K-modes Clustering Algorithm for Categorical Data

Neha Sharma
Samrat Ashok
Technological Institute
Department of Information
Technology, Vidisha, India

Nirmal Gaud
Samrat Ashok
Technological Institute
Department of Information
Technology, Vidisha, India

ABSTRACT

Partitioning clustering is generally performed using K-modes cluster algorithms, which work well for large datasets. A K-modes technique involve random chosen initial cluster centre (modes) as seed, which lead toward that problem clustering results be regularly reliant on the choice initial cluster centre and non-repeatable cluster structure may be obtain. K-Modes technique has been widely applied to categorical data a clustering in replace means through modes. The pervious algorithms select the attributes on frequency basis but not provided better result. Proposed algorithm select attributes on information gain basis which provide better result. Experimental results showing the proposed technique provided better accuracy.

Keywords

Clustering, Categorical data, K-mean algorithm, K-modes algorithm, Text mining

1. INTRODUCTION

Categorical data clustering is an important research problem in pattern recognition and data mining. Clustering is a broadly use method in which objects are partition into groups, in such a way that object in the same group (or cluster) are more similar among themselves than to those in other clusters[1,2]. K-Means is the widely used numerical clustering method where Euclidean distance is use as a distance measure. Categorical attribute are with small domains. Categorical data are the one which cannot be ordered. Example of categorical value is {male, female} and {low, medium, high}. It is not possible to find the distance between male and female. The geometric properties are not applicable to categorical data. K-Means does not guarantee single clustering since we get different results by random select initial clusters. K-means clustering technique fail to handle datasets with categorical attributes since it minimize the cost function by calculate means. The K-means based Partitioning clustering methods are used for process large numeric data sets for its simplicity and efficiency. Huang proposed a simple matching measure in K-Modes technique, toward cluster categorical data [3]. Clustering technique can be generally classified into two groups: hierarchical, partitioning clustering. Hierarchical algorithm can be further divided into bottom-up and top-down algorithms and partitioning clustering divided into k-mean and k-modes algorithms. The k-modes algorithm as an extension to k-means for categorical data, by replacing k-means with k-modes, introduce a different dissimilarity measure and update the modes with a frequency based method [4,5,6]. In its basic form the clustering problem is defined as the problem of finding homogeneous groups of objects in a given dataset. The K-means based partition clustering methods are used for processing large numeric data set for its simplicity and efficiency [4, 7].

2. LITERATURE SURVEY

In paper [8], method integrate distance density together select initial cluster centre and overcomes shortcoming the exits initialization method for categorical data.

$$Dens(X) = \frac{1}{|U|}\sum_{y\in u} d(x, y)$$

A new initialization method for categorical data clustering has been proposed by taking into account the distance between the objects and the density of the object and overcomes shortcomings of the existing initialization methods. Furthermore, the time complexity of the proposed method has been analyzed They proposed k-modes algorithm which evaluate cluster centre initialization algorithms categorical data use seven real world data set from UCI Machine learn store and experimental result has shown the proposed methods showing that initialization method in k-modes algorithms.

In paper [2], focal point soft subspace cluster in categorical data. Thus, the EBK-modes can be view soft subspace cluster algorithm. They algorithm using the k-means pattern to searching a partition of U into k cluster that minimize the objective task P (W, Z, A) with unfamiliar variables *W*, *Z* and A as follows:

$$\min_{W,z,A} P(W, Z, A) \sum_{l=1}^{k} \sum_{i=1}^{n} wli(xi, zi)$$

In proposed categorical data which evaluate entropy-base k-modes outperform the state-of-the-art algorithm.

In paper [3], offer a distance measure for K-Mode base on the cardinality of domain of attribute. They propose a distance evaluate for K-Modes base on the cardinality of domain of attributes. Objective function is same as K-Modes.

$$\text{Sim } (o_i, o_j) = 1 - \sum_{h=1}^{m} d(x_{h,} y_h)$$

Where

$$\text{d } (X_h Y_h) = \begin{cases} 1 & if X_h = Y_h \\ 0 & otherwise \end{cases}$$

In paper proposed categorical data algorithm evaluates distance measure base on a domain the new measures not only consider the frequency of attribute values but also consider the domain of the attributes.

