VOL. 13, NO. 5, MARCH 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

IMPROVING THE K-MEANS CLUSTERING USING VISUAL CORRELATION ANALYSIS

A. Suresh Babu[1] and B. Rama Subbaiah[2]
[1]Department of Computer Science Engineering, JNTUCE, Ananthapuramu, India
[2]Department of IT, RGMCET, Nandyal, India
E-Mail: asureshjntu@gmail.com

ABSTRACT

In this paper, we mainly focus on tweaking the performance of clustering by K-means for the given acquisition data. The data include a lot of attributes having different categories. Mainly the attributes are categorized into Numerical attributes and Categorical attributes. By using these attributes, data can be classified into a) Numerical data having only numerical attributes b) Categorical data having only categorical attributes c) Mixed data having both Categorical and numerical attributes. Initially, the Correlation Analysis is used for knowing the relationship among the attributes in the given data. It is exceptionally hard to discover Correlation Analysis for a tremendous measure of information. It may be conceivable of missing the traits with the tremendous measure of information. So in this paper, the correlation map is constructed for visualizing the correlated attributes by leaving irrelevant attributes among the given acquisition data. This correlated data available from the correlation map are used for tweaking the performance of K-means clustering results. For extracting the correlated data and tweaking the k-means clustering results, the Correlation based K-means Clustering (CBK) algorithm is proposed. In this paper, we mainly visualize the Clustering Accuracy and Normalized Mutual Information (NMI) among the attributes using K-means and future Correlation based k-means (CBK).

Keywords: correlation analysis, cluster analysis, CBK, NMI.

1. INTRODUCTION

The hasty enlargement of information machinery produces cosmic amounts of statistics by way of abundant attributes. These high-dimensional data sets bid fabulous opportunities in favor of analyzing behavioral patterns and additionally to use for forecasting desired result. Critical moving toward on numerous occasions comes as of cloud interrelationships that keep living in the midst of insight characteristics (or factors). For pattern, in brain science dive into, researchers attempt to see affiliations between brain power, propensity, and societal manner. In financial side, to capitalize on proceeds, economists seem to be in the assembly of variables that are recurrently interrelated to the communal backing. To finish in the societal and instinctive sciences, researchers try to read and elucidate the scenery of kindred amid confident phenomena. To formulate an enhancement in this ample assortment of areas, analysts require practical, susceptible, and perceptive utensils reveal these associations. Correspondence scrutiny is solitary such utensil. It looks meant for associations amid variables with the container is evidence for whether pairs of variables are associated along with how stalwartly. Association analysis has developed into further and further admired in loads of areas; together with psychology, instruction, sponsorship, merchandising, and climatology, very soon to bring up a little. Correlations, on the other hand, are easier said than ended to comprehend, supervise, and assessment some time ago a number of variables become even rather thick. Given D variables, there is O (D^2) connection pairs, which create composite associations easier said than ended to be acquainted from the columns of numbers unaided. Consequently, in that admiration is a lucid necessity for a valuable design boundary that allows analysts to (2) rapidly get a summing up of the taken as a whole connection associations in the data, and (3) effortlessly maneuver the information to expose secreted associations via diverse modes of communications, such as filtering, selection, bracketing, and clustering.

Correlation analysis is associated to deterioration psychotherapy. Whereas deterioration psychotherapy quantifies the linear association amid a reliant erratic and solitary or further sovereign variables, relationship psychotherapy makes rebuff dissimilarity amid autonomous in addition to reliant variables-it is barely a measure of linear involvement amid two variables. The potency of this linear involvement is gauged in the connection coefficient r. It is this coefficient with the intention of relative correlation as well as deterioration psychotherapy, for the reason that squaring r, or else r2 yields the coefficient of will power which is a determination of how able-bodied the deterioration procession represents the data. It is imperative to appreciate, on the other hand, that neither connection nor deterioration psychotherapy can institute cause-and-effect associations in the midst of the vacillating, These tins can barely subsist incidental with an individual being psychotherapist, along with this situation forms a focal enthusiasm of our exertion

In view of the fact that we regard as the variables the major actors, we seek a visual boundary that can best show how variables-unqualified and geometric-be active as one through solitary an added, by means of spatial propinquity brainwashing en route for putting into words the potency of the interactions. We entitle this illustration boundary the correlation map. It builds taking place our preceding stab reported in [4] everywhere we foremost useful a force-directed algorithm just before optimizing a correlation-centric 2D describe of every one of variables furthermore subsequently computed a cut-rate alleyway transversely it

toward agreeing on a high-quality ordering of axes in a comparable organize presentations. Users may well maneuver this conduit unswervingly inside the boundary with as a result adjusts the axes orderings spontaneously. Our current effort furthermore uses this outline, however, it does this pro a diverse rationale-interactive visual connection psychotherapy. on the way to facilitate this we boast introduced an innovative situate of relations with chart representations. Other concretely, we encompass devised

