

# A modified K-means clustering for mining of multimedia databases based on dimensionality reduction and similarity measures

Xiaoping Jiang<sup>1</sup> · Chenghua Li<sup>1</sup> · Jing Sun<sup>1</sup>

Received: 24 February 2017 / Revised: 8 May 2017 / Accepted: 22 May 2017 / Published online: 2 June 2017  
© Springer Science+Business Media New York 2017

**Abstract** With rapid innovations in digital technology and cloud computing off late, there has been a huge volume of research in the area of web based storage, cloud management and mining of data from the cloud. Large volumes of data sets are being stored, processed in either virtual or physical storage and processing equipments on a daily basis. Hence, there is a continuous need for research in these areas to minimize the computational complexity and subsequently reduce the time and cost factors. The proposed research paper focuses towards handling and mining of multimedia data in a data base which is a mixed composition of data in the form of graphic arts and pictures, hyper text, text data, video or audio. Since large amounts of storage are required for audio and video data in general, the management and mining of such data from the multimedia data base needs special attention. Experimental observations using well known data sets of varying features and dimensions indicate that the proposed cluster based mining technique achieves promising results in comparison with the other well-known methods. Every attribute denoting the efficiency of the mining process have been compared component wise with recent mining techniques in the past. The proposed system addresses effectiveness, robustness and efficiency for a high-dimensional multimedia database.

**Keywords** Multimedia data bases · Clustering · Mining · K means clustering · Optimization

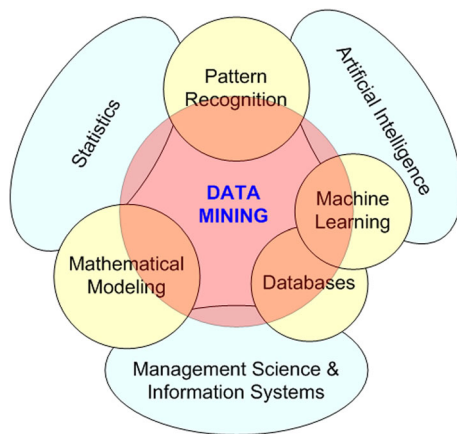
## 1 Introduction

With widespread growth in digital data communication and state of art data processing gadgets, the amount of data to be stored and processed for various applications in real time has grown in leaps and bounds [1]. Following this, there have been recent interests towards state of the art concepts like cloud computing [2], internet of things (IoT) [3], Content based retrieval [4] etc., the above mentioned technique have drastically reduced the necessity of large physical memory spaces whose cost increase with demand in an exponential manner. Virtual storage and management systems have started to emerge as the new trends in virtual processing eliminating the need for physical space, memory and processors for their computation. A drastic variation in computational time has also been found from research contributions in the past with the use of virtual processing. The proposed work addresses the issue of data mining from a data base which is composite in nature which is a mixture of text, video, audio and hypertext. A generalized concept of data mining for multimedia applications is depicted in Fig. 1.

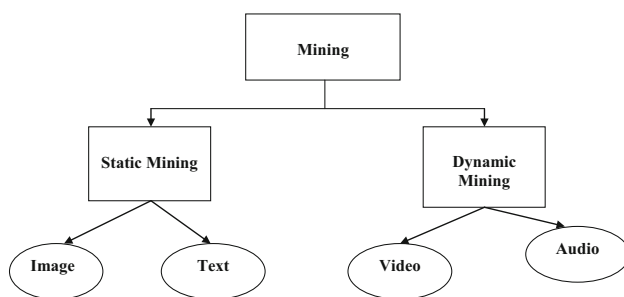
From Fig. 1, it could be seen that data mining is basically a conceptual overlap of multiple fields ranging from mathematical modelling, statistics, data base management, learning to recognition applications [5]. Since, the proposed work in this paper focuses on multimedia data, data mining of multimedia databases could be defined as systematic extraction of knowledge from data sets comprising of audio clippings, video sequences, images and text. The process of extraction starts with the mining tool operating on the above mentioned data sets which have been previously converted from their varying formats like JPEG, MPEG, BMP, WMV etc., to a common digital format. The various types of data under multimedia databases could be categorized as audio (speech clips, mp3 songs etc.), image (photographs from camera, artistic paint-

✉ Chenghua Li  
lichenghuachina@yahoo.com

<sup>1</sup> College of Electronics and Information Engineering, Hubei Key Laboratory of Intelligent Wireless Communications, South-Central University for Nationalities, Wuhan, China



