

5-2002

# An iterative initial-points refinement algorithm for categorical data clustering

Ying Sun

Qiuming Zhu

*University of Nebraska at Omaha*, [qzhu@unomaha.edu](mailto:qzhu@unomaha.edu)

Zhengxin Chen

*University of Nebraska at Omaha*, [zchen@unomaha.edu](mailto:zchen@unomaha.edu)

Follow this and additional works at: <https://digitalcommons.unomaha.edu/compscifacpub>

 Part of the [Computer Sciences Commons](#)

---

## Recommended Citation

Sun, Ying; Zhu, Qiuming; and Chen, Zhengxin, "An iterative initial-points refinement algorithm for categorical data clustering" (2002). *Computer Science Faculty Publications*. 26.

<https://digitalcommons.unomaha.edu/compscifacpub/26>

This Article is brought to you for free and open access by the Department of Computer Science at DigitalCommons@UNO. It has been accepted for inclusion in Computer Science Faculty Publications by an authorized administrator of DigitalCommons@UNO. For more information, please contact [unodigitalcommons@unomaha.edu](mailto:unodigitalcommons@unomaha.edu).



2           An iterative initial-points refinement algorithm for  
3                           categorical data clustering

4           Ying Sun, Qiuming Zhu \*, Zhengain Chen

5           *Department of Computer Science, Digital Imaging and Computer Vision Laboratory,*  
6           *University of Nebraska at Omaha, Omaha, NE 68182-0050, USA*

---

8 **Abstract**

9     The original  $k$ -means clustering algorithm is designed to work primarily on numeric data sets. This prohibits the  
10 algorithm from being directly applied to categorical data clustering in many data mining applications. The  $k$ -modes  
11 algorithm [Z. Huang, Clustering large data sets with mixed numeric and categorical value, in: Proceedings of the First  
12 Pacific Asia Knowledge Discovery and Data Mining Conference. World Scientific, Singapore, 1997, pp. 21–34] ex-  
13 tended the  $k$ -means paradigm to cluster categorical data by using a frequency-based method to update the cluster  
14 modes versus the  $k$ -means fashion of minimizing a numerically valued cost. However, as is the case with most data  
15 clustering algorithms, the algorithm requires a pre-setting or random selection of initial points (modes) of the clusters.  
16 The differences on the initial points often lead to considerable distinct cluster results. In this paper we present an ex-  
17 perimental study on applying Bradley and Fayyad’s iterative initial-point refinement algorithm to the  $k$ -modes clus-  
18 tering to improve the accurate and repetitiveness of the clustering results [cf. P. Bradley, U. Fayyad, Refining initial  
19 points for  $k$ -mean clustering, in: Proceedings of the 15th International Conference on Machine Learning, Morgan  
20 Kaufmann, Los Altos, CA, 1998]. Experiments show that the  $k$ -modes clustering algorithm using refined initial points  
21 leads to higher precision results much more reliably than the random selection method without refinement, thus making  
22 the refinement process applicable to many data mining applications with categorical data. © 2001 Published by  
23 Elsevier Science B.V.

24 **Keywords:** Data clustering; Pattern classification; Refinement algorithm; Data mining

---

25 **1. Introduction**

26     Partitioning a set of objects in a data collection  
27 of multiple attributes into homogeneous groups

(clusters) of certain intra-relations is a fundamen- 28  
tal operation in data mining. The most distinct 29  
characteristic of clustering operation in data min- 30  
ing is that the data sets often contain both numeric 31  
and categorical (symbolic) attribute values. This 32  
requires the clustering algorithms to be capable of 33  
dealing with the complexity of the intra- and inter- 34  
relations of the data sets expressed in different 35

---

\* Corresponding author. Tel.: +1-402-554-3685; fax: +1-402-554-3400.

*E-mail address:* zhuq@unomaha.edu (Q. Zhu).

36 types of the attributes, no matter numeric or cat-  
37 egorical (Michalski et al., 1998).

38 Among the clustering algorithms that have been  
39 developed, the  $k$ -means algorithm is the most  
40 popular one (Jain and Dubes, 1988). Many other  
41 clustering algorithms were derived from it, such as  
42 the fuzzy  $k$ -means algorithm, the ISODATA, the  
43  $k$ -modes algorithm (Huang, 1998), etc. The  $k$ -  
44 means algorithm is well known for its efficiency in  
45 clustering large data sets (MacQueen, 1967; An-  
46 derberg, 1973). However, the original  $k$ -means al-  
47 gorithm works only on numeric data because it  
48 aims at minimizing a cost function that is numer-  
49 ically measured. This prohibits the  $k$ -means algo-  
50 rithm from being directly used in applications  
51 where categorical data are involved, such as the  
52 data mining applications.

