Contents

Declaration 5

Acknowledgements 6

Abstract 7

List of Tables 11

List of Figures 12

Symbols 13

1 Introduction 1

1.1 Introduction 1

1.2 Aim of the Project 2

1.3 Objectives 2

1.4 Related Research 2

1.5 Motivation & Discussion 3

1.6 Application of Outlier Detection 4

1.7 Report Organization 5

2 Outlier Detection Approaches-a Survey 6

2.1 LOF- Identifying Density based Local Outlier 7

2.2 Mining Distance Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule- ORCA [3] 9

2.2.1 Experiments for scaling the performance 10

2.2.2 Analysis of time scaling 10

2.2.3 Limitations of ORCA 11

2.3 Locality Sensitive Hashing (LSH)-based Outlier Detection 12
2.3.1 Outlier Likelihood Ranking 12
2.3.2 Ranking Methodology 14
2.3.3 Outlier Detection 14
2.4 *OutRank*- A Graph-based Outlier Detection Framework using Random Walk [17] 14
2.4.1 Graph Representation 15
2.4.2 Markov Chain Model 15
2.4.3 Discussion 16

3 Robust Outlier Detection using Hybrid Approach 17
3.1 Background of the RODHA Algorithm 17
3.1.1 Outlier 18
3.1.2 Distance-based outlier 18
3.1.3 Density-based outlier 18
3.1.4 Entropy 19
3.2 Robust Outlier detection using Hybrid Approach 19
3.2.1 Clustering the Training Set 20
3.2.1.1 Initial Centroid Selection 20
3.2.2 Distance based Outlier Detection 21
3.2.3 Density based Outlier Detection 22
3.2.4 Entropy as a Parameter for Outlier Detection 24
3.2.5 Complexity Analysis 25

4 Implementation and Empirical Evaluation 28
4.1 Environment Used 28
4.2 Datasets Used 28
4.3 Experimental Results 30
4.4 Discussion 31
4.5 Entropy measure as an outlier detector 32
4.6 Sensitivity of the proposed algorithm 32
4.7 Selection of threshold ε in the proposed RODHA Algorithm 34
4.8 Selection of threshold τ in the proposed RODHA Algorithm 34
4.9 Selection of threshold k in LSH-based outlier detection technique . . 35

5 Conclusions **37**

5.1 Conclusion 37

5.2 Future Work 37

Bibliography **38**

List of Tables

2.1 Existing Outlier Detection Algorithms: A General Comparison 16

4.1 Basic information of the datasets used 29

4.2 Performance Evaluation of RODHA with other Competing Algorithms in terms of Detection Rate (DR) 31

List of Figures

2.1 *reach-dist(p_1,o)* and *reach-dist(p_2,o)*, for k=4 8

2.2 Calculation of Outlier Likelihood Rank 13

3.1 Outliers in two dimensional dataset 18

3.2 Illustrating distance-based outlier detection approach 22

3.3 Illustrating distance-based outlier detection approach fails in case of clusters being concave in nature 23

3.4 Illustrating density-based outlier detection approach 23

4.1 Performance of imputation techniques with 20% missing value on four datasets 30

4.2 A 2-Dimensional synthetic data set containing two clusters and outliers 32

4.3 The increase in entropy for addition of test dataset objects to the nearest cluster 33

4.4 Synthetic dataset for testing sensitivity of RODHA 33

4.5 Accuracy of outlier detection against threshold ε in RODHA 34

4.6 Variation of outlier detection rate against the user-defined threshold τ in RODHA 35

4.7 Accuracy of outlier detection against threshold k in LSH-OD 36

List of Symbols

H- Hash Family

R- R-neighborhood

Pr- Probability

$h()$- Hash Function

X- Random Number variable

U- Universe of Discourse

$l_p{}^d$- L$_p$ in d-dimension

WOS- Weakest Outlier Score

$\Re$- Real Space

d_{thres}- Distance Threshold

$H(C_k)$- Within-cluster entropy of cluster C_k