In paper [9], recover the process of k-mode, when allocate categorical object cluster, the number of every attribute items in cluster be update, so that the new modes of cluster can be compute after read the total dataset. The implementation of parallel *K*-modes has been released the open source community. Proposed method categorical data algorithm evaluates distance measure base on a domain the new measures not only consider the frequency of attribute values but also consider the domain of the attributes. Experimental result showing that proposed distance measure is more efficient than the K-Modes.

In paper [7], present an approach to compute initial modes for K-mode clustering algorithm to cluster categorical data sets. There are several possible ways to accumulate evidence in the context of unsupervised learning: (a) combine results of different clustering algorithms; (b) produce different results by re-sampling the data, such as in bootstrapping techniques (like bagging) and boosting; (c) running a given algorithm many times with different parameters or initializations. Proposed method computing initial modes for k-modes algorithms which evaluate evidence accumulation reduce time consuming

3. K-MODES ALGORITHM

The K-means cluster technique cannot cluster categorical data since of the different measure it using. The K-mode cluster algorithms is base on K-mean pattern other than remove the numeric data limitation even as preserve its effectiveness. This K-mode technique extend K-mean pattern to cluster categorical data through eliminate the limitation forced by K-means follow modification:

- Using simple match dissimilar evaluate or hamming distance used for categorical data object
- change means of cluster by modes

$$d(x, y) = \sum_{i=1}^{f} \delta(X_{j,}Y_j) \ldots\ldots \quad (3.1)$$

d (x, y) gives equal significance to every kind of an attribute. Let Z be a set of categorical data objects described by categorical attributes, A1, A2 . . . . Am. while the above is used because the dissimilarity determine for categorical data objects, the cost function become

$$C\ (Q) = \sum_{i=1}^{n} d(Z_i, Q_i) \qquad (3.2)$$

Where Z_i is the ith element and Q_i is the near cluster centre of Z_i. The K-modes technique minimizes the cost Function defined in Equation 3.2

The K-modes assumes that the information of number of probable group of data (i.e. K) is accessible and consists of the following steps: -

1. Generate K clusters by arbitrarily selecting data objects and choose K initial cluster center, one for every of the cluster.

2. Assign data object to the cluster whose cluster center is near toward it according to equation 3.2.

3. Update the K cluster base on allocation of data objects plus 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 fulfill.

3.1. Computing Initial Cluster Centre Using Multiple Attribute Clustering

3.1.1. Vanilla Attributes – A vanilla attributes is to consider every attributes present data in clustering.

3.1.2. Prominent Attributes – A prominent attributes is to consider important attributes present in data and clustering.

3.1.3. Significant Attributes – A prominent attributes is to consider further important attributes present in data and clustering.

4. PROPOSED ALGORITHMS

This complete algorithm can be described in following steps:

Step 1: calculate information gain matrix for given dataset D have target classes T as follows:

Information Gain: its represent the mutual information which can be gain by one variable by observing the other variable. The information gain about the variable Y by some other variable X is represented as Gain (X/Y) and can be calculated as

$$\text{Gain } (Y/X) = \text{H(X)} - \text{H } (X/Y) = \text{H(Y)} - \text{H } (Y/X) \ldots.\ (3.5)$$

Where H(Y/X) is the conditional entropy and interpreted as they remain entropy of variable Y if the value of another variable X is known, on the basic of probability it can be given as

$$\text{H } (Y/X) = -\sum_{x\in X} p(x) \sum_{y\in Y} p(y/x) log_2(y/x) \ldots\ (3.6)$$