53 Work on clustering data with categorical attri-  
54 butes has been done by several researchers. Ra-  
55 lambondrainy (1995) presented an approach by  
56 using the  $k$ -means algorithm to cluster categorical  
57 data. The approach is to convert multiple category  
58 attributes into binary attributes (using 0 and 1 to  
59 represent either a category absent or present) and  
60 to treat the binary attributes as numeric in the  $k$ -  
61 means algorithm. The main drawback of the ap-  
62 proach is that the cluster means, given by values  
63 between 0 and 1, often do not indicate the exact  
64 characteristics of the clusters. Gower and Diday  
65 (1991) used a similarity coefficient and other dis-  
66 similarity measures to process data with categori-  
67 cal attributes. However, the quadratic  
68 computational cost makes them unacceptable for  
69 clustering large data sets.

70 Conceptual clustering algorithms developed in  
71 machine learning were able to cluster data sets  
72 with categorical values (Michalski and Stepp,  
73 1983) and also produce conceptual descriptions of  
74 the clusters (Lebowitz, 1987; Fisher, 1987). Unlike  
75 statistical clustering methods, the algorithms are  
76 based on a search for objects, which carry the same  
77 or similar concepts. Therefore, their efficiency re-  
78 lies on good search strategies. For problems in  
79 data mining that often involve many concepts and  
80 very large object spaces, the concept-based search  
81 methods can become a potential handicap for  
82 these applications.

The  $k$ -modes algorithm (Huang, 1997) extends 83  
the  $k$ -means paradigm to cluster categorical data 84  
by using (1) a simple matching dissimilarity mea- 85  
sure for categorical objects, (2) modes instead of 86  
means for clusters, and (3) a frequency-based 87  
method to update modes in the  $k$ -means fashion to 88  
minimize the clustering cost function of clustering. 89  
Because the  $k$ -modes algorithm uses the same 90  
clustering process as  $k$ -means, it preserves the ef- 91  
ficiency of the  $k$ -means algorithm, which is highly 92  
desirable for data mining. A similar work that 93  
aims to cluster large data sets is the CLARA 94  
(abbreviated from Clustering LARGE Application) 95  
algorithm (Kaufman and Rousseeuw, 1990). 96  
CLARA is a combination of a sampling procedure 97  
and the clustering program Partitioning Around 98  
Medoids (PAM). Given a set of objects  $X$  and the 99  
number of clusters  $k$ , PAM clusters  $X$  by finding  $k$  100  
medoids (representative objects of clusters) that 101  
can minimize the average dissimilarity of objects to 102  
their closest medoids. Ng and Han (1994) have 103  
analyzed that the computational complexity of 104  
PAM in a single iteration is  $O(k(n-k)^2)$  where  $n$  is 105  
the number of objects in  $X$ . Obviously, PAM is not 106  
efficient when clustering large data sets. That 107  
makes CLARA inefficient in clustering large data 108  
sets. 109

As for the traditional clustering algorithms, 110  
most of the above-mentioned categorical data 111  
clustering algorithms, including the  $k$ -modes al- 112  
gorithm, require a random selection or setting up 113  
of initial data points in addition to a known or 114  
estimated number of clusters (also called starting 115  
conditions), before iteratively mapping the data 116  
records to separate clusters. This leads to the 117  
problem that the clustering results are often de- 118  
pendent on the selection of the initial points re- 119  
gardless of what measurement metric is used for 120  
the similarity (distance) evaluation operation. 121  
That is, the clustering solution is very much sen- 122  
sitive to the initial-point choices. An inappropriate 123  
setting up of initial points would lead to some 124  
unacceptable clustering results. For example, a 125  
large percent of data samples might be crowded 126  
into one or a few clusters with other clusters 127  
having only a few scarce samples, leaving users 128  
questioning its reality. Moreover, the clustering 129  
results often cannot be repetitively generated, 130

131 causing problems in the validation of the cluster-  
132 ing results.

133 The intrinsic problem of initial-point selection  
134 in clustering algorithms and the computation cost  
135 of the categorical data clustering call for an ap-  
136 proach that provides a better organized initial  
137 setting for improving the performance of cluster-  
138 ing processes. Hopefully, the improved initial-  
139 point sets would let the clustering algorithm con-  
140 verge with the global optimal or close to the op-  
141 timal solution more accurately and repetitively.  
142 That is, the selection of initial data points fits more  
143 appropriately and consistently with the nature and  
144 underlying distributions of the data sample sets.

145 In this paper we present an experiment on ap-  
146 plying the iterative refinement algorithm to the  
147 setting of the initial points so as to map the cate-  
148 gorical data sets to clustering results that have  
149 better consistency rates. This paper is organized as  
150 follows. Section 2 discusses the basics of the  $k$ -  
151 modes algorithms (Huang, 1997). Section 3 de-  
152 scribes Bradley and Fayyad's initial-points refine-  
153 ment algorithm and its principle (cf. Bradley and  
154 Fayyad, 1998). Section 4 presents our experimental  
155 results in applying the initial-points refinement to  
156 the  $k$ -modes algorithm for clustering categorical  
157 data samples. Section 5 concludes the presentation.

