Clustering Indus Texts using *K*-means

Nisha Yadav
Department of Computer Science, University of Mumbai, Santacruz, Mumbai – 400 098.
Department of Astronomy & Astrophysics, Tata Institute of Fundamental Research, Colaba, Mumbai- 400 005.

Ambuja Salgaonkar
Department of Computer Science, University of Mumbai, Santacruz, Mumbai – 400 098.

Mayank Vahia
Department of Astronomy & Astrophysics, Tata Institute of Fundamental Research, Colaba, Mumbai- 400 005.

ABSTRACT

One of the most important undeciphered scripts of the ancient world is the Indus script. Earlier studies had focused on the correlations between signs in the Indus texts using various statistical and computational techniques such as *N*-grams or Markov chains. In the present study, *K*-means clustering, an unsupervised machine learning technique is used to identify clusters of similar texts without making any assumptions about its content. The technique is effective in extracting significant clusters and patterns in the script. Nine clusters are extracted from this study. The texts in each cluster share a common set of structural elements and are more similar to each other than the texts in other clusters. The clusters, as extracted from the study, reveal inherent patterns due to adjacent and non-adjacent dependencies between signs in the Indus texts. These clusters have definitive patterns in the usage of the signs but are only weakly associated to any archaeological site or medium of writing. The characteristic signature features of each cluster are identified in the study. The study provides a good handle to extract the logic of writing in the Indus script.

General Terms

Machine learning, unsupervised learning, clustering, *K*-means

Keywords

Indus texts, ancient script, undeciphered script

1. INTRODUCTION

Unsupervised machine learning techniques such as clustering find widespread use in several scientific and commercial applications such as web search, information retrieval, image pattern recognition, business intelligence, marketing, biology and security [1-3]. These techniques are also used to cluster documents based on their topics. However, such studies have largely focused on known languages where other background information is available. In this study, this technique is used to cluster the undeciphered texts of the Indus valley civilization (ca. 2600 to 1900 BC).

The Indus valley civilization is the largest Bronze Age civilizations of the ancient world that flourished over an area of about a million square kilometers in the north western parts of the Indian subcontinent (see e.g. [4-7], for a detailed overview). The civilization has left about four thousand samples of its writing on various types of objects that include seals, sealings, miniature tablets (generally made of steatite and terracotta), copper tablets, bronze implements, stone or ivory objects, pottery shreds and other miscellaneous objects. Reasons that make the problem of the Indus script more challenging are the brevity of the Indus texts (average length of an Indus text ~ 5 signs), lack of definitive knowledge about the language(s) of the Indus Valley people, and absence of bilingual or multilingual inscriptions. Despite these hurdles, there have been numerous attempts in the past to decipher the Indus script but there is no universal consensus on any of the proposed interpretations [8-11].

Earlier studies have explored the sequential structure of the Indus script using various statistical and computational techniques [12-20], the design of Indus signs [21] and different types of patterns inscribed on these objects [22-24]. These studies do not make any assumptions about its nature, content or meaning. They have demonstrated that Indus texts have a rich syntax and an underlying logic in their structure. In the present study, the approach is extended to identify clusters of Indus texts based on contagious as well as non-contagious correlations between signs using the technique of unsupervised machine learning. The subsequent sections provide an overview of the methodology and the results followed by discussion and conclusion.

2. CLUSTERING INDUS TEXTS

Clustering partitions a dataset into subsets such that the elements within a subset (or cluster) share high level of similarity amongst themselves. Clustering algorithms can be broadly classified into partition-based clustering (such as *K*-means), hierarchical clustering (such as agglomerative clustering), and fuzzy clustering (such as fuzzy *K*-means) [1-3].

The choice of the clustering algorithm generally depends upon the application. Some clustering algorithms require specification of the number of clusters before the analysis is performed while in other cases the number of clusters is determined based on the results of the analysis. Optimal clustering results can be obtained by varying the distance measure or other clustering parameters. There are several methods to compute the distance between the elements of a dataset. Some of them include the Euclidean distance, the Manhattan distance, cosine measure and so on. *K*-means, one of the simplest yet a widely used clustering method is used here to cluster the texts in the corpus of the Indus script [1-3]. It explores the contagious and non-contiguous dependencies between the signs in the Indus texts and clusters the Indus texts based on their similarity in the usage of distinct signs.

