Appendix A
Online Resources for Outlier Detection

A.1 Open Source Tools/Packages

There are a number of data mining and machine learning libraries made available on the Internet by various working groups of professionals, for academic and non-profit usage. Typically, these libraries vary in terms of the following aspects.

- Coverage of algorithms (data analysis, visualization, predictive models, etc.)
- Implementation language (such as C, C++, Java, Python, etc.)
- Target platform for usage (like Windows, Linux, Mac OS, etc.)
- Mode of usage (such as online browser-based use, download and install, etc.)

Free and open source software tools for data mining are being developed for the past several years. The common objective of these tools is to facilitate the data analysis process in an effective manner and to offer all interested researchers a free alternative to the commercial solutions. This is typically done by employing integrated development environments, and implementation using standard programming languages.

As per the open source philosophy, the source codes of these libraries are also generally available for researchers working in this area. The idea behind this philosophy is that one can extend the functionality of these tools by incorporating algorithms corresponding to recent innovation and maintain the tools from time-to-time. Free and open source software are typically distributed following the GNU general public licensing method.

The objective here is to make young professionals and researchers planning to work in the field of data mining and outlier detection, get familiarized with the availability of various online resources such as different implementations of the machine learning algorithms, benchmark data sets to experiment, research forums to interact, etc. Typically, such resources are required for carrying out advanced research related to data mining. Details on a few of the well known data mining tools

N. N. R Ranga Suri et al., *Outlier Detection: Techniques and Applications*,
Intelligent Systems Reference Library 155,
https://doi.org/10.1007/978-3-030-05127-3

and libraries are furnished below for a quick reference. It is important to note that details regarding only a subset of the tools are presented here for brevity.

1. *Waikato Environment for Knowledge Analysis (WEKA)*: This is a collection of various machine learning algorithms for data analysis and predictive modeling, developed using Java programming language at the University of Waikato, New Zealand. This tool comprises necessary graphical user interfaces for easy and convenient use of the analysis functionality. The tool distribution supports multiple operating platforms for its use enabling cross-platform deployment.
   WEKA suite consists of the following functional modules offering a variety of analysis functionality on data. More details regarding this tool can be obtained from: 'https://www.cs.waikato.ac.nz/ml/weka'.

   - Explorer
   - Experimenter
   - KnowledgeFlow
   - Simple CLI

2. *Konstanz Information Miner (KNIME)*: This tool was developed at University of Konstanz as a proprietary product. The objective was to create a modular, highly scalable and open data processing platform for easy integration of different data loading, processing, transformation, analysis and visual exploration modules. This tool has been widely used in pharmaceutical research, customer data analysis, business intelligence and financial analysis.
   KNIME allows users to visually create data flows and selectively execute some of the analysis steps in the sequence. It is developed using Java language based on Eclipse IDE and allows plugins to provide additional functionality. The basic package includes various commonly used methods of statistics, data mining, analysis and text analytics. This tool supports big data analysis framework and also works with Apache Spark library. For more details in this regard, one may refer to: 'https://www.knime.com/knime-software/knime-analytics-platform'.
3. *Environment for DeveLoping KDD-Applications Supported by Index-Structures (ELKI)*: The initial version of this tool distribution contained several algorithms to perform cluster analysis, anomaly detection, along with other data mining tasks. As its name indicates, some spatial index structures are included in this tool with its focus primarily on subspace clustering and correlation clustering methods.
   It was developed at the Ludwig Maximilian University of Munich, Germany, based on various Java interfaces for use in teaching and research purposes. The visualization interface of ELKI uses a scalable graphics library and also supports loss less export of user interface displays to PostScript and PDF file formats. For the purpose of detecting outliers in data, this particular tool offers various algorithms based on different detection strategies such as distance-based outliers, LOF-based local outliers, and many others. Therefore, it can be used as a reference platform for exploring and understanding the research issues related to this problem space. Further details in this regard can be accessed at: 'https://elki-project.github.io'.

4. *Orange Software*: This is a component-based visual programming software package for data visualization, machine learning, and data analysis. The earlier versions of this software include core components written in C++ language and the wrappers using Python scripts. The advanced version makes use of various Python-based scientific computing libraries such as *numpy*, *scipy* and *scikit-learn* to accomplish the intended functionality. The graphical user interface is designed based on the Qt framework, for supporting portability across platforms.
   Orange software basically provides a visual programming front-end for explorative data analysis and interactive data visualization. Interactive analytics helps in better understanding of the analysis process with insights on the inherent structure of the data. This software can be installed on all major computing platforms supporting Python based development. More information regarding this tool can found at: 'https://orange.biolab.si'.
5. *PyTorch Library*: This is a python based open source machine learning library used as a computing platform for developing deep learning based applications. The workflow with this tool is similar to that of the python-based scientific computing library *numpy*. It basically supports two key functionalities

   - Tensor computations with multi-GPU support
   - Computations relating to deep neural networks

   PyTorch provides a framework for building computational graphs allowing the users to perform computations on graph components before the graph is built completely, and even change them during runtime. For more details on this library, one may refer to: 'https://pytorch.org/'.