## 158 2. The $k$ -modes algorithm for categorical data 159 clustering

160 Let  $S = \{X_1, X_2, \dots, X_n\}$  denote a set of  $n$  data  
161 objects, and  $X_i = [X_{i1}, X_{i2}, \dots, X_{id}]$ ,  $i = 1, 2, \dots, n$ ,  
162 be an object represented by  $d$  attribute values. Let  
163  $k$  be a positive integer. The objective of  $k$ -means  
164 clustering is to find a partition that divides object  
165 set  $S$  into  $k$  disjoint regions that meet certain cri-  
166 teria and constraints. For a given  $n$  and  $k$ , the  
167 number of possible partitions is definite but could  
168 be extremely large. A common way of solving it is  
169 to choose a clustering criterion that guides the  
170 search for an approximate solution. The most  
171 common criterion has been the minimization of  
172 the total distances of the data points to their  
173 cluster centers. Formulated as a mathematical  
174 programming problem  $P(W, Q)$ , the  $k$ -means  
175 clustering algorithm has been traditionally ex-

pressed as the following (Hartigan, 1975; Bo- 176  
browski and Bezdek, 1991): 177

$$\text{Minimize } P(W, Q) = \sum_{l=1}^k \sum_{i=1}^n w_{i,l} d(X_i, Q_l),$$

$$\text{Subject to } \sum_{l=1}^k w_{i,l} = 1, \quad 1 \leq i \leq n; \quad w_{i,l} \in \{0, 1\}, \\ 1 \leq i \leq n, \quad 1 \leq l \leq k,$$

where  $W$  is an  $n \times k$  partitioning matrix;  $Q =$  179  
 $\{Q_1, Q_2, \dots, Q_k\}$ , namely the  $k$ -means, is a set of 180  
objects in the same object domain;  $d(\cdot, \cdot)$  is the 181  
distance metric, e.g. a squared Euclidean as the 182  
most common one, between two objects. 183

Problem  $P(W, Q)$  is solvable by iteratively 184  
solving the following two sub-problems: 185

1. *Sub-problem  $P_1$* : Fix  $Q = \hat{Q}$  and solve the re- 186  
duced problem  $P(W, \hat{Q})$ . 187

2. *Sub-problem  $P_2$* : Fix  $W = \hat{W}$  and solve the re- 188  
duced problem  $P(\hat{W}, Q)$ . 189

Sub-problem  $P_1$  is solved by 190

$$w_{i,l} = 1 \quad \text{if } d(X_i, Q_l) \leq d(X_i, Q_t) \\ \text{for } 1 \leq t \leq k \quad \text{or} \quad w_{i,t} = 0 \quad \text{for } t \neq l;$$

and sub-problem  $P_2$  is solved by 192

$$q_{l,j} = \frac{\sum_{i=1}^n w_{i,l} X_{i,j}}{\sum_{i=1}^n w_{i,l}} \quad \text{for } 1 \leq l \leq k \quad \text{and} \quad 1 \leq j \leq m.$$

The basic algorithm to solve problem  $P(W, Q)$  194  
is given as follows: 195

1. Choose an initial  $Q^0$  and solve  $P(W, Q^0)$  to ob- 196  
tain  $W^0$ . Set  $t = 0$ . 197

2. Let  $\hat{W} = W^t$  and solve  $P(\hat{W}, Q)$  to obtain  $Q^{t+1}$ . 198  
If  $P(\hat{W}, Q^t) = P(\hat{W}, Q^{t+1})$ , output  $\hat{W}, Q^t$  and 199  
stop; otherwise, go to 3. 200

3. Let  $\hat{Q} = Q^{t+1}$  and solve  $P(W, \hat{Q})$  to obtain  $W^{t+1}$ . 201  
If  $P(W^t, \hat{Q}) = P(W^{t+1}, \hat{Q})$ , output  $W^t, \hat{Q}$  and 202  
stop; otherwise, let  $t = t + 1$  and go to 2. 203

The computational cost of the algorithm is 204  
 $O(Tkn)$ , where  $T$  is the number of iterations and  $n$  205  
the number of objects in the input data set. 206

In principle the formulation of problem  $P$  in the 207  
above is also valid for categorical and mixed-type 208  
data objects. The reason that the  $k$ -means algo- 209  
rithm cannot cluster categorical objects is its dis- 210  
similarity measure used to solve problem  $P_2$ . These 211