- new-fangled mechanisms so as to can lever in cooperation unconditional Furthermore geometric variables surrounded by an incorporated skeleton
- A multi-scale zooming come within reach of to accomplish scalability for huge facts of vacillating,
- Convertible techniques used for exploring the bang of rate grade on tie-in.
- A cemented hallucination of the vice- seats spanned by sets of interrelated vacillating.

2. RELATED WORK

2.1 Correlation

Correspondence is visually uttered while patterns the records forming the put on view. In 2D scatter plots, the further directly the point cloud adheres to an immediate procession up, the better-quality the linkage of the two vacillating. On the further tender, in an analogous harmonize demonstrate [1] a muscular affirmative correlation amid two variables is visually expressed by articulate bundles of position with comparable slopes. also, a solid apathetic connection makes these reliable packs frame a prototype attractiveness outline with a focusing hybrid point.

In the mutually 2D scatter plots and analogous coordinates contribute to one auxiliary inadequacy-not possible effortlessly be evidence for correlations to facilitate engross auxiliary than two vacillating.even though 2D scatter plots be able to be lengthened into scatter plot matrices (SPLOM) [6] Furthermore the axes in analogous coordinates can be re-ordered [5] to interpretation associations of diverse sets of vacillating, harmonize this incongruent in sequence crosswise what's more carpet or axes is complicated.

A substitute draws near is to envisage the connection atmosphere directly, provided that a rounded observation in excess of the patchy legroom. In this demonstration, every atmosphere cell denotes the connection of solitary up-and-down brace. The atmosphere scrutiny has started off an ample situate of applications. Seo and Shneiderman [7] exploit a matrix-based visualizer toward affording an indication of the standing of skin texture, whereas Henry and Fekete [8] incorporate the node-link plan through a matrix-based put on view to prop up the investigation of societal networks. a lot of these methods carry interactive filtering and clustering,and also atmosphere reordering [7]. In our casing, it can reveal clusters of allied vacillating.

The discernment of the clusters can be auxiliary superior by coding correlation potency to color, squashy a heat map [7]. on the other hand, From the fallout of Bertin's levels of association [2], vividness Furthermore blush are unfortunate visual variables on behalf of reckon-able in sequence-geographical variables such as magnitude along with closeness are far-flung superior choices.

Data integration

In this scheme harmonize the statistics addicted to a montage consequent commencing the arrangement of variables. This is one more rationale why matrix analysis is not a realistic opportunity pro our system-such an inspection would not tolerate a firm facts incorporation. At foremost momentary look the visualizations we fabricate fairly be similar to the hyper box [2], but we allow flee contrive carpet of superior two variables. to conclude, Claessen and van Wijk [4] illustrate an organism with the aim of uses strips of 2D spread plots and associations their axes by the analogous equivalent coordinates put on view segments.The strips in our skeleton are extra wide-ranging and are unswervingly connected at their collective axes.

Numeric variable

The morals of a geometric unpredictable are numbers. They can be added off the record into detached and unremitting mutable.

- Secluded geometric mutable

A mutable whose morals are complete statistics (counts) is called secluded. For exemplar, the quantity of stuff bought by a purchaser in a superstore is secluded.

- Uninterrupted geometric mutable

A mutable that possibly will enclose a few assessment contained by several series is called uninterrupted. For exemplar, the moment that the purchaser consumes in the superstore is uninterrupted.

geometric techniques that can be worn for constant factors are not for time everlasting legitimate for separated impermanent.

A geometric or uninterrupted mutable is may get hold of happening several assessments surrounded by a restricted or else inestimable hiatus two classifications of geometric mutable, hiatus, as well as proportion. A hiatus mutable has morals whose digressions are explicable, nevertheless it does not boast a spot on nil. A high-quality instance is a warmth in Centigrade degrees. statistics on a hiatus range can be further and subtracted, but cannot be emotively multiplied or alienated. For pattern, we can't articulate that single day is two times as hot as an extra day. In distinction, a fraction mutable has morals with an exact zero and can be extra, subtracted, multiplied or alienated (e.g., weight).

3. BACKGROUND WORK

The scheme obtainable for correlation analysis is based on their intention variables, i.e., methods pertinent only to geometric variables, such as Pearson's correlation coefficient.

