DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4301

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Review on K-Mode Clustering Algorithm

Manisha Goyal
Research Scholar: Department of Computer Science and Engineering
Sri Guru Granth Sahib World University
Fatehgarh Sahib, India

Shruti Aggarwal
Assistant Professor: Department of Computer Science and Engineering
Sri Guru Granth Sahib World University
Fatehgarh Sahib, India

Abstract: The main purpose of the process of data mining is to extract useful information from a huge amount of dataset. As one of the most important tasks in data mining, clustering is the process of grouping object attributes and features such that the data objects in one group are more similar than data objects in another group. It is a form of unsupervised learning that means how data should be grouped the data objects (similar types) together will be not known in advance. The algorithms used for clustering are k-means algorithm, k-medoid algorithm, k-nearest neighbour algorithm, k-mode algorithm etc. The K-Mode Algorithm is an eminent algorithm which is an extension of the K-Means Algorithm for clustering data set with categorical attributes and is famous for its simplicity and speed. The 'Simple Matching Dissimilarity' measure is used instead of Euclidean distance and the 'Mode' of clusters is used instead of 'Means'. In this paper, review on the K-Mode Algorithm is done.

Keywords: Data Mining, Clustering, K-Means Algorithm, K-Mode Algorithm

1. INTRODUCTION

Data mining may be defined as the task of process the data from different dimensions and in turn summarized it into the useful information. This process consisted of extraction, transformation, and loading of transaction data onto the data warehouse system, save and process the data in a multidimensional database system, give data access to business analysts and information technology professionals, check the data by application software, present the data in a useful format, such as a graph or table [2]. The datasets to be mined contain millions of objects described by tens, hundreds or even thousands of various types of attributes or variables. The accessed data can be stored in one or more operational databases, a data warehouse or a flat file. Major components of data mining technology have been under development such as statistics, artificial intelligence and machine learning in research areas [1]. The data mining operations and algorithms are required to deal with different types of attributes. In this sophisticated data analysis tools are used along with visualization techniques to segment the data. After this it probability of future events are evaluated [2]. It involves the anomaly detection, association rule learning, classification, regression, summarization and clustering.

In data mining the data is mined using two learning approaches i.e. supervised learning or unsupervised learning [5].

A. *Supervised learning*: In this learning, data includes together the input and the desired result. It is the fast and a perfect learning method. The accurate results are known and are given in inputs to the model during learning procedure. Neural network, Multilayer perception, Decision tree is supervised models.

B. *Unsupervised learning*: The desired result is not provided to the unsupervised model during learning procedure. This method can be used to cluster the input data in classes on the basis of their statistical properties only. These models are for various types of clustering, k-means, distances and normalization, self-organizing maps [3].

2. CLUSTERING USING K-MODE ALGORITHM

Clustering is one of the fundamental tools available, for understanding the nature of the dataset. It is the unsupervised learning that used to place data elements into related groups without advance knowledge of the group definitions [4]. It has alienated the large dataset into groups or clusters according to similarity of properties. From a practical perspective, it plays an outstanding role in data mining applications such as information retrieval and text mining, spatial database applications, Web analysis, marketing, medical diagnostics and many others [1].

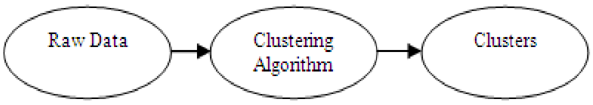

Figure 1. Stages of Clustering

Clustering algorithms have five categories like hierarchical based algorithms, partition-based algorithms, density-based algorithms and grid based algorithms.

Table 1. Types of Clustering

Type	***Description***	***Algorithms***
Hierarchical Based Clustering	It builds a cluster hierarchy and is based on the connectivity approach, based on the clustering algorithms. It uses the distance matrix criteria for clustering the data and constructs clusters step by step.	BIRCH, CURE etc.
Partitioning Based Clustering	The data objects are split into k partitions, where each partition represents a cluster and k<=n, where n is the number of data points. It is based	K-Means, K-Mediod, PAM etc.

	on the idea that a cluster can be represented by a centre point.	
Density Based Clustering	This method finds the cluster according to the regions which grow with high density.	DBSCAN, GDBSCAN, OPTICS etc.
Grid Based Clustering	This method maps all the objects in a cluster into a number of square cells, known as grids. It has a fast processing time that depends on the size of the grid instead of the data.	STING, CLIQUE etc.
Model Based Clustering	In this method, each of the clusters is best fitted to the given model. It may locate clusters by constructing a density function that reflects the space distribution of the data points.	Expectation-Maximization (EM) algorithm etc.