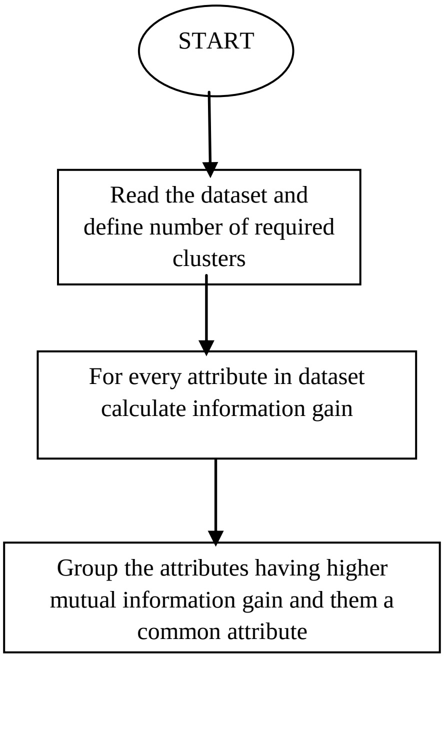
Where p (y/x) is conditional probability of occurrence of value y of the feature F_i with domain y together with occurrence of value x of the feature F_i with domain X. As the equation (3.6) show information gain is a symmetrical measure therefore $G(Y|X) = G(X|Y)$.

Step 2: resting on the basic of information gain value group the features having similar information gain value.

Step 3: apply agglomerative hierarchical cluster tree on the filtered dataset and calculate the cluster centroids.

Step 4: perform the K-modes clustering and find out the labels of each data provided by k-modes.

Step 5: compare the labels of each data provided by K-modes with the known one to test the accuracy of the method.


START
Read the dataset and define number of required clusters
For every attribute in dataset calculate information gain
Group the attributes having higher mutual information gain and them a common attribute

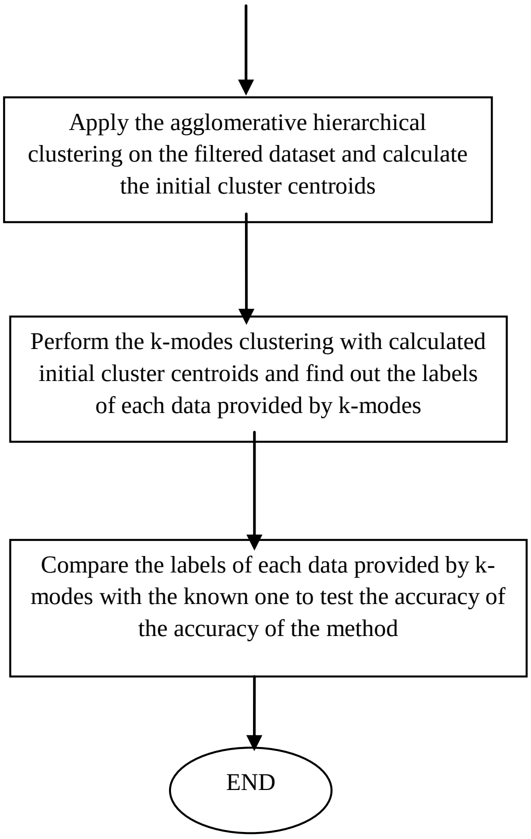

Apply the agglomerative hierarchical clustering on the filtered dataset and calculate the initial cluster centroids
Perform the k-modes clustering with calculated initial cluster centroids and find out the labels of each data provided by k-modes
Compare the labels of each data provided by k-modes with the known one to test the accuracy of the accuracy of the method
END

Fig 1: Flow Chart of Proposed Algorithm

5. EXPERIMENT RESULTS

To evaluate the efficiency of k-mode technique conduct experiment on four categorical attributes dataset. Here soybean dataset, lung cancer dataset, zoo dataset, dermatology dataset, congressional dataset. Every dataset be taken from UCI machine store. A select these dataset to experiment this technique since all attributes of the can be treated as categorical. Tables 1 summarize the categorical dataset that use in experiment.

