**Discussion: Data Clustering**

**Your Name**

**Instructor’s Name**

**Course**

**Date of Submission**

1. **What are the characteristics of data?**

Data refers to any form of information collected and organized in a meaningful format and can be processed to insight various decisions that promote the organization transformation. Currently, "information is power" is implemented since it is implemented as an indication of the current position of the organization and other critical areas. Therefore it is crucial to understand various data characteristics. The essential features of data include the accuracy of information. Data must be precise and maintain accurate information to enable easy understanding to the user. The second characteristic is that data must be relevant based on the needs of the user. Therefore the legitimacy of the data has to be enhanced before considering the usage of the data. The third characteristic of data is consistency and reliability. The provided must be reliable and consistent and avoid sharing false information (Kumar, 2004). The other characteristic is the relevance of the data; this is important in ensuring the quality of the data and promote its usefulness. The other feature of data is its quality, where high-quality information is unique to the user's need, easily accessible, and quickly processed.

1. **Compare the difference in each of the following clustering types: prototype-based, density-based, and graph-based.**

Clustering aims at grouping unblended statements in a given data set. Prototype-based clustering approaches effectively analyze robust statistics and can easily be used to identify the critical pattern in the presented data. Prototype-based clustering is based on K-means and its fuzzy C- means algorithms that enable practical data mining. Prototype-based clustering ensures each observation from data is linked to the nearest prototype cluster. Also, the prototype-based cluster consists of a set of the labelled prototype, which enables easy classification of data.

Destiny based clustering can be defined as organized learning systems that find distinctive groups of data. Destiny based clustering is based on the idea that clusters in a data environment are connecting sections of high region density separated from other cluster regions with low connections point destiny. Destiny-based clustering operates by providing essential principles that focus on finding the neighbourhoods of information clusters that exceed particular destiny thresholds. Destiny-based clustering does not need a priori classification of numbers of clusters. It is also effective in identifying noise while clustering.

Graphical-based clustering is based on a family of unorganized grouping algorithms that are designed to cluster various vertices and edges of the graph instead of objects in the presented feature space. Graphical based clustering is highly implemented in data mining in online social networks. Graph-based clustering focuses on implementing graphs in clustering. In this model, the clustering problem can be presented using a graph at which each element is classified as a node (Liu & Barahona, 2020). The space between the components is transformed through specific weights on the edges to connect each node.

1. **What is a scalable clustering algorithm?**

The scalable clustering algorithm is a clustering technique that helps in the optimization of data using an ant colony algorithm using swam intelligence. The scalable clustering approach is effective in obtaining evaluation results using synthetic datasets. Scalable clustering helps in recognizing common data characteristics in the specific type of data based on their size. The scalable clustering algorithms are effective in analyzing the distribution of data in large data sets.

1. **How do you choose the right algorithm?**

The practical step that I would follow when selecting the suitable algorithm includes the size of the training data. This is important in collecting adequate data to get reliable predictions at the presented data. Also, I would ensure the chosen is accurate and effective in interpreting the current output of data. The next important step in selecting the suitable algorithm is the speed of the training period; this is important in ensuring the successful implementation of the collected data. The other important factor that I could consider when choosing a suitable algorithm is linearity (Oyelade et al., 2019). Linearity in the algorithm is vital to assume practical analysis of data following a straight right. The other important factor to consider is the number of features. Since a data set may consist of many components, algorithms are effective in analyzing the significant elements.

**References**

Liu, Z., & Barahona, M. (2020). Graph-based data clustering via multiscale community detection. *Applied Network Science*, *5*(1). https://doi.org/10.1007/s41109-019-0248-7

Oyelade, J., Isewon, I., Oladipupo, O., Emebo, O., Omogbadegun, Z., & Aromolaran, O. et al. (2019). Data Clustering: Algorithms and Its Applications. *2019 19Th International Conference On Computational Science And Its Applications (ICCSA)*. <https://doi.org/10.1109/iccsa.2019.000-1>

Tan, Steinbach, Kumar, (2004) Introduction to Data Mining.