Partitioning Based Clustering is one popular approach of clustering, which transfer objects by moving them from one cluster to another cluster starting from a certain point. The amount of clusters for this technique should be predefined. The algorithms used in this approach are K-Means Algorithm, K-Medoid Algorithm, K-Nearest Neighbour Algorithm etc [4].

K-Means Algorithm is a partitioning based algorithm for clustering that creates clusters of the same type of data according to their closeness to each other based on the Euclidean distance [5]. It intends to partition the objects into a number of clusters in which each object belongs to that cluster with the nearest mean. This method produces exactly the different number of clusters of greater separation distance which is not known as a priori and must be computed from the data [6].

K-Mode Algorithm is an extension of K-Means Algorithm and is the partitioning based clustering algorithm. It uses simple matching dissimilarity function instead of using Euclidean distance. Modes are used to represent centroids and a frequency based method is used to find the centroids in each iteration of the algorithm [7].

Algorithm for K-Mode Clustering:

The steps for k-mode algorithm are as follow:

INPUT: Number of desired clusters K, Data objects D= {d1, d2…dn}

OUTPUT: A set of K clusters

1. *Generate K clusters arbitrarily by selecting the data objects and choose K initial cluster centre, one for every of the cluster.*
2. *Assign data object to the cluster whose cluster centre is near toward it according to Equation (1) and (2).*

$$d(X,Y)= \sum^{m}_{k=1} \delta(x_{i,}\ yi) \qquad (1)$$

$$\delta(x_{i,}\ yi)= \begin{cases} o, x_i = y_i \\ 1, otherwise \end{cases} \qquad (2)$$

3. *Update the K cluster base on allocation of data objects. Calculate K latest modes of every one clusters.*
4. *Repeat step 2 to 3 awaiting no data object has changed cluster relationship otherwise some additional predefined criterion is fulfil.*

K-Mode, an eminent algorithm, works well for categorical datasets whereas K-Means Algorithm does not work well for Categorical datasets. It is famous for simplicity, speed and is linearly scalable with respect to the dataset.

3. SURVEY ON THE VARIANTS OF K-MODE

The Survey on the variants of k-mode algorithm is divided into three sections.

A. First section discusses the existing ways to select initial centroids to improve the accuracy of the clusters in K-Modes algorithm.
B. Second section discusses the algorithm to find an appropriate dissimilarity measure for the dataset containing both numerical and categorical data.
C. Third section discusses the way to remove the dependency on specifying the number of clusters.

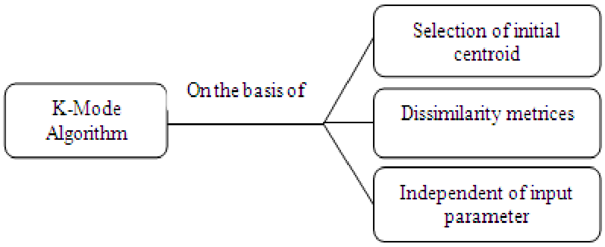

Figure 2. Division of survey on K-Mode Algorithm

A. Selection of Initial Centroids in K-Mode Algorithm

In this section, many possibilities are provided to improve the accuracy of the clusters by improving the selection of initial centroids of the cluster in the K-Mode Algorithm and discussed in the Table 2.

