliers using GLOSH scores

4.1 GLOSH scores at the "Best" min_{pts}

In this section we study the sorted distribution of the GLOSH scores obtained at the "best" min_{pts} value, m^*, estimated using Auto-GLOSH. In practice, we do not know n, the number of outliers, nor do we know a fixed threshold suitable to label inliers and potential outliers. To address this issue, we first need to understand the sorted sequence of the GLOSH scores obtained using the m^* value, and investigate the possibility of a pattern or gap in the GLOSH scores that can indicate a suitable threshold. In the context of GLOSH scores, a "gap" can be defined as a noticeable deviation or abrupt increase in an sorted sequence of GLOSH scores. Such a gap in the GLOSH scores may indicate a sudden decrease in the density of points which may indicate the possibility of outliers in a dataset.

As shown in the previous chapter, the m^* value yields the best precision@n for majority of the datasets. Thus, we expect GLOSH scores mostly of outliers to be at the end of the sorted sequence of GLOSH scores. In Figure 4.1, we present the sorted sequence of GLOSH scores obtained using the m^* value on the **Type 3** datasets. The x-axis arranges the datapoints x_i in ascending order of their GLOSH scores



(a) Anisotropic–Clumps (b) Banana–Clumps (c) Circular–Clumps

(d) Anisotropic–Local (e) Banana–Local (f) Circular–Local

(g) Anisotropic–Global (h) Banana–Global (i) Circular–Global

Figure 4.1: Sorted Sequence of GLOSH scores at $min_{pts} = m^*$

and the y-axis represents the corresponding GLOSH scores $\Gamma_{m^*}(x_i)$. The green dots represents the GLOSH scores Γ_{m^*} of inliers and red dots represents the scores of outliers. Similar to what we observed for the Outlier Rank Dissimilarity–Profile (ORD–Profile) in section 3.3, we observe an "knee like" structure, this time on the right side of the plots. Initially, the GLOSH scores increase following an "almost linear" trend and after a certain point, the GLOSH scores start deviating from this trend and show an accelerated growth. The GLOSH scores of points at the beginning of the sorted list are 0 (which is the lower bound of GLOSH scores). These scores are assigned to the points that are the densest points in their own clusters, and after that, since the GLOSH scores are arranged in an ascending order, they will follow an increasing trend. However, a more or less sudden, “knee like” increase in the GLOSH scores should only happen, when there are points with significantly different densities from the majority of points. Therefore, the GLOSH score at the “knee” may seem to be a good option for a threshold to label inliers and potential outliers. However, one can see that there are also a few inlier GLOSH scores after the “knee”. The inlier data may itself contain points that satisfy the properties of local outliers. For instance, if the inliers follow a Gaussian distribution, there can be datapoints that are, for instance, more than three standard deviations away and those can already behave like local outliers. In a strict sense, such datapoints are indeed local outliers, but they maybe labelled as inliers.

4.2 Automatically Labelling Potential Outliers

In this section we design “Potential Outlier Labelling AppRoach” (POLAR)—a fully unsupervised approach to label potential outliers by using the sorted sequence of GLOSH scores obtained at the “best” min_{pts} value, m^*, estimated using Auto-GLOSH. Our approach does not need to specify n or any other parameter.

As we observed in the previous section, the sorted sequence of GLOSH scores forms a “knee” when the GLOSH scores are sorted from smallest to largest. When the GLOSH scores are computed using a min_{pts} value that would give high precision@n, the majority of the outliers are then (necessarily) at the tail end of the sorted list. Taking the “knee” GLOSH score as a threshold has the potential to correctly label the outliers and most of the inliers. Therefore, in a first step, our approach finds the “knee” GLOSH score in the sorted sequence as the threshold. However, with this threshold, inliers that behave like local outliers may increase the number of false positives as their GLOSH scores may lie beyond the “knee”. Therefore, we propose an additional strategy to estimate a threshold beyond the “knee” to lower the chances of false positives. Taking a threshold value that is too high has a risk of labelling outliers as inliers. Therefore, we select the value so that it is only slightly greater than the “knee” GLOSH score.

4.2.1 Finding the “knee” GLOSH score

In this section, we use a similar strategy to find the “knee” GLOSH score as previously done in section 3.4 to locate the “elbow” in the ORD–Profile. As illustrated in Figure 4.2, we firstly compute the displacement vector $\overrightarrow{AB}$, as the difference between the last GLOSH score B and the first GLOSH score A in the sorted sequence.

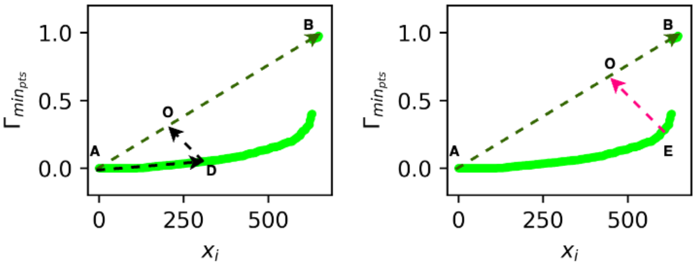


Figure 4.2: Banana Dataset with Outlier Clumps: Finding the “knee” point in the sorted sequence of GLOSH scores at $min_{pts} = m^*$. The GLOSH score E is the “knee” GLOSH score that has the maximum orthogonal distance to $\overrightarrow{AB}$.

