Comparative Analysis of Genetic k-means and Fuzzy k-modes Approach for Clustering Tweets

Akash Shrivastava
Computer Science Engineering Department
DIT University, Dehradun

M. L. Garg, PhD
Computer Science Engineering Department
DIT University, Dehradun

ABSTRACT

Social media plays a key role in decision making process. The challenge with the social media data is that it is highly categorical in nature. The classification of dataset into some prescribed format is really a tedious task. In this paper, the existing two clustering approaches is being experimented on the twitter datasets i.e. tweets to justify the fact that clustering is really an approach essentially utilized to classify the categorical dataset. Genetic k-means and fuzzy k-modes algorithm is tested on the tweets. Results shown that genetic k-means performs better for tweets classification.

Keywords

Genetic k-means, Fuzzy k-modes

1. INTRODUCTION

Segmentation analysis or taxonomy analysis is a way to create set of objects in databases into homogeneous groups or clusters is a fundamental operation in data mining. It is useful in a number of tasks, such as classification (unsupervised) aggregation and dissection [1] and has four design phases: data representation, modeling, optimization and validation [2]. The problem of clustering in general deals with partitioning a data set consisting of n points embedded in m-dimensional space into k distinct set of clusters, such that the data points within the same cluster are more similar to each other than to data points in other clusters. The three sub-problems [3] addressed by the clustering process are:

(i) Implementing an efficient algorithm to discover the clusters of most similar elements in an unsupervised way

(ii) Derive a description that can characterize the elements of a cluster in an epigrammatic manner and

(iii) Defining a similarity measure to judge the similarity (or distance) between different elements

Traditional clustering algorithms used Euclidean and Manhattan distance [7] measure to judge the similarity of two data elements [4] [5]. This works well when the defining attributes of a data set are purely numeric in nature. However, Euclidean and Manhattan distance measure fails to capture the similarity of data elements when attributes are categorical or mixed. Increasingly, the data mining community is besiege with a large collection of categorical data [6] Like those collected from banks, or health sector, web-log data and biological sequence data.

Banking sector or health sector data are primarily mixed data containing numeric attributes like age, salary, etc. and categorical attributes like sex, smoking or non-smoking, etc. In order to handle mixed or categorical data, some of the strategies that have been used are as under: [2]

(1) The data representation phase predetermines what kind of cluster structures can be identified in the data.

(2) On the basis of data representation, the modeling phase defines the notion of clusters and the criteria that separate the desired group structures from unpropitious ones.

(3) A quality measure which can be either optimized during the search for hidden structures in the data is produced.

(4) The validation phase is necessary to validate the results produced by the clustering algorithm.

(5) Another approach has been discretizing the data in the columns to enable the use of the algorithm to produce a mining model. It is the process of putting the values into buckets that there will be possible state

2. RELATED WORK

In general, clustering algorithm is classified into two ways: hard clustering algorithm and fuzzy clustering algorithm. In the hard clustering, each object belongs to one and only one clusters and in fuzzy clustering each object is allowed to have membership functions to all clusters. In terms of clustering we are interested in Genetic algorithms which can efficiently cluster large data sets containing mixed and categorical values because such data sets are frequently encountered in data mining application. Genetic algorithms (GA) were originally proposed by Holland [8].GA has been applied to many function optimization problems and is shown to be good in finding best and near optimal solutions. Their robustness of search in large search spaces and their domain independent nature motivated their application in various fields like pattern recognition, machine learning, VLSI design etc. [9]. Krishna and Murty proposed a new clustering method called genetic k-means algorithm (GKA) [4], which hybridizes a genetic algorithm with the k-means algorithm and genetic fuzzy k-modes algorithm. This approach combines the robust nature of the genetic algorithm with the high performance of the k-means algorithm. Lu et al [11] introduced fast genetic K-means cluster technique (FGKA).

FGKA did several improvements over GKA including an efficient evaluation of the objective value Total Within-cluster Variation (TWCV), avoiding illegal string elimination overhead, and a simplification of the mutation operator. These improvements result that FGKA runs 20 times faster than GKA [4] but FGKA suffers from a potential disadvantage. The cost of calculating centroids and TWCV from score can be much more expensive because if the mutation probability is small, then the number of allele changes will be small.

To overcome from this problem of FGKA has been proposed an incremental genetic k-means algorithm (IGKA) [5].When the mutation probability is small IGKA inherits all the advantage of FGKA including the convergence to the global optimum, and outperforms FGKA. The main idea of IGKA is to calculate the objective value TWCV and to cluster centroids incrementally. IGKA performs well when mutation probability is smaller than some threshold value but not when mutation probability is larger than some threshold value.