A.2 Benchmark Data Sets

Benchmark data sets facilitate a systematic evaluation of the machine learning algorithms based on some predefined application settings. These data sets can also be used for establishing a comparative view of various algorithms designed to accomplish a common computational task. Thus, use of appropriate benchmark data sets enables the research community for advancement of the data mining technology in a quantified manner. More specifically, the case of dealing with the outlier detection problem needs special attention in terms of identifying suitable data sets due to the required imbalance in class distribution. To this effect, certain pre-processing steps are performed on the original data, making it ready for outlier detection.

1. *UCI ML repository*: This is a collection of data sets popularly used by the machine learning community for empirical analysis of their algorithms. This repository has divided the data sets under different classifications for easy selection and usage, as per the details given below.

   - Learning task: classification, regression, clustering, etc.
   - Type of attribute: categorical, numerical, mixed.

- Data generation method: uni-variate, multivariate, sequential, time-series, etc.
- Application area: life sciences, physical sciences, computer science and engineering, social sciences, business, game data, etc.

The above data are available at: 'https://archive.ics.uci.edu/ml/datasets.html'.

2. *Stanford Large Network Dataset Collection*: This collection consists of many network data sets divided into 17 broad categories based on the nature of the data captured. The main categories include social networks, networks with ground-truth communities, communication networks, citation networks, collaboration networks, web graphs, autonomous systems, signed networks, temporal networks, etc. These data sets are made available as part of the Stanford Network Analysis Platform (SNAP) platform, which is a general purpose network analysis and graph mining library. These data sets can be accessed at the web page: 'http://snap.stanford.edu/data/index.html'.
3. *Outlier Detection Data Sets (ODDS)*: This collection provides open access to a number of data sets meant for outlier detection, with ground truth as per its availability. These data sets are divided into the following groups based on the specific data they contain and for enabling precise usage.

   - Multi-dimensional point data sets
   - Time series graph data sets for event detection
   - Time series point data sets (Multivariate/Univariate)
   - Adversarial/Attack scenario and security data sets
   - Crowded scene video data for anomaly detection

   The above data are available online at the URL: 'http://odds.cs.stonybrook.edu/'.

Glossary

Anomaly detection The process of identifying data objects that do not conform to the expected behavior of a given collection of data.

Autonomous systems A collection of routers whose prefixes and routing policies are under common administrative control, and they communicate using Border Gateway Protocol.

Bagging Bootstrap aggregating (bagging) is an ensemble method designed to improve the accuracy of the learning algorithms used for classification and regression.

Categorical variable A categorical variable (nominal variable) is one that doesn't have intrinsic ordering of its categories.

Centrality measure In graph theory and network analysis, a centrality measure identifies the most important vertices in a topological sense.

Cluster analysis An exploratory analysis tool that divides the data objects into various groups such that the degree of association between two objects is maximal within a group and minimal otherwise.

Community detection The process of dividing a network into groups of nodes with dense connections internally and sparser connections between groups.

Confusion matrix It contains information about the actual and the predicted classifications done by a classifier.

Curse of dimensionality The phenomena that arise when performing data analysis in high-dimensional spaces (typically hundreds of dimensions).

Data mining The process of identifying valid, novel, potentially useful and ultimately understandable patterns in large data sets.

Decision attribute In a decision tree, it is part of a node, so it performs as a consequent in the decision rules on the paths down the tree to the leaf node.

Dense subgraph A component of the graph with high edge density in comparison to all other sub-graphs of the graph.

Discernibility matrix A matrix in which the classes are indicies and the condition attributes which can be used to discern between the classes in the corresponding row and column are inserted.

N. N. R Ranga Suri et al., *Outlier Detection: Techniques and Applications*,
Intelligent Systems Reference Library 155,
https://doi.org/10.1007/978-3-030-05127-3

Egonet The ego-network of a node is the induced subgraph of its 1-step neighbors in the given graph.

Eigenvalues A special set of scalars associated with a linear system of equations that are sometimes also known as characteristic values, or latent roots.

Ensemble methods The methods that use multiple models to obtain better predictive performance over any of the constituent models.

Entropy The entropy of a random variable is a measure of the uncertainty associated with that random variable.

Evolving networks Networks that change as a function of time, either by adding or removing nodes or links over time.

Example-based outlier mining Given a set of outlier examples, find additional outliers from the data that exhibit similar outlier characteristics.