Next, we compute the orthogonal distances between each GLOSH score in the sorted sequence and $\overrightarrow{AB}$. To do so, for every GLOSH score D in the sequence, we first compute $D - A$ to estimate the displacement vector $\overrightarrow{AD}$ and subsequently, the orthogonal distance as $\frac{||\overrightarrow{AD} \times \overrightarrow{AB}||}{||\overrightarrow{AB}||}$. The GLOSH score in the sequence with the maximum orthogonal distance is identified as the “knee” GLOSH score.

4.2.2 Adjusting the Inlier Threshold

In this section we propose a strategy to find a threshold beyond the “knee” GLOSH score in the sorted sequence. To do so, we first aim to capture the progression of the inlier GLOSH scores in the sorted sequence. We use a simple linear regression model [37] to estimate the trend of the inlier GLOSH scores in the sequence. The idea here is to capture the “almost” linear trend of the GLOSH scores observed at the beginning of the sorted sequence for inliers.

The linear regression model estimates a quantitative value $\hat{Y}$ based on a predictor variable X, assuming there is approximately a linear relationship between X and Y. The linear regression model can be represented using the following equation:

$$\hat{Y} = \beta_0 + \beta_1 \times X \tag{4.1}$$

For a dataset with N points, the indexes of each GLOSH score in the sorted sequence can be represented as $[x_1, x_2, \ldots, x_N]$. We use the GLOSH scores until the "knee" GLOSH score in the sorted sequence to estimate the model coefficients β_0 and β_1. The optimal values for β_0 and β_1 are estimated by minimizing the Mean Squared Error (MSE) between the estimated $\hat{Y}$ and the known Y.

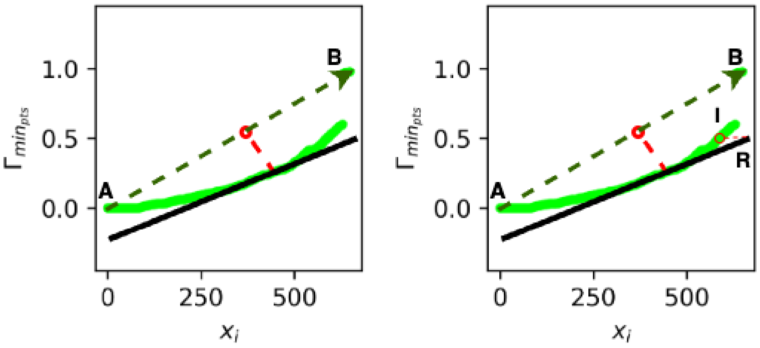


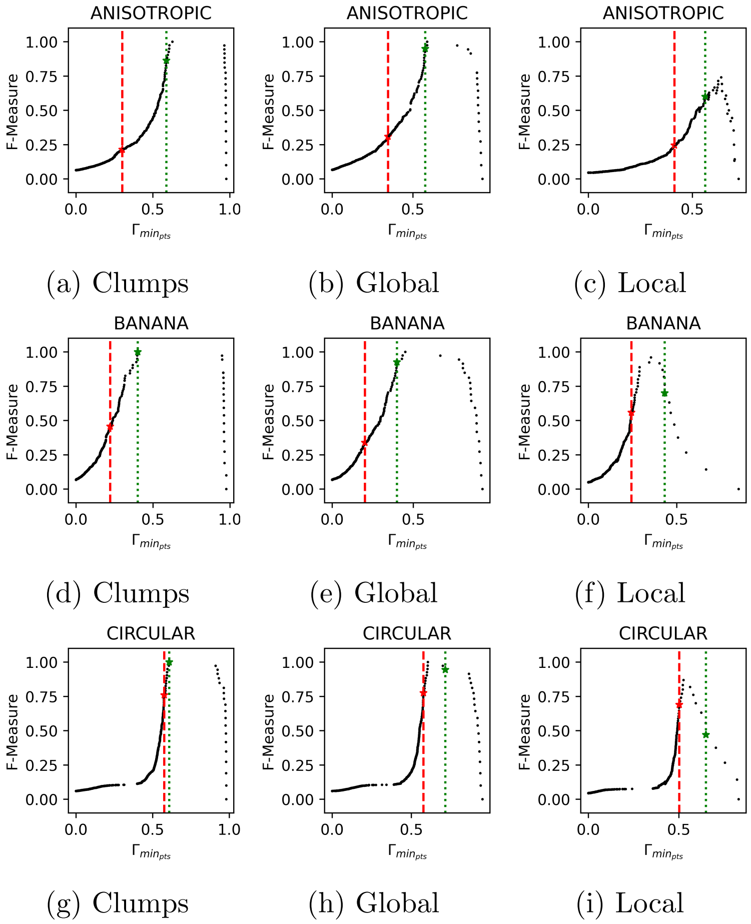
Figure 4.3: Illustration of the adjusted threshold estimation on Anisotropic dataset with Outlier Clumps at $min_{pts} = m^*$: Using a linear regression on the sorted sequence of GLOSH scores to estimate the highest GLOSH score R. Then, we find the GLOSH score I that is closest to R.

In Figure 4.3, the estimated regression line is represented as a black solid line. Using the regression model we estimate the highest GLOSH score R at index x_N where N is the total number of datapoints. Intuitively, this estimated GLOSH score R reflects the score that could have been reached if there were only inliers in the dataset and the scores in the sequence followed the trend that was observed up to the "knee". Then, we search the GLOSH score that is the most similar to R in the sorted sequence between the "knee" and x_N. As shown in Figure 4.3, for the plotted sequence of GLOSH scores, the score I is closest to R and is taken as the adjusted threshold.

