Chapter 21

Pre–Clustering Techniques for Healthcare System:
Evaluation Measures, Evaluation Metrics, Comparative Study of Existing vs. Proposed Approaches

Asadi Srinivasulu

Sree Vidyanikethan Engineering College, India

ABSTRACT

This chapter presents a comparative study of the proposed approaches (i.e., extended dark block extraction [EDBE], extended cluster count extraction [ECCE], and extended co-VAT approaches). This chapter evaluates pre-clustering and post-clustering algorithms on real-time data and synthetic datasets. Unlike traditional clustering algorithms, pre-clustering algorithms provide a prior clustering on different datasets. Simulation studies are carried out using datasets having both class-labeled and unlabeled information. Comparative studies are performed between results of existing pre-clustering and proposed pre-clustering approaches. A simulated RDI-based preprocessing method is also applied for data diversification. Extensive simulation on real and synthetic datasets shows that pre-clustering algorithms with simulated RDI-based pre-processing performs better compared to conventional post-clustering algorithms.

INTRODUCTION

This chapter presents a comparative study of the proposed approaches viz., Extended Dark Block Extraction (EDBE), Extended Cluster Count Extraction (ECCE), and Extended Co-VAT approaches. This chapter evaluates Pre-Clustering and Post-Clustering algorithms on real-time data and synthetic datasets. Unlike traditional clustering algorithms, Pre-Clustering algorithms provide a prior clustering on different datasets. Simulation studies are carried out using datasets having both class-labeled and unlabeled information. Comparative studies are performed between results of existing Pre-Clustering

and proposed Pre-Clustering approaches. A simulated RDI based preprocessing method is also applied for data diversification. Extensive simulation on real and synthetic datasets shows that Pre-Clustering algorithms with simulated RDI based pre-processing performs better compared to conventional Post-Clustering algorithms.

Comprehensive study has been conducted on cluster validation metrics for several Pre-Clustering approaches. For real time data, Pre-Clustering results are generated based on "dark blocks" concept with EDBE, ECCE and Extended Co-VAT approaches on *WINE* and *IRIS* data samples. It is important to compare various cluster validation metrics and select one that fits best with the "VAT" data distribution. Cluster validation is the process of assessing the quality and reliability of the cluster sets derived from various Pre-Clustering processes. A comparative study has been performed using four cluster validation metrics viz., clustering accuracy, clustering error, time complexity and cluster coefficient for assessing the quality of different pre-clustering approaches. The following Table 6.1 describes two datasets viz., *NUMERIC* and *IMAGE* dataset that consist of characteristics such as physical classes, attributes, size of the dataset and number of clusters. Pre-Clustering approaches are applied on different sizes of datasets in-order to detect the number of clusters.where $N_1 - N_4$ stands for *NUMERIC* dataset, $I_1 - I_4$ are *IMAGE* datasets.

The *NUMERIC* and *IMAGE* dataset characteristics i.e. size of datasets, physical classes, attributes, and number of clusters of EDBE, ECCE and Extended Co-VAT are summarized in Table 1.

Table 1. Datasets with Characteristics for the Proposed Approaches

Synthetic Dataset	Physical Classes	Attributes	Size (n)	Generated number of Clusters (C)		
				Extended Co-VAT	ECCE	EDBE
Numeric Dataset N_1	2	3	3*3	2	2	2
Numeric Dataset N_2	3	10	10*10	2	2	2
Numeric Dataset N_3	3	50	5000* 5000	3	3	3
Numeric Dataset N_4	5	100	100000* 100000	3	3	3
IMAGE Dataset I_1	2	3	3*3	2	2	2
IMAGE Dataset I_2	3	10	10*10	2	2	2
IMAGE Dataset I_3	3	50	5000* 5000	3	3	3
IMAGE Dataset I_4	5	100	100000* 100000	3	3	3

Evaluation Measures

One of the most important issues in cluster analysis is the evaluation of the clustering results. The evaluation of clustering results is the most difficult task within the whole clustering workflow.

The ways of evaluation are divided in two parts:

- Internal quality measures, and
- External quality measures.

Related Content

Features Selection Study for Breast Cancer Diagnosis Using Thermographic Images, Genetic Algorithms, and Particle Swarm Optimization

Amanda Lays Rodrigues da Silva, Maíra Araújo de Santana, Clarisse Lins de Lima, José Filipe Silva de Andrade, Thifany Ketuli Silva de Souza, Maria Beatriz Jacinto de Almeida, Washington Wagner Azevedo da Silva, Rita de Cássia Fernandes de Limaand Wellington Pinheiro dos Santos (2021). *International Journal of Artificial Intelligence and Machine Learning (pp. 1-18).*

www.irma-international.org/article/features-selection-study-for-breast-cancer-diagnosis-using-thermographic-images-genetic-algorithms-and-particle-swarm-optimization/277431

Automatic Multiface Expression Recognition Using Convolutional Neural Network

Padmapriya K.C., Leelavathy V.and Angelin Gladston (2021). *International Journal of Artificial Intelligence and Machine Learning (pp. 1-13).*

www.irma-international.org/article/automatic-multiface-expression-recognition-using-convolutional-neural-network/279275

Coffee Leaf Diseases Classification Using Deep Learning Approach

Sudhir Kumar Mohapatra, Anbesaw Belete, Ali Hussen, Abdelah Behari, Seid Huseenand Srinivas Prasad (2024). *Machine Learning Algorithms Using Scikit and TensorFlow Environments (pp. 91-111).*

www.irma-international.org/chapter/coffee-leaf-diseases-classification-using-deep-learning-approach/335185

An Integrated Process for Verifying Deep Learning Classifiers Using Dataset Dissimilarity Measures

Darryl Hond, Hamid Asgari, Daniel Jefferyand Mike Newman (2021). *International Journal of Artificial Intelligence and Machine Learning (pp. 1-21).*

www.irma-international.org/article/an-integrated-process-for-verifying-deep-learning-classifiers-using-dataset-dissimilarity-measures/289536

International Trade, Economic Growth, and Turkey

Aytaç Gökmen (2023). *Encyclopedia of Data Science and Machine Learning (pp. 640-650).*

www.irma-international.org/chapter/international-trade-economic-growth-and-turkey/317476