Assignment 2: Applications of Unsupervised Methods

Introduction

This assignment aims to:

- provide students the opportunity to apply the knowledge learned to conduct literature review and data analysis related to unsupervised methods, and
- assess students' learning outcomes of unsupervised methods and their applications.

This assignment is an individual assignment. Each student is required to:

- conduct a literature review on anomaly detection methods with the required scope,
- design an anomaly detection algorithm based on k-means as required, and
- apply the designed algorithm to detect anomalies in the given dataset, with the use of SAS Enterprise Miner.

The deliverables of the assignment comprise a paper, a set of slides presenting the paper, and a recording of your presentation, as required in the next section.

The assignment **is due on Sunday, 5 November 2023, 11:59pm Adelaide time.**

The weighting of this assignment is 35% of the total marks of the course.

Deliverables

1. **(85%) A paper** that contains the following main sections/parts: (a) (10%) An introduction presenting (but not limited to):
   - What anomaly detection is in your own words,
   - Motivation of anomaly detection by using some examples, and
   - Main categories of anomaly detection methods, including the basic idea of each type of methods and suitable scenarios to use (data types, dimensionality etc.).

   (b) (25%) A literature review of anomaly detection methods that are based on clustering techniques. This section must contain the following contents (as a minimal requirement):
   - An overview of the literature on the topic, i.e., clustering based anomaly detection methods. Group similar type of methods and organise the overview based on the types of methods grouped. Then provide an overview of the basic idea of each type of methods,
   - A review of each type of the methods, including the description and discussion on the representative method(s)/paper(s) of each type, and
   - A comparison of the different types of methods in terms of their pros and cons, and similarities and differences of the different types of methods.

   (c) (20%) A description of the k-means based anomaly detection algorithm designed by you. The description must include the following:
   - Definition of the anomaly score designed by you and a description regarding why and how the defined anomaly score can capture anomalousness of data objects.
     The score must be defined based on the information and results provided by k-means clustering on an input dataset (in which anomalies are detected), in addition to other information you want to use. Use examples and figures to help illustrate your design idea of the anomaly score.
   - The designed anomaly detection algorithm presented in the format of pseudo code, and a text description of the algorithm (how it works, the steps etc.) while referring to the pseudo code of the algorithm. The algorithm designed by you must use the anomaly

score defined above as the basis to detect anomalies. Use examples to help illustrate the steps of the algorithm and/or how it works.

(d) (30%) An application of the k-means based anomaly detection algorithm designed by you to the given dataset (`a2-housing.csv`) to detect anomalies in the dataset.

This dataset is a modified version of the original UCI Data (download the names file at https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names for some background information about the original dataset, particularly the meaning of the attributes). Each record of the given dataset contains the information related to housing in a suburb in Boston area.

Note that when applying your algorithm (i.e. following the steps of your algorithm) to detect anomalies in the given dataset, you must use the Cluster node of SAS EM to do the k-means clustering step of your algorithm. Other steps/parts of your algorithm can be done with the given dataset using any tools or ways you prefer.

In this part of your paper, you must:

- describe how you follow the steps of your designed algorithm to detect anomalies in the given dataset, including the steps done using SAS EM Cluster node (for k-means clustering) and possibly other SAS EM nodes, and other tools or ways,
- present the detected anomalies, and discuss the results, e.g., why or why not an anomaly detected by your algorithm makes sense, and
- justify if any pre-processing of the given dataset is needed, and if yes, describe how the data pre-processing is done.

Notes:

- Write the paper as if it is to be submitted to KDD, one of the top conferences in the field of data analytics (https://www.kdd.org/kdd2023/).
- Use the template file (`kdd-template.docx`) provided to typeset your paper. The **page limit of the paper is 7 pages (excluding references)** using the template provided, and you are not allowed to modify the font size or layout of the given template.
- The paper also needs to have an abstract and a conclusion section. Although no marks are allocated to the abstract or conclusion, marks may be deducted if your paper does not contain a well-written abstract or conclusion section.

2. **(15%) A presentation (slides & recording)** of your paper, pretending that your paper has been accepted by the conference, and you need to give the presentation at the conference.

   The number of slides must not be over 12 and the duration of the presentation must not be over 10 minutes. The submission of presentation must contain the slides and an audio or video file of your presentation.

Instructions/Hints

About the literature review:

1. Be clear with the topic and focus of your review, which will help you determine the keywords used for literature search. Note that the review needs to be focused on anomaly/outlier detection methods that are based on clustering.
2. You may use search engines such as Google as the starting point of your search, but you are strongly encouraged to search the technical databases for related literature, especially research papers, such as IEEEXplore and ScienceDirect available from UniSA library website.

3. It will be useful to find a couple of recent high-quality review/survey papers on the topic, and then expand your search based on the references included in the survey papers. However, please make sure that you do your own search and review of related papers.
4. After some quick reading of the papers found, select, say, 8-10 mostly related papers (may include a couple of survey papers), read them carefully, and base your literature review section mainly on the selected papers.

