ISSN: 0976-3031

Available Online at http://www.recentscientific.com

International Journal of Recent Scientific Research
Vol. 8, Issue, 1, pp. 15443-15447, January, 2017

International Journal of Recent Scientific Research

Research Article

NEW ATTRIBUTE CONSTRUCTION IN MIXED DATASETS USING CLUSTERING ALGORITHMS

***Sagunthaladevi.S[1] and Bhupathi Raju Venkata Rama Raju[2]**

[1]Department of Computer Science, Mahatma Gandhi University, Meghalaya-793101, India
[2]Department of Computer Science, IEFT College of Engineering, Villupuram-605108, Tamilnadu, India

ARTICLE INFO

Article History:

Received 10th October, 2016
Received in revised form 14th November, 2016
Accepted 08th December, 2016
Published online 28th January, 2017

Key Words:

Clustering, Classification, Prediction, Clustering Algorithm for Mixed Dataset (CAMD), Clustering Algorithm for Categorical Dataset (CACD), Clustering Algorithm for Numerical Dataset (CAND).

ABSTRACT

Classification is a challenging task in data mining technique. The main aim of Classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then group the similar data into number of classifiers and it assigns items in a collection to target categories or classes. Finally classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. Various classification algorithms have been developed to group data into classifiers. However, those classification algorithms works effectively either on pure numeric data or on pure categorical data and most of them performs poorly on mixed categorical and numerical data types. Previous classification algorithms do not handled outliers perfectly. To overcome those disadvantages, this paper presents Clustering Algorithm for Mixed Dataset (CAMD) Algorithm for clustering. CAMD algorithm divides into CACD and CAND algorithms for Numerical and Categorical datasets separately to improve the performance of clustering.

INTRODUCTION

Classification is one of the important data mining techniques. Learning from a given data set to build a classification model becomes difficult when available sample size is small. As databases are rich with hidden information which can be used for intelligent decision making, Classification can be used to extract models describing important data classes. Such analysis can helps us to understand large amount of data. Classification predicts categorical (discrete, unordered) labels. To build a correct classification model sufficient amount of training data is required. Classification techniques have better capability to handle a wider variety of datasets than regression. It can also be described as a supervised learning algorithm in the machine learning process. The objective of the supervised classification models is to build a concise classifier model, minimize the error rate and improve the efficiency of the learning process. People often make mistakes during data analysis and while establishing new relationships between multiple features. Thus, the inconsistency in dataset creates a difficulty in finding a solution to certain problems. This can be successfully solved using supervised machine learning techniques. To evaluate the derived classification model, preprocessed and dimension reduced dataset is partitioned into training data and testing data. It curtails the data discrepancies in selected datasets.

Basically classification and prediction analyze class labeled data objects, but clustering analyze the data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attribute values describing the objects. Often, distance measures are used. Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A

**Corresponding author: Sagunthaladevi.S*
Department of Computer Science, Mahatma Gandhi University, Meghalaya-793101, India

cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and it can be considered as a form of data compression. Although classification is an effective technique for distinguishing groups or classes of objects, it requires costly collection and labeling of a large set of training tuples or patterns, which the classifier uses to model each group. It is often more desirable to proceed in the reverse direction: First partition the set of data into groups based on data similarity (e.g., using clustering), and then assign labels to the relatively small number of groups. Additional advantages of such a clustering-based process are that it is adaptable to changes and helps single out useful features that distinguish different groups.

Mixed Dataset Clustering

Data Clustering is a group of objects into classes of similar objects, which are grouped into one cluster; dissimilar objects are grouped into another cluster. In existing, clustering algorithms are only works on either numeric dataset or categorical dataset. Those existing works cannot consider the concept of mixed dataset. In proposed, mixed dataset used with the base of classification and clustering. Here, Clustering Algorithm for Mixed Dataset (CAMD) Algorithm is used for clustering the constructed attributes. Basically classification predicts categorical data which is unordered or discrete. In order to process numerical data which is present in mixed datasets, clustering concepts is proposed.

METHODOLOGY

Mixed Datasets

Mixed Attribute types dataset contains both numerical and categorical types of attributes. As mixed attribute type datasets are common in real life, clustering and classification techniques for mixed attribute type datasets is required in various informatics fields such as bio informatics, medical informatics, geo informatics, information retrieval, to name a few. These mixed attribute datasets provide challenges in clustering and classification because there exist many attributes in both categorical and numerical forms so mixed attribute type should be considered together for more accurate and meaningful clustering and classification. It is a small dataset, so three mixed small data sets are explained in this section namely Contact Lenses Data Set, Hayes Roth Data Set and Statlog Data Set. These 3 data sets are downloaded from the UCI Machine Learning Repository database.