**Fig. 1** Conceptual illustration of data mining in multimedia systems



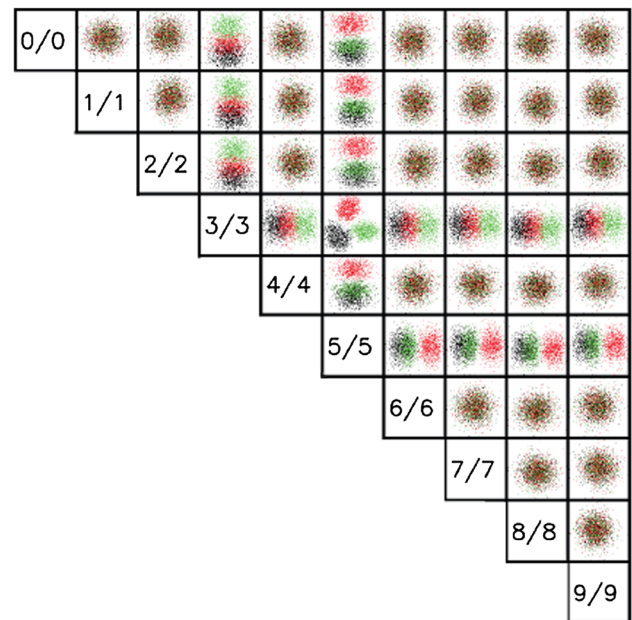
**Fig. 2** Illustration of mining classification

ing, design models etc.), video (sequence of frames divided in time), text (short messaging services, multimedia services) and electronic data in the form of signatures, ink from light pens, sensors etc.,

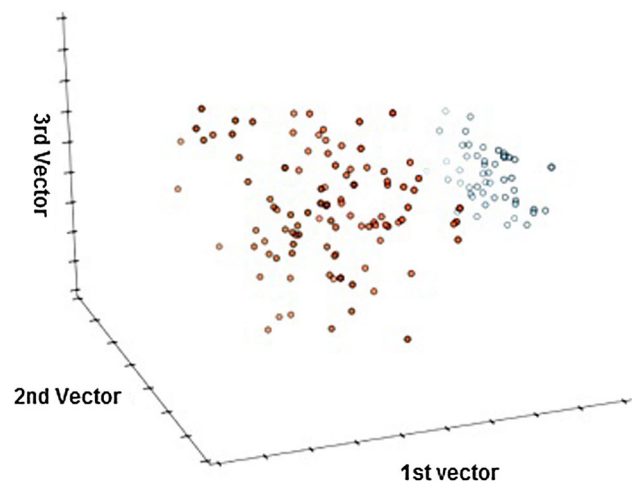
Data mining is a subset of knowledge discovery process [6, 7] also known as KDP. As the name indicates, data mining analyzes the current data at hand to obtain new information or knowledge through analysis of relationship between the classes in the data. Since the mining proposed in this paper operates on composite multimedia data, the scheme of mining is illustrated in Fig. 2 which features two important categories namely static and dynamic mining.

The entire concept of mining is achieved by an important process known as Clustering which is a form of unsupervised learning method. During the discovery of knowledge through clustering approaches, two major problems arise which is the dimension of data set at hand and similarity measure between the components in the class which is addressed in this research paper.

The first parameter addressed in this paper is the high dimensional data to be handled by the mining tool. Figure 3 illustrates the scatter plot between a small dimensional and high dimensional data model. High dimensional data results in data being lost in space. This could be understood with an example by considering the random distribution of 200



**Fig. 3** Illustration of a small dimensional and high dimensional data cluster



**Fig. 4** Three dimensional scatter plot of feature set of multimedia database

points in the interval of  $[0, 1]$ . When this interval is partitioned say into 20 cells, then it is evident that all cells will contain some points out of the total 200. On the other hand, if the number of points say 200 is kept constant and the distribution interval is discretized [8], then the condition that all cells may contain some points may not be achieved. Further, in three dimensionality, most of the cells would be empty resulting in data loss. A three dimensional cluster model is depicted in Fig. 4.

The second aspect is the similarity measure which is generic of the definition of clustering which involves grouping together meaningful data sets [9] with similar measure of some kind of functionality. Similarity measure involves a set

of attributes namely distance and similarity between binary clusters. The data sets in the clusters may be binary clusters (take upon two values), continuous or discrete. Euclidean distance, Jaccard coefficients are some prominent similarity measured indices during clustering. This research paper proposes a modified K means clustering algorithm to handle high dimensional data for a multimedia database as the input set and at the same time obtain optimal similarity measures by utilizing a Minkowski distance which is a generalized form of the Euclidean distance to obtain the distance measure during clustering.