4.3 Experimental Analysis

4.3.1 Setup

To evaluate our approach, POLAR, we use recall, F-measure, and Kubat's G-Mean metric [38, 39]. Recall measures the fraction of outliers labelled correctly out of all the outliers. F-measure is the harmonic mean of recall and precision. Precision is computed as the fraction of correctly predicted outliers out of all the instances that are labelled as outliers. Kubat's G-Mean metric computes a geometric mean between recall and True Negative Rate (TNR). TNR is the fraction of inliers labelled correctly out of all inliers. The G-Mean metric is particularly useful when the datasets are imbalanced, a common occurrence in outlier detection studies [40, 41]. It is useful because its considers the performance in detecting both outlier and inlier samples and provides a balanced measure. For the current investigation, we use a total of 84 datasets, including **Type1**, **Type 2** and **Type 3** datasets (as described in section 3.1.3). The **Type 1** datasets are real one-class classification datasets. The datasets used are: MVTec-AD_zipper, HEPATITIS, LETTER, PIMA, STAMPS, VERTEBRAL, VOWELS, WDBC WINE, WPBC, YEAST, BREASTW, CARDIO, CARDIOTOCOGRAPHY, and 20news_3. The **Type 2** datasets are the datasets with real inliers and synthetically generated different kinds of outliers (local, global, and outlier clumps). To create these datasets, we use the inlier class samples from the one-class classification datasets. The **Type 3** datasets are synthetic datasets obtained from [28], which are: Anisotropic, Banana, and Circular. Similar to what we did for **Type 2** datasets, we generate different kinds of outliers (local, global, or outlier clumps) using the inlier samples provided in the dataset.



(a) Clumps (b) Global (c) Local

(d) Clumps (e) Global (f) Local

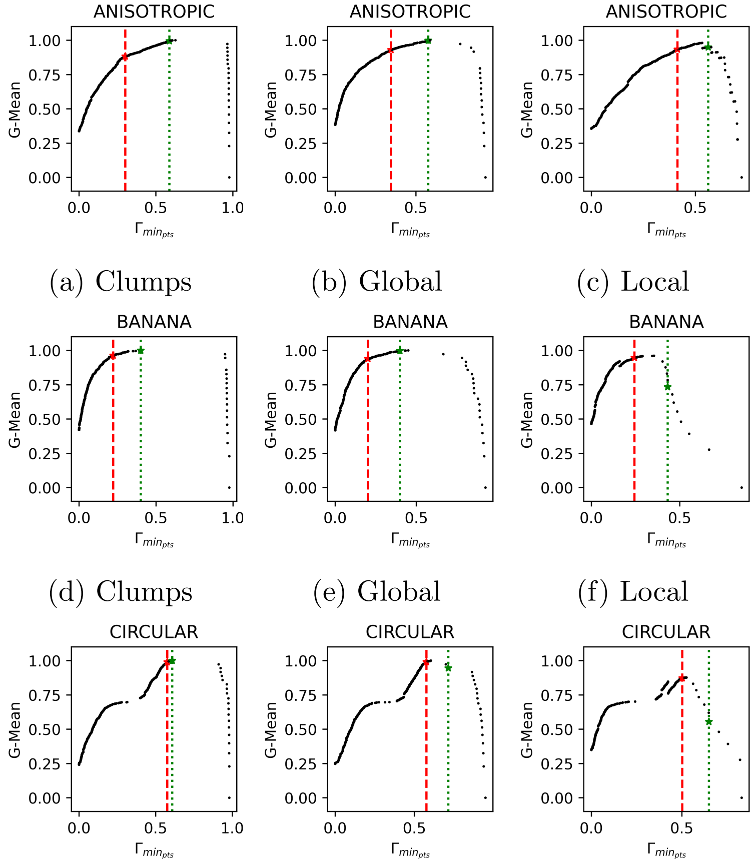
(g) Clumps (h) Global (i) Local

Figure 4.4: F-Measure obtained on Type 3 Datasets across different thresholds at $min_{pts} = m^*$: The red dashed line signifies the F-measure obtained when using the knee GLOSH score as the threshold, while the green dotted line signifies the F-measure obtained when using the adjusted threshold.

4.3.2 Results

In the following figures we plot the value of each measure as a function of every possible threshold on a dataset. Each GLOSH score in the sorted sequence of GLOSH scores is taken as a threshold and points with a GLOSH score below the threshold are labelled as inlier. The X-axis represents the GLOSH scores taken as thresholds, while the Y-axis represents the value of a measure obtained with that threshold.

Figure 4.4 shows the result for the F-Measure on Type 3 datasets. Firstly, one can see that the "perfect" threshold—where the F-Measure is 1 or close to 1—varies across datasets. For example, in the Anisotropic dataset with outlier clumps (Figure



(a) Clumps (b) Global (c) Local

(d) Clumps (e) Global (f) Local

(g) Clumps (h) Global (i) Local

Figure 4.5: G-Mean obtained on Type 3 Datasets across different thresholds at $min_{pts} = m^*$: The red dashed line signifies the G-Mean obtained when using the knee GLOSH score as the threshold, while the green dotted line signifies the G-Mean obtained when using the adjusted threshold.