Table 2. Selection of initial centroids in k-mode algorithm

S. No.	*Algorithm Name*	*Description*	*Limitation*
1	K-Mode Algorithm (1997) [8]	This algorithm uses mode instead of calculating means, and a frequency based method is used to update modes in the clustering to deal with categorical attributes.	Initial Modes are chosen randomly.
2	K-Prototype Algorithm (1998) [9]	This algorithm integrates the dissimilarity measure in the K-Means and K-Mode algorithms for clustering objects having mixed numeric and categorical.	The relative frequencies of attribute values are not taken into account in the cluster centroids.
3	Iterative K-Mode Algorithm (2002) [10]	He introduced an initialization method based on Bradley’s iterative initial-point refinement algorithm to the K-Modes clustering.	Many parameters have to be asserted in advance and it takes more time to compute.
4	COOLCAT Algorithm (2002) [11]	This algorithm is able to deal with clustering of data streams and is based on the notion of entropy.	It depends on inputting the parameter m that represents the size of the smallest cluster.
5	Distance based K-mode (2009) [12]	This algorithm proposed initialization method for categorical data and the distance between objects was calculated based on the frequency of attribute values.	The subsample is selected randomly and the single clustering result cannot be guaranteed.
6	Cluster Center Initialization based K-mode (2013) [13]	Some objects whose features are very similar to each other are introduced to this algorithm and have same cluster membership irrespective of the choice of initial cluster centres.	The accuracy of the clusters produced is not better than other algorithms.

7	Entropy based K-Mode Algorithm (2015)[14]	This algorithm improves the cluster accuracy with the analyses of its time complexity while retaining the scalability of the K-Mode Algorithm.	This algorithm can be improved by some other optimization algorithm while retaining its scalability.

K-Mode Algorithm is an extension of K-Means Algorithm as the mode is calculated instead of calculating the mean value in order to find the accurate clusters to avoid overlapping by Haung in 1997. K-Prototype Algorithm proposed by Haung in 1998 results in allocation of less similar objects in a cluster. Iterative K-Mode proposed by Sun et al. in 2002 results in accurate number of clusters. COOLCAT Algorithm proposed by Barber in 2002 is based on the notion of entropy. Distance based k-mode proposed by Cao et al. in 2009 calculates the densities of all the objects for categorical data. Cluster centre initialization based k-mode algorithm proposed by Cao et al. in 2013 generates accurate clusters using prominent attributes. Entropy based k-mode algorithm proposed by Ravi Shankar et al. in 2015 which improves the accuracy of the clusters.

B. Dissimilarity Metric based K-Mode Algorithm

In this section, the amount of work carried out in developing a dissimilarity measure to deal with the datasets containing categorical data and mixed data. Some of the work is discussed in the Table 3.

The dissimilarity is measured in terms of distance function in order to provide the goodness of the cluster. An appropriate metric is used in order to achieve the best clustering because it directly influences the shape of clusters. Improved K-Mode Algorithm proposed by Deng et al. observed that this similarity is directly proportional to the sum of relative frequencies of the common values in the mode in 2005. K-Mode based upon distance metrics proposed by Ahmad in whom the similarity of two attribute values is dependent on their relationship with other attributes in 2007. K-Mode based on cost function proposed by Ahmad improves the Haung's algorithm by cost function in 2007. Dissimilarity based k-mode proposed by Ng et al utilizes some theorems to update the mode of the cluster in 2007. DVD based k-mode algorithm proposed by Lee et al. suggested a new measure called Domain Value Dissimilarity in 2009. DILCA Algorithm proposed by Ienco et al. proposed a method called distance learning for categorical attributes. DISC algorithm proposed by Desai et al. suggested the method Data-Intensive Similarity Measure for Categorical Data in 2011. Biological and Genetic taxonomy information based k-mode proposed by Cao et al. improves the accuracy of the clusters.

C. K-Mode Algorithm independent of input parameter

In this section there will be discussion of two algorithms that deals with the limitation of inputting the value of k to improve the accuracy of the clusters in the k-mode algorithm in which k is the number of clusters formed. These algorithms are discussed with its advantage as well as limitations in the Table 4.