5.1. Performance Parameter

Consider measures: accuracy (AC), precision (PE), recall (RE). Object in an ith cluster are assumed to be classified each correctly or incorrectly through respect to a known class of object. Let the number of suitably classified object be a1, let the numeral of incorrectly classified object be b1, and let the number of object in a given class but not in a cluster be c1. The clustering accuracy, recall, and precision and F-measure are defined as follows.

5.2. Results Evaluation

Perform statistical test to see whether the results produced by k-modes algorithms for categorical data. In paper algorithm use attributes on information gain basis provide better accuracy. Results of AC, PC, RC, F-measure, are showing here table-1-5.

$$\text{Average Accuracy} = \frac{\sum_{i=1}^{k} e_i}{N}$$

$$\text{Where, Precision} = \frac{\sum_{i=1}^{k}(\frac{e_i}{e_i+b_i})}{K}$$

$$\text{Recall} = \frac{\sum_{i=1}^{k}(\frac{e_i}{e_i+c_i})}{K}$$

$$\text{F-measure} = \frac{(\beta^2+1)*Precision*Recall}{\beta^2 Precision+Recall}$$

Table1: Soybean Dataset

	TPR	TNR	FPR	FNR	Accuracy	Precision	Recall	F-measure
Vanilla	0.8693	0.8898	0.162	0.0986	0.8844	0.6611	0.8982	0.7616
Prominent	0.9404	0.9073	0.0766	0.1122	0.9161	0.817	0.8934	0.8535
Significance	0.9516	0.9352	0.0958	0.0971	0.9396	0.7832	0.9075	0.8407
Paper	0.9602	0.94	0.0838	0.055	0.9454	0.8064	0.9458	0.8705
Proposed	0.957	0.9256	0.0447	0.0314	0.934	0.8862	0.9683	0.9254

Table2: Lung Cancer

	TPR	TNR	FPR	FNR	Accuracy	Precision	Recall	F-measure
Vanilla	0.0185	0.012	0.9353	0.0159	0.0159	0.0283	0.0189	0.0227
Prominent	0.3178	0.3537	0.6535	0.3323	0.3323	0.4174	0.3131	0.3578
Significance	0.4243	0.4322	0.5225	0.4275	0.4275	0.5447	0.425	0.4774
Paper	0.5097	0.5105	0.4525	0.51	0.51	0.624	0.5136	0.5635
Proposed	0.554	0.5721	0.4139	0.5613	0.5613	0.6635	0.5673	0.6117

Table 3: Zoo Dataset

	TPR	TNR	FPR	FNR	Accuracy	Precision	Recall	F-measure
Vanilla	0.7473	0.7222	0.2279	0.2865	0.7282	0.5097	0.7229	0.5978
Prominent	0.8416	0.8247	0.2075	0.1915	0.8288	0.5624	0.8146	0.6654
Significance	0.8326	0.8651	0.1633	0.1366	0.8573	0.6178	0.859	0.7187
Paper	0.8733	0.8966	0.1034	0.1062	0.891	0.728	0.8916	0.8015
Proposed	0.8704	0.9052	0.1286	0.0913	0.8968	0.6821	0.905	0.7779

Table4: Dermatology Dataset

	TPR	TNR	FPR	FNR	Accuracy	Precision	Recall	F-measure
Vanilla	0.5563	0.5532	0.4467	0.4487	0.5538	0.2388	0.5535	0.3337
Prominent	0.7111	0.6543	0.3438	0.3446	0.6658	0.3426	0.6736	0.4542
Significance	0.761	0.7124	0.2827	0.2318	0.7222	0.4041	0.7666	0.5292
Paper	0.7884	0.7638	0.2294	0.2122	0.7687	0.464	0.7879	0.5841
Proposed	0.8207	0.7783	0.1805	0.1899	0.7868	0.5339	0.8121	0.6442

