IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 7, Issue 6, Dec - Jan, 2020
**ISSN: 2320 – 8791 (Impact Factor: 2.317)**
**www.ijreat.org**

# HORIZONTAL PARTITIONING BASED APPROACH OF CLUSTERING PROPOSED FOR BROWSING BEHAVIOUR DATA

Kavita Das[1], Prof. O. P. Vyas[2]
[1]Information Technology, IIIT-A, Allahabad, India.
[2]Guest lecturer (CS), Govt. D.B. Girls' PG (Autonomous) College, Raipur, India,

## Abstract

This work aims in exploring the types of web browsing behaviours of surfers using unsupervised approach. The dataset is collected using an online newspaper. A simple and weighted k-modes algorithm have been applied to the dataset. Due to the characteristics of the dataset, the algorithm could not generate reliable result. Hence, a modification has been introduced to the used algorithm, to make them suitable to the dataset. The modified approach not only increased the purity of clusters but it could generate the actual number of clusters, independent of the number of initial modes given to it. Although the dataset was collected for 4 types of behaviours, our approach shows that there are 6 types of behaviours, at 5% threshold of cluster size to total dataset size. Also the clusters reveal that page-view-duration, number-of–times-of–visit-to-start-page and number-of-news-categories-visited attributes play important role to predict the type of browsing behavior of a surfer. Further, Classification Rule Mining is applied on the original as well as clustered data to find the effect of cluster-classification rule mining and get an insight into the utility of the results.

## I. INTRODUCTION

Clustering of a large dataset is a technique of unsupervised learning by which the dataset can be categorized into sub-groups of similar elements. Each group represents a model with special characteristics. Browsing behaviour of surfers is a new area which provides important data for web personalization. Since its characteristics are wide and not known completely, it is felt justified to apply clustering technique on it to discover models out of it.

The aim of web personalization is to free the surfer from the burden of too much searching the pool of items in the web [7], [8], [14]. The ways of supporting the surfing activities are many. Each approach uses a basic set of information about the surfers to support them. This may include static or dynamic information about the surfers such as user-profiles, webpage-navigation-pattern, browser-usage, browsing-behaviour.

Various approaches of learning the interest of surfers have been adopted. Each approach uses a basic set of information about the surfers to know their interest of information and item. They are as follows-

1. Visitor's profile [9] is created from their history of accessed information and items, and their personal details such as hobby, job. Next prediction was made based on user's profile.
2. Sequences of web pages visited by surfers are observed to find the popular web page navigation paths [2]-[4] among them.
3. Topic of interest of surfers is learned at any time based on pattern of accessing web page [11], [15].

The learning of surfing activities can be supervised and unsupervised. Supervised approach such as classification, requires prior knowledge of types of activities and then learning of their characteristics. But when knowledge about types of activities is not precise, it is possible to learn their charactieristics by generalized groupings i.e. unsupervised approach like clustering.

In this paper, browsing behaviour data i.e. the data related to the browsing activities at the browser has been used. The data has been collected related to four types of tasks – Information Gathering, Just Browsing, Fact finding and Transaction. We applied an efficient k-modes algorithm on the browsing behaviour dataset. This provided insight into the structure of dataset and the problems in applying the clustering algorithm. Hence, we modified the clustering algorithm that is suitable to the dataset and overcome the problems faced by the original algorithm. This technique improved the qualities of the clusters of the data.

Further, this paper is organized as follows. Section II decribes the Dataset. Section III describes the Problem Domain. Section IV decribes with the K-Modes Clustering and the problems faced on our data . Section V describes the improvement made on the Clustering algorithm. Section VI presents the experiments and results. In the following sections, this paper is concluded and made a discussion.

## II. THE DATASET

The dataset is related to the browsing activities of surfers on the browser while fulfilling different tasks. It is a real life dataset of size 80, collected on an online newspaper. The browsing activities are related to tasks such as Information Gathering, Just Browsing, Fact Finding, Transaction, etc. There are 80 records in the data set.

It has 6 attributes – Task Duration, Page View duration, PageViewsPerMinute, NumberOfStartpages, TimeOnStartpage, and NoOfNewsCategories.

The original data, which was numeric, is discretized into two values corresponding to higher and lower values from the mode of each attribute as 'high' and 'low'.

## III. PROBLEM DOMAIN

Browsing Behaviour refers to the way of browsing over a webpage in the browser by a surfer. The pattern of browsing may provide useful insight about the type of task being fulfilled by the surfer and what could be his next preferred item. Exploring the browsing behaviour is a novel approach that can give useful information by applying various data mining techniques on it. In order to find general characteristics from the records of browsing behaviour, K-Modes [12] Clustering technique has been used for exploration of groups of behaviours from the data beyond the known types in the data.