2.1 Data

The EBUDS corpus of the Indus script is used in the present study which contains 1548 texts with 7000 sign occurrences [12]. EBUDS is a filtered corpus created from Mahadevan's concordance [25] after removal of duplicates and ambiguous texts. The number of distinct signs in EBUDS is 377. As a convention followed in the present study, the texts depicted as strings of sign images are to be read from right to left, whereas the texts represented as strings of sign numbers are to be read from left to right.

2.2 Method

In order to implement the *K*-means clustering algorithm, a term by document matrix was created for the texts in the EBUDS corpus of the Indus script. Terms are the individual signs of the Indus script and documents are the distinct texts of the Indus script. The term by document matrix for Indus texts was then subjected to *K*-means clustering with cosine as a measure of similarity. *K*-means clustering routine permits creation of an arbitrary number of clusters (*K*). The clusters were analyzed by varying the *K*-value from 2 to 20, and the content of the clusters was evaluated in each case. The clustering with *K* = 9 optimizes the content of each cluster and the boundary separating each cluster is found to be distinct with respect to its constituent texts in this case. The resulting nine clusters were used for further analysis.

3. RESULTS

3.1 Basic characteristics of the clusters

General characteristics of the nine clusters C1 to C9 are given in Table 1.

Table 1: General characteristics of the clusters C1 to C9

	Number of texts	Number of distinct signs	Number of sign occurrences
EBUDS	1548	377	7000
C1	187	157	622
C2	70	97	328
C3	141	112	709
C4	276	197	1571
C5	352	201	1556
C6	105	120	514
C7	151	163	484
C8	138	139	561
C9	128	150	655

As can be seen from Table 1, the size of the nine clusters vary, with the smallest cluster C2 having about 5 percent of the texts in EBUDS and it is almost 20% of the size of the largest cluster C5. The largest cluster C5 accounts for about 23% of the texts in EBUDS. In general, the number of signs used in the cluster is related to the size of the cluster.

3.2 Comparison of text lengths

The average length of a text in EBUDS is ~4.5 signs with the shortest text of 1 sign and longest text of 14 signs. In Figure 1, the cumulative percentage of texts of various lengths in EBUDS is compared with the nine clusters.

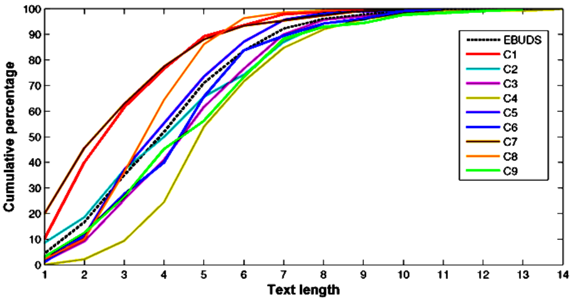


Figure 1: Cumulative distribution of text lengths in EBUDS in comparison to the nine clusters C1 to C9.

It can be seen from Figure 1 that

1) The distribution of texts of various lengths in clusters C1 to C9 is distinct.

2) Cluster C4 has highest average text length and clusters C1 and C7 have lowest average text length.

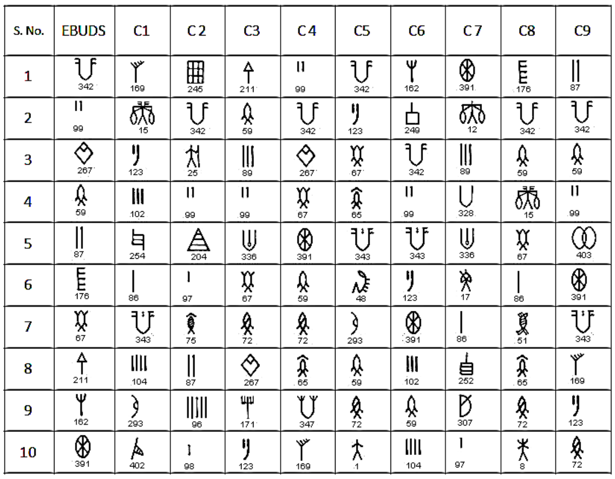
3) About 75% texts in C4 and 60% of texts in C6 are of length 5 or more compared to 48% in EBUDS.