Table 3. Dissimilarity metrices based k-mode algorithm

S. No.	*Algorithm Name*	*Description*	*Limitation*
1	Improved K-Mode Algorithm (2005) [15]	This algorithm proposed a dissimilarity measure based on the similarity between a data object and cluster mode.	It carried forward the same weakness as in K-Modes of choosing the initial modes randomly.
2	K-Mode based upon distance metrics (2007) [16]	This algorithm proposed a dissimilarity measure based on the distance between two attribute values of the same attribute.	It is not suitable for noisy and high dimensional datasets.
3	K-Mode based on cost function (2007) [17]	The proposed cost function added weight for numeric attributes computed from the dataset and all numeric attributes were normalized and discretised to do the calculations.	This algorithm can be improved further by improving the discretising methods for numeric valued attributes.
4	Dissimilarity based k-mode (2007) [18]	This algorithm proposed a new dissimilarity measure in which the modes of clusters were updated in each iteration.	It takes more time to compute.
5	DVD based K-Mode (2009) [19]	The information about distribution of data correlated to each categorical value was used to define the dissimilarity measure.	It takes more computation time and the memory.
6	DILCA Algorithm (2011) [20]	The distance between two values of a categorical attribute was determined by the way in which the value of the other attributes was distributed in the dataset.	The performance depends on some input parameter of this algorithm.
7	DISC Algorithm (2011) [21]	This algorithm suggested this measure didn't require any domain knowledge to understand the dataset.	It requires feedback from a classifier for more accurate results.
8	Biological and Genetic taxonomy information based k-mode (2012) [22]	This algorithm suggested a new dissimilarity measure based on the idea of biological and genetic taxonomy and rough membership function.	It takes more computation time than that of the K-Modes with Huang's measure.

Table 4. K-Mode algorithm independent of input parameter

S. No.	*Algorithm Name*	*Description*	*Limitation*
1	K-mode based upon cluster centres (2004) [23]	This algorithm used a regularization parameter to control the number of clusters in the clustering process.	It takes more memory.
2	K-Mode based upon unified similarity metrics (2013) [24]	This algorithm penalized competitive learning algorithm and these algorithm required some initial value of k which should be greater than the original value of k.	This algorithm can be further improved by better optimization algorithm in terms of accuracy.

K-mode based upon cluster centre algorithm proposed by San et al. in which a suitable value of regularization parameter was chosen to find the most stable clustering results in 2004. K-Mode based upon unified similarity metrics algorithm proposed by Cheung et al. in which the resulting clusters are more accurate than the original K-Mode Algorithm. Both the algorithm provides the better result than the original k-mode algorithm and provides accurate number of clusters.

In this section, previous work done by the researcher in the k-mode clustering is reviewed. Clustering categorical data is an important research topic in data mining. There is the list of the different optimized k-mode algorithm in order to obtain the accurate result in all the above three tables.

4. COMPARATIVE ANALYSIS

The comparison of the various k-mode algorithms is discussed in the Table 4 based on different output parameters. The comparison of the different k-mode algorithms such as k-prototype, modified k-mode, COOLCAT, DVD based k-mode, DILCA, DISC based k-mode algorithm etc. is discussed. The output parameters that are used in Table 5 are discussed as follows:

- *True Positive Rate (TP):* A true positive test result is one that detects the condition when the condition is present.
- *True Negative Rate (TN):* A true negative test result is one that does not detect the condition when the condition is absent.
- *False Positive Rate (FP):* A false positive test result is one that detects the condition when the condition is absent.
- *False Negative Rate (FN):* A false negative test result is one that does not detect the condition when the condition is present.

The output parameters used in the table are defined as follow:

1. *Accuracy:* The Accuracy is the total number of module that is predicted correctly.

$$\text{Accuracy} = (TP+TN) / (TP+FP+TN+FN) \tag{3}$$

2. *Precision:* Precision is the measure of exactness i.e. what percentage of tuples labeled as positive that are actually such.

$$\text{Precision} = (TP) / (TP+FP) \tag{4}$$

3. *Recall:* Recall is the measure of completeness i.e. what percentage of positive tuples did the classifier labelled as positive.

$$\text{Recall} = (TP) / (TP+FN) \tag{5}$$

4. *Execution Time:* The execution time is defined as the time spent by the system executing the task.