Feature selection The process of identifying and removing irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model.

Frequent subgraph Discovering subgraphs that occur frequently over the entire set of graphs.

Fuzzy set Any set that allows its members to have different grades of membership (membership function) in the interval [0,1].

Game theory The study of mathematical models of strategic interaction between rational decision-makers.

Genetic algorithm It is a heuristic search and optimization technique inspired by natural evolution.

Granular computing It is a nature inspired computing paradigm that performs computation and operations on information granules.

Graph mining Extracting patterns (subgraphs) of interest from graphs, that describe the underlying data.

Hamming distance Hamming distance between two strings of equal length is the number of positions at which the corresponding characters differ.

High dimensional data Data with large number of features, attributes or characteristics that lead to curse of dimensionality.

Imbalanced data A data set in which one of the two classes has more samples than the other, resulting in skewed distribution of data.

Information theory The mathematical study of the coding of information in the form of sequences of symbols, impulses, etc.

Joint probability A statistical measure that calculates the likelihood of two events occurring together and at the same point in time.

Knowledge discovery An interdisciplinary area focusing on methodologies for extracting useful knowledge from data.

Labeled data A group of data samples that have been tagged with one or more labels.

Likelihood Typically refers to events that have a reasonable probability of occurring, but are not definite or may be influenced by factors not yet observed or measured.

Machine learning This is an application of Artificial Intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

Marginal probability The probability of any single event occurring independent of other events.

Matrix algebra The process of performing common algebraic operations (such as addition, subtraction, multiplication, determining rank, etc.) on matrix variables.

Modularity It measures the strength of division of a network into various modules/communities, such that there are dense connections between the nodes within communities but sparse connections across communities.

Mutual information According to information theory, the mutual information of two random variables is a measure of the mutual dependence between them.

Nearest neighbor search A form of proximity search, trying to find the object in a given data set that is closest to a given object.

Network regularization In the field of machine learning, regularization is a process of introducing additional information during the learning phase in order to prevent over fitting.

Non-negative matrix factorization A technique for obtaining low rank representation of matrices with non-negative or positive elements.

Outlier An observation which deviates so much from the other observations as to arouse suspicion that it was generated by a different mechanism.

Outlier detection The process of finding out objects that are considerably dissimilar, exceptional and inconsistent with respect to the majority objects in a data set.

Outlying subspaces The outlying subspace of a data object is the subspace in which the object displays outlier characteristics.

Overlap measure For two multivariate categorical data points, the overlap measure indicates the similarity between them by taking the number of attributes in which they match.

Proximity measure This measures how alike objects are to one another (object similarity) or how unlike they are (object dissimilarity).

Random projection A technique that allows one to substantially reduce the dimensionality of a problem while still retaining a significant degree of the structure of the data set.

ROC curve A receiver operating characteristic (ROC) curve is a 2-dimensional plot showing the performance of a classification model at all classification thresholds.

Rough clustering In contrast to conventional data clustering, the definite membership of some objects cannot be determined when performing rough clustering.

Rough set An approximation of a crisp set in terms of a pair of sets representing the lower and the upper approximation of the original set.

Route views A mechanism that allows Internet users to view global BGP routing information.

Scale-free networks Those information networks that satisfy a power law degree distribution.

Semi-supervised learning A machine learning paradigm that uses a large amount of unlabeled data, together with a small amount of labeled data, to build better classifiers.

Sliding window The basis for how we can turn any time series data set into a supervised learning problem.

Social network It is a web service that allows people with similar interests to come together and share information, photos and videos.

Soft computing A computing method that deals with approximate models and gives solutions to complex real-life problems. It is tolerant of imprecision, uncertainty, partial truth, and approximations.

Sparse data A data set in which a relatively high percentage of the attribute values are unspecified.

Snapshot graph The snapshots of an evolving graph taken periodically form a sequence of snapshot graphs.

Stochastic network A deterministic network that is modeled randomly due to its complexity.

Streaming data The data that is generated continuously by thousands of data sources and sent to the collection system simultaneously, in small sizes.

Subspace The space spanned by a sub-set of features/attributes of a data set.

Traffic patterns Typical traffic patterns over a communication network include traffic volume distribution, application usage, and application popularity.

Unsupervised learning A machine learning paradigm that draws inferences from data sets without class labels.

Volatile graph A stream of duration-stamped edges with potentially infinite number of nodes, and edges may appear and disappear over time.

Web service It is a standardized medium to propagate communication between the client and server applications on the World Wide Web.

Index

N. N. R Ranga Suri et al., *Outlier Detection: Techniques and Applications*,
Intelligent Systems Reference Library 155,
https://doi.org/10.1007/978-3-030-05127-3