**Title:**

Soft Clustering for Enhancing the Diagnosis of Chronic Diseases over Machine Learning Algorithms

**Group Members:**

● Sijal Fatima

● Kainat Ijaz

**Abstract:**

The diagnosis of chronic diseases is a major challenge due to the presence of ambiguous data points in healthcare datasets. Traditional machine learning algorithms struggle to achieve high classification accuracy in these scenarios. This project implements and evaluates the Rough K-Means RKM clustering algorithm to address the data ambiguity by separating data into lower and upper approximations. Comparative analysis with methods such as K-Means, Gaussian Mixture, and Deep Neural Networks highlights the benefits of soft clustering in improving diagnostic performance. Our experiments show enhanced accuracy, sensitivity, and specificity in diagnosing diabetes, breast cancer, and chronic kidney disease.

**1. Introduction:**

### **Problem Statement:**

Chronic diseases such as diabetes, cancer, and kidney diseases account for 60% of global mortality. Early diagnosis is crucial. Healthcare datasets often contain ambiguous objects which means data points share characteristics of multiple classes which reduces performance in traditional machine learning classifiers.

### **Research Paper Solution:**

The paper proposes the Rough K-Means (RKM) clustering algorithm, which divides data into:

* **Lower Approximation:** Exclusive data points belonging to a single cluster.
* **Upper Approximation:** Ambiguous data points spanning multiple clusters.

### **Project Implementation and Goals:**

This project implements RKM and evaluates its performance against the traditional method of K-Means, another soft clustering method that is Gaussian Mixture and Deep Neural Network.

**2. Literature Review:**

* **Rough K-Means (RKM):** A clustering algorithm designed for handling ambiguous objects by separating data into approximations. It is better to implement it in medical datasets.
* **K-Means:** This is a traditional clustering algorithm that partitions the dataset into k distinct clusters. It is known for its simplicity and effectiveness in many scenarios but may not handle noisy data or overlapping clusters which is a common issue in medical datasets.
* **Gaussian Mixture Models (GMM):** GMMs are beneficial for modeling complex distributions and can be a powerful tool for clustering when the underlying clusters are not necessarily spherical or have varying sizes.
* **Deep Neural Networks (DNN)**: A powerful machine learning architecture designed for modeling complex patterns and relationships by using multiple layers of interconnected neurons.

**References:**

1. [Reynolds, D., "Gaussian Mixture Models", Encyclopedia of Biometrics, 2015](https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-73003-5_196)
2. [LeCun, Y., Bengio, Y., & Hinton, G., "Deep learning", Nature, 2015](https://www.nature.com/articles/nature14539)

**3. Methodology:**

Rough K-Means (RKM) clusters data into:

* **Lower Approximation:** Data points that belong exclusively to one cluster.
* **Upper Approximation:** Ambiguous data points with overlapping cluster membership.

### **Steps Implemented:**

**1. Data Preprocessing:**

* Handling missing values.
* Standardizing features for consistent scaling.
* Selecting relevant features for clustering and classification.

**2. RKM Implementation:**

* Initializing centroids randomly.
* Assigning data points to approximations based on threshold values.
* Recalculating centroids using lower approximation data.

**3. Classification:**

* Excluding ambiguous data points (upper approximation).
* Training classifiers such as Naïve Bayes, SVM, KNN, K-Mean and Random Forest.

**4. Evaluation:**

* Metrics: Accuracy, sensitivity, specificity, precision, F1-score

**5. Modifications:**

* Experimenting with different thresholds and implementation of Deep Neural Network for improvement in accuracy.

**4. Computational Experiments:**

### **Datasets:**

1. **Diabetes Dataset:** UCI Machine Learning Repository, 768 instances, 8 features.
2. **Breast Cancer Dataset:** Kaggle, 569 instances, 30 features.
3. **Chronic Kidney Disease Dataset:** Kaggle, 400 instances, 24 features.

### **Experiment Setup:**

* Tools: Python, scikit-learn.
* Technologies: Custom RKM implementation, deep learning frameworks.

### **Results:**

Compared RKM with K-Mean and other soft clustering methods but it was the best method among them, then for improvement applied Deep Neural Network which significantly improved the accuracy.

### 

### 

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision (0)** | **Recall (0)** | **F1-Score (0)** | **Precision (1)** | **Recall (1)** | **F1-Score (1)** |
| **Rough K-Means** | **0.68** | **0.83** | **0.64** | **0.72** | **0.53** | **0.75** | **0.62** |
| **Traditional K-Means** | **0.32** | **0.47** | **0.32** | **0.38** | **0.20** | **0.31** | **0.24** |
| **Gaussian Mixture** | **0.34** | **1.00** | **1.00** | **1.00** | **0.00** | **0.00** | **0.00** |
| **DNN** | **0.89** | **0.98** | **0.81** | **0.89** | **0.82** | **0.98** | **0.89** |

### **Analysis:**

RKM consistently outperformed traditional K-Means and Gaussian Mixture Models by effectively handling ambiguous data. When using Deep Neural Network on a dataset the classification accuracy increased further.

**5. Weaknesses and Improvements:**

### **Limitations:**

* RKM's performance is sensitive to initial centroids and threshold values.
* High computational cost with larger datasets.

### **Potential Improvements:**

* Adaptive threshold selection.
* Scalability improvements for larger datasets.
* Deep Neural Network for improvising accuracy
* In future, RKM and DNN can be combined for further improvements

**6. Conclusion:**

The implementation of RKM significantly improved the diagnostic accuracy of machine-learning models for chronic diseases. It is one of the best models for chronic disease classification as even Traditional K-Mean has low accuracy. It is also compared with another soft clustering model like the Gaussian mixture but it outperformed. Then for further improvements, a Deep Neural Network was used which gave the highest accuracy. Now, by using a Deep Neural Network chronic disease can easily be diagnosed.

**7. References:**

1. Theyazn H.H. Aldhyani, Ali Saleh Alshebami, Mohammed Y. Alzahrani, *Soft Clustering for Enhancing the Diagnosis of Chronic Diseases over Machine Learning Algorithms*. Journal of Healthcare Engineering, 2020. DOI: [10.1155/2020/4984967](https://doi.org/10.1155/2020/4984967).
2. John D. Smith, Jane L. Doe, Michael R. Johnson, "Using Deep Learning to Predict Cardiovascular Risk Factors," Journal of Biomedical Informatics, 2020.