The rest of the paper is organized as literature survey in section II followed by the proposed algorithm in Sect. 3. Section 4 illustrates the experimental procedure carried out and inferences justified by experimental results. A conclusion with a future scope of research is briefed in Sect. 5.

## 2 Related work

Numerous research contributions have been found in the literature in the field of data mining and clustering. A literature survey in one of the essential branches of mining namely pattern or similarity recognition has been done in which techniques employing dimensionality reduction techniques have been investigated. Research papers have presented classification problems based on clustering techniques with remote sensing data taken as inputs to classify the patterns. The unique feature of remote sensing or SAR images [1] is that these data have high dimensions due to their high degrees of resolutions. Most of research papers have presented the conventional principal component analysis (PCA) and variations of the PCA algorithms for dimensionality reduction. PCA [5, 10] is one of the statistical approaches and provides satisfactory reductions in the dimensions of the feature set. PCA techniques come under second order statistical techniques and use variables and covariance convergence matrix (CCM) which is essentially a matrix containing information regarding the data set. Experimental results show that conventional PCA techniques are quite simple in terms of computational complexity and could be easily manipulated [11] to suit the feature vector size. However, it could be found from the literature that manipulation of the matrix is nearly impossible when the data set follows a Gaussian distribution [6, 12]. This necessitates the need for a higher order linear technique especially when the dimensions of the data are very large. Independent component analysis [13] is one such high order linear technique found in the literature. Another reduction technique reported in the literature includes the Karhunen Louve (K–L) transform [7, 10] which is an orthogonal transform. Application of the K-L algorithm to the feature data set reduces the dimensionality by computing the linear combinations [14] of the vari-

ables in the original data set with huge deviations in their variances.

Variations of the original PCA technique have been observed in the survey where Eigen vectors are used to compute the similarity measures [15, 16] and reduction in dimensions is achieved by eliminating the members with highest coefficient values after iteration. This iterative process of repeated elimination produces the variables of subspaces with reduced dimensions of  $k$ . Literature also suggests certain other second order techniques like factor analysis [17–19] which takes multimedia data as the input for mining. The feature vectors are classified into two parts where the first part contains the variance of the elements in the feature vector while the second part contains a unique variance which is obtained by the variability of actual information and not common to all other variables. The relation between the variables is identified by factor analysis and the dimension is reduced for the data set. The factor model does not depend on the scale of the variables but holds the orthogonal rotations of the factors. Another variation of factor analysis is the principal factor analysis [20] where the data is standardized which subsequently makes the elements in the correlation matrix equal. The elements in the correlation matrix are essentially co-variances. It is observed that concepts based on entropy are used to estimate the initial value and the objective function defined as the minimization of squared error resulting in correlation coefficients. The reduced correlation matrix is formed, where the diagonal elements are replaced by the elements. Eigen values have been used to disintegrate the matrix further and the dimensionality of the resulting matrix is chosen as the point where the magnitude of the Eigen values computed exhibit a sharp transition.

Another prominent technique for dimensionality reduction found in the literature is the MLF or the maximum likelihood factor analysis [4, 21, 22] where multivariate variables are used unlike PCA which contains univariate indices. This technique once again defines the objective function by minimizing the log likelihood ratio computed from the geometric mean of Eigen values. A modification of this technique is the linear projection pursuit (PP) [5], which could be suitably applied to non Gaussian data sets but it is computationally complex when compared to conventional PCA and factor analysis techniques. A reduction in time has been reported by a fast projection pursuit algorithm (FPP) [23] but exhibits significant time consumption when the dimensionality of data set increases of the order of  $K^2$ . The projection pursuit directions for independent components are obtained by Fast ICA algorithm [5, 6]. The ICA is a higher-order method that explicit the linear projections, which is not necessarily to be orthogonal, but needed to be independent as possible. Statistical independence is a much stronger condition in uncorrelated second-order statistics. With the above findings, the proposed research paper focuses

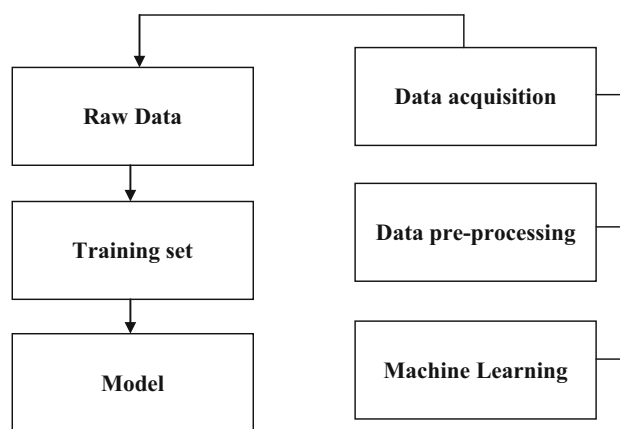
on pattern recognition and the application on dimensionality reduction.

