ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/259174077

Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms

Conference Paper · December 2013

DOI: 10.1007/978-3-319-03095-1_15

CITATIONS
24

READS
7,259

1 author:

M. Ramakrishna Murty
Anil Neerukonda Institute of Technology and Sciences
26 PUBLICATIONS **280** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

cluster analysis View project

All content following this page was uploaded by M. Ramakrishna Murty on 01 December 2014.

The user has requested enhancement of the downloaded file.

Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms

R. Madhuri[1], M. Ramakrishna Murty[1], J.V.R. Murthy[2],
P.V.G.D. Prasad Reddy[3], and Suresh C. Satapathy[4]

[1] Dept. of CSE, GMR Institute of Technology, Rajam, Srikakulam(Dist.) A.P., India
{ravi.madhuri5,ramakrishna.malla}@gmail.com
[2] Dept. of CSE, JNTUK-Kakinada, A.P., India
mjonnalagedda@gmail.com
[3] Dept. of CS&SE, Andhra University, Visakhapatnam, A.P., India
prasadreddy.vizag@gmail.com
[4] Dept. of CSE, ANITS, Visakhapatnam, A.P., India
sureshsatapathy@gmail.com

Abstract. The k-means algorithm is well-known for its efficiency in clustering large data sets and it is restricted to the numerical data types. But the real world is a mixture of various data typed objects. In this paper we implemented algorithms which extend the k-means algorithm to categorical domains by using Modified k-modes algorithm and domains with mixed categorical and numerical values by using k-prototypes algorithm. The Modified k-modes algorithm will replace the means with the modes of the clusters by following three measures like "using a simple matching dissimilarity measure for categorical data", "replacing means of clusters by modes" and "using a frequency-based method to find the modes of a problem used by the k-means algorithm". The other algorithm used in this paper is the k-prototypes algorithm which is implemented by integrating the Incremental k-means and the Modified k-modes partition clustering algorithms. All these algorithms reduce the cost function value.

Keywords: Cluster, K-means, K-modes, K-prototypes, mixed data.

1 Introduction

The most diverse characteristic of data mining is that it deals with very large and complex data sets.[2] The datasets to be mined often contain millions of objects described by tens, hundreds or even thousands of various types of attributes or variables. This requires the data mining operations and algorithms to be scalable and capable of dealing with different types of attributes. In terms of clustering, we are interested in algorithms which can efficiently cluster large data sets containing both numeric and categorical values because such data sets are frequently encountered in data mining applications.

S.C. Satapathy et al. (eds.), *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of CSI - Volume II*, Advances in Intelligent Systems and Computing 249,
DOI: 10.1007/978-3-319-03095-1_15, © Springer International Publishing Switzerland 2014

In this paper, we presented three new algorithms that uses the incremental k-means paradigm to cluster data having numerical data, Modified k-modes paradigm to cluster data having categorical data and k-prototypes paradigm to cluster mixed data i.e., categorical and the numerical data. [8]The Modified k-modes algorithm extended the k-means paradigm to cluster categorical data by using (1) a new matching dissimilarity measure for categorical objects, (2) modes instead of means for clusters as centroids and (3) a frequency-based method to provide initial modes to minimize the clustering cost function. The k-prototypes algorithm in general integrates the k-means and k-modes to cluster data with mixed numeric and categorical values. So here in our paper, we used Incremental k-means and Modified k-modes paradigms to get integrated for implementing the k-prototypes algorithm.

The clustering process [4] of the k-prototypes algorithm is similar to the k-means algorithm except that it uses k-modes approach to provide initial modes for the categorical attributes of cluster prototypes. [7]Because these algorithms use the same clustering process as k-means and they preserve the efficiency of the k-means algorithm which is highly desirable for data mining. In this paper we deal with the following partitioning clustering methods, namely K-means, K-modes, K-mediods and K-prototype.

2 Our Proposed Work

We compared and implemented three algorithms in this paper, namely incremental k-means, Modified k-modes and K-prototype algorithms with different combinations of real world data sets and found that incremental K-means provide better results than simple K-means for numeric data, Modified K-modes is better than K-modes for categorical data and K-prototype is useful for mixed data clustering. We also observed K-prototypes paradigm is the combination of the K-means and the K-modes paradigms.

The number of iterations required to obtain the effective clustering results gets reduced. The cost function or the dissimilarity rate of the clustering ultimately obtained is comparatively low. Since the number of iterations converges, the time complexity is also reduced.