Pearson's correlation coefficient [6] is the majority admired procedures used for important linear associations amid two variables:

$$r(x, y) = \frac{\sum_{i=1}^{N}(x_i - \mu(x))(y_i - \mu(y))}{\sqrt{\sum_{i=1}^{N}(x_i - \mu(x))^2}\sqrt{\sum_{i=1}^{N}(y_i - \mu(y))^2}} \quad (1)$$

Here x,y two vectors of the equivalent extent μ *(x),μ (y)* are their significant means, and *N* is the total data points. The association, *r*, ranges start from -1 to +1. The adjoining *r* is to -1 or +1, the closer the two variables are linearly associated, where *r* equal to 0 revenue that there is no linear relationship amid the two variables.

Cluster Analysis (CA)

Cluster scrutiny [10] is the mission of consortium a deposit of bits and pieces in such an approach that substance in the matching assembly (called a cluster) are added analogous to every other than to folks in supplementary groups (clusters). It is a significant task of examining data mining, and a widespread procedure on behalf of geometric statistics scrutiny, used in lots of fields, along with machine learning, pattern recognition, image analysis, information retrieval, and Bioinformatics. Cluster investigation itself is not one explicit method, but the sweeping mission to be solved. It can be attained by an assortment of algorithms that fluctuate extensively in their view of what constitutes a cluster and how to proficiently come across them. fashionable accepted wisdom of clusters consists of groups with small distances in the midst of the cluster members, opaque areas of the data space, intervals or fastidious geometric distributions. Clustering can consequently be formulated as a multi-objective optimization dilemma. The suitable clustering algorithm and constraint settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the personage data set and projected exercise of the consequences. Cluster analysis is not an ordinary mission, but an iterative process of acquaintance sighting or intuitive multi-target advancement that includes experimenting with and crash. It will repeatedly be obligatory to amend data preprocessing and reproduction parameters until the result achieves the desired properties.

Normalized Mutual information (NMI)

In prospect hypothesis and information premise, the mutual information (MI) [9] or (formerly) transformation of two arbitrary variables is a determination of the variables' communal faith. Not incomplete to real-valued haphazard variables like the correlation coefficient, MI is more wide-ranging and determines how comparable the joint distribution p(X, Y) is the harvest of factored insignificant distribution p (*x*) p (*y*). MI is the probable value of the pointwise mutual information (PMI) [9]. The most frequent unit of measurement of mutual information is the bit.

$$I(X;Y) = \sum_{x,y} p(x, y) \ln \frac{p(x, y)}{p(x)\,p(y)} \quad (2)$$

The information goes beyond between X and Y is 0 when the two variables are self-determining, as p(*x*)p(*y*) = p(x; y). When X determines Y,
I(X; Y) = H(Y),
where H(Y) is the entropy of, or lack of information about, Y, de_ned as:

$$H(Y) = -\sum_{y} p(y) \ln p(y). \quad (3)$$

When X and Y are absolutely correlated (they determine each other), I(X; Y) reaches its maximum of H(X) = H(Y) = H(X; Y), where H(X; Y) is the joint entropy of X and Y, which we get by replacing the insignificant distribution in (3)with the joint distribution p(x; y).

In this paper, we use the seven dissimilar types of data sets.

Table-1. Different types of datasets.

S. No.	Name of the dataset	Number of attributes	Number of records
1	Abalone	8	4177
2	Auto mpg	7	398
3	Data user	5	258
4	Iris	4	150
5	Seed	7	210
6	Vote	16	435
7	Wine	13	178

The correlation between attributes of datasets is analyzed from the following sample correlation matrices.

Table-2(a). Correlation matrix for 'Abalone Dataset'.

0.00	**0.99**	**0.83**	**0.93**	**0.90**	**0.90**	**0.90**	**0.56**
0.99	0.00	0.83	0.93	0.89	0.90	0.91	0.57
0.83	0.83	0.00	0.82	0.77	0.80	0.82	0.56
0.93	0.93	0.82	0.00	0.97	0.97	0.96	0.54
0.90	0.89	0.77	0.97	0.00	0.93	0.88	0.42
0.90	0.90	0.80	0.97	0.93	0.00	0.91	0.50
0.90	0.91	0.82	0.96	0.88	0.91	0.00	0.63
0.56	0.57	0.56	0.54	0.42	0.50	0.63	0.00

Table-2b. Correlation matrix for 'Auto mpg Dataset'.