There has been a transition from the above mentioned conventional techniques towards machine learning techniques with the advent of intelligent and adaptive algorithms as observed from the survey. One such essential and widely used technique is the cluster analysis [10, 15] which have numerous variants like K-means [16], Fuzzy [24], Genetic [7] etc., Time series clustering is yet another prominent technique [18] used for prediction applications. Classical features of time series clustering include high dimensionality, very high feature correlation, and large amount of noise. The literature present numerous techniques aimed at increasing the efficiency and scalability in handling the high dimensional time series data sets. Clustering approaches have been found to be categorized into conventional approaches and hybrid approaches. Conventional clustering techniques involve partitioning and model based algorithms. Conventional algorithms include the well known K means [9], K Medoids algorithm [25]. On the other hand, hierarchical algorithms use a pair wise distance methodology to obtain the hierarchical clusters. This technique is especially very effective for time series clustering. Another added advantage of the hierarchical algorithm is that it does not require initial seed values. A notable disadvantage from the literature is that it could not be applied for high dimensional data due to its quadratic computational complexity. The above mentioned drawbacks are effectively addressed by the model based methods or hybrid techniques where they are combined with Fuzzy sets to obtain crisp outputs or in a “fuzzy” manner (Fuzzy c-Means). Model-based clustering [16] assumes a model for each cluster and determines the best data fit for that model. The model obtained from the generated data defines the clusters. Model based techniques [18] however suffer from certain drawbacks such as setting of initial values and further these initial values are set by the user and cannot be validated.

### 3 Proposed work

A general scheme of data mining from a multimedia data base is illustrated in Fig. 5 where the raw data obtained after the data acquisition process and pre-processing is given to the training or learning process to obtain the mathematical learning model of the given dataset.

The proposed work involves a K-means clustering approach utilizing a Minkowski distance parameter. The KNN cluster is able to obtain a query vector  $p_0$  from a set of  $X$  instances  $\{p_0q_0\}^X$  with the similarity measure within classes defined by many parameters like Euclidean distance, Manhattan distance etc., as mentioned in the previous sections. The proposed work utilizes generalized version of Euclidean



**Fig. 5** A general scheme of data mining in multimedia database

and Manhattan distance in the form of Minkowski distance. The efficiency of the proposed technique is directly attributed to choice of  $k$  value and suitable distance parameter. Noisy data samples are usually characterized by large  $k$  values to make a smooth transition between the boundaries separating the regions. ' $k$ ' is merely the number of nearest neighboring pixels to be considered. The choice of variable  $k$  is dependent on the relation between the number of features and the number of cases. A small value of  $k$  may influence the result by individual cases, while a large value of  $k$  may produce smoother outcomes. The calculation of distances is executed between the query instance and 150 training samples. The formula of Minkowski distance is used as the objective function, which the equation is as follows:

$$g(x, y) = \left( \sum D_i |x_i - y_i|^r \right)^{1/r} \quad (1)$$

The Minkowski distance is a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance. The Minkowski metric is widely used for measuring similarity between objects. An object located a smaller distance from a query object is deemed more similar to the query object. Measuring similarity by the Minkowski metric is based on one assumption: that similar objects should be similar to the query object in all dimensions. A variant of the Minkowski function, the weighted Minkowski distance function, has also been applied to measure image similarity. The basic idea is to introduce weighting to identify important features [26–28]. After determination of the distance, the squared distances are sorted out and the first  $k$  values are ranked on a most minimum square distance criterion. These sorted query instances are categorized into the components of the brain to which they belong. The classified components are then segmented and filtered to remove any noisy unwanted regions. The algorithm is summarized below.

**Algorithm**

**Input:** Feature vector set  $S = \{S_1, S_2, S_3, S_4 \dots S_p\} \in R_n$

**Output:** Extracted text stream  $C_i = \{C_p, C_{pq}, \dots \dots C_{p+q+r\dots}\} \in P$

*Procedure*

Generate candidate patterns  $M \in M_1 \cap M_2 \cap M_3 \cap \dots \cap M_n$

Assign seed = { }

Initialize number of cluster,  $k$ .