Table 5. Comparison of various k-mode algorithms

S. No.	*Algorithm Name*	*Datasets*	*Accuracy*	*Execution Time*	*Precision*	*Recall*
1	K-Mode Algorithm	Soyabean	Better than k-means	Better than k-means	-	-
2	K-Prototype Algorithm	Soybean, Credit approval	Better than k-mode	More than k-mode	-	-
3	Iterative K-Mode Algorithm	Soybean disease	Better than k-mode	-	Better than k-mode	-
4	COOLCAT algorithm	Congressional votes, Synthetic datasets	-	Better than rock algorithm	-	-
5	Distance based K-mode	Soyabean, Zoo, Breast cancer, Mushroom	Better than k-mode	-	Better than k-mode	Better than k-mode
6	Cluster Center Initialization based K-mode	Soyabean, Breast cancer, Zoo, Dermatology, Mushroom, Vote	Better than density based k-mode and distance based k-mode	Similar to distance based k-mode and better than density based k-mode	Better than density based k-mode and distance based k-mode	Better than density based k-mode and distance based k-mode
7	Entropy based K-Mode Algorithm	Synthetic datasets	Better than K-Mode	Similar to K-Mode	-	-
8	Improved K-Mode Algorithm	Congressional vote, Mushroom	Better than k-mode	-	-	-
9	K-Mode based upon distance metrics	Iris, Vote, Heart disease, Australian credit data,	Better than k-means and k-mode	-	-	-
10	K-Mode based on cost function	Wisconsin breast cancer, Letter, Coral data	Better than k-mode	-	-	-
11	Dissimilarity based k-mode	Soyabean	Better than k-mode	More than k-mode	Better than k-mode	Better than k-mode
12	DVD based K-Mode	Synthetic dataset, Wisconsin breast cancer, Letter recognition	Better than k-mode	More than k-mode	-	-
13	DILCA Algorithm	Audiology, Vote, Mushroom, Soyabean, Car evaluation, Adult, Dermatology, Synthetic dataset	-	Better than LIMBO, Delta algorithm	-	-
14	DISC Algorithm	Iris, Breast cancer, Car evaluation, Balance, Hayes roth, Limphography.	Better than k-mode	-	-	-
15	Biological and Genetic taxonomy information based k-mode	Lung cancer, Breast cancer, Zoo, Mushroom, Nursery, Synthetic dataset	Better than k-mode and fuzzy k-mode	-	-	-
16	K-mode based upon cluster centres	Soyabean, Nursery	Better than k-mode	-	-	-

17	K-Mode based upon unified similarity metrics	Soyabean, Breast cancer, Zoo, Vote	Better than k-mode and k- prototype	Better than k-mode and k-prototype.	-	-

The comparison of the improvements done in the k-mode algorithm is shown in the Table 5. This table describes the comparison based on the output parameters such as accuracy, precision, recall and execution time performed on different datasets by researchers.

5. CONCLUSION

The determination of grouping in a set of unlabelled information on the basis of its features is the main objective of clustering. This review work discussed most of the k-mode clustering technique with different approaches. From the discussion, it may be analyzed that there is not any absolute best criterion which can be independent of the final aim of the clustering. This paper presents the analysis of k-mode clustering with their limitations which helps the researcher to select the one according to their need. Some limitations of existing algorithm will be eliminated in the future. This technique will be useful in extraction of useful information using cluster from large data set.

6. REFERENCES

[1]. Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan, “Classification and prediction based data mining algorithms to predict slow learners in education sector”, 3rd International Conference on Recent Trends in Computing, Elsevier, Vol. 57, pp. 500-508, 2015.

[2]. Jeyhun Karimov, Murat Ozbayoglu, “Clustering Quality Improvement of k-means using a Hybrid Evolutionary Model”, Conference Organized by Missouri University of Science and Technology, San Jose, Science Direct, Vol. 61, pp. 38-45, 2015.

[3]. Rui Xu, “Survey of Clustering Algorithms”, IEEE Transactions on Neural Networks, Vol. 16, pp. 645-678, May 2005.

[4]. Han, J. and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, 3rd Edition, India, 2011.

[5]. Farhi Marir, Huwida Said, Feras Al-Obeidat, “Mining the Web and Literature to Discover New Knowledge about Diabetes”, The 3rd International Workshop on Machine Learning and Data Mining for Sensor Networks, Elsevier, Vol. 83, pp. 1256-1261, 2016.