Therefore, a hybrid genetic k-means algorithm (HGKA) is proposed. HGKA combines the benefits of FGKA and IGKA and did well in smaller and larger mutation probability.

Mathematically, a fuzzy clustering problem can be shown as an optimization problem [18]:

Min Y(X, Z) = $\sum_{i=1}^{k}\sum_{i=1}^{n} \mathrm{x}_{li}^{\alpha}\, e(\mathrm{z_l}, \mathrm{q_i})$

Such that

$0 \leq w_{li} \leq 1, 1 \leq l \leq k, 1 \leq i \leq n,$ (1)

$\sum_{l=1}^{k} w_{li} = 1, 1 \leq i \leq n,$ (2)

$0 < \sum_{i=1}^{n} w_{li} < n,\ 1 \leq l \leq k,$ (3)

Where n is the number of objects in the data set under consideration, k is the number of clusters, E=$\{j1, j2, \ldots j3\}$ is a set of n objects each of which is described by d attributes, Z= $\{z_{1,} z_{2,} \ldots z_k\}$ is a set of k clusters centers ,X=(w_{li}) is a k x n fuzzy membership matrix, α € [1,∞] is a weighting exponent , and e ($z_{i,}$ q_i) is a certain distance measure between cluster center z_l and the object e_i .

A well-known fuzzy clustering algorithm is the fuzzy k-Means algorithm due to [12] [10]. The fuzzy k-Means algorithm starts with an initial value of X and then repeatedly iterates between estimating cluster centers Z given X and estimating the membership matrix X given Z until two successive values of X or Z are equal. Since the fuzzy k-Means algorithm works only on numeric values, a fuzzy k-Modes algorithm [13] has been developed for the purpose of clustering categorical data sets. A known problem associated with both the fuzzy k-Means algorithm and the fuzzy k-Modes algorithm is that they may only stop at local optima of the optimization problem, since the function F (X,Z) is non-convex in general [14].

To find a global solution of the optimization problem, GA [15] and tabu-search (TS) based techniques [16] are applied. The genetic k-means algorithm [17] for example, integrates the k-means algorithm and the genetic algorithm so as to find the globally optimal solutions. In order to find the globally optimal solution for the fuzzy k-Modes algorithm, Ng and Wong introduced tabu-search based fuzzy k-Modes algorithm [14].

3. GENETIC K-MEANS CLUSTERING ALGORITHM FOR MIXED NUMERIC AND CATEGORICAL DATA

In this section we will describe proposed genetic k-means clustering algorithm for mixed numeric and categorical data.

3.1 Objective Function

The data for clustering consists of N genes and their corresponding N patterns. Each pattern is a

Vector of Z dimensions recording the expression levels of the genes under AGKA each of the Z

Monitored conditions or at each of the Z time points. The goal of AGKA algorithm is to partition the N patterns into user-defined K groups. The Total Within-Cluster Variation (TWCV) is used to minimize for clustering in GKA, FGKA and IGKA. It can define as Eq. (4).Let *X1, X2,…, XN* be the *N* patterns, and *Xnd* denotes the *zth* feature of pattern *Xn* (*n*=1…*N*).

Each partitioning is represented by a string, a sequence of numbers *a1…aN*, where *an* takes a value from {1, 2,…, K} representing the cluster number that pattern *Xn* belongs to. Let *Gk* denote the *kth* cluster and *Dk* denote the number of patterns in *Gk*.

The Total Within-Cluster Variation (*TWCV*) is define as [5]

TWCV=$\sum_{n=1}^{N} \quad \sum_{z=1}^{Z} X_{nz}^{2}$ - $\sum_{k=1}^{K} \frac{1}{D_k} \sum_{z=1}^{Z} SF_{kz}^{2}$ (4)

the numeric attribute. Here we are using modified cost function specified in Eq. (5s), which is to be minimized for clustering mixed data sets has two distinct components, one for handling numeric attributes and another for handling categorical attributes. The cost function can define for clustering mixed data sets with n data objects and m attributes (mr numeric attributes, mc categorical attributes, m = mr + mc) as

Ψ=$\sum_{i=1}^{n} V(z_{i,}C_j)$ (5)

Where V(Zi, Cj) is the distance of a data object di from the closest cluster center Cj. V(Zi, Cj) is defined as Eq.(3)

