Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Applied Sciences
2(2), 2014, 128-136

ISSN 2321 – 7456

Original Research Article

AN EFFECTIVE VARIABLE CLUSTERING METHOD FOR CHOOSING RIGHT NUMBER OF CLUSTERS

S. M. A. Khaleelur Rahman[1], M. Mohamed Sathik[1], K. Senthamarai Kannan[2]

[1]Dept. of Computer Science, Sadakathullah Appa College, Tirunelveli, Tamilnadu, India
[2] Dept. of Statistics , Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu,

Abstract

Data mining techniques have been used for finding useful patterns in data. Sometimes these patterns are visible easily. Many times, there are lack of patterns or very excessive patterns. Real world data usually contains more complex structures so that predicting useful pattern is not easy for even a sophisticated data mining technique. Clustering provides a method to break a larger database in to meaningful pieces to describe each piece simply. Once the proper clusters have been identified and defined then it is possible to find patterns within each cluster. In this paper , a proposed method based on variable clustering method is presented. We first perform VARCHA approach. This is a bottom up hierarchical agglomerative approach relies on popular K-MEANS algorithm. A top down method VARCLUS is also used .Choosing right number of cluster is an important criterion in clustering problems. Required number of clusters is decided by calculating Variation and Proportion. Experimental results clearly indicated that our methods are effective for choosing number of clusters. The test results suggested our proposed method can be applied to different data sets for effective results.

Key words: Clustering, Cluster correlation, Inter-cluster correlation, Intra-cluster correlation, Linear correlation

1. Introduction

Clustering is the process of grouping data in to different classes or cluster [1]. Objects within a cluster have high similarity in comparison to one another and dissimilar to objects in other clusters. The quality of clustering can be assessed by measuring dissimilarity of objects. Dissimilarities are assessed by different attribute values of the objects [2]. Cluster analysis is an important day-to-day activity . Clustering is used in data segmentation applications. For outlier detection clustering can also be used. In recent days huge amounts of

For Correspondence:
smakrahmanATgmail.com
Received on: June 2014
Accepted after revision: June 2014

data collected so that cluster analysis tool can be used to gain knowledge about the distribution of data.

In order to compute similarities between examples Euclidean distance, Manhattan distance or Minkowski distance is used[14]. It has been used widely by many researchers that classification is an effective form for distinguishing classes of objects or clusters. But it requires collection of a large set of trained patterns. Groups are formed on the classifier and to model each group. Supervised learning methods are suitable for discrete unordered labels. Clustering is an unsupervised learning and does not rely on predefined classes and trained examples[1].In data mining, active researches are going on for clustering mixed numerical and categorical data in large data sets.

2. Related Work

Cluster analysis is an important human activity. Clustering is also called as data segmentation. In data mining, clustering is used as a tool to know about distribution of data. In machine learning, clustering is used for unsupervised learning.

Many clustering approaches have been proposed. These approaches can be classified into seven major categories based on the techniques used (Zhang and Wang, 2006), which are: partition-based, hierarchical-based, density-based, grid-based, model-based, clustering high-dimensional data and constraint-based clustering. For categorical data, k-modes algorithm is used. K-prototype algorithms were proposed for hybrid data.

Agglomerative hierarchical clustering algorithms were discussed by Day and Edelsbrunner. Kaufman [15] and Rousseeuw introduced both Agglomerative hierarchical clustering AGNES and divisive hierarchical clustering, DIANA [5]. Clustering quality of hierarchical clustering can be improved by integrate distance-based iterative relocation. Hierarchical clustering also performed by linkage analysis, transformation or nearest neighbour such as CURE, ROCK.

The quality of unsupervised clustering techniques can be significantly improved by pairwise constraints. Basu, Bilenko and Mooney proposed a framework for semi-supervised clustering.

3. Proposed Work

It has been argued that dissimilarity between objects described by continuous or ratio-scaled variables can be computed by using correlation coefficients also known as the squared correlation coefficient. For group interpretation, we use positive or negative correlation on the latent factors. Both bottoms-up and top-down method are used.

In this paper, we proposed a method which divides a set of numeric variables into either disjoint or hierarchical clusters. This procedure divides a set of variables in to non overlapping and unidimensional cluster. This procedure tries to maximize the sum across clusters. The correlation coefficient can be analyzed. In correlations, all variables are treated as equally important. If correlations are used, variables with larger variances have more importance in the analysis [10].

In an attempt to create the initial clusters, if we use non-hierarchical clustering methods which would spread the outliers across all clusters. Moreover, most of those methods depend on the initialization of the clusters, to be given by the user [10]. This requires domain knowledge. This might be an unstable approach. Therefore, we use hierarchical clustering methods, which are not dependent on the initialization of the clusters.