New Attributes Construction

To Evaluate the Performance of the New Attributes Construction, two parameters were considered, those two parameters are,

1. New Attributes Construction Time
2. Number of new Attributes Constructed.

New Attributes Construction Time

Before new attributes construction, note the current time in Milliseconds (Starting Time). Then construct the new attributes. After new attributes construction, note the current time (Ending Time) once again. For new attributes construction time, subtract both time.

New Attributes Construction Time = Ending Time – Starting Time

Number of New Attributes Constructed

Before new attributes construction, note the no of attributes exists (Available attributes before construction). Then construct the new attributes. After new attributes construction, note the no of attributes exists (Available attributes after construction) once again. For newly constructed attributes, subtract both results.
Number of New Attributes Constructed = Available attributes after construction – Available attributes before construction.

Camd (Clustering Algorithm For Mixed Dataset) Algorithm

Input: The Mixed Dataset, k
Output: Clusters

1. Splitting the Mixed Dataset into Numerical Dataset (NDS) and Categorical Dataset (CDS).
2. Clustering the Categorical Dataset with k value using Clustering Algorithm for Categorical Dataset (CACD) Algorithm.
3. Clustering the Numeric Dataset with k value using Clustering Algorithm for Numerical Dataset (CAND) Algorithm.

Camd Architecture

As mentioned in previous section, MDCA algorithm split the given input of mixed dataset into numerical and categorical. For numerical datasets CAND algorithm is going to apply and for categorical datasets CACD algorithm is chosen.

Table 1 Dataset Characteristics

S.No	Characteristics	Contact Lenses	Hayes – Roth	Statlog
1	Data Set Characteristics	Multivariate	Multivariate	Multivariate
2	Number of Instances	24	160	270
3	Area	N/A	Social	Life
4	Attribute Characteristics	Categorical, Integer	Categorical, Integer	Categorical, Real
5	Number of Attributes	4	5	13
6	Associated Tasks	Classification	Classification	Classification
7	Missing Values	No	No	No

Table 1 lists the characteristics of input mixed datasets with their assigned values. Each dataset have some different behavior in their character.

Cand (Clustering Algorithm for Numerical Dataset)

CAND is a clustering algorithm used for numerical datasets. It adopts a middle ground between centroid based mostly and all-points approach. It is belongs to bottom-up approach.

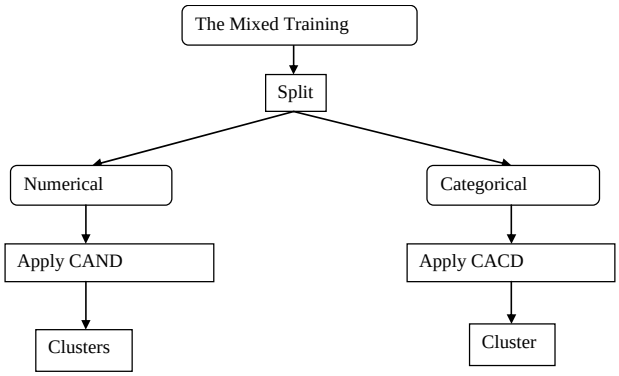

Figure 1 Clustering Algorithm for Mixed Dataset Architecture

Here it represents each cluster by a fixed number of points which are generated by selecting well scattered points from the cluster. Then shrink the points towards the cluster centre by a specified faction. The advantage of CAND is shrinking helps to intense the effects of outliers. Set a target i.e., representative point number c, for each of the clusters select c well scattered points attempting to capture the physical shape and geometry of the cluster. The chosen scattered representative points are then finally shrunk towards the centroid in a fraction of a where 0<= a<=1.

CAND employs a hierarchical clustering algorithm that adopts a middle ground between the centroid based and all point extremes for avoid the problems with non-uniform sized or shaped clusters. In CAND, a constant number c of well scattered points of a cluster are chosen and they are shrunk towards the centroid of the cluster by a fraction α. The scattered points after shrinking are used as representatives of the cluster. The clusters with the closest pair of representatives are the clusters that are merged at each step of CAND's hierarchical clustering algorithm. This enables CAND to correctly identify the clusters and makes it less sensitive to outliers.