[6]. Preeti Arora, Deepali, Shipra Varshney, “Analysis of K-Means and K-Medoids Algorithm For Big Data”, International Conference on Information Security & Privacy, India, Science Direct, Vol. 78, pp. 507-512, 2016.

[7]. Feng Jiang, Guozhu Liu, Junwei Du, Yuefei Sui, “Initialization of K-modes clustering using outlier detection techniques”, Information Sciences, Science Direct, Vol. 332, pp. 167-183, 2016.

[8]. Z. Huang, “A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining”, In proceeding SIGMOD workshop research issues on data mining and knowledge discovery, pp.1–8, 1997.

[9]. Z. Huang, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values”, ACM Transaction on Data Mining and Knowledge Discovery, Vol. 2, pp. 283–304, 1998.

[10]. Y. Sun, Q. Zhu, Z. Chen, “An iterative initial-points refinement algorithm for categorical data clustering”, Pattern Recognition Letters, Elsevier, Vol. 23, Issue. 7, pp. 875–884, 2002.

[11]. D. Barbara, J. Coute, Yi Li, “COOLCAT: An entropy based algorithm for categorical clustering”, Proceedings of the eleventh international conference on Information and knowledge management, USA, ACM, pp. 582-589, 2002.

[12]. F. Cao, J. Liang, L. Bai, “A new initialization method for categorical data clustering”, Expert Systems with Applications, Science Direct, Vol. 36, pp. 10223-10228, 2009.

[13]. S. S. Khan, A. Ahmad, “Cluster Center Initialization for Categorical Data Using Multiple Attribute Clustering”, Expert Systems with Applications, Elsevier, Vol. 40, pp. 7444–7456, 2013.

[14]. R. S. Sangam, H. Om, “The k-modes algorithm with entropy based similarity coefficient”, 2nd International Symposium on Big Data and Cloud Computing, Procedia Computer Science, Elsevier, Vol. 50, pp. 93-98, 2015.

[15]. Z. He, S. Deng, X. Xu, “Improving K-Modes Algorithm Considering Frequencies of Attribute Values in Mode”, Computational Intelligence and Security, Springer, pp. 157-162, 2005.

[16]. Amir Ahmad, Lipika Dey, “A K-Mean Clustering Algorithm for Mixed Numeric and Categorical Data”, Data & Knowledge Engineering, Science Direct, Vol. 63, pp. 503–527, 2007.

[17]. Amir Ahmad, Lipika Dey, “A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set”, Pattern Recognition Letters, Science Direct, Vol. 28, Issue. 1, pp. 110–118, 2007.

[18]. M. K. Ng, M. J. Li, J. Z. Huang, “On the Impact of Dissimilarity Measure in K-Modes Clustering Algorithm”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, Issue. 3, pp. 503-507, 2007.

[19]. J. Lee, Y. J. Lee, M. Park, “Clustering with Domain Value Dissimilarity for Categorical Data”, Advances in Data Mining, Applications and Theoretical Aspects, Lecture Notes in Computer Science, Springer, Vol. 5633, pp. 310-324, 2009.

[20]. D. Ienco, R. G. Pensa, R. Meo, “From Context to Distance: Learning Dissimilarity for Categorical Data Clustering”, ACM Transactions on Knowledge Discovery from Data, pp.1-22, 2011.

[21]. A. Desai, H. Singh, V. Pudi, “DISC: Data Intensive Similarity Measure for Categorical Data”, Proceedings of Advances in Knowledge Discovery and Data Mining – 15th Pacific Asia Conference, Springer, pp. 469 – 481, 2011.

[22]. F. Cao, J. Liang, D. Li, L. Bai, C. Dang, “A dissimilarity measure for the k-modes clustering algorithm”, Knowledge-Based Systems, Elsevier, Vol. 26, pp. 120–127, 2012.

[23]. O. M. San, V. Hyunh, Y. Nakamori, “An Alternative Extension of the k-Means Algorithm for Clustering Categorical Data”. International Journal Applied Math and Computer Science, Vol.14, pp. 241–247, 2004.

[24]. Y. M. Cheung, H. Jia, “Categorical and numerical attribute data clustering based on a unified similarity metric without knowing cluster number”, Pattern Recognition, Elsevier, Vol. 46, pp. 2228–2238, 2013.