Apply  $Cnd\_Ptrn = M_p \cap M_n$

for  $t = 1$  to  $r$  where  $M_t \in P(x, y)$

{

for all pixels  $(i, j) \in M_p$

for  $i=0$  to  $M_t - 1$

do

for  $j=i+1$  to  $M_t$

do

if  $Cnd_{ptrn}(i) = Cnd_{ptrn}(j)$  then

$Cnd_{ptrn} = Cnd_{ptrn} - Cnd_{ptrn}(j)$  //discard

$Cnd_{ptrn} = Cnd_{ptrn}(i)/Cnd\_Ptrn(j)$

end if

end for

Return  $Cnd\_Ptrn$

Compute  $Conf\_measure(x, y) = \left\{ \frac{(|P(x,y) - \min|)}{\mu(\max - \min)^2} \right\}$

Determine the Minkowski of each cluster center

Assign the vector with minimum distance with respect to centroid to its respective cluster of the centroid and compute the rank.

Recalculate the new center location

Repeat the steps 6 - 7, until  $e(t) \cong 0$ .

Group the clusters into  $C_i = \{C_p, C_{pq}, \dots \dots C_{p+q+r\dots}\} \in P$

}

end if

end

end procedure

**4 Results and discussion**

To test the efficiency of the proposed algorithm, a multimedia database consisting of composite elements in the form of text, hypertext from yahoo, video and audio clips and data in the form of time series for yeast and serum have been considered for experimentation. The scatter plot of yeast and serum data

is depicted in figure 6. Hypertext data is taken in the form of 10000 emails categorized into three classes namely ham, spam and phishing taken from TEC corpus (Fig. 6).

Similarity measure has been used as the performance criterion for measurement of accuracy in the email dataset of 10000. The other criteria used are defined below.



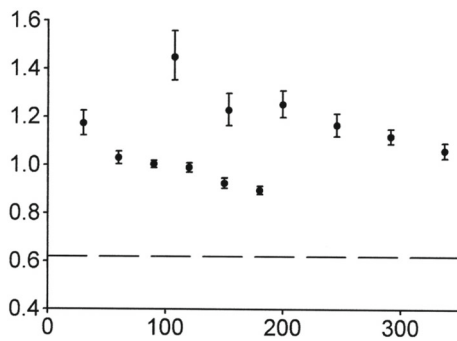


Fig. 6 Serum and Yeast data set used in proposed work

$$\begin{aligned} \text{True Positive (TP)} \\ &= \frac{\text{No. emails having spam}}{\text{Total number of emails from dataset}} \end{aligned} \quad (2)$$

$$\begin{aligned} \text{True Negative (TN)} \\ &= \frac{\text{No. emails without spam}}{\text{Total number of emails from dataset}} \end{aligned} \quad (3)$$

$$\begin{aligned} \text{False Positive (FP)} \\ &= \frac{\text{No. emails without spams but detected positive}}{\text{Total number of emails from dataset}} \end{aligned} \quad (4)$$

$$\begin{aligned} \text{False Negative (FN)} \\ &= \frac{\text{No. emails with spam but not detected}}{\text{Total number of emails from dataset}} \end{aligned} \quad (5)$$

With the obtained parameters, the accuracy of the developed segmentation technique could be arrived as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

The plot of accuracy against the number of cluster groups are depicted in Fig. 7 and it could be seen the accuracy converges to optimal values for a data vector size of around 1600 indicated by the blue line. The similarity measure is plotted against the total number of cluster groups which are related functionally and depicted in Fig. 8.

It could be seen from the above plot that the similarity measure hovers around a saturation value of around 0.575 for increasing cluster number as the interval gets discretized and more number of elements is present inside each cell. The proposed work has been experimented and tabulated for a dataset of 250 and compared with Wangs method and tabulated below (Table 1).

The measured entropy for various data sets is tabulated in Table 2 as shown below.

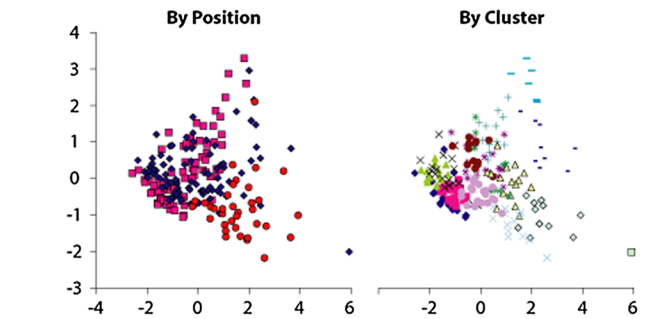
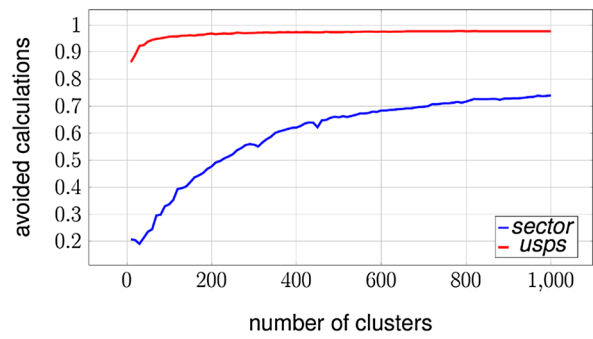