1.00	-0.78	-0.80	-0.75	-0.83	0.42	0.58
-0.78	1.00	0.95	0.82	0.90	-0.51	-0.35
-0.80	0.95	1.00	0.87	0.93	-0.54	-0.37
-0.75	0.82	0.87	1.00	0.84	-0.67	-0.41
-0.83	0.90	0.93	0.84	1.00	-0.42	-0.31
0.42	-0.51	-0.54	-0.67	-0.42	1.00	0.29
0.58	-0.35	-0.37	-0.41	-0.31	0.29	1.00

Table-2c. Correlation matrix for 'Data user Dataset'.

1.00	0.08	0.04	0.10	0.21
0.08	1.00	0.08	0.10	0.18
0.04	0.08	1.00	0.04	0.12
0.10	0.10	0.04	1.00	-0.27
0.21	0.18	0.12	-0.27	1.00

Table-2d. Correlation matrix for 'Iris Dataset'.

1.00	-0.14	0.87	0.83
-0.14	1.00	-0.43	-0.37
0.87	-0.43	1.00	0.96
0.83	-0.37	0.96	1.00

4. EXPERIMENTS AND RESULT ANALYSIS

In this dissertation, we accepted out the investigational results using seven real-time datasets. Table-3 shows the clustering exactness for k-means and proposed correlation based k-means (CBK).

Table-3. Clustering Accuracy (CA).

Datasets	K-means	CBK
Abalone	0.5133	0.5284
Auto mpg	0.4548	0.4548
Data user	0.593	0.5659
Iris	0.9133	0.9133
Seed	0.8952	0.8524
Vote	5241	0.5241
Wine	0.7022	0.7022

Table-4. Normalized Mutual Information (NMI).

Datasets	K-means	CBK
Abalone	0.121	0.1551
Auto mpg	0.1885	0.1885
Data user	0.4111	0.3902
Iris	0.7733	0.7733
Seed	0.6935	0.6508
Vote	0.323	0.0323
Wine	0.4287	0.4287

Figure-1 and Figure-2 are shows the bar charts of clustering accuracy (CA) and NMI for the methods k-means and CBK. These bar charts demonstrate that the CBK is more efficient.

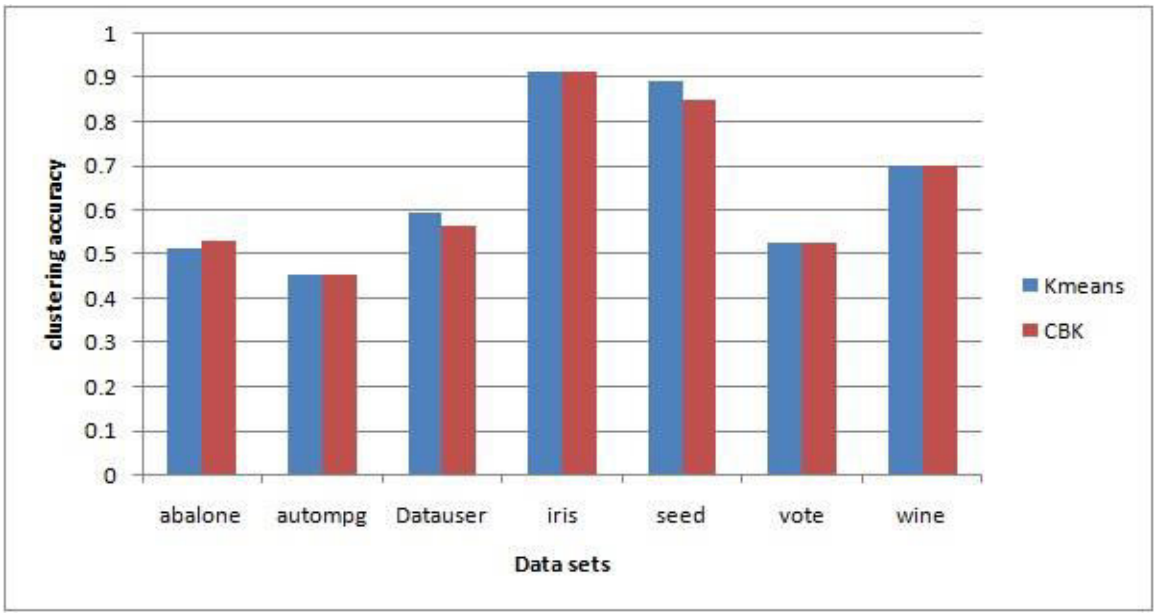

Figure-1. Clustering accuracy for k-means and CBK for real-time data sets.

