

Sanjay Chakraborty • SK Hafizul Islam  
Debabrata Samanta

# Data Classification and Incremental Clustering in Data Mining and Machine Learning

# Contents

|          |                                                                       |           |
|----------|-----------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction to Data Mining and Knowledge Discovery.</b>           | <b>1</b>  |
| 1.1      | Introduction                                                          | 1         |
| 1.2      | Architecture of Data Mining System                                    | 1         |
| 1.2.1    | Knowledge Discovery Process (KDD)                                     | 3         |
| 1.2.2    | Nature of Data                                                        | 3         |
| 1.2.3    | Data Mining Techniques                                                | 3         |
| 1.2.4    | Anomaly or Outlier Detection                                          | 7         |
| 1.3      | Regression Analysis and Prediction                                    | 8         |
| 1.3.1    | Sequence Analysis                                                     | 9         |
| 1.3.2    | Limitations of Data Mining                                            | 9         |
| 1.4      | Applications of Data Mining                                           | 10        |
| 1.5      | Incremental Data Mining                                               | 11        |
| 1.5.1    | Benefits of Incremental Data Mining                                   | 12        |
| 1.6      | Regression Methods                                                    | 12        |
| 1.6.1    | Illustration                                                          | 13        |
| 1.6.2    | Assumptions of Linear Regression                                      | 15        |
| 1.7      | Evaluating Model Performances                                         | 17        |
| 1.7.1    | Conclusion                                                            | 19        |
| 1.8      | Exercise                                                              | 19        |
| 1.9      | Interview Questions                                                   | 20        |
|          | References                                                            | 20        |
| <b>2</b> | <b>A Brief Concept on Machine Learning</b>                            | <b>23</b> |
| 2.1      | Introduction                                                          | 23        |
| 2.1.1    | Case Study                                                            | 23        |
| 2.1.2    | Interesting Definitions                                               | 24        |
| 2.2      | Artificial Intelligence (AI) vs. Data Mining vs. ML vs. Deep Learning | 24        |
| 2.2.1    | An Illustration                                                       | 25        |
| 2.3      | How to use ML?                                                        | 26        |
| 2.4      | Types of Data and ML Algorithms                                       | 26        |

- 2.4.1 Supervised Learning . . . . . 27
    - 2.4.2 Learning Without Supervision . . . . . 28
    - 2.4.3 Learning That Is Semi-supervised . . . . . 29
  - 2.5 Conclusion . . . . . 30
  - 2.6 Exercise . . . . . 30
  - References . . . . . 30
- 3 **Supervised Learning-Based Data Classification and Incremental Clustering** . . . . . 33
  - 3.1 Introduction . . . . . 33
    - 3.1.1 Case Study . . . . . 33
  - 3.2 Classification . . . . . 34
    - 3.2.1 K-nearest Neighbour Classification . . . . . 34
    - 3.2.2 Probabilistic Learning with Naïve Bayes Classification . . . . . 39
    - 3.2.3 Divide and Conquer Classification . . . . . 46
  - 3.3 Decision Tree Pruning . . . . . 56
  - 3.4 Postpruning Operations . . . . . 56
  - 3.5 Random Forests Classification . . . . . 57
  - 3.6 Support Vector Machine Classification . . . . . 60
  - 3.7 Multi-class SVM . . . . . 65
  - 3.8 R Framework . . . . . 67
  - 3.9 Conclusion . . . . . 67
  - 3.10 Exercises . . . . . 68
  - References . . . . . 70
- 4 **Data Classification and Incremental Clustering Using Unsupervised Learning** . . . . . 73
  - 4.1 Introduction . . . . . 73
  - 4.2 Literature Review . . . . . 73
  - 4.3 Types of Clustering . . . . . 82
  - 4.4 Popular Clustering Techniques . . . . . 83
    - 4.4.1 Flowchart of K-Means Clustering . . . . . 83
    - 4.4.2 Applications of K-Means Clustering . . . . . 84
    - 4.4.3 K-Medoids Clustering . . . . . 84
  - 4.5 Flowchart of DBSCAN Clustering Algorithm . . . . . 87
  - 4.6 Illustrations . . . . . 89
  - 4.7 Hierarchical vs. K-Means Clustering . . . . . 96
  - 4.8 Outlier Analysis . . . . . 96
  - 4.9 Conclusion . . . . . 96
  - References . . . . . 97
- 5 **Research Intention Towards Incremental Clustering** . . . . . 101
  - 5.1 Introduction . . . . . 101
  - 5.2 Problem Definition . . . . . 101
  - 5.3 Solution Approach . . . . . 102
  - 5.4 Incremental Clustering . . . . . 103