212 barriers can be removed by making the following  
 213 modifications:  
 214 1. Using a simple matching dissimilarity measure  
 215 for categorical objects.  
 216 2. Replacing means of clusters by modes.  
 217 3. Using a frequency-based method to find the  
 218 modes to solve problem  $P_2$ .  
 219 Let  $X, Y$  be two categorical objects described by  
 220  $m$  categorical attributes. The dissimilarity measure  
 221 between  $X$  and  $Y$  can be defined by the total  
 222 mismatches of the corresponding attribute cate-  
 223 gories of the two objects. This measure is often  
 224 referred to as simple matching (Kaufman and  
 225 Rousseeuw, 1990). Formally, we have

$$d_1(X, Y) = \sum_{j=1}^m \delta(x_j, y_j), \quad \text{where}$$

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j), \\ 1 & (x_j \neq y_j). \end{cases}$$

227 Let  $X$  be a set of categorical objects described  
 228 by categorical attributes,  $A_1, A_2, \dots, A_m$ , a mode of  
 229  $X = \{X_1, X_2, \dots, X_n\}$  is a vector  $Q = [q_1, q_2, \dots, q_m]$   
 230 that minimizes

$$D(X, Q) = \sum_{i=1}^n d_1(X_i, Q).$$

232 Here,  $Q$  is not necessarily an element of  $X$ . Let  $n_{c_{k,j}}$   
 233 be the number of objects having the  $k$ th category  
 234  $c_{k,j}$  in attribute  $A_j$ , and  $f_r(A_j = c_{k,j}|X) = (n_{c_{k,j}}/n)$   
 235 the relative frequency of category  $c_{k,j}$  in  $X$ . The  
 236 function  $D(X, Q)$  is minimized iff

$$f_r(A_j = q_j|X) \geq f_r(A_j = c_{k,j}|X) \quad \text{for } q_j \neq c_{k,j},$$

for all  $j = 1, \dots, m$ .

238 The above expression defines a way to find  $Q$   
 239 from a given  $X$ , and therefore is important because  
 240 it allows the  $k$ -means paradigm to be used to  
 241 cluster categorical data. The expression implies  
 242 that the mode of a data set  $X$  is not unique. For  
 243 example, the mode of set  $\{[a, b], [a, c], [c, b], [b, c]\}$   
 244 can be either  $[a, b]$  or  $[a, c]$ .

245 When the above is used as the dissimilarity  
 246 measure for categorical objects, the cost function  
 247 becomes

$$P(W, Q) = \sum_{l=1}^k \sum_{i=1}^n \sum_{j=1}^m w_{i,l} \delta(x_{i,j}, q_{l,j}), \quad \text{where}$$

$$w_{i,l} \in W \quad \text{and} \quad Q_l = [q_{l,1}, q_{l,2}, \dots, q_{l,m}] \in Q.$$

That is, to minimize the cost function, the  $k$ -modes  
 algorithm proceeds by: (1) using the simple  
 matching dissimilarity measure to solve the prob-  
 lem  $P_1$ , and (2) using modes for clusters instead of  
 means and selecting modes to solve the problem  
 $P_2$ .

### 3. Initial points refining algorithm for data cluster- ing

As mentioned above, both  $k$ -means and  $k$ -  
 modes algorithms draw an initial estimation (ap-  
 proximation model) of the clusters represented by  
 $Q^0$  from a randomly selected subset of the data  
 points in the constraint object space. The algo-  
 rithms then subsequently extend the model as-  
 ymptotically to the whole data set by gradually  
 adjusting the  $k$ -means or modes as more data  
 points are accumulatively included. The process is  
 iterated until a complete solution (which may be  
 optimal or sub-optimal) of a clustering model is  
 obtained.

The idea of applying a refinement procedure to  
 the initial-point selection for obtaining a better  
 approximation of the true clusters in the set-up  
 stage was proposed by Bradley et al. (1998). The  
 heuristics behind the idea is that every data cluster  
 has an underlying model (or distribution) that  
 governs the positioning of the data samples (Ah-  
 rens and Dieter, 1973). This underlying model  
 behaves on both large and small sets of data  
 samples, except that it is more precisely presented  
 in larger data sets and less precisely in smaller data  
 sets. If one can draw a sufficiently precise model  
 from the smaller data sets, then the model can be  
 used to describe, or guide the description of, the  
 larger data set. In clustering, it means that if one  
 can make a proper modeling on the subset of the  
 data samples, then this model can be used to  
 quickly and accurately derive the underlying clus-  
 ters of the larger (or the whole) data set.