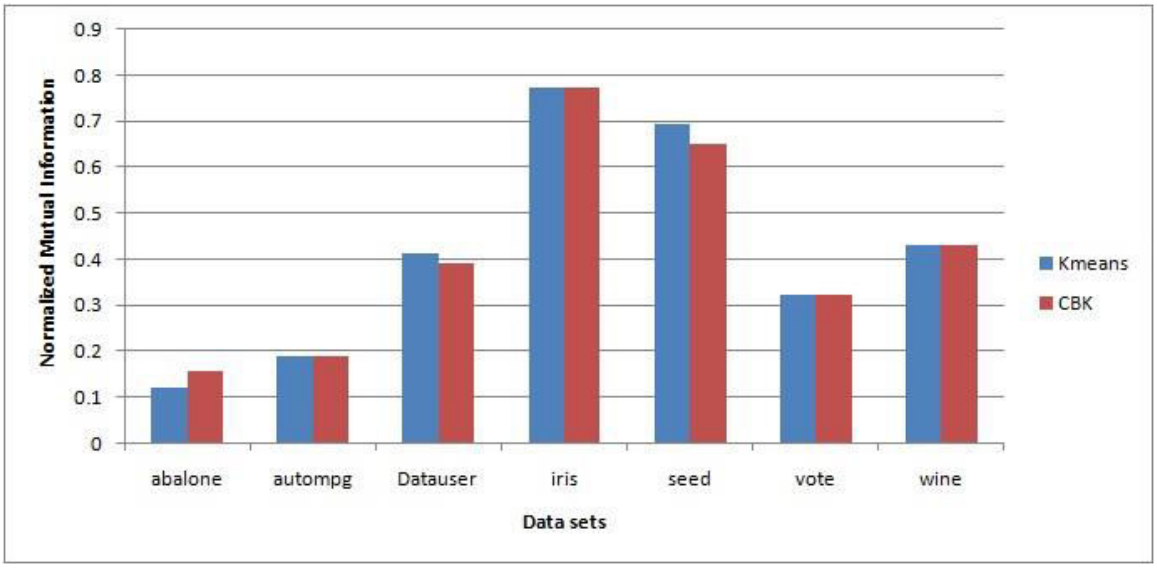

Figure-2. Clustering accuracy for k-means and CBK for real-time data sets.

5. CONCLUSIONS

The k-mean is an effective clustering algorithm for information grouping. However, some datasets consist of more number of attributes which incorporates corresponded and interrrelated qualities. We need to extract the correlated attributes for improving the results of k-means clustering. For this aspect, the correlation based k-means clustering a(CBK)algorithm is proposed. In the proposed algorithm, initially, the correlation map is constructed for visualizing the correlated attributes. The correlated data is the relevant data i.e., which excludes the irrelevant attributes data. This correlated attributes data would be advantageous in improving the clustering results. Hence, the proposed CBK is a more efficient clustering method. The future scope of the work is to an extent on image data for improving the image clustering results.

REFERENCES

[1] Zhiyuan Zhang. 2015. Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map. IEEE Transactions on Visualization and Computer Graphics. 21(2).

[2] B. Alpern and L. Carter. 1991. The hyper box. in Proc. IEEE Conf. Vis. pp. 133-139.

[3] J. Bertin, Semiology of Graphics: Diagrams, Networks, Maps. Madison.

[4] Z. Zhang, K. T. McDonnell, and K. Mueller. 2012. A network-based interface for the exploration of high-dimensional data spaces. in-proc. IEEE Pac. Vis. pp. 17-24.

[5] B. Ferdosi and J. Roerdink. 2011. Visualizing high-dimensional structures by dimension ordering and filtering using subspace analysis. Comput. Graph. Forum. 30(3): 1121-1130.

[6] J. Hartigan. 1972. Direct clustering of a data matrix. J. Amer. Statist. Assoc. 67(337): 123-129.

[7] J. Seo and B. Shneiderman. 2005. A rank-by-feature framework for interactive exploration of multidimensional data. Inf. Vis. 4(2): 96-113.

[8] N. Henry and J.-D. Fekete. 2006. Matrix Explorer: A dual-representation system to explore social networks. IEEE Trans. Vis. Comput. Graph. 12(5): 677-684.

[9] Gerlof Bouma. Normalized (Pointwise) Mutual Information in Collocation Extraction" Department Linguistik, Universit□at Potsdam.

[10] C. Chen, C. Wang, K. L. Ma, and A. Wittenberg. 2011. Static correlation visualization for large time-varying volume data. in Proc. IEEE Pac. Vis. pp. 27-34.