2.1 Incremental K-Means Paradigm

In the incremental k-means, after assigning each object to any cluster, the mean of that cluster is immediately updated.[5] So the next object comparison does with all the updated means. Thus incremental k-means is more appropriate and does clustering effectively with less number of iterations.

Algorithm

Step.1: Specify the number of clusters

Step.2: Select initial centroids randomly based on number of clusters specified.

Step.3: The centoids can be updated incrementally after each assignment of a data object to a cluster.

Step.4: Update that particular cluster's mean as:

$$mean_{jf} = mean_{jf} + a_{if} \quad (a_{if} \text{ is the data object })$$

Step.5: Repeat the Step.3to Step.4 with the updated means until all the instances' are assigned to clusters.

2.2 Modified K-Modes Paradigm

Clustering and other data mining applications frequently involve categorical data. The traditional approach of converting categorical data into numerical ones does not necessarily produce meaningful results. Thus, handling such data is a very important research topic in data mining. The simple k-modes algorithm proposed by Haung uses a simple dissimilarity measure [4] only. So, in this paper, we are going to use the "Modified k-modes algorithm" by avoiding too many iterations using frequency methodology [2] to select the initial centroids (modes) for clustering.

Algorithm

Step.1: Start

Step.2: Select the dataset used for clustering.

Step.3: Choose attributes in the dataset for clustering.

Step.4: Sort in descending order each and every field from the data set according to the most frequent number of values present in the dataset.

Step.5: Select the required number of clusters and choose the appropriate initial centroids(modes).

Step.6: Perform the concordance-discordance test [3]. Here the difference between each object $y_i \ (i \in n)$ and mode $x_j \ (j \in k)$ for each attribute using the formula:

$$D(x_{j,f}, y_{i,f}) = \frac{(m_{x_{j,f}} + m_{y_{i,f}})}{(m_{x_{j,f}} \times m_{y_{i,f}})} \times \delta(x_{j,f}, y_{i,f}) \tag{1}$$

where, $x_{j,f}$ =value of mode x_j on attribute a_f

$y_{i,f}$= value of object y_i on attribute a_f

$m_{x_{j,f}}$=number of times $x_{j,f}$appears in the set of modes on attributes a_f

$m_{y_{i,f}}$=number of times $y_{i,f}$ appears in the set of modes on attributes a_f

and $\delta(x_{j,f}, y_{i,f}) = \begin{cases} 0 \ if \ x_{j,f} = y_{i,f} \\ 1 \ if \ x_{j,f} \neq y_{i,f} \end{cases}$

Step.7: Assign instance or object to the cluster to which the above dissimilarity difference measure is low.

Step.8: Repeat the same process from step.7 and step.8 until the entire object's assignments is completed.

Step.9: Calculate the cost function [6] for each such iteration to find the dissimilarity rate obtained after the clustering process is finished. It is calculated using the formula as shown in Eqn. (2):

$$C(Q) = \sum_{j=1}^{k} \sum_{i=1}^{n} \sum_{f=1}^{F} \delta(x_{j,f}, y_{i,f}) \tag{2}$$

Where, k is number of clusters, n is number of elements present in each cluster, F is number of attributes and $\delta(x_{j,f}, y_{i,f}) = \begin{cases} 0 \; if \; x_{j,f} = y_{i,f} \\ 1 \; if \; x_{j,f} \neq y_{i,f} \end{cases}$

2.3 K-Prototypes Paradigm

K-prototypes algorithm is a combined approach of the k-means and the k-modes algorithm. Here, we are incorporating "Incremental k-means" to cluster numerical data and the "Modified k-modes" algorithm to cluster categorical data in the mixed datasets.

Algorithm

Step.1: Select the dataset containing both numerical and the categorical data for clustering process to start.

Step.2: Choose or take the required fields or attributes to start the clustering as per requirements.

Step.3: Sort each and every taken field as:

a) The numerical data fields in the dataset are sorted in the ascending order.

b) The categorical data fields present in the dataset should be taken an be sorted by taking the most frequent values in each field and should be arranged in the descending order and those values taken should be distinct.

Step.4: Choose the number of clusters to perform.

Step.5: Choose the centroids for those clusters i.e., means [as chosen in the incremental k-means procedure] and the modes [as chosen in the Modified k-modes procedure] for the clusters taken.