CAND Algorithm
CAND (no. of points, k)
Input: A set of points S
Output: k clusters

Step 1: For each cluster u, in u.mean & u.rep store the points mean in the cluster. A set of c points of the cluster (initially c = 1 since each cluster has one data point). Also u.closest stores the cluster closest to u. Then all the input points are inserted into a k-d tree T

Step 2: Now, Treat each input point as separate cluster, compute u.closest for each u. Then insert each cluster into the heap Q. (clusters are arranged in increasing order of distances between u and u.closest). While size(Q) > k

Step 3: Then Remove the top element of Q(say u). Merge it with its closest cluster u.closest(say v). Compute the new representative points for the merged cluster w. Remove u and v from T and Q.

Step 4: For all the clusters x in Q, update x.closest and relocate x. Insert w into Q
Step 5: repeat

CACD (Clustering Algorithm for Categorical Dataset)

CACD is also one of the Hierarchical clustering that deals with concepts of links i.e., the number of common neighbours between two objects for data with non-numeric which is categorical attributes. The formal clustering algorithms for clustering data with Boolean and categorical values use distance functions such as Euclidean and Manhattan distance. However these distance functions doesn't leads to high quality clusters when clustering categorical data.

CACD follows two steps those are Random Sampling, Clustering with links,

Random Sampling

1) Usually it is a large number of data
2) Enables CACD to reduce the number of points considered to reduce complexity
3) Clusters generated by the sample points
4) With appropriate sample size, the quality of clustering is not affected

Clustering with links

It determines the best pairs of clusters to merge at each step of CACD's hierarchical clustering algorithm. For a pair of clusters Ci and Cj, link [Ci, Cj] stores the number of cross links between the clusters Ci and Cj i.e., after drawing a random sample from the database, a hierarchical clustering algorithm that employs links is applied to the sampled points. Finally, the clusters involving only the sampled points are used to assign the remaining data points on disk to the appropriate clusters.

Algorithm for Computing Links

Procedure ComputeLinks(S)
Begin
Step 1: Compute nbrlist[i] for every point i in S. Set link[i,j] to be zero for all i,j
Step 2: for i:=1 to n do {
Step 3: N := nbrlist[i]
Step 4: for j:=1 to |N|-1 do
Step 5: for l:=j+1 to |N| do
Step 6: link|N[j],N[l]] := link[N[j],N[l]]+1
Step 7: }
End

Algorithm for CACD Clustering Algorithm

Procedure cluster(S,k)//Set on n Sample points and k number of clusters
begin
Step 1: link:= ComputeLinks(S) // Compute links
Step 2: for each s E S do
Step 3: q[s]:= build_loacl_heap(link,s) //link[s; j]>0; Q=build_global_heap(S,q) // g(s, max(a[s])
Step 4: while size(Q)>k do{ // iterates until k clusters are left
Step 5: u:= extract_max(Q); v:= max(q[u]) // find the best clusters to merge
Step 6: delete(Q,v); w:= merge(u,v)
Step 7: for each x E q(u) U a{v} do{// Update the q for each x
Step 8: link[x,w]:= link[x,u]+link[x,v]; delete (q[x].u): delete(q[x], v); Insert (q[x],w,g(x,w)); insert(q[w],x,g(x,w
Step 9: update (Q,x,q[x])
Step 10: }

Step 11: insert (Q, w, q[w]) //Update Q
Step 12: deallocate(q[u]);deallocate (q[v])
Step 13:}
End

RESULTS AND DISCUSSION

This section documents the results of constructing a new attributes with clustering from mixed dataset. Two steps are followed for clustering in mixed datasets. First is to apply Constructing a new Attributes for small dataset. The experiment constructs a new attributes for both numerical and categorical datasets to improve the performance. The Second experiment is applying the mixed dataset clustering (CAMD) Algorithm for cluster similar objects. It contains different set of parameters to be followed for cluster. CAMD presents tow different set of

MIXED DATASET CLUSTERING

To Evaluate the Performance of mixed dataset clustering, two parameters were considered. Those two parameters are,

1. Clustering Time
2. Each Cluster Size

Clustering Time

Before clustering, note the current time in Milliseconds (Starting Time). Then execute CAND & CACD Algorithms. After clustering, note the current time (Ending Time) once again. For clustering time, subtract both time. i.e., Clustering Time = Ending Time – Starting Time

Each Cluster Size

After Clustering, each cluster has many similar objects. So Cluster Size considered is an important parameter.