V$(z_i C_j)$=$\sum_{u=1}^{m_{r-1}} w_u(z_{iu}^{r} - C_{ju}^{r}))^2$ + $\sum_{u=1}^{m_{c-1}} \Omega(z_{iu,}^{c} C_{ju}^{c})^2$ (6)

Where $\sum_{u=1}^{m_{r-1}} w_u(z_{iu}^{r} - C_{ju}^{r}))^2$ denotes the distance of object z_i from its closest cluster center $C_{j,}$ for numeric attributes only, w_u denotes the significance of the uth numeric attribute, which is to be computed from the data set $\sum_{u=1}^{m_{c-1}} \Omega(z_{iu,}^{c} C_{ju}^{c})^2$ denotes the distance between data object z_i and its closest cluster center C_j in terms of categorical attributes only.

3.2 The Selection Operator

Proportional selection is used for the selection operator in which, the population of the next Generation is determined by *N* independent random experiments. Each experiment randomly selects a solution from the current population {X1, X2 ,…, *Xz*} according to the probability distribution {*p1, p2,…, pz*} defined by [5].

$p_n = \frac{F(X_n)}{\sum_{n=1}^{N} F(X_n)}$(n=1,2,.......N) (7)

F (Xn) denotes the fitness value of solution *Xn*. In our context, the objective is to minimize the V which can obtain from equation (6). Therefore, solutions with smaller *V*x should have higher probabilities for survival and should be assigned with greater fitness values. In addition, illegal

strings are less desirable and should have lower probabilities for survival, and thus should be assigned with lower fitness values. We define F (Xn) as follows,

F(X_n)=$\begin{cases} 1.5 * V_{max} - V(X_n), if\ X_n\ is\ leagal \\ e(X_n) * F_{min}, otherwise \end{cases}$ (8)

Where *Vmax* is the maximum *V* that has been encountered till the present generation, *Fmin* is the smallest fitness value of the legal strings in the current population if they exist, otherwise *Fmin* is defined as 1

3.3 The Mutation Operator

The mutation operator performs the function of shaking the algorithm out of a local optimum, and moving it towards the global optimum. During mutation, we replace *an* by *an*' for *n=1,…,N* simultaneously. *an*' is a cluster number randomly selected from {1,…,K} with the probability distribution {*p1,p2,…,pK*} defined by

$$p_k = \frac{1.5 * d_{max}(X_n) - d\left(X_{n,C_k}\right) + 0.5}{\sum_{k=1}^{K}(1.5 * d_{\max}(X_n) - d\left(X_{n,Ck}\right) + 0.5)} \qquad (9)$$

where *d(Xn,ck)* is the Euclidean distance between pattern *Xn* and the centroid *ck* of the *k*th cluster, and $d_{max}(X_n) = \{d\left(X_{n,C_k}\right)\}$. If the *k*th cluster is empty, then *d (Xn,ck)* is defined as 0. The bias 0.5 is introduced to avoid divide by-zero error in the case that all patterns are equal and are assigned to the same cluster in the given solution.

3.4. The k-Means Operator

In order to speed up the convergence process, one step of the classical K-means algorithm, which we call *K-means operator (KMO)* is introduced. Given a solution that is encoded by *a1…aN*, we replace *an* by *an*' for *n=1,….N* simultaneously, where *an*' is the number of the cluster whose centroid is closest to *Xn* in Euclidean distance.

To account for illegal strings, we define *d(Xn, ck)* = +∞ if the *k*th cluster is empty. This definition is different from section 3.2, in which we defined *d (Xn, ck)* = 0 if the *k*th cluster is empty. The motivation for this new definition here is that we want to avoid reassigning *all* patterns to empty clusters. Therefore, illegal string will remain illegal after the application of KMO.

4. FUZZY K-MODES

To describe the fuzzy k-Modes algorithm [13], let us begin with some notations.

Let E={Q_1,$Q_{2,}$.........Q_3} be a categorical data set with n objects each of which is described by e categorical attributes $A_{1,}$$A_2$,........$A_{3...........}$$A_{e..}$ Attribute A_j (1 ≤ j ≤ e) has n_j categories, i.e. DOM(A_j)={a_{j1},a_{j2}.......a_{jn}}.let the cluster centers be represented by z_l = ($z_{l1,}$$z_{l2,........}$$z_{le}$) for 1≤ l ≤ k,