4) In C1 and C7, texts of length less than or equal to two signs account for about 40% and 45% of the constituent texts respectively, compared to 17% in EBUDS.

3.3 Comparison of frequent signs

In Table 2, the ten most frequent signs in the clusters C1 to C9 are listed.

Table 2: Ten most frequent signs (listed in descending order of frequency) in clusters C1 to C9.



S. No.	EBUDS	C1	C2	C3	C4	C5	C6	C7	C8	C9
1	342	169	245	211	99	342	162	391	176	87
2	99	15	342	59	342	123	249	12	342	342
3	267	123	25	89	267	67	342	89	59	59
4	59	102	99	99	67	65	99	328	15	99
5	87	254	204	336	391	343	343	336	67	403
6	176	86	97	67	59	48	123	17	86	391
7	67	343	75	72	72	293	391	86	51	343
8	211	104	87	267	65	59	102	252	65	169
9	162	293	96	171	347	72	59	307	72	123
10	391	402	98	123	169	1	104	97	8	72

As can be seen from Table 2, the most prominent signs in the clusters vary significantly. Some of the important conclusions from Table 2 are:

1) The most frequent sign in each cluster is a terminal sign (text beginner or text ender [12]). Either the most frequent or the second most frequent sign in each cluster is a frequent text ender.

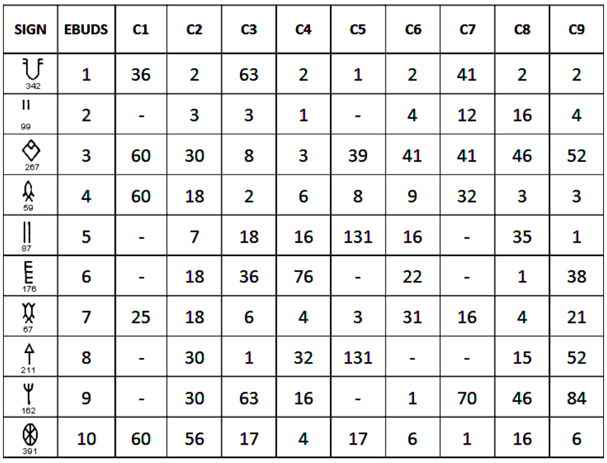
2) While a text ender or a text beginner may appear as the most frequent sign in a cluster, the set of prominent terminal signs in each cluster is unique. These set of terminal signs constitute a signature pattern for each cluster.

3) The signs other than the terminal signs in the clusters have much lower frequency in EBUDS. This suggests a specific associative structure where relatively lower frequency signs show significant affinity to a specific set of terminal signs.

4) The most frequent text beginner in EBUDS, sign number 267, is conspicuous by its absence in the list of most frequent signs of the nine clusters.

The ranks of the ten most frequent signs in EBUDS and the nine clusters C1 to C9 are compared in Table 3.

Table 3: Comparison of ranks of ten most frequent signs in EBUDS and clusters C1 to C9. Dash indicates that the sign does not appear in that cluster



SIGN	EBUDS	C1	C2	C3	C4	C5	C6	C7	C8	C9
342	1	36	2	63	2	1	2	41	2	2
99	2	-	3	3	1	-	4	12	16	4
267	3	60	30	8	3	39	41	41	46	52
59	4	60	18	2	6	8	9	32	3	3
87	5	-	7	18	16	131	16	-	35	1
176	6	-	18	36	76	-	22	-	1	38
67	7	25	18	6	4	3	31	16	4	21
211	8	-	30	1	32	131	-	-	15	52
162	9	-	30	63	16	-	1	70	46	84
391	10	60	56	17	4	17	6	1	16	6

Table 3 shows that there is significant difference in the frequency of occurrence of signs in various clusters. For example, sign number 342 which has rank 1 in EBUDS retains its rank only in C5 while it is ranked 63 in C3 and 41 in C7.

3.4 Text beginner-ender asymmetry

There exists an asymmetry in the usage of text beginners and text enders in EBUDS [14]. The number of signs required to account for 80% of the text beginners, text enders and all sign occurrences in EBUDS and the nine clusters is given in Table 4.