4.4c), a threshold of 0.5 results in a very low F-measure (less than 0.4). Whereas, for the banana dataset with global outliers (Figure 4.4e), a threshold of 0.5 results in a F-Measure of 1. This shows that one cannot assume a fixed threshold that will be applicable across all datasets. Across most of the Type 3 datasets, one can see that the adjusted threshold estimated using our method, POLAR, consistently leads to an F-Measure of 1 or close to that. Taking the unadjusted knee GLOSH score as the threshold results in a lower F-Measure in most of the datasets. As discussed earlier, this happens because there may exist inliers with GLOSH scores beyond the knee, leading to an increase in false positives.

Dataset	**POLAR**								***Best**	
	Knee				**Adjusted**					
	Recall	**TNR**	**F-Measure**	**G-Mean**	**Recall**	**TNR**	**F-Measure**	**G-Mean**	**F-Measure**	**G-Mean**
MVTec-AD_zipper	1.0	0.89	0.49	0.94	1.0	0.98	0.87	0.99	1.0	1.0
HEPATITIS	1.0	0.97	0.75	0.98	1.0	1.0	1.0	1.0	1.0	1.0
LETTER	1.0	0.93	0.61	0.96	1.0	0.99	0.94	0.99	1.0	1.0
PIMA	1.0	0.93	0.61	0.96	1.0	0.99	0.96	0.99	1.0	1.0
STAMPS	1.0	0.86	0.41	0.92	1.0	0.97	0.79	0.98	1.0	1.0
VERTEBRAL	1.0	0.83	0.37	0.91	1.0	0.96	0.71	0.98	1.0	1.0
VOWELS	1.0	0.88	0.46	0.94	1.0	0.98	0.87	0.99	1.0	1.0
WDBC	1.0	0.79	0.33	0.89	1.0	0.98	0.83	0.99	1.0	1.0
WINE	1.0	0.85	0.41	0.92	1.0	0.99	0.92	0.99	1.0	1.0
WPBC	1.0	0.79	0.34	0.89	1.0	0.98	0.89	0.99	1.0	1.0
YEAST	1.0	0.78	0.31	0.88	0.40	0.96	0.40	0.62	0.56	0.95
BREASTW	1.0	0.64	0.21	0.80	1.0	0.88	0.46	0.94	1.0	1.0
CARDIO	1.0	0.83	0.37	0.91	0.71	0.99	0.76	0.83	0.85	0.99
CARDIOTOCOGRAPHY	1.0	0.92	0.55	0.95	1.0	0.99	0.96	0.99	1.0	1.0
20news_3	1.0	0.87	0.43	0.93	1.0	0.97	0.78	0.98	1.0	1.0

Table 4.1: Evaluating POLAR on Type 2 datasets with outlier clumps at $min_{pts} = m^*$: We compare the performance achieved by POLAR with the best performance that can be achieved across all possible thresholds.

In Figure 4.5 we plot the G-Mean for every possible threshold on a dataset. Figure 4.5 shows the result for the G-Mean Measure on Type 3 datasets. One can observe that for most of the Type 3 datasets, the adjusted threshold estimated by POLAR consistently yields a G-Mean of 1 or close to 1 in most cases. This implies that the adjusted threshold is often able to separate most of the inlier and outlier GLOSH scores. For datasets such as Circular and Banana with local outliers, one can see that with the adjusted threshold there is a drop in the G-Mean value compared to the threshold placed at the "knee" GLOSH score. This happens when certain inliers get GLOSH scores very close to the local outliers. In those cases, the "knee" GLOSH score tends to perform better as a threshold.

In Table 4.1 we show the recall, true negative rate, F-measure, and G-Mean obtained using the knee GLOSH score and the adjusted threshold on the Type 2 datasets with outlier clumps. Additionally, we also show the F-measure and G-mean obtained using the best possible threshold. One can see that for each of the datasets, the adjusted threshold estimated by our approach, POLAR, leads to a better F-measure

Dataset	POLAR								*Best	
	Knee				Adjusted					
	Recall	TNR	F-Measure	G-Mean	Recall	TNR	F-Measure	G-Mean	F-Measure	G-Mean
MVTec-AD_zipper	1.0	0.93	0.60	0.96	1.0	1.0	1.0	1.0	1.0	1.0
HEPATITIS	1.0	0.95	0.66	0.97	1.0	1.0	1.0	1.0	1.0	1.0
LETTER	1.0	0.93	0.61	0.96	1.0	0.99	0.98	0.99	1.0	1.0
PIMA	1.0	0.95	0.70	0.97	1.0	0.99	0.94	0.99	1.0	1.0
STAMPS	1.0	0.81	0.34	0.90	1.0	0.96	0.73	0.98	1.0	1.0
VERTEBRAL	1.0	0.81	0.34	0.90	1.0	0.95	0.68	0.97	0.95	0.99
VOWELS	1.0	0.96	0.73	0.98	1.0	0.99	0.95	0.99	1.0	1.0
WDBC	1.0	0.82	0.36	0.90	1.0	0.98	0.85	0.99	1.0	1.0
WINE	1.0	0.87	0.44	0.93	1.0	1.0	1.0	1.0	1.0	1.0
WPBC	1.0	0.82	0.18	0.90	1.0	0.96	0.54	0.98	1.0	1.0
YEAST	1.0	0.79	0.32	0.89	0.75	0.98	0.69	0.86	0.73	0.96
BREASTW	1.0	0.67	0.23	0.82	1.0	0.90	0.51	0.95	1.0	1.0
CARDIO	1.0	0.86	0.42	0.93	1.0	0.99	0.94	0.99	1.0	1.0
CARDIOTOCOGRAPHY	1.0	0.94	0.65	0.97	1.0	0.99	0.96	0.99	1.0	1.0
20news_3	1.0	0.87	0.45	0.93	1.0	0.97	0.79	0.98	1.0	1.0