### *A. The K-Modes Approach*

The K-Modes algorithm is similar to K-Means paradigm to cluster categorical dataset, by using simple dissimilarity measures, using modes instead of means of dataset and a frequency based updating of the modes. It preserves the efficiency of k-means algorithm as it follows the same clustering process as the latter.

The dataset and the number of clusters, k, are given as input to the K-Modes algorithm. The initial modes are selected as either k number of distinct objects or as tuple of most frequently occurring attribute values.

It uses a dissimilarity measure between two categorical objects, also called distance measure. The optimization problem for partitioning a set of n objects having m categorical attributes into k clusters is to minimize the total dissimilarities of all the clusters taken together. This is called Cost function.

## IV. THE ORIGINAL APPROACH

For clustering our data, we used the K-Modes technique from [12].

### *A. The Original K-Modes Algorithm Used*

To minimize the cost function, the k-modes algorithm [12] uses the following procedure as given by -

1) For k clusters, input k initial modes.
2) Apply the dissimilarity measure between each object and each mode. And allocate an object to the cluster measuring least dissimilar mode. Similarly allocate all objects to their respective clusters.
3) After each object has been allocated to a cluster, update the mode of each obtained cluster.
4) Retest the objects with new modes. If an object is found to be nearest to mode belonging to another cluster rather than its current one, reallocate the object to that cluster.
5) Repeat steps 3 and 4 as long as an object has changed cluster within a full cycle test of the whole dataset.

### *B. The Dissimilarity Measure Used*

1) *Chi-Square Distance:* This is a frequency-based dissimilarity measure [12].

Here, Dissimilarity is given by

$$d(X,Y) = \sum_{j=1} w_j . \delta(x_j, y_j) \quad \text{,according to [12]}$$

where,

$w_j = ( n_{xj} + n_{yj}) \div (n_{xj.}\ n_{yj})$ ;

$n_{xj,}\ n_{yj}$ are number of objects in the dataset that have values $x_j$ and $y_j$ for attribute j

$$\delta(x_j, y_j) = 0 \text{ if } x_j = y_j \text{ and}$$
$$\delta(x_j, y_j) = 1 \text{ if } x_j \neq y_j$$

2) *Purity of Cluster:* A pure cluster is the one in which all the objects exactly belong to the same cluster, and not to any other cluster. We propose to measure the purity of a cluster having categorical values only as-

Impurity = (number of mismatching attribute values of all objects to the values in mode) ÷ (number of objects in the cluster)

Since, in this dataset there are 6 attributes, maximum impurity can be 6.

Hence, Purity = 6 – impurity

### *C. Experiment and Results on the Dataset*

We apply the k-modes algorithm with k=6 initial modes which is more than the number of types in the dataset. The dataset was collected with a view of recording 4 types of browsing tasks.

The clusters obtained are described in the following tables. It contains the modes, their average dissimilarities, number of objects, and number of dissimilarities.

TABLE I
CHI-SQUARE DISSIMILARITY MEASURE OF CLUSTERS

| Modes | Avg $d(X,Q_i)$ | No. Of $X_i$ | No of dissimilarities | Impurity |
|---|---|---|---|---|
| $Q_0$ | 0.1618 | 19 | 06 | 0.3157 |
| $Q_1$ | 0.094 | 20 | 37 | 1.85 |
| $Q_2$ | 0.061 | 20 | 24 | 1.2 |
| $Q_3$ | 0.0619 | 05 | 06 | 1.2 |
| $Q_4$ | 0.0 | 06 | 00 | 0.0 |

| | | | | |
|---|---|---|---|---|
| $Q_5$ | 0.0352 | 10 | 07 | 0.7 |

Total number dissimilarities in table I = 80
Average impurity in table I = 0.878
Average Purity in table I = 5.122

### *D. The Problems*

While applying the algorithms on the dataset, we faced following problems-

1) Each cluster contained many objects which were same as the initial modes i.e. having dissimilarity=0. Hence, the algorithms could run only one iteration, providing no updation of modes or improvements over initial modes.
2) If modes were updated using only objects having non-zero dissimilarities, the iterations for updation of modes went into infinite loop. This happened because after a few iterations, the last two set of modes interchange each other continuously.
3) The k-modes algorithms use user-specified number of modes k. So the clusters and their qualities obtained are sensitive to these initial modes. Hence, the solutions are locally optimal.