Table 5 Congressional vote Dataset

	TPR	TNR	FPR	FNR	Accuracy	Precision	Recall	F-measure
Vanilla	0.7028	0.744	0.339	0.2938	0.7251	0.6384	0.7052	0.142647
Prominent	0.7536	0.8053	0.2153	0.236	0.7815	0.7487	0.7615	0.054856
Significance	0.8439	0.8051	0.1938	0.1569	0.823	0.7876	0.8432	0.018336
Paper	0.8513	0.8515	0.1501	0.1316	0.8514	0.8284	0.8661	0.077998
Proposed	0.8612	0.8495	0.1323	0.127	0.8549	0.8472	0.8715	0.377878

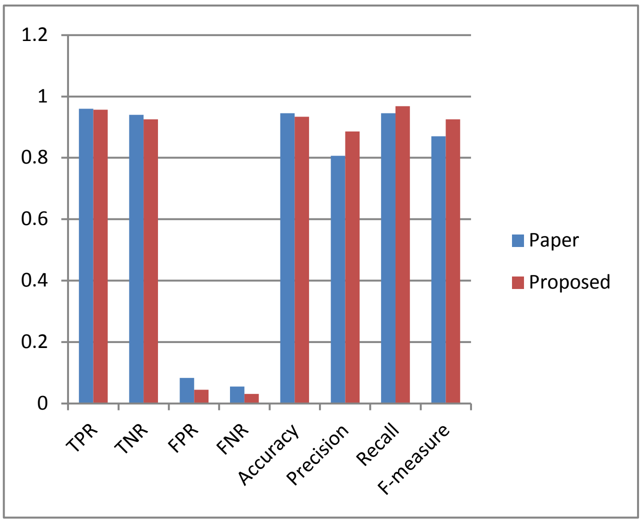


Fig.1: Soybean Dataset

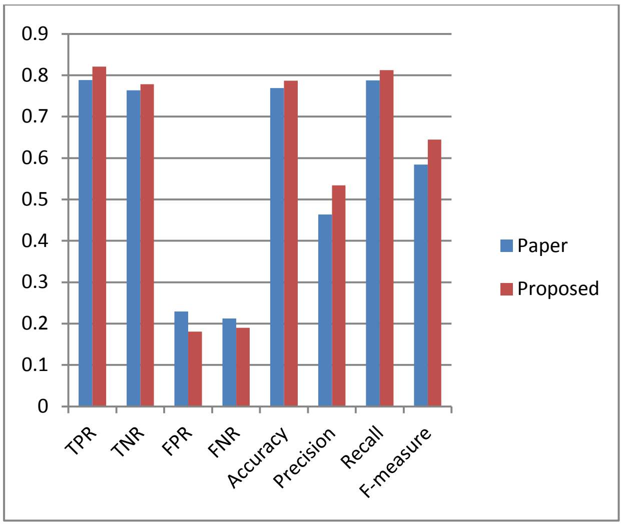


Fig.2: Dermatology Dataset

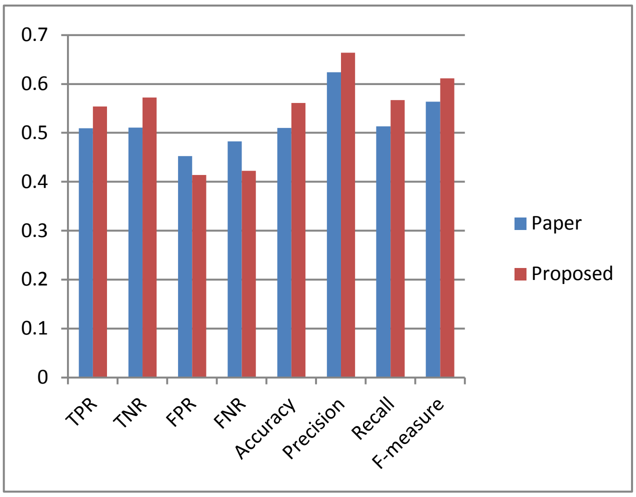


Fig.3: Lung Cancer Dataset

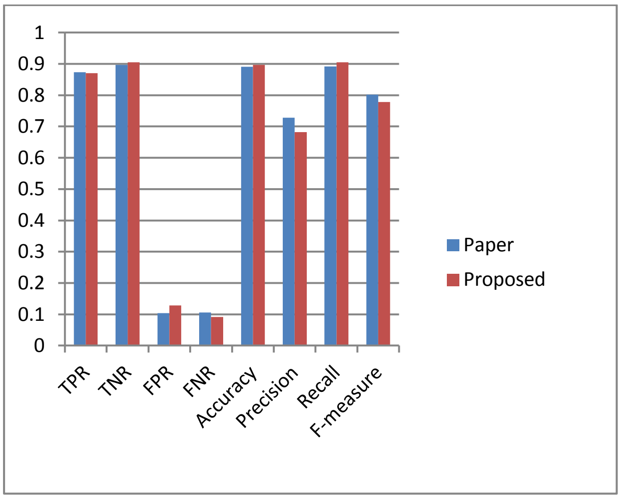


Fig.4: Zoo Dataset

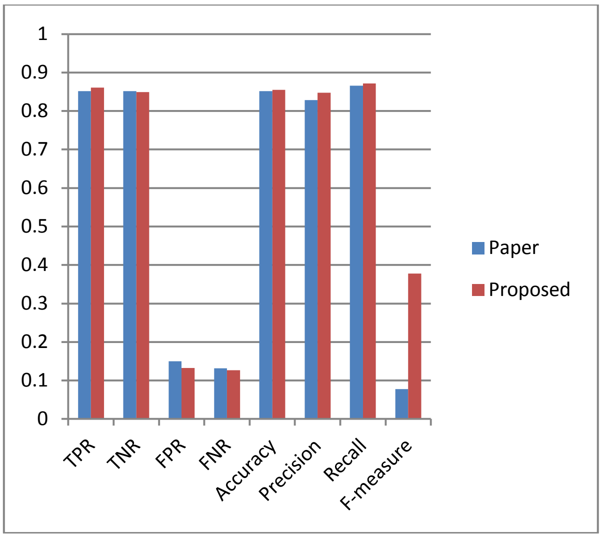


Fig5: Congressional Vote Dataset

6. CONCLUSION

K-mode cluster algorithms is working to partition categorical data into pre-define k cluster, however the cluster in result essentially depend on the selection of random initial cluster centre, that can cause non-repeatable result as well as generate improper cluster structures. There is several possible research directions may be worked on in the future to further extend and enhance the work presented in this paper. In propose algorithm for categorical data technique. This technique used as seed to k-modes cluster algorithm, improve result, computation time.