Table 4.2: Evaluating POLAR on Type 2 datasets with global outliers at $min_{pts} = m^*$: We compare the performance achieved by POLAR with the best performance that can be achieved across all possible thresholds.

than taking the knee GLOSH score. Although we get an increased F-measure, one can see that for datasets such as YEAST and CARDIO, the recall drops from what we could get with the knee GLOSH score as the threshold. This shows that the adjusted threshold increases the precision (reduces false positives) at the cost of recall (increases false negatives) in those datasets. Therefore, one can see a drop in the G-Mean with the adjusted threshold when compared to the "knee" GLOSH score. However, across most of the datasets in Table 4.1, one can see that with the adjusted threshold we achieve F-Measure and G-Mean scores that are close to the best that can be achieved if one somehow knows the best possible threshold.

In Table 4.2 we show the recall, true negative rate, F-measure, and G-Mean obtained using the knee GLOSH score and the adjusted threshold on the Type 2 datasets with global outliers. We also show the F-measure and G-mean obtained using the best possible threshold. Overall, one can see that even for global outliers, the adjusted threshold tends to be a better choice than the knee. With the adjusted threshold, one can see that except for the YEAST dataset, the recall does not drop for any of the

Dataset	POLAR								*Best	
	Knee				Adjusted					
	Recall	TNR	F-Measure	G-Mean	Recall	TNR	F-Measure	G-Mean	F-Measure	G-Mean
MVTec-AD_zipper	1.0	0.94	0.27	0.97	1.0	0.98	0.66	0.99	1.0	1.0
HEPATITIS	1.0	0.32	0.11	0.57	0.0	1.0	0.0	0.0	1.0	1.0
LETTER	1.0	0.95	0.65	0.97	0.85	0.99	0.91	0.92	0.92	0.98
PIMA	1.0	0.78	0.29	0.88	0.90	0.99	0.88	0.95	0.93	0.99
STAMPS	1.0	0.86	0.40	0.93	0.35	0.98	0.41	0.59	0.53	0.95
VERTEBRAL	1.0	0.82	0.35	0.91	0.8	0.96	0.61	0.87	0.77	0.96
VOWELS	1.0	0.93	0.35	0.96	0.96	0.98	0.74	0.97	0.92	0.99
WDBC	1.0	0.82	0.08	0.91	1.0	0.96	0.3	0.98	0.85	0.99
WINE	1.0	0.90	0.26	0.95	1.0	0.96	0.5	0.98	1.0	1.0
WPBC	1.0	0.80	0.25	0.89	1.0	0.96	0.66	0.98	0.88	0.99
YEAST	0.97	0.80	0.33	0.88	0.97	0.97	0.77	0.97	0.91	0.98
BREASTW	1.0	0.50	0.16	0.71	0.72	0.85	0.31	0.78	0.36	0.83
CARDIO	1.0	0.84	0.39	0.92	0.98	0.96	0.73	0.97	0.88	0.98
CARDIOTOCOGRAPHY	1.0	0.89	0.48	0.94	0.79	0.99	0.84	0.88	0.91	0.98
20news_3	1.0	0.89	0.43	0.94	1.0	0.97	0.75	0.98	0.97	0.99

Table 4.3: Evaluating POLAR on Type 2 datasets with local outliers at $min_{pts} = m^*$: We compare the performance achieved by POLAR with the best performance that can be achieved across all possible thresholds.

other datasets when compared to the knee. This means that among the "potential outliers" labelled using the adjusted threshold, all the "true outliers" have a larger GLOSH score than the threshold most of the times. Overall, one can see that with the adjusted threshold, we achieve high values for F-measure and G-Mean that are close to the best possible values across many datasets. However, for both global outliers and clumps, choosing the "knee" GLOSH score as the threshold always results in high recall.

In Table 4.3 we show the recall, true negative rate, F-measure, and G-Mean obtained using the knee GLOSH score and the adjusted threshold on the Type 2 datasets with local outliers. We also show the F-measure and G-mean scores that is achievable using the best possible threshold. The local outliers are expected to pose a challenge as some datapoints that are labelled as inliers may behave like local outliers. Therefore, these inlier points end up with GLOSH scores very close to scores of "true local outliers" and some may even have a higher score. In such cases, when we care more about finding true positives than avoiding false positives, it maybe better to take the