Table 4: Number of signs constituting 80% of text beginners, text enders and all signs in EBUDS and clusters C1 to C9.

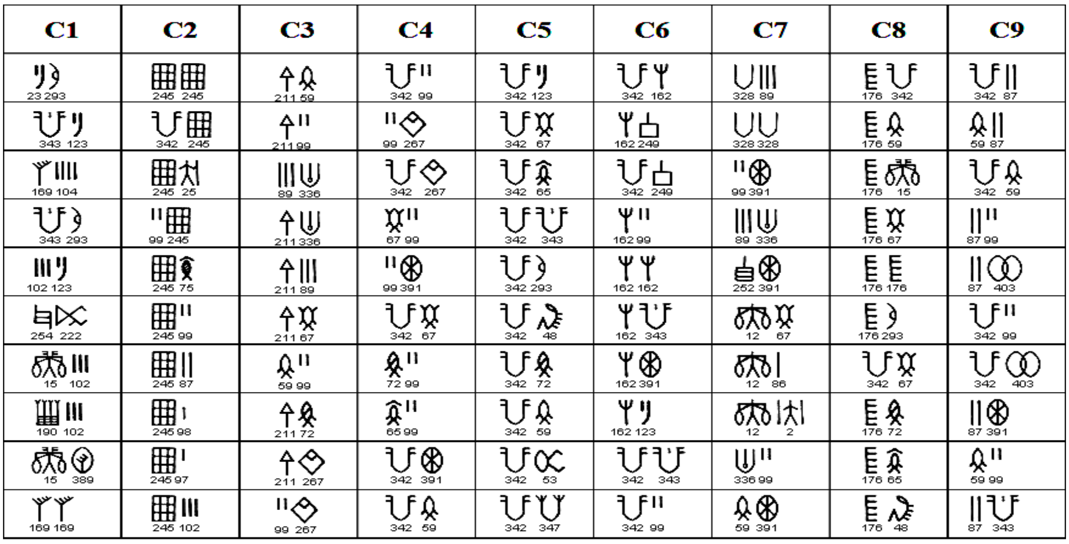
	No. of texts	No. of text beginners (B)	No. of text enders (E)	All signs	E/B
EBUDS	1548	82	23	69	0.3
C1	187	55	22	58	0.4
C2	70	22	13	43	0.6
C3	141	35	3	28	0.1
C4	276	21	10	49	0.5
C5	352	57	1	52	0.0
C6	105	32	3	38	0.1
C7	151	55	44	67	0.8
C8	138	52	1	49	0.0
C9	128	36	5	50	0.1

The usage pattern of text beginners and text enders for EBUDS and the nine clusters is compared in Table 4. The number of text enders required to account for 80% of the texts varies from 1 for clusters C5 and C8 to 44 for C7. The variation in the number of text beginners is from 21 for C4 to 57 for C5. All this suggests that each cluster has affinity to a very restricted number of text enders. This is highlighted in the last column of Table 4.

3.5 Comparison of sign combinations

The most frequent sign pairs and sign triplets in clusters C1 to C9 are listed in Tables 5 and 6 respectively. Note that the sign combinations may not always be contiguous in the texts.

Table 5: Most frequent sign pairs in C1 to C9.



C1	C2	C3	C4	C5	C6	C7	C8	C9
23 293	245 245	211 59	342 99	342 123	342 162	328 89	176 342	342 87
343 123	342 245	211 99	99 267	342 67	162 249	328 328	176 59	59 87
169 104	245 25	89 336	342 267	342 65	342 249	99 391	176 15	342 59
343 293	99 245	211 336	67 99	342 343	162 99	89 336	176 67	87 99
102 123	245 75	211 89	99 391	342 293	162 162	252 391	176 176	87 403
254 222	245 99	211 67	342 67	342 48	162 343	12 67	176 293	342 99
15 102	245 87	59 99	72 99	342 72	162 391	12 86	342 67	342 403
190 102	245 98	211 72	65 99	342 59	162 123	12 2	176 72	87 391
15 389	245 97	211 267	342 391	342 53	342 343	336 99	176 65	59 99
169 169	245 102	99 267	342 59	342 347	342 99	59 391	176 48	87 343

Table 6: Most frequent sign triplets in C1 to C9.