About designing and presenting the algorithm:

1. The algorithm designed by you does not need to be complicated, and you do not need to consider algorithm efficiency (time complexity). A workable solution will be fine.
2. Refer to the slides of the anomaly detection module for the problem definition and general scheme for an anomaly detection task, which can be helpful with setting up the outline/steps of your algorithm.
3. It is important to devise the anomaly score (and other criteria and procedures if needed) to be used by your algorithm to measure the anomalousness of a data point and to determine if a data point is an anomaly or not.
4. For the algorithm, in addition to its steps, clearly describe its input (e.g. dataset, user specified parameter or threshold values) and output (e.g. a set of data points that are anomalies).
5. A key step of the algorithm must be k-means clustering. Besides, there should be other steps such as those for calculating anomaly scores of data objects and using the calculated score and other criteria or procedures to detect outliers.
6. **When designing the algorithm:**
    a. As the k-means clustering step will be done by using SAS EM for the given dataset, your algorithm design must be based on (or restricted by) the information available in the results generated by SAS EM.
    b. Take into consideration the issues of k-means, e.g. number of clusters to use, and how your algorithm copes with the issues.
7. There is no strict format requirement on the pseudo code representing the algorithms. The following example should give you some idea about how normally an algorithm is presented using pseudo code.

Algorithm: DBSCAN: a density-based clustering algorithm.

Input:

- D: a data set containing n objects,
- ϵ: the radius parameter, and
- *MinPts*: the neighborhood density threshold.

Output: A set of density-based clusters.

Method:

```
(1)  mark all objects as unvisited;
(2)  do
(3)      randomly select an unvisited object p;
(4)      mark p as visited;
(5)      if the ε-neighborhood of p has at least MinPts objects
(6)          create a new cluster C, and add p to C;
(7)          let N be the set of objects in the ε-neighborhood of p;
(8)          for each point p' in N
(9)              if p' is unvisited
(10)                 mark p' as visited;
(11)                 if the ε-neighborhood of p' has at least MinPts points,
                     add those points to N;
(12)             if p' is not yet a member of any cluster, add p' to C;
(13)         end for
(14)         output C;
(15)     else mark p as noise;
(16) until no object is unvisited;
```

(Taken from J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques. 3rd edition. Morgan Kaufmann, 2012)

About the experiments with the given dataset:

1. When applying the designed algorithm to the given dataset, you need to follow the steps of the algorithm to detect anomalies in the dataset. Note that:
   a. For the k-means clustering step, you must use SAS EM.
   b. For the steps other than the k-means clustering step in your algorithm, you may use other tools, e.g. Excel for calculating anomaly scores for data objects based on your definition of anomaly scores and using the information obtained from the k-means clustering.
2. Note that you are not required to write a program to implement the designed algorithm.
3. To make your discussion of the anomalies detected more insightful, you need to examine the anomalies detected to see which attributes have played important roles in causing them to be anomalies, and you could discuss the practical meaning of the anomalies detected to justify why they are considered as anomalies. It may be useful to look for and use references to support your discussions. It may also be possible that some of the anomalies detected by your algorithm do not make good sense. In this case, you also need to discuss why, e.g., it could be a limitation of your algorithm. Note that no algorithms are perfect, so it is ok for your designed algorithm to miss some anomalies or output anomalies which are not meaningful, but it is important to understand and discuss the limitations of your algorithm.

Marking of the assignment

Marks allocated to each part of the assignment is listed in the Deliverables section.

Marks for the paper will be awarded mainly based on:

- understanding of the basics of anomaly detection,
- quality of literature review, including quality of the literature found, understanding of the literature, and insightful analysis of the literature,

- whether the algorithm designed is workable,
- whether reasonable results have been obtained with the given dataset, and whether insightful justification and discussion of the results have been provided. and
- quality of writing and a paper over page limit or in wrong format may incur mark deduction.

Marks for the presentation (slides and recording submitted) will be awarded mainly based on:

- whether key points of the paper have been included,
- balanced contents within required time, • logical flow of the presentation, and
- clarity of presentation.

Deadline and submission

- The assignment **is due on Sunday, 5 November 2023, 11:59pm Adelaide time.**
- The following three files must be submitted separately, and zip files are not accepted.
    1. The paper as a MS word document or a pdf file using the given template,
    2. Presentation slides, as power point slides, and
    3. Presentation recording (audio or video).
- An extension to the assignment will only be granted under unexpected circumstances (e.g., illness). In this case, a student must provide supporting documents and submit their extension request (via Learnonline) at least a day before the assignment is due. A request of extension made on or after the due date of the assignment will not be considered. Any late submission of the report without a pre-arranged extension, a penalty of 20% of the assessed mark per day (including weekend) will be incurred. For example, if a student's submission is 2 days late and the originally assessed mark is 80 out of 100, then after the late penalty is applied, the actual mark the student is awarded will be 48 out of 100.