Dataset	POLAR								*Best	
	Knee				Adjusted					
	Recall	TNR	F-Measure	G-Mean	Recall	TNR	F-Measure	G-Mean	F-Measure	G-Mean
MVTec-AD_zipper	1.0	0.91	0.66	0.95	1.0	1.0	1.0	1.0	1.0	1.0
HEPATITIS	1.0	0.97	0.85	0.98	1.0	0.98	0.92	0.99	1.0	1.0
LETTER	1.0	0.94	0.78	0.97	1.0	1.0	1.0	1.0	1.0	1.0
PIMA	1.0	0.88	0.65	0.94	1.0	0.898	0.67	0.94	1.0	1.0
STAMPS	1.0	0.86	0.61	0.92	1.0	0.88	0.66	0.94	1.0	1.0
VERTEBRAL	1.0	0.83	0.55	0.91	1.0	0.96	0.85	0.98	0.97	0.99
VOWELS	1.0	0.90	0.66	0.95	1.0	1.0	1.0	1.0	1.0	1.0
WDBC	1.0	0.82	0.46	0.90	1.0	0.93	0.68	0.96	0.98	0.99
WINE	1.0	0.84	0.51	0.91	1.0	1.0	1.0	1.0	1.0	1.0
WPBC	1.0	0.79	0.45	0.89	1.0	0.96	0.83	0.98	1.0	1.0
YEAST	1.0	0.72	0.43	0.85	0.83	0.96	0.77	0.89	0.77	0.95
BREASTW	1.0	0.98	0.93	0.99	1.0	0.88	0.65	0.94	1.0	1.0
CARDIO	1.0	0.83	0.56	0.91	0.91	0.99	0.92	0.95	0.94	0.99
CARDIOTOCOGRAPHY	1.0	0.83	0.56	0.91	1.0	0.99	0.98	0.99	1.0	1.0
20news_3	1.0	0.87	0.61	0.93	1.0	0.98	0.91	0.99	0.99	0.99

Table 4.4: Evaluating POLAR on Type 2 datasets with mixed outliers at $min_{pts} = m^*$: We compare the performance achieved by POLAR with the best performance that can be achieved across all possible thresholds.

unadjusted threshold. In Table 4.3, one can see that there exists multiple datasets for which taking the knee GLOSH score as the threshold yields a better recall. However, in many of the datasets one can still get a high recall with the adjusted threshold estimated using POLAR. One can see that with the adjusted threshold, one gets a higher true negative rate (TNR). With any kind of outliers, the false positives tends to be low using the adjusted threshold. Therefore, in most of the datasets using the adjusted threshold yields a higher F-measure than choosing the knee. Even for local outliers, across many datasets, taking the adjusted threshold for labelling is able to yield a performance that is close to the best performance that can be achieved if one knows the best possible threshold.

In Table 4.4 we show the results obtained on the Type 2 datasets with mixed outliers. Overall, one can see that for mixed outliers, the adjusted threshold is a better choice than taking the knee GLOSH score as the threshold. Although, the adjusted threshold yields a lower recall for datasets such as YEAST and CARDIO, its still records a higher F-Measure and G-Mean on both the datasets compared to

Dataset	POLAR								*Best	
	Knee				Adjusted					
	Recall	TNR	F-Measure	G-Mean	Recall	TNR	F-Measure	G-Mean	F-Measure	G-Mean
MVTec-AD_zipper	0.44	0.94	0.56	0.64	0.0	1.0	0.0	0.0	0.64	0.74
HEPATITIS	1.0	0.28	0.35	0.53	1.0	0.29	0.35	0.54	0.41	0.68
LETTER	0.27	0.9	0.19	0.49	0.0	1.0	0.0	0.0	0.2	0.62
PIMA	0.12	0.94	0.2	0.34	0.13	0.932	0.21	0.35	0.61	0.68
STAMPS	0.64	0.87	0.44	0.75	0.58	0.87	0.41	0.71	0.5	0.89
VERTEBRAL	0.03	0.78	0.02	0.16	0.0	0.96	0.0	0.0	0.22	0.46
VOWELS	0.94	0.83	0.27	0.88	0.0	1.0	0.0	0.0	0.55	0.89
WDBC	1.0	0.82	0.24	0.9	1.0	0.94	0.51	0.97	0.58	0.98
WINE	0.0	0.8	0.0	0.0	0.0	1.0	0.0	0.0	0.17	0.48
WPBC	0.17	0.8	0.18	0.36	0.0	0.97	0.0	0.0	0.42	0.56
YEAST	0.14	0.79	0.19	0.34	0.01	0.97	0.02	0.11	0.5	0.45
BREASTW	0.91	0.95	0.91	0.93	0.97	0.95	0.94	0.96	0.94	0.96
CARDIO	0.56	0.74	0.28	0.65	0.01	1.0	0.02	0.1	0.29	0.66
CARDIOTOCOGRAPHY	0.25	0.7	0.22	0.42	0.0	1.0	0.0	0.0	0.35	0.47
20news_3	0.23	0.93	0.18	0.46	0.06	0.99	0.11	0.25	0.25	0.67

Table 4.5: Evaluating POLAR on Type 1 datasets at $min_{pts} = m^*$: We compare the performance achieved by POLAR with the best performance that can be achieved across all possible thresholds.

choosing the GLOSH score at the knee as the threshold. One can also see across many datasets, the adjusted threshold estimated using POLAR yields a performance that is close to the best possible performance.

In Table 4.5 we present the results in the real one-class classification datasets. Overall, one can see that taking the “knee GLOSH score” as the threshold or the adjusted threshold, the recall is low in most datasets. As we showed earlier in Figure 3.11, in these datasets, the GLOSH–Profiles of the datapoints labelled as outliers show very close or even lower GLOSH scores than those of the datapoints labelled as inliers. Therefore, many of the labelled outliers get a GLOSH score before the knee GLOSH score in the sorted sequence. This impacts the overall performance as reported in Table 4.5 and it shows once again that instances of downsampled classes from classification datasets are not necessarily behaving according to the common definitions of outliers in the given feature spaces.