288 One practical problem of a simple sub-sampling  
289 approach is that severely sub-sampling the data  
290 will naturally bias the samples to representatives  
291 “near” the tails (or edges) of the data sets, while it  
292 makes nonsense to sub-sample a sufficiently large  
293 subset that is close to the actual data set. Fig. 1  
294 shows a data set drawn from a mixture of two  
295 Gaussian models (clusters) in 2-D with means at  
296 [1.5, 1.5] and [4, 4], respectively. A small sub-sample  
297 set is shown in Fig. 2, which was expected to  
298 provide information on the models of the joint  
299 probability density function of the original data set  
300 of Fig. 1. Each of the points on the Fig. 2 may be

thought of as a “guest” at the possible location of 301  
a model in the underlying distribution. It is seen 302  
that the sub-sample set exhibits certain “expected” 303  
behavior of the original data set. Worthy of note 304  
here is that the sub-sampling points are fairly 305  
spread out over the distribution region of the 306  
original data set. This observation indicates that 307  
the solutions obtained by clustering over a small 308  
sub-sample may not provide good initial estimates 309  
of the true means, or centroids, in the data. The 310  
simple sub-sampling method often produces noisy 311  
estimates due to single small sub-samples, especially 312  
in skewed distributions and high dimensions. 313

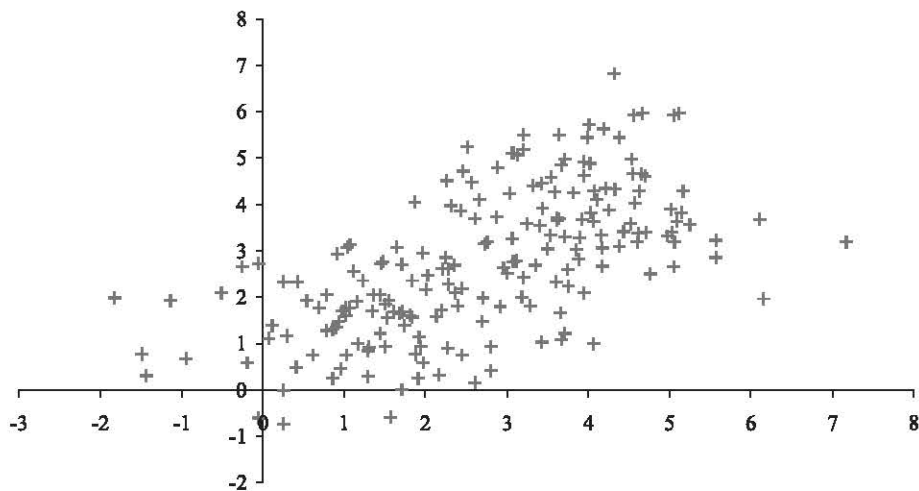


Fig. 1. Samples of Gaussian distribution in 2-D.

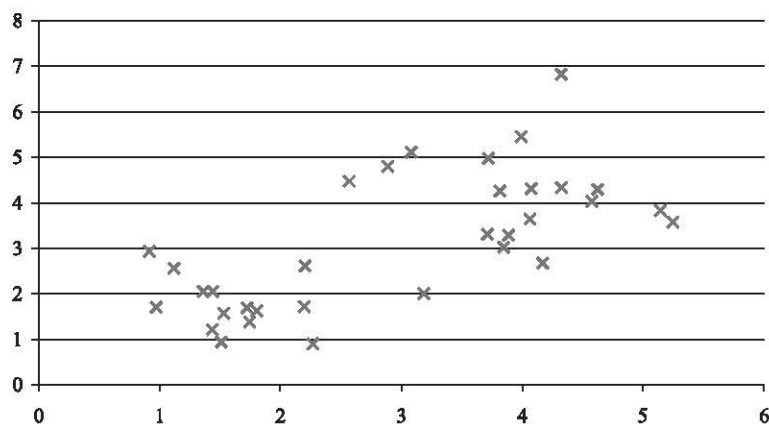


Fig. 2. Subset sampling of the Gaussian distribution of Fig. 1.

314 In general, one cannot guard against the pos-  
 315 sibility of having points from the tails appearing in  
 316 the sub-samples. However, if we sub-sample the  
 317 data set enough times, the tailing will disappear  
 318 and the combination of the sub-samples will rep-  
 319 resent the actual data set in certain precision.  
 320 Thus, in Bradley's random sampling refinement  
 321 procedure, the  $k$ -means algorithm is first applied  
 322 to a small percentage of the samples randomly  
 323 selected under the assumption that the smaller  
 324 initialization set has the same distribution as the  
 325 full sample set (Bradley and Fayyad, 1998). The  
 326 refinement algorithm is featured with an iterative,  
 327 multiple subset sampling and refinement process to  
 328 derive a proper initial-point set for the clustering  
 329 algorithms. The  $k$ -means obtained from these  
 330 random samples are then used as initial points for  
 331 a full scale conducting of the  $k$ -means algorithm  
 332 on the entire data set.