C1	C2	C3	C4	C5	C6	C7	C8	C9
343 123 293	245 245 25	211 89 336	342 99 267	342 123 293	342 162 249	328 328 12	176 342 67	342 59 87
15 389 178	245 25 98	211 99 267	342 67 99	342 67 65	162 162 162	328 342 48	176 342 48	342 87 99
190 102 123	245 25 345	211 59 99	342 99 391	342 123 150	162 249 343	162 328 48	176 342 8	342 87 403
02 123 293	245 25 97	211 59 267	342 347 99	342 89 336	162 99 391	162 328 386	176 342 65	59 87 99
342 169 249	245 98 99	211 336 99	342 8 99	343 123 293	342 162 343	162 328 12	176 342 193	342 59 99

3.6 Sensitivity of the clusters to sites and type of object

The sensitivity of the nine clusters to the archaeological sites of occurrence and the medium of writing is illustrated in Figures 2 and 3 respectively. No signification correlation between the clusters and the archaeological sites or medium of writing is observed. Only cluster C8 has a large fraction of texts on sealings (impressions of seals) from the site of Harappa.

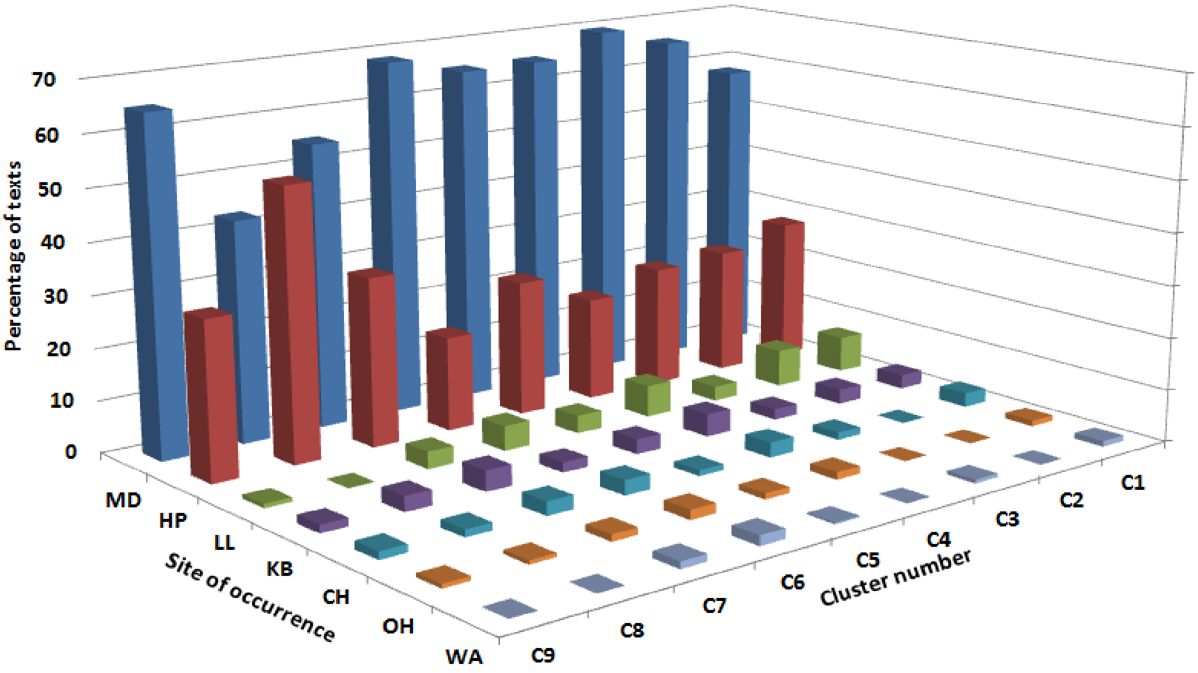


Figure 2: Percentage contribution of different archaeological sites to clusters C1 to C9 (MD: Mohenjodaro, HP: Harappa, LL: Lothal, KB: Kalibangan, CH: Chanhudaro, OH: Other Harappan sites, WA: West Asian sites). The data from each cluster is normalized to 100% to compare the relative contribution of various sites to different clusters.