- 5.5 Incremental K-Means Clustering . . . . . 103
  - 5.5.1 Proposed Incremental K-Means Clustering Algorithm . . . 103
  - 5.5.2 Proposed Model of Incremental K-Means Clustering . . . 104
  - 5.5.3 Illustrative Examples of Incremental K-Means Clustering . . . . . 105
  - 5.5.4 Benefits and Applications of Incremental K-Means Clustering . . . . . 108
- 5.6 Incremental DBSCAN Clustering . . . . . 109
  - 5.6.1 Explanation of Pseudocode . . . . . 109
  - 5.6.2 Proposed Model of Incremental DBSCAN Clustering . . . 109
  - 5.6.3 Benefits of Incremental DBSCAN Clustering . . . . . 109
- 5.7 Effects of Cluster Metadata on Incremental Clustering . . . . . 110
- 5.8 Experiment and Result Analysis . . . . . 112
  - 5.8.1 Air Pollution Database . . . . . 112
  - 5.8.2 Java Platform . . . . . 112
  - 5.8.3 Weka Framework . . . . . 113
  - 5.8.4 Explanation of the Typical K-Means Algorithm Code . . . 115
  - 5.8.5 Explanation of the Incremental K-Means Algorithm Code . . . . . 117
  - 5.8.6 Result Analysis . . . . . 118
  - 5.8.7 Clustering vs. Incremental Clustering . . . . . 122
- 5.9 Conclusion . . . . . 123
- 5.10 Exercise . . . . . 124
- References . . . . . 124
- 6 Real-Time Application with Data Mining and Machine Learning . . . . . 129**
  - 6.1 Introduction . . . . . 129
    - 6.1.1 Data Analysis in Finance . . . . . 129
    - 6.1.2 Industry of Retail . . . . . 129
    - 6.1.3 Telecommunications Sector . . . . . 130
    - 6.1.4 Analysing Biological Data . . . . . 130
    - 6.1.5 Additional Scientific Uses . . . . . 130
  - 6.2 Detection of Intruders . . . . . 131
    - 6.2.1 Choosing a DM Methodology . . . . . 131
  - 6.3 Applications of Machine Learning . . . . . 132
  - 6.4 Incremental Clustering . . . . . 132
  - 6.5 Supervised Learning . . . . . 133
    - 6.5.1 Input Dataset Description . . . . . 133
    - 6.5.2 Classification Algorithms with Pseudocode on Python Environment . . . . . 133
    - 6.5.3 Output . . . . . 135
  - 6.6 Classification on Iris Dataset in R Platform . . . . . 135
    - 6.6.1 Input Dataset Description . . . . . 135
    - 6.6.2 Pseudocode . . . . . 135