Three indicators R^2 with own cluster, R^2 with the nearest cluster and 1- R^2 ratio are used. Small value of this in this ratio means good clustering . If this value is more than 1, it suggests that the particular variable has more correlation with another cluster than the members in its own group[11].

By this procedure , we begin all variables in a single cluster and repeat the following steps:

1. Options are specified for splitting the clusters. Depending on the options, the selected cluster has either the smallest percentage of variation explained by its cluster component or the largest eigen value associated with the second principal component . The selected

cluster is split into two clusters by finding the first two principal components, and assigning each variable to the component with which it has the higher squared correlation[12].

2. Variables are reassigned iteratively to clusters so that the variance is maximized and counted for the cluster components. Through this reassignment a hierarchical structure may be maintained.

We can stop the procedure when each cluster satisfies a user-specified criterion. Each cluster has only a single Eigen value greater than one. This is very basic criterion for determining the single factor dimension. Variable are assigned to clusters in two ways. First in each iteration, the cluster components are computed and each is assigned to the component with which it has the highest squared correlation. In the second phase each variable is assigned to a different cluster and test whether it increases the amount of variance explained. Before testing the next variable for reassignment, the components of the two clusters involved are recomputed .

4. Results and Discussion

Now, we investigate our proposed method by using a synthetic data set. This is an artificial data set with two dimension and values are entered to demonstrate our clustering algorithm. This data set has 23 attributes with 5000 samples. All 21 attributes are continuous attributes.

Table 1 displays the Cluster characterization. It displays the number of clusters, the number of members (attributes) in each cluster, value and proportion explained in each cluster. The variation explained by the variables in each cluster is the contribution of variables only from the variables in the cluster but not from all variables of other clusters. In this data set only three clusters are detected . Cluster 1 has 9 members , 2 has 3 members and 3 has 8 members . The last column 'Proportion' represents the total explained variation divided by the sum of cluster variation. This value, 0.5510, informs us that about 55% of the total variation in the data can be considered for three clusters .

Cluster	# Members	Variation Explained	Proportion Explained
1	9	4.9051	0.5450
2	3	1.5399	0.5133
3	8	4.5749	0.5719
Total		11.0198	0.5510

Table 1 Cluster characterization Cluster summary

Table 2 shows how the variables are clustered. The first cluster represents variable V6,V7,V14,V5,V12,V4,V12,V3 ,and V2,the second cluster contains V10,V11 and V1 and the last ,third cluster contains the variables V15,V6,V8,V17,V9 and V18. We use three indicators to demonstrate correlation with other clusters.

For each cluster, two squared correlations are calculated. The third column in Table 2 'own cluster' gives the squared correlation to the variable with its own cluster. This value must be higher than the squared correlation with any other clusters. It is better to have a larger squared correlation. The next 'Next Closest' column values must be low if the clusters are well separated. The next column 1-R^2 Ratio with small values also indicate about good clustering. If this value is greater than 1 then we can conclude that this particular attribute is more correlated with another group than this group. If the clusters are well separated then value for a variable with the nearest cluster should be low. Small value in the ratio column indicates good clustering. The highest 1 - R^2 ratio is 0.9913 for the attribute V01 in the second cluster. This is demonstrated in Table 2. It is clearly displayed that in cluster 1 the variable V02, in cluster 2 variable V01 and in cluster 3 variable V20 have a weaker relation to its own cluster than other variable .

Displayed Output

Cluster	Members	Own Cluster	Next Closest	1-R^2 ratio
1	V06	0.7188	0.3611	0.4402
	V07	0.7153	0.4894	0.5576
	V14	0.6338	0.4614	0.6800
	V05	0.6875	0.2239	0.4027
	V13	0.6748	0.2201	0.4169
	V04	0.5488	0.1607	0.5376
	V12	0.4263	0.0092	0.5790
	V03	0.3608	0.1020	0.7118
	V02	0.1390	0.0343	0.8916
2	V10	0.7651	0.3322	0.3518
	V11	0.7660	0.0856	0.2558
	V01	0.0087	0.0000	0.9913
3	V15	0.7489	0.4490	0.4556
	V16	0.7336	0.3191	0.3912
	V08	0.6633	0.4386	0.5998
	V17	0.6824	0.1888	0.3915
	V09	0.6698	0.1782	0.4019
	V18	0.5627	0.1450	0.5114

	V19	0.3691	0.0838	0.6886
	V20	0.1451	0.0305	0.8818

Table 2
Cluster members and R-square values

Table 3 shows the cluster structure. How different variables are correlated to each cluster. Here we use a threshold value of 0.7 which produces better results. If we use a threshold value of 0.5, it also produces same number of clusters and each cluster has same number of attributes as it members. For example , cluster 2 has only three members V10,V11 and V1 for both threshold values. Size of the cluster is the important factor for observing behavior of one cluster from the remaining set of data. In the membership column, attribute with this threshold alone counted and considered. We can now say that variables are well clustered because no attribute is correlated to more than one cluster. If each variable is associated to only one cluster then we can say that variables are well clustered. It suggests that all variables are well clustered. High correlations between the variables and their own cluster component are displayed in Table 3. The correlations between the variables and the opposite cluster component are all moderate.