333 The iterative refinement algorithm has three  
 334 major steps. In the first step, a number  $J$  of small  
 335 sub-samples of data,  $S_i$ ,  $i = 1, \dots, J$ , are chosen  
 336 randomly from the whole data set. The  $J$  is se-  
 337 lected for the purpose of avoiding solutions "cor-  
 338 rupted" by outliers included in the sub-sample  $S_i$ ,  
 339 commonly,  $J$  ranges from  $0.1(n/k)$  to  $0.5(n/k)$ ,  
 340 depending on the data set size. The sub-samples  
 341 are clustered via a selected clustering algorithm ( $k$ -  
 342 modes in our case) using randomly determined  
 343 initial points. The sets  $CM_i$ ,  $i = 1, \dots, J$  are these  
 344 clustering solutions (cluster means) over the sub-  
 345 sample  $S_i$ . Let  $CM$  be the union of  $CM_i$ ,  
 346  $CM = \bigcup_{i=1}^J CM_i$ .

347 In the second step, where the refinement actu-  
 348 ally takes place, the  $CM$  is treated as the data set  
 349 and clustered via the selected clustering algorithm  
 350 again. The  $CM$  is clustered  $J$  times using  $CM_i$ ,  
 351  $i = 1, \dots, J$  as the initial points. Each clustering  
 352 solution over  $CM$  forms the sets  $FM_i$ ,  $i = 1, \dots, J$ .  
 353 Let  $FMS$  be the union of sets  $FM_i$ ,  $FMS = \bigcup_{i=1}^J FM_i$ .  
 354 The  $FMS$  represents a candidate set of the refined  
 355 initial points. The candidate initial-point sets  $FM_i$ ,  
 356  $i = 1, \dots, J$ , in  $FMS$  are further evaluated by a  
 357 distortion measurement function  $Distortion$   
 358 ( $FM_i, CM$ ) for the selection of the initial points to  
 359 be used in the clustering process.

In the distortion measurement, the function  
 $Distortion(FM_i, CM)$  is simply the summation of the  
 distance between the data items in  $CM$  and the  
 point  $FM_i$ . A smaller value for the distortion  
 measure indicates that the model parameters (i.e.  
 initial points  $FM_i$ ) are a better fit to the whole data  
 set. The  $FM_i$  that has the minimal distortion over  
 the set  $CM$  then is selected as the initial points for  
 the clustering algorithm.

The refinement algorithm takes these param-  
 eters as input:

- $Data$  – the data set to be clustered;
- $K$  – the number of desired clusters, and
- $J$  – the number of small sub-samples to be taken  
 from  $Data$ .

The algorithm is described as follows:

**Algorithm. Iterative Initial-Points-Refinement**  
 ( $Data, K, J$ )

*Step 1: //Sub-sampling*

1.0  $CM = 0$

1.1 For  $i = 1, \dots, J$

1.1.1. Let  $S_i$  be a small random sub-sample  
 set of  $Data$

1.1.2. Let  $SP_i$  be a randomly selected  $K$   
 sample from  $S_i$

1.1.3.  $CM_i = Clustering(SP, S_i, K)$

1.1.4.  $CM = CM \cup CM_i$

*Step 2: //Refinement*

2.0  $FMS = 0$

2.1 For  $i = 1, \dots, J$

2.1.1. Let  $FM_i = Clustering(CM_i, CM, K)$

2.1.2. Let  $FMS = FMS \cup FM_i$

*Step 3: //Selection*

3.1. Let  $FM = ArgMin_{FM_i} \{Distortion(FM_i,$   
 $CM)\}$

3.2. Return ( $FM$ )

The refinement algorithm has a computational  
 complexity of  $O(JK(\|S_i\|))$ , where  $K(\|S_i\|)$  is the  
 computation needed for clustering  $\|S_i\|$  number of  
 data points into  $k$  clusters.

The iterative initial-point refinement algorithm  
 has been applied successfully in clustering numer-  
 ically valued data sets. We present our testing and  
 experimental results of the algorithm on the cate-  
 gorical data clustering in the following section.





Table 4

The description of D1 obtained by conceptual clustering, described by a plant pathologist, and obtained by *k*-modes algorithm

Variable	Range determined by conceptual clustering	Range determined by plant pathologist	Range determined by <i>k</i> -modes with initial-point refinement
Precipitation	Above normal	Normal or above normal	Above normal
Temperature	Normal	Normal or above normal	Normal
Stem cankers	Above second node	Above second node	Above second node
Canker lesion color	Brown or n.a.	Brown	Dark brown/black
Fruiting bodies	Present	Present	Present
Condition of fruit pods	Normal	Normal	Normal
Time of occurrence	July–October	August–September	September
No. yrs. crop repeated	Several years	Several years	
Plant stand	Normal		Normal
External decay of stem	Firm and dry		Firm and dry
Int. discolor of stem	None		None
Sclerotia int. or ext.	Absent		Absent
Condition of roots	Normal	Not present	Normal
Damaged areas	Scattered areas or low areas	In expert	Low areas
Severity	Potentially severe or severs	Description	Pot-severe
Leaf spots	Absent		Absent
Shotholing/shreading	Absent		Absent
Leaf malformation	Absent		Absent
Leaf mildew growth	Absent		Absent
Condition of stem	Abnormal		Abnormal
Plant height	Abnormal		Abnormal
Condition of leaves	Abnormal		Abnormal
Mycelium on stem	Absent		Absent
Condition of seed	Normal		Normal
Seed treatment	None or fungicide		Fungicide