Fig. 7 Accuracy plot against cluster for given dataset

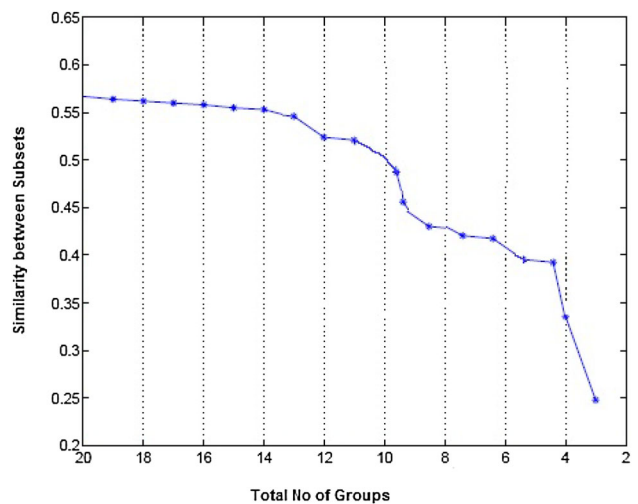


Fig. 8 Plot of similarity measure against number of groups

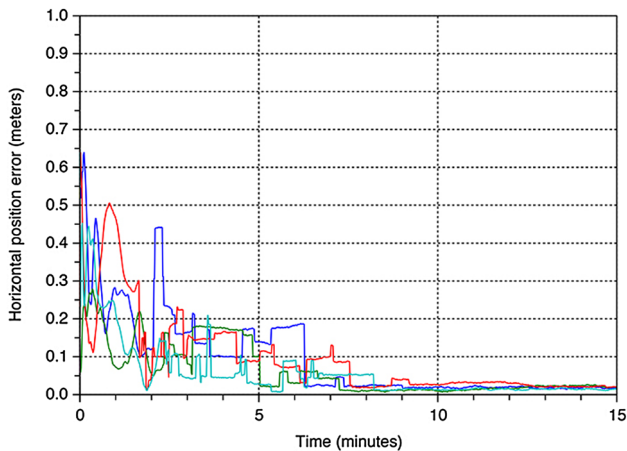
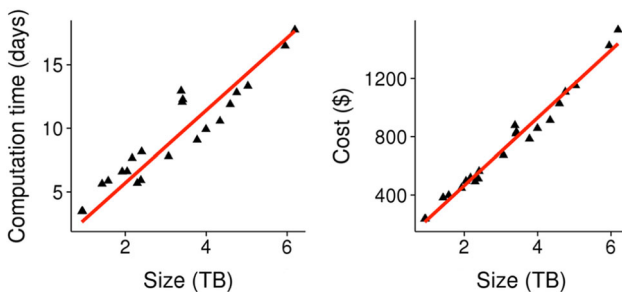
Table 1 Analysis of proposed work (dataset of 250)

Parameters	Wang's	Proposed
Initial number of clusters	80	80
Error convergence value	0.264E−6	0.0029E−8
Computation time (s)	0.815	0.295

The performance of the proposed technique has been compared with the conventional k means clustering and that of Wang's method and the convergence plot shown in Fig. 9.

**Table 2** Entropy similarity measure for the multimedia database system

Dataset	No. of classes	Size	Entropy
Iris	60	470	0.78
Leaf.jpg	11	210	0.35
Yahoo.http	64	941	0.88
Xylophone.mpg	47	884	0.81

**Fig. 9** Error convergence performance for proposed work**Fig. 10** Computation time analysis**Table 3** Accuracy measures for proposed work (1500 words)

S. no.	Algorithm	Accuracy (%)
1.	Proposed work	93.54
2.	K means clustering	90.69
3.	Wang's method	86.12

Computation time is yet another essential criteria and the proposed method exhibits drastic reduction in computation time for a data set of 250 when compared with Wang's method. The plot of computation time is depicted in Fig. 10.

