MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

USING CLUSTER ENSEMBLES FOR DETECTION AND ADDRESSING OF CONCEPT DRIFTS IN DATA STREAMS

Sriharsha Daparti 140953194
Visal Kancharla 140953148

Under the Guidance of
Mr. Nirmal Kumar Nigam
Asst. Professor – Sr. Scale

INTRODUCTION

- Data mining vs Data analysis
- Reasons for choosing clustering
- Clusters vs Cluster Ensembles
- Ensemble techniques - Basic
- Data streams
- Concept drifts

OBJECTIVE

- To identify (or develop) clustering algorithms for detection of concept drifts in data streams

METHODOLOGY

1. Dataset Acquisition
2. About the dataset
3. Use of database
4. Mixed Attributes
5. Handling mixed data
6. Medium of processing & clustering
7. Algorithms – Trails and assumptions
8. Algorithms – Final selection & reason
9. Available ensemble techniques
10. Consensus clustering – Not the best fit
11. Stacking implementation
12. Validation - Silhouette

VALIDATION MEASURE - SILHOUETTE

- What is Silhouette Coefficient?
- Cohesion
- Separation
- Calculation
- Reasons for selecting silhouette

PERFORMANCE ANALYSIS

No of Clusters(k)	Distance Measure	Cluster Sizes(no of data points /cluster)
2	Euclidean	96,4
2	Manhattan	4,96
2	Least Squares	8,92
3	Euclidean	4,82,14
3	Manhattan	4,82,14
3	Least Squares	4,82,14
4	Euclidean	66,10,3,21
4	Manhattan	20,3,70,7
4	Least Squares	4,64,22,10
5	Euclidean	12,58,7,4,19
5	Manhattan	9,18,59,4,10
5	Least Squares	10,2,22,2,64

Table1. Table showing Clusters with various k value and distance metrics

PERFORMANCE ANALYSIS

K clusters with specified metrics*	Average Silhouette Width		
	Independent	Self-Ensemble#	Cross Ensemble∞
2 - Euclidean	0.84	0.83	0.72
2 - Manhattan	0.83	0.84	0.83
3 – Euclidean	0.72	0.69	-
3 – Manhattan	0.69	0.72	-
4 – Euclidean	0.61	0.58	-
4 – Manhattan	0.58	0.62	-
5 – Euclidean	0.50	0.37	-
5 – Manhattan	0.47	0.59	-
2 - CLARA – Euclidean	0.72	0.71	0.84
2 - CLARA - Manhattan	0.68	0.72	0.83

* Default clustering algorithm is k-means
# If the initial clustering is done with Euclidean, ensembling is done with Manhattan (using the same algorithm) and vice-versa
∞ If the initial clustering is done with Euclidean, ensembling is also done with the same measure but using a different clustering. Here the cross was done between CLARA and k-means

Table2. Comparison using Silhouette - I

PERFORMANCE ANALYSIS

K clusters with specified metrics*	Average Silhouette Width		
	Independent	Self-Ensemble#	Cross Ensemble∞
5 – Euclidean	0.55	0.51	0.53
5 – Manhattan	0.51	0.55	0.36
10 – Euclidean	0.40	0.35	0.38
10 – Manhattan	0.37	0.38	0.34
20 – Euclidean	0.32	0.30	0.37
20 – Manhattan	0.34	0.35	0.29
5 - CLARA – Euclidean	0.53	0.38	0.55
5 - CLARA - Manhattan	0.37	0.47	0.51
10 - CLARA – Euclidean	0.37	0.34	0.40
10 - CLARA - Manhattan	0.33	0.37	0.37
20 - CLARA – Euclidean	0.33	0.30	0.36
20 - CLARA - Manhattan	0.29	0.36	0.31

Table3. Comparison using Silhouette - II

K (Used only with k-means)	K-means on HDBSCAN	HDBSCAN on K-means	HDBSCAN (Independent)
5	0.44	0.37	0.37
10	0.41	0.38	0.37
20	0.35	0.37	0.37

CONCEPT DRIFT

- Concept drifts are posed by data streams.
- Lack of a live data stream
- Static data manipulation
- Application of Cluster Ensembles
- Observation and detection

CONCLUSION

The aim of the project was to construct cluster ensembles and help formulate a comparison between regular clustering techniques and cluster ensembles. It was also intended to detect concept drifts of any kind in the data streams or large datasets (which can practically be assumed as data streams).

It constitutes a successful implementation of data type conversion i.e., from categorical to numerical, for working with the partitioning algorithms and hence be able to process and cluster any data with mixed data types.

The project demonstrates a way to design an ensemble using clustering algorithms with the help of ensemble techniques.