|          |                                                                                                                       |            |
|----------|-----------------------------------------------------------------------------------------------------------------------|------------|
| 6.6.3    | Process .....                                                                                                         | 136        |
| 6.6.4    | Output .....                                                                                                          | 137        |
| 6.7      | Unsupervised Learning .....                                                                                           | 138        |
| 6.7.1    | Partitional Clustering with K-Means Clustering .....                                                                  | 138        |
| 6.7.2    | Pseudocode of K-Means (DB,k) .....                                                                                    | 139        |
| 6.7.3    | Cluster Means .....                                                                                                   | 140        |
| 6.7.4    | Elbow Method to Find the Optimum Value of K. ....                                                                     | 140        |
| 6.8      | Density-Based Clustering .....                                                                                        | 141        |
| 6.8.1    | Pseudocode: DBSCAN (DB, epsilon, Minpts) .....                                                                        | 142        |
| 6.8.2    | Output .....                                                                                                          | 142        |
| 6.9      | Hierarchical Clustering Using Agglomerative Clustering .....                                                          | 144        |
| 6.9.1    | Output of Complete Linkage Method .....                                                                               | 145        |
| 6.10     | Grid-Based Clustering .....                                                                                           | 146        |
| 6.10.1   | Coding Snapshots of the Experiment. ....                                                                              | 146        |
| 6.11     | Explanation of the Typical K-Means Algorithm Code .....                                                               | 147        |
| 6.12     | Applications of Classifiers for Protein-Protein Prediction<br>Between SARS-CoV-2 and Human .....                      | 151        |
| 6.12.1   | Pseudocode (Python Platform) .....                                                                                    | 151        |
| 6.13     | Deep Multilayer Perceptron (DMLP) for Protein-Protein<br>Prediction Between SARS-CoV-2 and <i>Homo sapiens</i> . .... | 153        |
| 6.13.1   | Pseudocode (Python Platform) .....                                                                                    | 153        |
| 6.14     | Conclusion .....                                                                                                      | 154        |
|          | References. ....                                                                                                      | 154        |
| <b>7</b> | <b>Feature Subset Selection Techniques with Machine Learning .....</b>                                                | <b>159</b> |
| 7.1      | Introduction .....                                                                                                    | 159        |
| 7.2      | Irrelevant Variants (Without Variant Selection): A Problem .....                                                      | 160        |
| 7.3      | Algorithms for Machine Learning Were Utilized<br>in This Research. ....                                               | 162        |
| 7.4      | Methodology for Feature Selection .....                                                                               | 162        |
| 7.5      | Approaches for Reducing Dimension .....                                                                               | 164        |
| 7.5.1    | Variant Picking. ....                                                                                                 | 165        |
| 7.5.2    | Random Forest (RF) .....                                                                                              | 165        |
| 7.5.3    | Naïve Bayes .....                                                                                                     | 166        |
| 7.6      | Supervised Filter Model Based on Relevance and Redundancy ...                                                         | 167        |
| 7.7      | Wrapper Model Under Supervision. ....                                                                                 | 168        |
| 7.8      | Future Research on Feature Selection .....                                                                            | 169        |
| 7.8.1    | Machine Learning with Extreme Data. ....                                                                              | 170        |
| 7.8.2    | Variable Selection via the Internet. ....                                                                             | 170        |
| 7.8.3    | Deep Learning (DL) and Variable Selection .....                                                                       | 171        |
| 7.8.4    | Selection Properties of Variant Variants. ....                                                                        | 171        |
| 7.8.5    | Searching Engine Optimisation .....                                                                                   | 172        |
| 7.9      | Classification Accuracy .....                                                                                         | 172        |
| 7.10     | Conclusion .....                                                                                                      | 173        |
|          | References. ....                                                                                                      | 173        |

**8 Data Mining-Based Variant Subset Features . . . . . 177**

8.1 Introduction . . . . . 177

8.2 Review of the Literature. . . . . 178

8.3 Extraction of Variants. . . . . 179

8.4 Variant Selection and Extraction Approaches . . . . . 180

8.5 Choosing of Variant Subset Features . . . . . 181

8.6 Unsupervised Variable Selection . . . . . 185

8.7 Model of an Unsupervised Filter . . . . . 185

8.8 Unsupervised Wrapper Approach . . . . . 186

8.9 Selection of Semi-supervised Variants. . . . . 187

8.10 Evaluation of Variants . . . . . 188

8.10.1 Specification for Stopping . . . . . 188

8.10.2 Experimenting with Variable Selection Techniques. . . . . 188

8.10.3 I-RELIEF . . . . . 189

8.10.4 Interactive. . . . . 189

8.10.5 Genetic Algorithm (GA) . . . . . 190

8.11 Conclusion . . . . . 190

References. . . . . 191

**Index. . . . . 195**