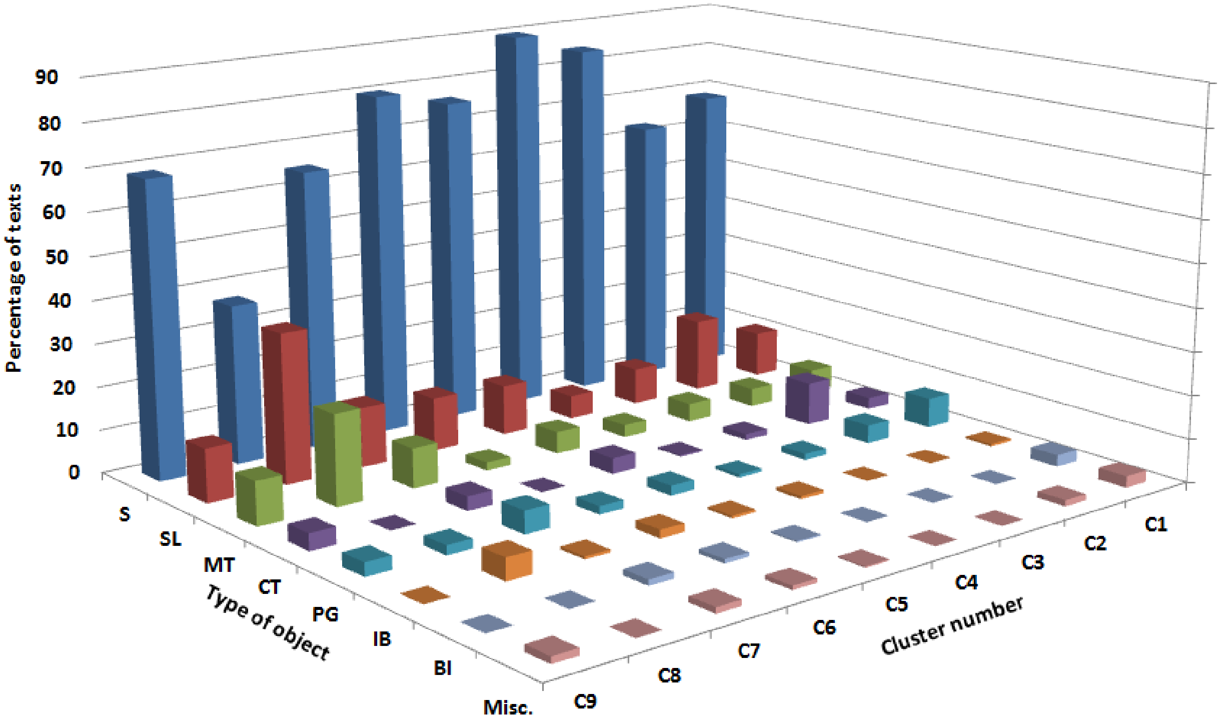


Figure 3: Percentage contribution of different type of objects to clusters C1 to C9 (S: Seals, SL: Sealings, MT: Miniature Tablets, CT: Copper Tablets, PG: Pottery Graffiti, IB: Ivory or Bone rods, BI: Bronze Implements, Misc.: Miscellaneous objects). The data from each cluster is normalized to 100% to compare the relative contribution of distinct types of objects to different clusters.

4. DISCUSSION

In the present study an unsupervised machine learning technique is used to identify texts with similar usage of signs in the corpus of Indus script. Using *K*-means, the clustering is found to be optimal when the texts are divided into nine clusters C1 to C9. The broad features of these clusters are given in Table 1. The distribution of text lengths in different clusters is given in Figure 1. It is clear from Table 1 and Figure 1 that the distribution of texts in various clusters is not uniform and the clusters differ not only in their size but also in the distribution of their text lengths. Clusters C4 and C6 consist of a large fraction of longer texts whereas clusters C1 and C7 consist of a large fraction of shorter texts.