Clustering Time

Table 2 Clustering Time (in ms)

S.No	Algorithm	Clustering Time (in ms)
1	CAND	241
2	CACD	33

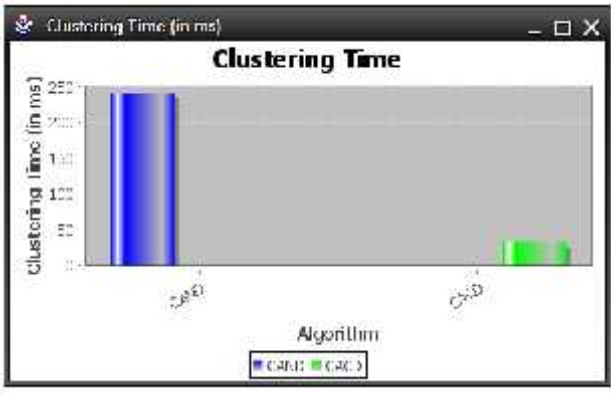

Chart 1 Clustering Time (in ms)

Table 2 shows clustering time of CACD and CAND algorithms, and the results shoes CAND algorithm takes long time for clustering. Likewise CACD Algorithm takes less time. Chart 1 is designed based on table 2 data.

Table 3 Each Cluster Size

S.No	Algorithm	Cluster	Size
1	CAND	Cluster – 1	46
2	CAND	Cluster – 2	48
3	CAND	Cluster – 3	30
4	CACD	Cluster – 1	24
5	CACD	Cluster – 2	35
6	CACD	Cluster – 3	65

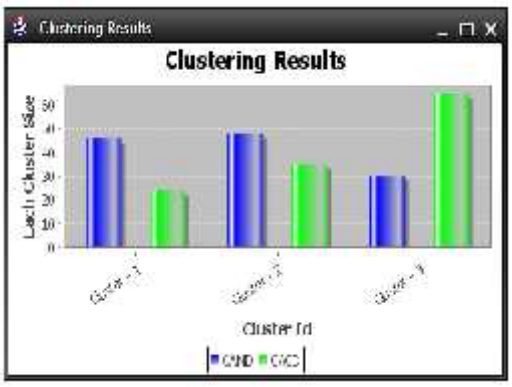

Chart 2 Each Cluster Size

Each Cluster Size

Below table contains cluster value and size of CAND and CACD algorithms. Values are obtained through the calculating parameters and three clusters are taken as samples.

Table 3 shows cluster values of CAND and CACD algorithms and also mention the results of each clusters size in both algorithms. Chart 2 is designed based on the values coated in the table 3. Three clusters are taken for both algorithms and the result shows both algorithms have same average. This is an evaluated results of clustering and it can be further proceed to classification process.

CONCLUSION

Classifying from small mixed dataset is fundamentally difficult, because it is having only few attributes in it. Mixed datasets contains both numerical and categorical attributes. This insufficient data will not lead to a robust classification performance. So this research constructs a new attributes for both numerical and categorical attributes to improve the performance of the classification and clusering. Experimental Results shows, Compared with Numerical Dataset, Categorical Dataset takes minimum time for new attributes construction. In existing clustering algorithms only works on either numeric dataset or categorical dataset and it cannot consider about mixed dataset. In proposed, mixed dataset used. For better classification given dataset must be grouped either numerical or categorical. So Clustering Algorithm for Mixed Dataset (CAMD) Algorithm is used for clustering unordered dataset. The CAMD Algorithm first split the Mixed Dataset into Numerical and Categorical dataset same as mixed dataset classification. Then it applied CAND Algorithm for Numerical Dataset and CACD Algorithm for Categorical Dataset. Experimental Results shows, Compared with CAND Algorithm, CACD Algorithm takes too minimum time for clustering.