Table 5

Discriminate characteristics for clusters of soybean disease cases produced by conceptual clustering algorithm

Variable	Cluster 1 – Diaporthe stem canker	Cluster 2 – Charcoal rot	Cluster 3 – Rhizotonia root rot	Cluster 4 – Phytophthora rot
Precipitation	Above normal	Below normal	Above normal	Normal/above
Temperature	Normal	Normal/above	Below normal	Normal/below
Stem cankers	Above second node	Absent	Below soil line	Below or slightly above soil line
Canker lesion color	Brown or n.a.	Tan	Brown	Dark brown/black
Fruiting bodies	Present	Absent	Absent	Absent
Condition of fruit pods	Normal	Normal	Few/none	Irrelevant
Plant stand	Normal	Normal	Irrelevant	Less than normal
External decay of stem	Firm and dry	Absent	Firm and dry	Absent/firm and dry
Int. discolor of stem	None	Black	None	None
Sclerotia int. or ext.	Absent	Present	Absent	Absent
Condition of roots	Normal	Normal	Normal/rotted	Rotted
Damaged areas	Scattered areas or low areas	Whole fields, upland areas	Low area	Whole fields, low area

416 The Michalski problem is to reconstruct a  
 417 classification of selected soybean diseases. Given in  
 418 the data set are 47 cases of soybean diseases each

characterized by 35 multi-valued variables. These 419  
 cases are drawn from four populations – each 420  
 population representing one of the following soy- 421

Table 6

Discriminate characteristics for clusters of soybean disease cases produced by  $k$ -modes cluster algorithm with initial-point refinement

Variable	Cluster 1 – Diaporthe stem canker	Cluster 2 – Charcoal rot	Cluster 3 – Rhizotonia root rot	Cluster 4 – Phytophthora rot
Precipitation	Above normal	Less than normal	Above normal	Normal
Temperature	Normal	Normal	Below normal	Normal
Stem cankers	Above second node	Absent	Below soil line	Below soil line
Canker lesion color	Brown or n.a.	Tan	Brown	Dark brown/black
Fruiting bodies	Present	Absent	Absent	Absent
Condition of fruit pods	Normal	Normal	Few	Irrelevant
Plant stand	Normal	Normal	Less than normal	Less than normal
External decay of stem	Firm and dry	Absent	Firm and dry	Absent/firm and dry
Int. discolor of stem	None	Black	None	None
Sclerotia int. or ext.	Absent	Present	Absent	Absent
Condition of roots	Normal	Normal	Rotted	Rotted
Damaged areas	Scattered areas or low areas	Upland areas	Low area	Low area

422 bean diseases: D1 – Diaporthe stem canker, D3 –  
 423 Rhizoctonia root rot, D2 – Charcoal rot, and D4 –  
 424 Phytophthora rot. Table 1 shows the 35 variables  
 425 to categorize these diseases. Ideally, a clustering  
 426 method should partition these given cases into  
 427 four groups corresponding to the diseases.

428 We run the program that implements the  $k$ -  
 429 modes with iterative initial-point refinement algo-  
 430 rithm 20 times and compared the results with that  
 431 of non-refinement initializations. The results are  
 432 evaluated using clustering an accuracy rate  $r$  de-  
 433 fined as

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

435 where  $a_i$  is the number of instances occurring in  
 436 both cluster  $i$  and its corresponding class,  $k$  is the  
 437 number of clusters (4 in this case), and  $n$  is the  
 438 number of instances in the data set (47 in this  
 439 case). The clustering error is defined as  $e = 1 - r$ .  
 440 The 20 results are summarized in Table 2.

441 The experiment results show that 70% (14 cases)  
 442 of the results has accuracy of 0.98 (only miss one  
 443 case) for refinement initializations. But only 45%  
 444 of the results have accuracy of 0.89 or above for  
 445 non-refinement initializations. This demonstrates  
 446 that the refinement initialization algorithm yields  
 447 better clustering results than non-refinement ini-  
 448 tialization methods in clustering categorical data  
 449 sets.