Step.6: Take each object or instance and perform the assignment to the appropriate cluster based on the difference and dissimilarity measures as:

a) For the numerical data typed object, we use the Euclidean distance measure i.e.,

$$d(i,j) = \sqrt{\sum_{f=1}^{F} (a_{if} - mean_{jf})^2} \tag{3}$$

where, j∈ k (k=number of clusters)
i∈ number (number=total number of instances in the dataset)
f∈ F (F=number of attributes)

b) For the categorical data typed attributes, we use the dissimilarity difference measure as:

$$\mathrm{D}(x_{j,f}, y_{i,f}) = \frac{(m_{x_{i,f}} + m_{x_{i,f}})}{(m_{x_{i,f}} \times m_{x_{i,f}})} \times \delta(x_{j,f}, y_{i,f}) \text{ and } \delta(x_{j,f}, y_{i,f}) = \begin{cases} 0 \; if \; x_{j,f} = y_{i,f} \\ 1 \; if \; x_{j,f} \neq y_{i,f} \end{cases}$$

Step.7: Measure or calculate $d(i,j) + D(x_{j,f}, y_{i,f})$ with every cluster (k) and assign that object to whichever cluster the overall difference is low.

Step.8: Repeat the same process for step.6 and step.7 until all the objects' assignments is completed.

Step.9: Calculate the dissimilarity rate obtained after the clustering process is finished for each iteration. It is calculated using the formulae as shown in Eqn. (4):

Cost function for categorical attributes:

$$C(Q) = \sum_{j=1}^{k} \sum_{i=1}^{n} \sum_{f=1}^{F} \delta\left(x_{j,f}, y_{i,f}\right) \tag{4}$$

$$\delta\left(x_{j,f}, y_{i,f}\right) = \begin{cases} 0 \; if \; x_{j,f} = y_{i,f} \\ 1 \; if \; x_{j,f} \neq y_{i,f} \end{cases}$$

Sum of squared errors calculation for numerical attributes as shown in Eqn. (5):

$$D(i,j) = \sqrt{\sum_{f=1}^{F} \left(a_{if} - mean_{jf}\right)^2} \tag{5}$$

Step.10: Stop

3 Data Set Analysis

The data sets for performing clustering have been taken from the UCI machine repository. Three types of data sets have been taken to apply for Incremental k-means, modified k-modes and k-prototypes paradigms.

3.1 Data Sets for Incremental k-Means

The data sets taken for implementing Incremental k-means algorithm are given below.

***Iris Data Set*:** Iris data set consists of 4 numerical attributes and 155 instances. The attributes are "petal length", "sepal length", "petal width" and the "sepal width".

***Cholesterol Data Set*:** Cholesterol data set consists of 2 numerical attributes and 250 instances. The attributes are "Item Number" and the "Fat content". Using this data, we are going to group the persons having the similar cholesterol levels.

3.2 Data Sets for Modified k-Modes

The data sets taken for implementing Modified k-modes algorithm are given below.

Contact-Lens Data Set: Contact-lens data set consists of 5 categorical attributes and 24 instances. The attributes are "age", "spectacle", "astigmatism", "tearrate" and the "contact lenses". Using this data, we are going to group the persons having the similar eye sights and their presence of contact lenses.

***Post-operative Data Set*:** Post-operative data set consists of 7 categorical attributes and 190 instances. The attributes are "lcore", "lsurf", "lo2", "lbp", "surface stability",

"core stability" and the "BP stability". Using this data, we are going to group the persons having the similar body temperatures and reactions.

3.3 Data Sets for k-Prototypes

The data sets taken for implementing k-prototypes algorithm are the mixed data sets (both numeric and categorical data sets) are discussed below.

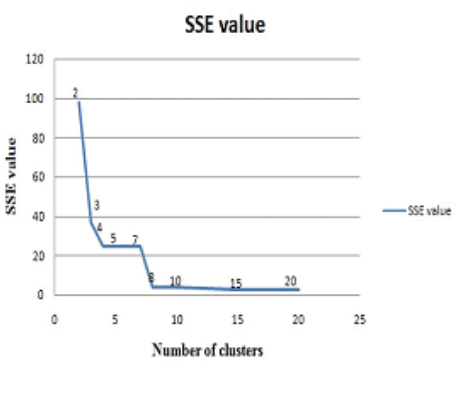
***Blood Information Data Set*:** Blood Information data set consists of 5 attributes wherein 3 are of numerical attributes and two are of categorical attributes and 200 instances. The attributes are "name", "blood content", "plasma content", "hemoglobin in cc" and the "color". Using this data, we are going to group the persons having the similar blood structures, levels and the groups.