Where k is the number of clusters. The simple matching distance measure between m and p in E is defined as

$$e_c(m,p)=\sum_{j=1}^{e} \delta(m_j,p_j), \qquad (10)$$

where m_j and p_j are the jth components of m and p, respectively, and

$$\boldsymbol{\delta}(m_j,p_j) = \begin{cases} \mathbf{0} \text{ if } m_j = p_j \\ \mathbf{1} \text{ if otherwise} \end{cases}$$

then the objective of the fuzzy k-modes clustering is to find X and Z that minimize

$$Y_c(X,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n} x_{li}^{\alpha}\, e_c\left(q_i,z_l\right) \qquad (11)$$

Subject to 1, 2, 3 where α > 1 is the weighting component,e_c(..) is defined in Eq(10)

X = (x_{li}) is the k x n fuzzy membership matrix, and Z ={$z_{1,}$$z_{2....}$$z_k$} is the set of clusters centers. Note that α = 1 gives the hard k-modes clustering i.e. the k-modes algorithm.

To update the cluster centers given the estimate of X, Huang and Ng [13] proved the following theorem

Theorem 1. The quantity Y_c(X, Z) defined in Eq. 11 is minimized if and only if $z_{lj} = a_{jr}$ ε DOM(A_j) where

$$r = \arg \overset{max}{1 \le t \le n_j} \sum_{i,q_{ij}=a_{jt}} x_{li}^{\alpha}$$

$$\text{i.e., } \sum_{i,q_{ij}=a_{jr}} x_{li}^{\alpha} \ge \sum_{i,q_{ij}=a_{jt}} x_{li}^{\alpha}, \qquad 1 \le t \le n_j$$

for 1 ≤ j ≤ d and 1 ≤ l ≤ k.

To update the fuzzy membership matrix X given the estimate of Z, the following theorem is also presented in [13].

Theorem 2 let Z= $\{z_1, z_{2,....}z_k\}$ be fixed, then the fuzzy membership matrix X which minimizes the quantity Y_c (X,Z) defined in eq (11) subject to 1,2,3 is given by

$$X_{li} = \begin{cases} 1 \quad if\ q_{i=}z_l\ ; \\ 0 \quad if\ q_{i=}z_{h\ ,h \ne\ l}\ ; \\ \dfrac{1}{\sum_{h=1}^{k}\left[\frac{d(q_i,z_l)}{d(q_i,z_h)}\right]^{\frac{1}{\alpha-1}}} \end{cases} \quad \text{if otherwise, } 1 \le l \le l \le k\ ,\ l \le i \le n$$

Based on the two theorems described above, the fuzzy k-modes algorithm can be implemented recursively (see algorithm 1.)

Algorithm 1 fuzzy k-modes algorithm, n is the maximum number of iterations

```
choose starting point $a_0$ ε $N^{mk}$;
Find $X_0$ $such\ that\ the\ cost\ function\ Y(X_0Z_0)\ is\ minimized$
for t= 1 to n do
    Find $Z_1$such that the cost function $Y(X_0Z_1) is\ minimized;$
    if Y($X_0Z_0$) = Y($X_0Z_1$) then
        stop;
    else
        Find $X_1$ such that the cost function Y($X_1Z_1$) is minimized;
        if Y($X_1Z_1$) = Y($X_0Z_1$) then
            stop;
        else
            $X_0 \Leftarrow X_1$ ;
        endif
    endif
end for
```

5. EXPERIMENT AND RESULTS

The algorithms defined in the paper have been implemented on twitter dataset. The twitter dataset is highly categorical dataset in the context of dimensions. It contains high dimensional attributes and it has no structure. The clustering algorithms are used to apply over numerical dataset. But here, it has been experimented over categorical dataset which is twitter in this case. The real time streaming tweets have been retrieved by applying the code developed in python script. The experiment runs on the machine having on Intel core i7, 3.9 GHz, 64GB RAM computer.

The results derived by comparing the values of AC, PR and RE which has been defined in Yang's accuracy measure [18].

The AC , PR and RE values defined as follows:

The definitions of accuracy (AC), precision (PR), and recall (RE) are given as follows [18]:

$$AC= \frac{\sum_{i=1}^{k} a_i}{n}$$

$$PR= \frac{\sum_{i=1}^{k}\frac{a_i}{a_i+b_i}}{k}$$

$$RE= \frac{\sum_{i=1}^{k}\frac{a_i}{a_i+c_i}}{k}$$

Where,

$a_{i\ =}$ The number of data objects that are correctly allocated to class C_i,

b_i = The number of data objects that are incorrectly allocated to class C_i,

c_i = The number of data objects that are incorrectly denied from class C_i,

k = The total number of class contained in a dataset, and

n = The total number of data objects in a dataset,

The higher the values of these parameters the better clustering results it considers.