Attribute	# membership	Cluster 1	Cluster 2	Cluster 3
V01	0	-0.0081	0.0935	-0.0032
V02	0	0.3728	-0.0732	-0.1851
V03	0	0.6007	-0.1210	-0.3194
V04	1	0.7408	-0.1644	-0.4008
V05	1	0.8291	-0.1800	-0.4732
V06	1	0.8478	-0.0424	-0.6010
V07	1	0.8458	0.0592	-0.6996
V08	1	0.6623	0.2677	-0.8144
V09	1	0.4222	0.4929	-0.8184
V10	1	0.0657	0.8747	-0.5764

V11	1	-0.3210	0.8752	-0.2925
V12	0	-0.6529	0.4548	0.0960
V13	1	-0.8215	0.1780	0.4691
V14	1	-0.7961	-0.1077	0.6793
V15	1	-0.6701	-0.3308	0.8654
V16	1	-0.5649	-0.4060	0.8565
V17	1	-0.4345	-0.4824	0.8261
V18	1	-0.3808	-0.4344	0.7502
V19	0	-0.2894	-0.3470	0.6075
V20	0	-0.1746	-0.1799	0.3809

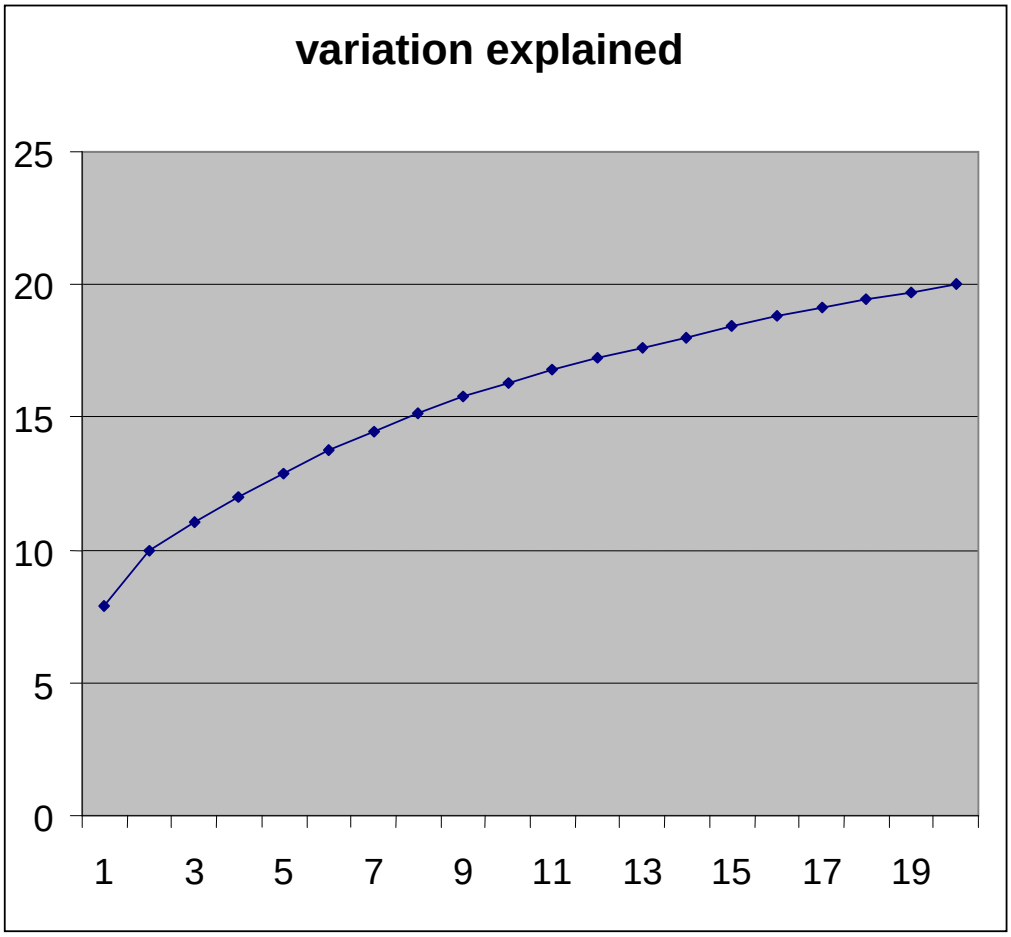
Table 3 Cluster correlations -- Structure

From the result of Table 4, we can infer that the right number of clusters related to area with a large value of the angle. The factor maximum cluster number does not affect the process of deciding suitable number of clusters because it will limit the search process for maximum number of clusters[7]. Here the best selection is 3 in green and the next possible are 2 and 4 in grey. Our method proposes optimal number of clusters 3, 4 and 2. This can be checked with the domain knowledge.