***Weather Data Set*:** Weather data set consists of 4 attributes in total wherein 2 are of categorical attributes and the rest are of numerical attributes and there are 350 instances. The attributes are "outlook", "temperature", "humidity" and the "windy nature". Using this data, we are going to group the similar weather reports.

4 Experimental Results

A common data set named "post operation" is taken to analyze the results obtained for all the three algorithms. "Post operation" data set consists of 8 attributes wherein seven are of numerical attributes and one is of categorical attribute.

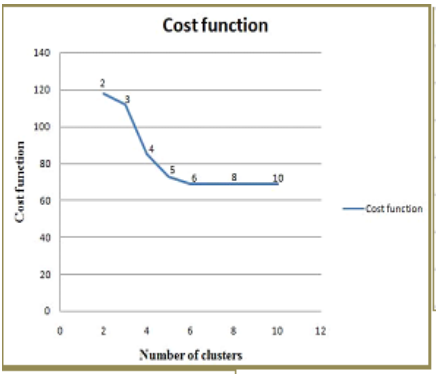
4.1 Analysis for Incremental K-Means



Number of clusters	SSE value
2	98
3	37
4	25
5	25
6	25
7	25
8	4
9	4
10	4
15	3
20	3
25	3

Fig. 1. Dissimilarity in incremental k-means to the number of clusters when clustering 90 records of the post operation data set

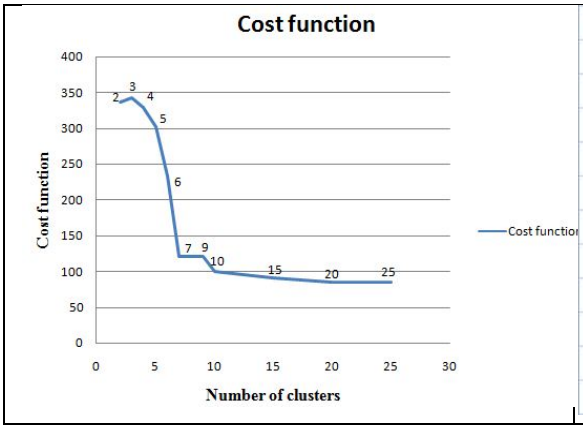
4.2 Analysis for Modified K-Modes



Number of clusters	Cost function
2	118
3	112
4	85
5	73
6	69
8	69
10	69

Fig. 2. Dissimilarity of Modified k-modes to the number of clusters when clustering 90 records of the post operation data set

4.3 Analysis for K-Prototypes



Number of clusters	Cost function
2	336
3	342
4	328
5	302
6	233
7	121
9	121
10	100
15	91
20	85
25	85

Fig. 3. Dissimilarity of k-prototypes to the number of clusters when clustering 90 records of the post operation data set

5 Conclusion

The real world data is becoming huge day-by-day with even growing data typed objects. The different data types included in the real world are categorical, numerical, scaled, Boolean etc and Sometimes there will be mixed data (combination of numerical and categorical). Clustering such different data sets as per requirements is a difficult task. To make the task easier and effective, the above three partition clustering algorithms, namely “Incremental k-means”, “Modified k-modes” and “k-prototypes” are implemented.

By using the “Incremental k-means” and “Modified k-modes” independently, we have reduced the number of iterations. The dissimilarity rate i.e., the SSE value (Sum of Squared Errors) in case of Incremental k-means and the Cost function value in case of Modified k-modes paradigm can also be reduced.

References

1. Haung, Z.: Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Canberra, ACT 2601, Australia (1998)
2. He, Z., Deng, S., Xu, X.: Improving K-Modes Algorithm Considering Frequencies of Attribute Values in Mode. Harbin Institute of Technology, China (2005)
3. Haung, Z.: A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining
4. Sayal, R., Vijay Kumar, V.: A Novel Similarity Measure for Clustering Categorical Data Sets. International Journal of Computer Applications (2011)
5. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques (2011)
6. Mastrogiannis, N., Giannikos, I., Boutsinas, B., Antzoulatos, G.: CL.E.KMODES: A modified k-modes clustering algorithm. University of Patras, Greece (2009)
7. Khan, S.S., Kant, S.: Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation (2007)
8. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson education (2006)
9. He, Z.: Approximation Algorithms for K-Modes Clustering. Harbin Institute of Technology, China (2006)
10. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Elsevier (2006)

View publication stats