References

1. Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.
2. Data Mining Concepts and Techniques, Third Edition ISBN: 978-0-12-381479-1, Morgan Kaufmann Publishers, 225Wyman Street, MA 02451 (USA), 2012.
3. Mrs.Sagunthaladevi.S and Dr.Bhupathi Raju Venkata Rama Raju, "Classification Techniques in Data Mining: An Overview", *Global Journal of Engineering Science and Researches*, Volume *3*, Issue 7, July 2016.
4. Mrs.Sagunthaladevi.S and Dr.Bhupathi Raju Venkata Rama Raju, "Novel Method to Construct a New Attributes for Classification with Mixed Dataset Using Cure and Rock Algorithms", *International Journal of Innovative Research in Computer and Communication Engineering*, Volume 4, Issue 9, September 2016.
5. Mrs.Sagunthaladevi.S and Dr.Bhupathi Raju Venkata Rama Raju, "Performance Analysis Of Cure And Rock Algorithms On Constructing A New Attribute With Mixed Datasets", *International Journal of Innovative Research in Science, Engineering and Technology*, Volume 6, Issue 1, January 2017.
6. Mrs.Sagunthaladevi.S and Dr.Bhupathi Raju Venkata Rama Raju, "New Attribute Construction in Mixed Datasets Using Classification Algorithms", International *Journal of Engineering Sciences & Research Technology*, Volume 6, Issue 2, February 2017.
7. Muhammad Husnain Zafar and Muhammad Ilyas, "A Clustering Based Study of Classification Algorithms", *International Journal of Database Theory and Application* Vol.8, No.1, pp.11-22, 2015.
8. Yogita Rani, Manju & Harish Rohil, "Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9", The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 2, No. 1, January-February 2014.
9. Mierswa, I, "Evolutionary learning with kernels: a generic solution for large margin problems", In Proceedings of the 8th annual conference on Genetic and evolutionary computation, ACM, New York,pp. 1553-1560, 2006.
10. Sivaramakrishnan K.R, Karthik K. and Bhattacharyya, "Kernels for Large Margin Time-Series Classification, International Joint Conference on Neural Networks, pp. 2746-2751, 2007.
11. Hofmann T, Schölkopf B, and Smola A.J, "Kernel Methods in Machine Learning, the Annals of Statistics", Volume 36, pp. 1171-1220, 2008.
12. Kuo-Ping Wu and Sheng-De Wang, "Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space, Pattern Recognition", Volume 42, Issue 5,pp. 710-717, ISSN 0031-3203, 2009.
13. Y.Muto and Y.Hamamoto, "Improvement of the Parze n Classifier in Small Training Sample Size Situations," Intelligent Data Analysis, vol. 5, no. 6, pp. 477-490, 2001.
14. D.C. Li and C.W. Liu, "A Neural Network Weight Determination Model Designed Uniquely for Small Data Set Learning," Expert Systems with Applications, vol. 36, pp. 9853-9858, 2008.
15. K. Saravanan and S. Sasithra, "Review on classification based on artificial neural networks" International Journal of Ambient Systems and Applications (IJASA) Vol.2, No.4, December 2014.
16. Qasem A. Al-Radaideh, Eman Al Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance", *International Journal of Advanced Computer Science and Applications*, Vol. 3, No. 2, 2012.
17. Mohd. Mahmood Ali, Mohd. S. Qaseem, Lakshmi Rajamani, A. Govardhan, "Extracting Useful Rules through Improved Decision Tree Induction Using Information Entropy", *International Journal of Information Sciences and Techniques* (IJIST) Vol.3, No.1, January 2013.
18. Andreas G.K. Janecek, Wilfried N. Gansterer, "On the Relationship between Feature Selection and Classification Accuracy", JMLR: Workshop and Conference Proceedings 4: 90-105, 2008.
19. P.Niyogi, F.Girosi, and P.Tomaso, "Incorporating Prior Information in Machine Learning by Creating Virtual Examples," Proc. IEEE, vol. 86, no. 11, pp. 2196-2209, Nov. 1998.
20. Limère A, Laveren E, and Van Hoof, K. "A classification model for firm growth on the basis of ambitions, external potential and resources by means of decision tree induction", Working Papers 2004 027, University of Antwerp, Faculty of Applied Economics.
21. Hoi, S. C., Lyu, M. R, Chang, E. Y. (2006). "Learning the unified kernel machines for classification, In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 187-196.
22. Xu, J. W., Paiva, A. R., Park, I., and Principe, J. C. (2008). A reproducing kernel Hilbert space framework for information-theoretic learning, IEEE Transactions on Signal Processing, Volume 56, Issue 12, pp.5891-5902.
23. Shilton, A., and Palaniswami, M. (2008). "A Unified Approach to Support Vector Machines", In B. Verma, & M. Blumenstein (Eds.), Pattern Recognition Technologies and Applications: Recent Advances, pp. 299-324.
24. Seema Maitrey, C. K. Jha, Rajat Gupta, Jaiveer Singh, "Enhancement of CURE Clustering Technique in Data Mining", National Conference on Development of Reliable Information Systems, Techniques and Related Issues (DRISTI), Proceedings published in *International Journal of Computer Applications* (IJCA), 2012.
25. Raj Kumar, Dr. Rajesh Verma, "Classification Algorithms for Data Mining: A Survey", *International Journal of Innovations in Engineering and Technology* (IJIET), Vol. 1 Issue 2 August 2012, ISSN: 2319 – 1058, pg: 7-14
26. C. Kim and C.H. Choi, "A Discriminant Analysis Using Composite Features for Classification Problems," Pattern Recognition, vol. 40, no. 11, pp. 2958-2966, 2007.