In Table 2 the ten most frequent signs in each cluster are listed. Table 2 gives a clear indication that the nine clusters have their own unique identity in terms of their sign preferences. It can be seen from Table 2 that the most frequent sign in clusters C1 to C9 is different for each cluster. Moreover, each of these signs have a rank between 1 to 10 in EBUDS except for sign numbers 169 and 245 for clusters C1 and C2 that have ranks 17 and 18 respectively in EBUDS (Table 3). In the largest cluster C5, sign number 342 (the most frequent sign as well as the most frequent text ender in EBUDS) retains its rank 1. However the second most frequent sign in EBUDS (sign number 99) does not appear in cluster C5 at all. Cluster C5 is also characterized by very rare or no occurrence of other frequent text enders in EBUDS such as sign numbers 176 and 211. Similarly, C3 is characterized by the high frequency of sign number 211 (rank 1 in C3) but the rank of sign number 342 in C3 is 63, significantly lower than that in EBUDS (rank 1). Such differences in the occurrence pattern of different signs in clusters C1 to C9 illustrate distinct properties of each cluster.

The statistics of the usage of signs that begin and end texts as well as the signs that appear at any location in the texts of each cluster is given in Table 4. The total number of signs constituting 80% of text beginners, text enders and all signs, in EBUDS and clusters C1 to C9, vary significantly. As in the case of EBUDS, there exists an asymmetry in the usage pattern of text beginners and text enders in all the clusters with far fewer signs ending the texts than beginning them [14]. However, the absolute number of signs that can begin or end texts varies for each cluster (Table 4). In the extreme case, the number of signs required to end the texts in C7 (44) is more than that in EBUDS (23). Clusters C5 and C8 require just one text ender (sign number 342) to account for 80% of its text enders. In contrast, cluster C7 with 55 text beginners and 44 text enders is at the other extreme and seems to hold miscellaneous texts. The clusters seem to have a small ender to beginner ratio suggesting that they are internally homogenous sets of texts, with their own set of preferred signs while conforming to the accepted grammatical style.

In order to identify the most prominent structures in each cluster, the most frequent contiguous and non-contiguous sign pairs and sign triplets are listed in Tables 5 and 6 respectively. Each cluster has an affinity to specific signs and sign groups. For example, while sign number 342 is amongst the frequent signs in the clusters C2, C4, C5, C6, C8 and C9 (rank 1 to 3, see Table 3), the sign often associated with it in the most frequent sign pairs in each cluster is different (Table 5). The current analysis demonstrates that there are different substructures in Indus writing that distinguish apparently similar texts by their association to specific set of signs.

Analysis of the association of these clusters to different archaeological sites of occurrence and types of objects is detailed in Figures 2 and 3. There is no definitive affinity of the clusters to any archaeological site or medium of writing. However, a weak affinity of the clusters to different sites and media can be seen.

5. CONCLUSION

In the present study, *K*-means, an unsupervised machine learning technique, is used to identify substructures in the Indus writing. The study suggests that the Indus texts can be optimally divided into nine clusters, each with its own characteristic usage of signs and contiguous or non-contiguous sign pairs and sign triplets. Each cluster has affinity to specific signs and sign groups. This suggests that the writing in the Indus valley civilization had several distinct patterns in the usage of signs.

The analysis presented here has significant implications on any models of decipherment of the Indus script. Considering that the texts can be divided into nine clusters suggests that there are clusters of texts which primarily use different sets of signs from the general pool. The absence of any significant correlation with the archaeological sites or medium of writing suggests a high level of uniformity of writing over the entire stretch of the Indus civilization. This also indicates a high level of standardization and agreement in the function of each sign.

These and other studies [12-20], therefore provide a unique set of constraints on any interpretation of this undeciphered script. It suggests a high level of standardization, nine clusters of texts each with its own preferred set of signs and sign sets, a uniformity of grammar and an agreed function of each sign across the culture. In another study, based on the design of the Indus signs, it has been noted that the number of basic signs in the sign list of Indus script is 154 (about one-third of the total set of 417 signs) while other signs seem to be derived from these signs [21]. All these aspects provide crucial insights to understand the structure and usage of the Indus script.

6. REFERENCES

[1] Jain, A. K. and Dubes, R. C. 1988 Algorithms for Clustering Data. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

[2] Han, J., Kamber, M., and Pei, J. 2011 Data Mining: Concepts and Techniques. San Francisco, California: Morgan Kaufmann Publishers.

[3] Myatt, G. J. and Johnson, W. P. 2009 Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications. New Jersey: John Wiley and Sons, Inc.