Table 1 The AC of the three algorithms on the Twitter dataset

Algorithms	AC		
	Best	Avg	Std
Fuzzy k-modes	0.8231	0.8124	0.0107
Genetic k-means	0.8014	0.7697	0.0040

Table 2 The PR of the three algorithms on the Twitter dataset

Algorithms	PR		
	Best	Avg.	Std.
Fuzzy k-modes	0.8432	0.8329	0.0082
Genetic k-means	0.7213	0.7863	0.0071

Table 3 The RE of the three algorithms on the Twitter dataset

Algorithms	RE		
	Best	Avg.	Std.
Fuzzy k-modes	0.7823	0.7213	0.0032
Genetic k-means	0.7235	0.6921	0.0063

6. CONCLUSION

The experiment carried out on the twitter datasets which is highly categorical in nature. The results show that clustering approaches better can be utilized to cluster the unstructured datasets in the format which may be used to implement for decision-making process. Twitter data can be proven as an asset for so many organizations. The genetic k-means algorithm found to be better in comparison with the fuzzy k-modes algorithm to cluster categorical datasets like tweets. In future, hybrid approaches may be developed through the inclusion of existing evolutionary clustering algorithm.

7. REFERENCES

[1] G. Gan, Z. Yang, and J. Wu (2005), A Genetic k-Modes Algorithm for Clustering for Categorical Data, ADMA , LNAI 3584, pp. 195–202.

[2] G. Gan, J. Wu, Z. Yang A genetic fuzzy k-Modes algorithm for clustering categorical data * Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada M3J 1P3.

[3] A. Ahmad and L. Dey, (2007), A k-mean clustering algorithm for mixed numeric and categorical data', Data and Knowledge Engineering Elsevier Publication, vol. 63, pp 503-527.

[4] J. Z. Haung, M. K. Ng, H. Rong, Z. Li (2005) Automated variable weighting in k-mean type clustering, IEEE Transaction on PAMI 27(5).

[5] K. Krishna and M. Murty (1999), 'Genetic K-Means Algorithm'**,** IEEE Transactions on Systems, Man, and Cybernetics vol. 29, NO. 3, pp. 433-439.

[6] Jain, M. Murty and P. Flynn (1999), 'Data clustering: A review', ACM Computing Survey., vol.31, no. 3, pp. 264–323.

[7] Hui Ding. Goce Trajcevski Peter Scheuermann xiaoyue Wang. Eamonn Keogh. Proceedings of the VLDB endowment VLDB endowment, Querying and mining of time series data: vol.1, issue 2, august 2008.

[8] A.Ahmad and L.Dey,(2007),A K-means clustering algorithm for mixed and categorical data,Data and Knowledge Engineering Elsevier Publication,vol.63,pp.503-527.

[9] Dharmendra K Roy and Lokesh K Sharma,Genetic K-means Clustering Algorithm For Mixed and Categorical data,Department of Information Technology and MCA,Rungta college Of Engineering and technology,Bhilai(CG)-India,International journal of Artifical Intelligence & Applications(IJAIA) vol 1,no 2,april 2010.

[10] E. R. Ruspini, "A new approach to clustering," *Inform. Contr.*, vol. 19, pp. 22–32, 1969.

[11] S.Guha,R.Rastogi and K.Shim (2000). Rock: A robust clustering algorithm for categorical attributes, Information System,vol 25 no 5,pp345-366.

[12] Bezdek, J. (1974). Fuzzy mathematics in pattern classification, Ph.D. thesis, Ithaca, NY: Cornell University (April).

[13] Huang, Z., & Ng, M. (1999). A fuzzy k-modes algorithm for clustering categorical data. IEEE Transactions on Fuzzy Systems, 7(4), 446–452.

[14] Ng, M., & Wong, J. (2002). Clustering categorical data sets using tabu search techniques. Pattern Recognition, 35(12), 2783–2790.

[15] Davis, L. (1991). Handbook of genetic algorithms. New York, USA: van Nostrand Reinhold.

[16] Glover, F., & Laguna, M. (1997). Tabu search.Boston: Kluwer Academic Publishers.

[17] Krishna, K., & Narasimha, M. (1999). Genetic k-means algorithm. IEEE Transactions on Systems, Man and Cybernetics, Part B, 29(3), 433–439.

[18] Dunn, J. (1974). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics, 3(3), 32–57.