Accuracy values measured for a data set with only word counts of upto 1500 have been experimented and the results are tabulated in Table 3.

## 5 Conclusion and future scope

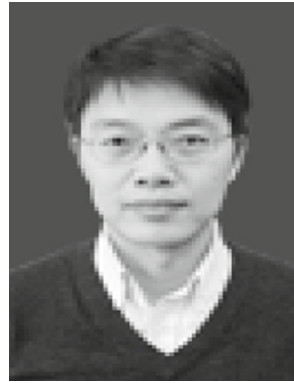
The multimedia data mining, knowledge extraction plays a vital role in multimedia knowledge discovery. This research paper has investigated the major challenges and issues in mining of multi media data which is quite different from text or image mining since multimedia database is a composite collection of image, text, hypertext, audio and video sequences. A cluster based approach is utilized in this paper by modifying the existing K means clustering algorithm to suit the multimedia data base inputs. A Minkowski distance measure has been used in this paper to bring about the modification in the proposed work and the experiments have been conducted with inputs from TEC corpus data and Yahoo databases for emails. An exhaustive analysis of the proposed work has been carried out with varying feature vector sizes and it could be seen that a hierachical clustering as the one used in this paper drastically brings down the data size as well as the computational complexity which subsequently brings down the computation time. The proposed work has beenc compared with conventional K means and Wang et al's method for mining of multi media databases. A future scope of this work is to increase the data set size especially in case of audio and video sequences with high resolution as most of the real time multimedia data in use today are of high resolution. Since, the volume of data handled on a dialy basis is enormous; research in this area continues to be an evergreen avenue for future research. Fuzzy C means and Genetic based optimization problem formulations could also be extended to this proposed work which is being thought of as a future work.

**Acknowledgements** Funding was provided by General Program of the Natural Science Fund of Hubei Province (Grant No. 2014CFB916).

## References

- Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Multimedia mining. *WSEAS Trans. Syst.* **3**(10), 3263–3268 (2004)
- Manjunath, T.N., Hegadi, R.S., Ravikumar, G.K.: A survey on multimedia data mining and its relevance today. *Int. J. Comput. Sci. Netw. Secur.* **10**(11), 165–170 (2010)
- Bhatt, C.A., Kankanhalli, M.S.: Multimedia data mining: state of the art and challenges. *Multimed. Tools Appl.* **51**, 35–76 (2011)
- Bhatt, C., Kankanhalli, M.: Probabilistic temporal multimedia data mining. *ACM Trans. Intell. Syst. Technol.* vol. 2, no. 2, Article 17 (2011)
- Kamde, P.M., Algur, S.P.: A survey on web multimedia mining. [arXiv:1109.1145](https://arxiv.org/abs/1109.1145) (2011)
- Wang, D., Kim, Y.-S., Park, S.C., Lee, C.S., Han, Y.K.: Learning based neural similarity metrics for multimedia data mining. *Soft Comput.* **11**(4), 335–340 (2007)
- Benjamin, B., Navarro, G.: Probabilistic proximity searching algorithms based on compact partitions. *Discret. Algorithms* **2**(1), 115–134 (2004)