7. REFERENCES

[1] Shehroz S. Khan, Amir Ahmad, Cluster center initialization algorithm for K-modes clustering, Expert Systems with Applications 40 (2013) 7444–7456.

[2] Joel Luis Carbonera, Mara Abel, An entropy-based subspace clustering algorithm for categorical data, 2014 IEEE 26th International Conference on Tools with Artificial Intelligence.

[3] S.Aranganayagi, K.ThangaveI, S.Sujatha, New Distance Measure based on the Domain for Categorical Data.

[4] Zhexue Huang, A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining.

[5] J.L.Carbonera, and M. Abel, "Categorical data clustering: a correlation-based approach for unsupervised attribute weighting," in *Proceedings of ICTAI 2014*.

[6] D. Bacciu, I.H. Jarman, T.A. Etchells and P.J.G. Lisboa. Patient Stratification with competing risks by multivariate Fisher distance. International Joint Conference on Neural Networks, 14-19th June 2009, pp 213-220.

[7] Rishi Syal, Dr V.Vijaya Kumar, Innovative Modified K-Mode Clustering Algorithm www.ijera.com Vol. 2, Issue 4, July-August 2012, pp.390-398

[8] Liang Bai, Jiye Liang, Chuangyin Dang , Fuyuan Cao, A cluster centers initialization method for clustering categorical data, Expert Systems with Applications 39 (2012) 8022–8029

[9] Guo Tao, Ding Xingu, Li Yefeng, Parallel *k*-modes Algorithm based on MapReduce.