## V. THE IMPROVED APPROACH

We suggested a way of improving the quality of clusters on the dataset using horizontal partitioning of the dataset as described in the following sections.

### *A. K-Modes with Horizontal Partitioning (HP-KModes)*

Although the quality of clusters using Chi-square distance is good, the issue of finding the exact number of clusters in a dataset still remains to be solved. Traditionally when we input the value of k, it produces k clusters. It is needed that if the actual number of clusters in a dataset is n, the input k should be able to expand or shrink up to n. The qualities of the clusters are also locally optimal.

The improved algorithm is proposed as follows-

1) Input k initial modes, with wide different categorical values manually.
2) Run single iteration of k-modes algorithm
3) In each cluster
   a) If the size of the cluster is less than 5% of the dataset size, the cluster is not significant.
   b) count the ratio of values present in each attribute
   c) If an attribute is having number of non-matching values with the value in the mode, more than one-third of the size of cluster, creates another mode with this non-matching value in the attribute.
   d) If an attribute is having non-matching values to the value in the mode, more than two-third of the size of cluster, create another mode with this non-matching value in the attribute and delete the current one for the next run.
4) Repeat steps 1, 2 and 3, with new set of modes until there is no updation in modes.

Hence, this approach creates new clusters by partitioning an existing cluster based of frequency of attribute values.

## VI. EXPERIMENTS AND RESULTS

### *A. Experiment*

Algorithms used are K-modes with Chi-Square measure and the Improved Algorithm. Table I shows the clusters obtained from Chi-Square K-Modes algorithm. Table II shows the clusters obtained from the Improved Algorithm with k=3 as input. Table III shows the clusters obtained from the Improved Algorithm with k=7 as input.

TABLE II
CLUSTERS USING IMPROVED ALGORITHM WITH K=3.

| Modes | Avg $d(X,Q_i)$ | No. Of $X_i$ | No of dissimilarities | Impurity |
|---|---|---|---|---|
| $Q_0$ | 0.05028 | 2 | 1 | 0.50 |
| $Q_1$ | 0.2845 | 18 | 10 | 0.55 |
| $Q_2$ | 0.0341 | 18 | 12 | 0.66 |
| $Q_3$ | 0.2189 | 21 | 09 | 0.428 |
| $Q_4$ | 0.01885 | 8 | 03 | 0.375 |
| $Q_5$ | 0.03154 | 8 | 03 | 0.375 |
| $Q_6$ | 0.02021 | 5 | 02 | 0.4 |

Total no. of dissimilarities = 39 (excluding $Q_0$)
Average Impurity of clusters = 0.464
Average Purity = 5.535

TABLE III
CLUSTERS USING IMPROVED ALGORITHM WITH K=7.

| Modes | Avg $d(X,Q_i)$ | No. Of $X_i$ | No of dissimilarities | Impurity |
|---|---|---|---|---|
| $Q'_0$ | 0.02875 | 23 | 13 | 0.565 |
| $Q'_1$ | 0.0427 | 24 | 19 | 0.79 |
| $Q'_2$ | 0.01269 | 08 | 02 | 0.25 |
| $Q'_3$ | 0.02795 | 02 | 01 | 0.50 |
| $Q'_4$ | 0.03022 | 10 | 05 | 0.50 |
| $Q'_5$ | 0.0 | 06 | 0 | 0 |
| $Q'_6$ | 0.05060 | 07 | 07 | 1 |

Total no. of dissimilarities = 46 (excluding $Q_3$)
Average Impurity of clusters = 0.3105
Average Purity = 5.482

The modes obtained corresponding to table II and table III are given in table IV and table V respectively, along with the classification rules obtained in corresponding clusters. The classification rules are obtained using benchmark C4.5 approach. The classification rules of the original dataset are shown in Fig. 1.