# clusters	Var.Expl.	Dif.	Cos	Angle	Moving Avg.
1	7.9039	0.0000	0.0000	0.0000	0.0000
2	10.0062	2.1023	0.9445	0.3347	0.1143
3	11.0198	1.0136	1.0000	0.0082	0.1339
4	12.0168	0.9970	0.9983	0.0588	0.0230

5	12.9030	0.8862	1.0000	0.0018	0.0604
6	13.7859	0.8829	0.9927	0.1207	0.0418
7	14.4739	0.6879	1.0000	0.0028	0.0532
8	15.1577	0.6838	0.9994	0.0360	0.0441
9	15.7900	0.6323	0.9956	0.0934	0.0474
10	16.2985	0.5085	0.9999	0.0129	0.0434
11	16.7908	0.4923	0.9997	0.0238	0.0353
12	17.2539	0.4631	0.9976	0.0691	0.0322
13	17.6355	0.3816	1.0000	0.0038	0.0249
14	18.0214	0.3859	1.0000	0.0017	0.0026
15	18.4054	0.3840	1.0000	0.0023	0.0172
16	18.7867	0.3813	0.9989	0.0476	0.0184
17	19.1143	0.3277	1.0000	0.0051	0.0289
18	19.4363	0.3220	0.9994	0.0340	0.0149
19	19.7212	0.2849	1.0000	0.0056	0.0132
20	20.0000	0.2788	0.0000	0.0000	0.0000

Table 4
Detailed Results



Variation Explained and Determination of the Number of Clusters

5. Conclusion

Structure of the complex database can be divided in to simpler clusters by using undirected automatic clustering. By using different distance measures, automatic clustering can be applied on any kind of data. In this paper, a new method is proposed based on clustering algorithms for detecting number of clusters. Perform hierarchical clustering algorithm. This procedure divides a set of variables in to non overlapping and unidimensional cluster . The test results show that the proposed approach gave a clear idea to the user to create right number of clusters automatically. Methods proposed are not dependent on the initialization of the clusters. The test results show that the proposed approach gave effective results when applied to different data sets and if we run any number of times it will give the same results.

References

[1] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques ,2nd ed. Elsevier

[2] Michael J.A.Berry, Gordon S. Linoff, . Data Mining Techniques For Marketing, Sales and CRM ,2nd ed., Wiley

[3] Jain, A. and R. Dubes, 1988. Algorithms for Clustering Data. Prentice-Hall.

[4] Loureiro, A., L. Torgo and C. Soares, 2004. Outlier Detection using Clustering Methods: a Data Cleaning Application, in Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.

[5] Kaufman, L. and P. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.

[6] Han, J. and M. Kamber , 2006. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2nd ed.

[7] Takamasa Yokoi, Wtaru Ohyama, Tetsushi Wakabayashi and Fumitaka Kimaru, Joint IAPR International Workshops, SSPR 2004 and SPR 2004, Lisbon, Portugal, August 2004 Proceedings, Eigen Space method by Associative Networks for Object Recognition : 95 – 103

[8] Glenn W. MilliganLane, , Clustering and Classification, World Scientific Pub.River Edge,NJ,1996. 75-80.

[9] Knorr, E. and R. Ng, Algorithms for Mining Distance-based Outliers in Large Data Sets, 1998. Proc. the 24th International Conference on Very Large Databases (VLDB), pp. 392-403.

[10] P.Arabie, L.J.Hubert,G.De Soete , Clustering and Classification, World Scientific Pub. Co. Pte Ltd,Rep.1999, pg. 341-340.

[11] Jain, A. and R. Dubes, 1988. Algorithms for Clustering Data. Prentice-Hall.

[12] Joseph Lucas. A Dimension Reduction Technique for Local Linear Regression.pp 273-75

[13] MacQueen, J.,1967. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. and Prob, pp. 281-97.

[14] Wai-Ki Ching, Michael Kwok-Po Ng,World Scientific 2003.

[15] Kaufman, L. and P. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.

[16] Francisco de A.T. de Carvalho, Yves Lechevallier and Renata M.C.R. de Souza ,
A Dynamic Clustering Algorithm Based on Lr Distances for Quantitative Data , Proceedings of the International Federation of Classification Societies