4.4 Conclusion

This chapter successfully answers the second research question of this thesis: Given a min_{pts} value that assigns most, if not all, “true outliers” higher GLOSH scores than the “true inliers”, how to select a threshold to label inliers and “potential outliers”? In this chapter we develop an unsupervised strategy, POLAR, to estimate a threshold for labelling inliers and “potential outliers” using the “best” min_{pts} value, m^*. Across many datasets, the results indicate that the threshold estimated using POLAR closely approximates the best achievable performance w.r.t. f-measure and G-Mean.

Chapter 5

Conclusion & Future Work

5.1 Conclusions

This thesis addresses two key challenges: (i) the selection of a min_{pts} value that yields the best or nearly the best performance for GLOSH, and (ii) the subsequent challenge of determining a suitable threshold for labelling inliers and "potential outliers" once the best or nearly the best min_{pts} value is known.

To address the first challenge, we propose using GLOSH–Profiles, which that capture GLOSH scores at a range of different min_{pts} values. Through a comprehensive study on different datasets, we observe different behaviors in GLOSH–Profiles for different outlier types. One major observation revealed that at the min_{pts} value where GLOSH yields the best or nearly the best performance in terms of precision@n, the GLOSH scores in the profiles start to change at a similar rate. This observation served as a key to develop an unsupervised approach, Auto-GLOSH, to estimate the min_{pts} value where the GLOSH in the profiles scores start to change at a similar rate —what we call as the "best" min_{pts} value, m^*. The experimental analysis across a range of datasets showed that the m^* value determined using Auto-GLOSH is able to yield the best or nearly the best results for GLOSH.

To address the second challenge, we firstly investigate the behavior of the sorted sequence of the GLOSH scores w.r.t. the m^* value. Our investigation revealed a distinct pattern: the scores initially follow an "almost linear" trend followed by a "knee," indicating a deviation where the outlier GLOSH scores lie. This observation served as a key to develop an unsupervised strategy, POLAR, for automatic labeling "potential outliers" and inliers in a dataset based on the distribution of the sorted sequence of GLOSH scores at the "best" min_{pts} value, m^*. POLAR undertakes a parameter-less estimation of a threshold in the GLOSH scores for labelling. Across various datasets, the results indicate that the threshold estimated using POLAR closely approximates the best achievable performance when one knows the perfect threshold for labelling.

5.2 Future Work

There are a few avenues that can be explored as future work in this area of research. Similar to GLOSH, other nearest neighbor based outlier detection methods such as KNN and LOF estimates outlier scores based on the deviation of a datapoint from its nearest neighborhood, defined by a parameter k. Given the sensitivity of these algorithms to the choice of k, our proposed strategy to find the "best" min_{pts} value, may hold potential for also identifying the "best" k value for those methods. Considering the commonality in their underlying principles, it is plausible that outlier scores generated by these methods may exhibit a "knee"-like structure, similar to our observations with GLOSH. Therefore, our threshold identification strategy, POLAR, may be applicable to those methods. We plan to investigate the feasibility of applying our strategies, on algorithms that estimate outlier scores based on the nearest neighborhood of datapoints.

In our investigation, we saw that the GLOSH–Profiles of different outliers behave differently. Therefore, we also plan to investigate strategies that can capture the behavior of GLOSH–Profiles to automatically identify the types of different outliers in a dataset.

Bibliography

[1] D. M. Hawkins, *Identification of outliers*. Springer, 1980, vol. 11.

[2] A. J. Hall, N. Pitropakis, W. J. Buchanan, and N. Moradpoor, "Predicting malicious insider threat scenarios using organizational data and a heterogeneous stack-classifier," in *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, pp. 5034–5039.

[3] M. Riazi, O. Zaiane, T. Takeuchi, A. Maltais, J. Günther, and M. Lipsett, "Detecting the onset of machine failure using anomaly detection methods," in *Big Data Analytics and Knowledge Discovery: 21st International Conference, DaWaK 2019, Linz, Austria, August 26–29, 2019, Proceedings 21*, Springer, 2019, pp. 3–12.

[4] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-Gonzalez, and A. López-Cuevas, "A one-class classification approach for bot detection on twitter," *Computers & Security*, vol. 91, p. 101 715, 2020.

[5] P. Jain, S. Jain, O. R. Zaïane, and A. Srivastava, "Anomaly detection in resource constrained environments with streaming data," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 3, pp. 649–659, 2021.

[6] C. Bellinger, S. Sharma, and N. Japkowicz, "One-class versus binary classification: Which and when?" In *2012 11th International conference on machine learning and applications*, IEEE, vol. 2, 2012, pp. 102–106.

[7] H. O. Marques, L. Swersky, J. Sander, R. J. Campello, and A. Zimek, "On the evaluation of outlier detection and one-class classification: A comparative study of algorithms, model selection, and ensembles," *Data Mining and Knowledge Discovery*, pp. 1–45, 2023.

[8] R. J. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical density estimates for data clustering, visualization, and outlier detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 10, no. 1, pp. 1–51, 2015.

[9] A. Cavalcante Araujo Neto, *A Framework for Hierarchical Density-Based Clustering Exploration*. Ph.D. thesis, University of Alberta, 2021.