[4] Kenoyer, J. M. 1998 Ancient Cities of the Indus Valley Civilization. Oxford: Oxford University Press.

[5] Possehl, G. L. 2002 The Indus Civilization: A Contemporary Perspective. New Delhi: Vistaar Publications.

[6] Wright, R. P. 2010 The Ancient Indus – Urbanism, Economy and Society. New York: Cambridge University Press.

[7] Vahia, M. N. and Yadav, N. 2011. Reconstructing the History of Harappan Civilisation. Journal of Social Evolution and History. 10, 67 - 86.

[8] Possehl, G. L. 1996 Indus Age: The Writing System. New Delhi: Oxford & IBH Publishing Co. Pvt. Ltd.

[9] Mahadevan, I. 2002. Aryan or Dravidian or Neither? A Study of Recent Attempts to Decipher the Indus Script (1995-2000). Electronic Journal of Vedic Studies. 8.

[10] Parpola, A. 1994 Deciphering the Indus Script. Cambridge: Cambridge University Press.

[11] Parpola, A. 2005. Study of the Indus Script. In Proceedings of the International Conference of Eastern Studies,Tokyo: The Tôhô Gakkai, , 28-66.

[12] Yadav, N., Vahia, M. N., Mahadevan, I. and Joglekar, H. 2008. A Statistical Approach for Pattern Search in Indus Writing. International Journal of Dravidian Linguistics. vol. XXXVII, pp. 39-52.

[13] Yadav, N., Vahia, M. N., Mahadevan, I. and Joglekar, H. 2008. Segmentation of Indus Texts. International Journal of Dravidian Linguistics. vol. XXXVII, pp. 53-72.

[14] Yadav, N., Joglekar, H., Rao, R. P. N, Vahia, M. N., Adhikari, R. and Mahadevan, I. 2010. Statistical Analysis of the Indus Script Using n-grams. PLoS ONE. vol. 5.

[15] Yadav, N. and Salgaonkar, A. 2012. Statistical Studies of the Indus Script. Man and Environment. vol. XXXVII pp. 1-7.

[16] Yadav, N., Salgaonkar, A. and Vahia, M. N. 2014. Computational Techniques for Inferring the Syntax of Un-Deciphered Scripts. International Journal of Computer Science and Applications. Vol. 11, No. 2, pp. 50-61.

[17] Yadav, N. 2013. Sensitivity of Indus Script to Site and Type of Object. Scripta, vol. 5, pp. 67-103.

[18] Rao, R. P. N., Yadav, N., Vahia, M. N., Joglekar, H., Adhikari, R. and Mahadevan, I. 2009. A Markov Model of the Indus Script. Proceedings of the National Academy of Sciences. vol. 106, pp. 13685-13690.

[19] Rao, R. P. N., Yadav, N., Vahia, M. N., Joglekar, H., Adhikari, R. and Mahadevan, I. 2009. Entropic Evidence for Linguistic Structure in the Indus Script. Science. vol. 324, p. 1165.

[20] Rao, R. P. N., Yadav, N., Vahia, M. N., Joglekar, H. Adhikari, R. and Mahadevan, I. 2010. Entropy, the Indus Script and Language: A Reply to R. Sproat. Computational Linguistics. vol. 36, pp. 795-805.

[21] Yadav, N. and Vahia, M. N. 2011. Indus Script: A Study of its Sign Design. Scripta, vol. 3, pp. 133-172.

[22] Vahia, M. N. and Yadav, N. 2010. Harappan Geometry and Symmetry: A Study of Geometrical Patterns on Indus Objects. Indian Journal of History of Science. vol. 45, pp. 343-368.

[23] Yadav, N. and Vahia, M. N. 2011. Classification of Patterns on Indus Objects. International Journal of Dravidian Linguistics. vol. 40, pp. 89-114.

[24] Sinha, S., Yadav, N. and Vahia, M. N. 2011. In Square Circle: Geometric Knowledge of the Indus Civilization. In Math Unlimited: Essays in Mathematics, R. Sujatha, H. N. Ramaswamy and C. S. Yogananda, Eds., ed Enfield: Science Publishers. pp. 451-462.

[25] Mahadevan, I. 1977. The Indus Script: Texts, Concordance and Tables. New Delhi: Archaeological Survey of India.