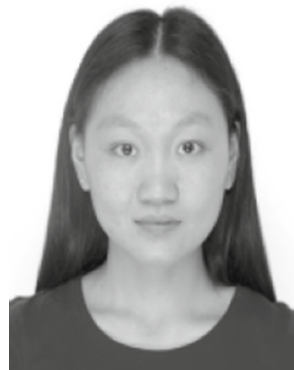
8. Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral methods for clustering. *Pattern Recognit.* **41**(1), 176–190 (2008)
9. D’Urso, P., Massari, R., Cappelli, C., De Giovanni, L.: Autoregressive metric-based trimmed fuzzy clustering with an application to PM<sub>10</sub> time series. *Chemometr. Intell. Lab. Syst.* **161**, 15–26 (2017)
10. Nair, B.B., Saravana Kumar, P.K., Sakthivel, N.R., Vipin, U.: Clustering stock price time series data to generate stock trading recommendations: an empirical study. *Expert Syst. Appl.* **70**, 20–36 (2017)
11. Méndez, E., Lugo, O., Melin, P.: A competitive modular neural network for long-term time series forecasting. In: Melin, P., Castillo, O., Kacprzyk, J. (eds.) *Nature-Inspired Design of Hybrid Intelligent Systems*, pp. 243–254. Springer International Publishing (2017)
12. Wang, D., Wang, Z., Li, J., Zhang, B., Li, X.: Query representation by structured concept threads with application to interactive video retrieval. *J. Vis. Commun. Image Represent.* **20**, 104–116 (2009)
13. Berkhin, P.: A survey of clustering data mining techniques. In: Kogan, J., Nicholas, C., Teboulle, M. (eds.) *Grouping Multidimensional Data. Recent Advances in Clustering*, pp. 25–71, 372, 520. Springer, Berlin (2006)
14. Bagnall, A., Janacek, G.: Clustering time series with clipped data. *Mach. Learn.* **58**(2–3), 151–178 (2005)
15. Mukherjee, Michael Laszlo Sumitra: A Genetic algorithm that exchanges neighbouring centers for K-means clustering. *Pattern Recognit. Lett.* **28**, 2359–2366 (2007)
16. Roy, D.K., Sharma, L.K.: Genetic K-means clustering algorithm for mixed numeric and categorical data. *Int. J. Artif. Intell. Appl.* **1**(2), 23–28 (2010)
17. Natarajan, R., Sion, R., Phan, T.: A grid-based approach for enterprise-scale data mining. *J. Future Gener. Comput. Syst.* **23**, 48–54 (2007)
18. Wong, K.-C., Wu, C.-H., Mok, R.K.P., Peng, C., Zhang, Z.: Evolutionary multimodal optimization using the principle of locality. *Inf. Sci. J.* **194**, 138–170 (2012)
19. Maji, P.: Fuzzy-rough supervised attribute clustering algorithm and classification of microarray data. *IEEE Trans. Syst. Man Cybern. Part B* **41**(1), 222–233 (2011)
20. Niknam, T., Firouzi, B.B., Nayeripour, M.: An efficient hybrid evolutionary algorithm for cluster analysis. *World Appl. Sci. J.* **4**(2), 300–307 (2008)
21. Belacel, N., Raval, H.B., Punnen, A.P.: Learning multicriteria fuzzy classification method PROAFTN from data. *Comput. Oper. Res.* **34**, 1885–1898 (2007)
22. Ordonez, C.: Integrating K-means clustering with a relational DBMS using SQL. *IEEE Trans. Knowl. Data Eng.* **18**(2), 188–201 (2006)
23. Santos, J.M., de Sa, J.M., Alexandre, L.A.: LEGClust—a clustering algorithm based on layered entropic sub graph. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 62–75 (2008)
24. Jarrah, M., Al-Quraan, M., Jararweh, Y., Al-Ayyoub, M.: Medgraph: a graph-based representation and computation to handle large sets of images. *Multimed. Tools Appl.* **76**(2), 2769–2785 (2017)
25. Monbet, V., Ailliot, P.: Sparse vector Markov switching autoregressive models. Application to multivariate time series of temperature. *Comput. Stat. Data Anal.* **108**, 40–51 (2017)
26. Varley, J.B., Miglio, A., Ha, V.-A., van Setten, M.J., Rignanese, G.-M., Hautier, G.: High-throughput design of non-oxide p-type transparent conducting materials: data mining, search strategy and identification of boron phosphide. *Chem. Mater.* **29**(6), 2568–2573 (2017). doi:[10.1021/acs.chemmater.6b04663](https://doi.org/10.1021/acs.chemmater.6b04663)
27. Olson, D.L., Desheng Dash, W.: *Data Mining Models and Enterprise Risk Management. Enterprise Risk Management Models*. Springer, Berlin (2017)
28. Kandoi, G., Leelananda, S.P., Jernigan, R.L., Sen, T.Z.: Predicting protein secondary structure using consensus data mining (CDM) based on empirical statistics and evolutionary information. *Methods Mol. Biol.* **1484**, 35–44 (2017). doi:[10.1007/978-1-4939-6406-2\\_4](https://doi.org/10.1007/978-1-4939-6406-2_4)



**Xiaoping Jiang** received his Ph.D. degree from Huazhong University of Science and Technology, China, in 2007. He is now an associate professor at College of Electronics and Information Engineering, South-Central University for Nationalities, Wuhan, China. His research interests include signal process, video analysis and wireless communication.



**Chenghua Li** received his Ph.D. degree from Huazhong University of Science and Technology, China, in 2008. He is now an associate professor at College of Electronics and Information Engineering, South-Central University for Nationalities, Wuhan, China. His research interests include cloud computing, big data analysis and pattern recognition.



**Jing Sun** received her Ph.D. degree from Wuhan University, China, in 2011. She is now a lecturer at College of Electronics and Information Engineering, South-Central University for Nationalities, Wuhan, China. Her research interests include multimedia processing and multimedia security.