[10] Z. Zuo, Z. Li, P. Cheng, and J. Zhao, "A novel subspace outlier detection method by entropy-based clustering algorithm," *Scientific Reports*, vol. 13, p. 15 331, 2023.

[11] A. C. A. Neto, M. C. Naldi, R. J. Campello, and J. Sander, "Core-sg: Efficient computation of multiple msts for density-based methods," *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 951–964, 2022.

[12] G. O. Campos *et al.*, "On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study," *Data mining and knowledge discovery*, vol. 30, no. 4, pp. 891–927, 2016.

[13] E. M. Knorr and R. T. Ng, "A unified notion of outliers: Properties and computation.," *KDD*, vol. 97, pp. 219–222, 1997.

[14] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 427–438, 2000.

[15] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

[16] E. Schubert, A. Zimek, and H.-P. Kriegel, "Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection," *Data mining and knowledge discovery*, vol. 28, pp. 190–237, 2014.

[17] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, 1996, pp. 226–231.

[18] L. Swersky, H. O. Marques, J. Sander, R. J. Campello, and A. Zimek, "On the evaluation of outlier detection and one-class classification methods," in *2016 IEEE international conference on data science and advanced analytics (DSAA)*, IEEE, 2016, pp. 1–10.

[19] L. Swersky, *A study of unsupervised outlier detection for one-class classification.* M.Sc. thesis, University of Alberta, 2018.

[20] Y. Zhao, R. Rossi, and L. Akoglu, "Automatic unsupervised outlier model selection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4489–4502, 2021.

[21] A. C. A. Neto, J. Sander, R. J. Campello, and M. A. Nascimento, "Efficient computation and visualization of multiple density-based clustering hierarchies," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3075–3089, 2019.

[22] H. O. Marques, R. J. Campello, A. Zimek, and J. Sander, "On the internal evaluation of unsupervised outlier detection," in *Proceedings of the 27th international conference on scientific and statistical database management*, 2015, pp. 1–12.

[23] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao, "Adbench: Anomaly detection benchmark," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 142–32 159, 2022.

[24] G. Steinbuss and K. Böhm, "Benchmarking unsupervised outlier detection with realistic synthetic data," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 4, pp. 1–20, 2021.

[25] R. C. Prati, G. E. Batista, and M. C. Monard, "Class imbalances versus class overlapping: An analysis of a learning system behavior," in *Third Mexican International Conference on Artificial Intelligence*, Springer, 2004, pp. 312–321.

[26] B. Krawczyk, I. Triguero, S. García, M. Woźniak, and F. Herrera, "Instance reduction for one-class classification," *Knowledge and Information Systems*, vol. 59, pp. 601–628, 2019.

[27] S. D. Walter, "The partial area under the summary roc curve," *Statistics in medicine*, vol. 24, no. 13, pp. 2025–2040, 2005.

[28] P. Koncar, *Synthetic dataset for outlier detection*, version 1.0, Feb. 2018. DOI: 10.5281/zenodo.1171077. [Online]. Available: https://doi.org/10.5281/zenodo.1171077.

[29] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel, "On evaluation of outlier rankings and outlier scores," in *Proceedings of the 2012 SIAM international conference on data mining*, SIAM, 2012, pp. 1047–1058.

[30] H. O. Marques, R. J. Campello, J. Sander, and A. Zimek, "Internal evaluation of unsupervised outlier detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 14, no. 4, pp. 1–42, 2020.

[31] H. Anton and C. Rorres, *Elementary linear algebra: applications version.* John Wiley & Sons, 2013.

[32] D. C. Lay, *Linear algebra and its applications.* Pearson Education India, 2003.

[33] T. Jayasinghe *et al.*, "The milky way project second data release: Bubbles and bow shocks," *Monthly Notices of the Royal Astronomical Society*, vol. 488, no. 1, pp. 1141–1165, 2019.

[34] G. Jarry, D. Delahaye, F. Nicol, and E. Feron, "Aircraft atypical approach detection using functional principal component analysis," *Journal of Air Transport Management*, vol. 84, p. 101 787, 2020.

[35] H. O. Marques, A. Zimek, R. J. Campello, and J. Sander, "Similarity-based unsupervised evaluation of outlier detection," in *International Conference on Similarity Search and Applications*, Springer, 2022, pp. 234–248.

[36] P. Yin and M. Peng, "Station layout optimization and route selection of urban rail transit planning: A case study of shanghai pudong international airport," *Mathematics*, vol. 11, no. 6, p. 1539, 2023.

[37] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Linear regression," in *An Introduction to Statistical Learning: With Applications in Python*, Springer, 2023, pp. 69–134.

[38] M. Kubat, S. Matwin, *et al.*, "Addressing the curse of imbalanced training sets: One-sided selection," in *Icml*, Citeseer, vol. 97, 1997, p. 179.

[39] Y. S. Aurelio, G. M. De Almeida, C. L. de Castro, and A. P. Braga, "Learning from imbalanced data sets with weighted cross-entropy function," *Neural processing letters*, vol. 50, pp. 1937–1949, 2019.

[40] A. Luque, A. Carrasco, A. Martín, and A. de Las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019.

[41] L. I. Kuncheva, Á. Arnaiz-González, J.-F. Díez-Pastor, and I. A. Gunn, "Instance selection improves geometric mean accuracy: A study on imbalanced data classification," *Progress in Artificial Intelligence*, vol. 8, pp. 215–228, 2019.