TABLE IV
MODES CORRESPONDING TO CLUSTERS IN TABLE II

| | Task Dur | PV Dur | PV perMin | No ofSP | TimeOn SP | News Cat |
|---|---|---|---|---|---|---|
| $Q_0$ | NA since size < 5 | | | | | |
| $Q_1$ | Low | Small | Few | High | Low | Small |
| $Q_2$ | Low | Small | Few | Low | High | Small |
| $Q_3$ | High | Big | Many | High | High | Big |
| $Q_4$ | High | Big | Many | Low | Low | Big |
| $Q_5$ | High | Big | Many | Low | High | Big |
| $Q_6$ | High | Big | Few | Low | High | Big |

The classification rules obtained are as follows-

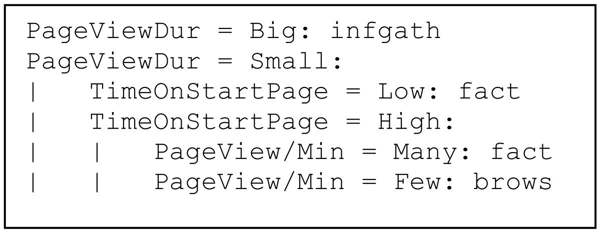
$Q_0$.NA
$Q_1$.NewsCat=Big: infgath; NewsCat=Small: fact
$Q_2$.PVperMin=Many: fact; PVperMin=Few: brows
$Q_3$.Infgath
$Q_4$.Infgath
$Q_5$.PVDur=Big: infgath; PVDur=Small: fact
$Q_6$.Infgath

TABLE V
MODES CORRESPONDING TO CLUSTERS IN TABLE III

| | Task Dur | PV Dur | PV perMin | No ofSP | Time On SP | News Cat |
|---|---|---|---|---|---|---|
| $Q'_0$ | High | Big | Many | High | High | Big |
| $Q'_1$ | Low | Small | Few | Low | High | Small |
| $Q'_2$ | Low | Small | Many | Low | High | Small |
| $Q'_3$ | NA since size < 5 | | | | | |
| $Q'_4$ | High | Big | Many | Low | High | Big |
| $Q'_5$ | High | Big | Many | Low | Low | Big |
| $Q'_6$ | High | Big | Few | Low | Low | Big |

The classification rules obtained are as follows-

$Q'_0$.Infgath
$Q'_1$.TimeOnSP=High:brows; TimeOnSP=Low: fact
$Q'_2$.Fact
$Q'_3$.NA
$Q'_4$.Infgath
$Q'_5$.Infgath
$Q'_6$.PVDur=Big: infgath; PVDur=Small: fact



```
PageViewDur = Big: infgath
PageViewDur = Small:
|   TimeOnStartPage = Low: fact
|   TimeOnStartPage = High:
|   |   PageView/Min = Many: fact
|   |   PageView/Min = Few: brows
```

Fig. 1. Rule set of the original dataset

### B. *Result*

Comparing the result from original clustering algorithm shown in table I with the results from the improved algorithm shown in table II and table III, it is found-

a) Average dissimilarity in clusters using HP-KModes algorithm < the average dissimilarity in clusters using original algorithm.
b) Number of dissimilarities in clusters using HP-KModes algorithm < Number of dissimilarities in clusters using original algorithm. Hence, purity of clusters improved using the improved algorithm.
c) Irrespective of the given number of initial number of modes k, it generated 6 clusters.
d) Out of the 6 clusters, 4 clusters are the same in both sets of clusters as given in tables IV and table V.

## VII. CONCLUSION

In order to get insight into browsing behaviour of surfers, we choose to apply K-modes technique of Clustering on our data. As a result, we found the following facts and problems regarding our data-

a) As there are many objects exactly same as the modes in our dataset, the existing algorithms does not iterate to update the modes. It is producing clusters in one run.
b) If we update the modes, setting aside the records that are exactly same as initial modes, the algorithm goes into infinite loop of modes updation. This happened because after a few iterations, the last two set of modes interchange each other continuously.
c) The clusters are initial modes sensitive, in terms of number of modes and values in clusters, and are locally optimal.

The improved K-Modes clustering algorithm adopted the way of horizontal partitioning of initial obtained clusters. This algorithm is capable of finding the number of clusters actually present in the dataset. Multiple tests of the algorithm on our dataset provided following results-

a) There are 6 clusters in the browsing behaviour dataset used.
b) The clusters are more pure than the clusters generated by original algorithm.
c) Out of the 6 clusters, 4 clusters are the same in multiple runs with different initial modes.

So the algorithm has removed the dependency of number of clusters on the given initial number of clusters, k. It also reduced the local optimization of clusters depending on initial conditions.

## VIII. DISCUSSION

Our proposed algorithm based on Horizontal Partitioning is different from Hierarchical partitioning algorithm because after updating of modes after each iteration, all the objects from the dataset are redistributed among the new updated modes.

Page-View-Duration can categorize the tasks of surfers into two major groups. Further, number-of-visits-to-start-page, and number-of-news-categories play decisive roles in deciding the task and mood of the surfers. By predicting the tasks, surfer's next preferred news items can be predicted and automatically generated.

When Classification Rule Mining (CRM) is applied on the whole dataset, 7 rules of length three were produced whereas CRM on each cluster has produced one rule of length one which are smaller parts of the previous set of rules. The aim was not to increase the efficiency of data mining, but to learn the qualities of the Browsing Behaviour data. This outcome is appreciable that by clustering there is no loss of rules. Instead the smaller rules are less complex and easy to implement in real situation.

For example, in the rule

```
PageViewDur = Small:
   |   TimeOnStartPage = High:
    |   |   PageView/Min = Many: fact
```

Implementing the rule cannot be implemented in the same sequence as the parameters in the rule; instead they are more easily calculated in the reverse order.

Also the number of rules is reduced with clustering.

Landscape mining proposes to explore issues that are general instead of the next data set that we come across. For all the landscapes generated, the general principle of landscape mining follows as –

- First observe the pattern of data.
- Cluster the dataset and then apply a classifier so that special characteristics of each cluster can be learned.

This approach of cluster-classification may benefit in the outlier analysis in the data from smaller clusters as well as in the study of different characteristics of different clusters.

## ACKNOWLEDGMENT

The dataset has been obtained from Indo-German DST-BMBF project, "Web Optimization by Data Mining in Browser Event Stream Data", dated 06/02/2008. We extend our gratitude to Prof Clemens H. Cap and Anne Gutschmidt from Rostock University, Germany, for making extensive studies and preparations in collecting Browser Events Stream data of surfers and churning out various characteristics related to their browsing behaviour.

## REFERENCES

[1] J. Han, M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers. Second Edition. 2001.

[2] J. Srivastava, R. Cooley, M. Deshpande and P.N.Tan, "Web Usage Mining: Discovery and application of usage patterns from web data," *ACM SIGKDD explorations*, 2000

[3] R. Cooley, P. N. Tan, and J, "Srivastava. Discovery of interesting usage patterns from web data," *Technical Report TR 99-022, University of Minnesota*, 1999.

[4] S. Gunduz, M. Tmerozsu, "A Web Page Prediction Model Based on Click stream Tree Representation of User Behaviour," *SIGKDD'03*, 2003

[5] P. Batista, M. J. Silva, "Mining online Newspaper for Web Access Log," in *Proc. of 2nd Intl' Workshop on Recommendation and Personalization in eCommerce (RPeC'02)* (in conjunction with AH 2002).

[6] S. Shinde, C. Sun, and H. Guo, 'Web Usage Mining with Fine-Grained Browsing Activity Tracking," *Knowledge and Information Systems, 1(1)*, 1999

[7] M. Kellar, C. Watters, M. Shepherd, "A Goal-based Classification of Web Information Tasks." in *Proc. of ASIS&T*, 2006

[8] M. Kellar, M. Shepherd, C. Waters, "A Field Study Characterizing Web based Information Seeking Tasks", *University Ave, Canada*, 2005.

[9] M. Eirinaki, M. Vazirgiannis, "Web Mining for Web Personalization," *ACM transactions on Internet Technology*, Vol. 3, No. 1, Feb 2003 pp. 1-27

[10] Norguet, Zimany, Steinberger, "Improving Web Sites with Web Usage Mining, web content Mining and semantic analysis," *SOFSEM 2006: Theory And Practice Of Computer Science, Lecture Notes in Computer Science*, Vol. 3831, 2006, pp. 430-439.

[11] K. Das, O. P. Vyas, C. H. Cap, and A. Gutschmidt, "Suitability of Web Usage Mining for Web Content Syndication," in *Proceedings of National Conference; INDIACom-2009 Computing for Nation Development*, Feb 26-27 2009.

[12] Z. Huang, "A fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining," *In Research Issues in Data Mining and Knowledge Discovery, 283-304* . Sep 1998

[13] S. Aranganayagi, K. Thangavel, "Improved K-Modes for Categorical clustering Using Weighted Dissimilarity Measure," *International Journal of Information and Mathematical Sciences 5:2*, 2009.

[14] A. Gutschmidt, C. H. Cap, F.W. Nerdinger, "Paving the Path to Automatic User Task Identification," in *Workshop on Common Sense Knowledge and Goal-Oriented Interfaces on the International Conference on Intelligent User Interfaces, 2008*

[15] A. Gutschmidth, "An approach to situational market segmentation on on-line newspapers based on current tasks," *in proceedings of fourth ACM conference on Recommender Systems*, Sep 2010, pp.321-324.