Next we compare our cluster centroids (cluster 450  
 modes) obtained from the  $k$ -modes algorithm to 451  
 refined initialization with Michalski's results (Mi- 452  
 chalski and Stepp, 1983). Michalski used concep- 453  
 tual clustering algorithm to cluster the same 454  
 soybean disease data set. We use the cases that 455  
 have 0.98 accuracy in our algorithm to compare 456  
 with Michalski's results. Table 3 listed the *modes* 457  
 of the four clusters from  $k$ -modes algorithm. 458

Table 4 represents the complete complex for 459  
 cluster D1 – Diaporthe stem canker. The first 460  
 column is the name of the 25 attributes used to 461  
 describe the characteristics of cluster D1. The 462  
 second column contains the values for the 25 463  
 variables from Michalski's conceptual clustering 464  
 algorithm. The third column represents values of 465  
 variables used by an expert plant pathologist to 466  
 describe the same disease for diagnosis. We re- 467  
 constructed the values of all attributes for cluster 468  
 D1 and listed them in the fourth column of the 469  
 table. As we can see from the table that the de- 470  
 scription of the disease determined by  $k$ -modes 471  
 algorithm contains all the symptoms of the disease 472  
 specified by Michalski's conceptual clustering al- 473  
 gorithm and by the plant pathologist. Table 5 474  
 shows the values of discriminate variables for each 475  
 cluster derived from conceptual clustering algo- 476  
 rithm. In Table 6 we show the same values derived 477  
 from the  $k$ -modes with initial-point refinement 478

479 algorithm. Once again the two set results match  
480 very well.

## 481 5. Conclusion

482 In this paper, an experiment on an iterative  
483 initial-point refinement process to  $k$ -modes clus-  
484 tering algorithm for clustering data set containing  
485 categorical (symbolically valued) values is pre-  
486 sented. The procedure is motivated by the obser-  
487 vation that sub-sampling can provide guidance  
488 regarding the location of the data modes governed  
489 by a joint probability density function assuming to  
490 have generated the data. The refinement algorithm  
491 operates over small sets of sub-samples of a given  
492 data set, hence requiring a small portion of the  
493 total memory needed to store the full data and  
494 making this approach very appealing for large-  
495 scale clustering problems. By initializing a general  
496 clustering estimation near the true modes, the true  
497 clusters are discovered more often in the repetitive  
498 applications of the program. However, more study  
499 is needed on the scalability and adaptiveness of the  
500 algorithms for much larger and complicated dis-  
501 tributed data sets. On the other hand, the tech-  
502 nique dealt with in this research is independent of  
503 the data set size in terms of algorithmic analysis of  
504 the technique presented. Therefore it can be ex-  
505 pected that the algorithm is to perform equally  
506 well on other data sets in principle.

## 507 Acknowledgements

508 This work described in this paper was partly  
509 supported by AFOSR Grant no: F49620-99-1-  
510 0211. The authors thank the anonymous reviewers  
511 for their valuable comments that helped improve  
512 the presentation.

## References

- Ahrens, J., Dieter, U., 1973. Extensions of Forsythe's method for random sampling from the normal distribution. *Math. Comput.* 27 (124), 927–937. 514  
515  
516  
Anderberg, M., 1973. *Cluster Analysis for Applications*. Academic Press, New York. 517  
518  
Bobrowski, L., Bezdek, J., 1991.  $C$ -means clustering with the  $l_1$  and  $l_\infty$  norms. *IEEE Trans. Systems Man Cybernet.* 21 (3), 519  
520  
521  
Bradley, P., Fayyad, U., 1998. Refining initial points for  $k$ -means clustering. In: *Proc. 15th Internat. Conf. on Machine Learning*. Morgan Kaufmann, Los Altos, CA. 522  
523  
524  
Bradley, P., Fayyad, U., Reina, C., 1998. Refining initialization of clustering algorithms. In: Ahsl, A. (Ed.), *Proc. 4th Internat. Conf. on Knowledge Discovery and Data Mining*. AAAI Press, New York. 525  
526  
527  
528  
Fisher, D., 1987. Knowledge acquisition via incremental conceptual clustering. *Machine Learning* 2 (2), 139–172. 529  
530  
Gower, J., Diday, E., 1991. Symbolic clustering using a new dissimilarity measure. *Pattern Recognition* 24 (6), 567–578. 531  
532  
Hartigan, J., 1975. *Clustering Algorithms*. Wiley, New York. 533  
Huang, Z., 1997. Clustering large data sets with mixed numeric and categorical value. In: *Proc. First Pacific Asia Knowledge Discovery and Data Mining Conf.* World Scientific, Singapore, pp. 21–34. 534  
535  
536  
537  
Huang, Z., 1998. Extensions to the  $k$ -means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery II*, 283–304. 538  
539  
540  
Jain, A., Dubes, R., 1988. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ. 541  
542  
Kaufman, L., Rousseeuw, P., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York. 543  
544  
Lebowitz, M., 1987. Experiments with incremental concept formation. *Machine Learning* 2 (2), 103–138. 545  
546  
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observation. In: *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability*, pp. 281–297. 547  
548  
549  
550  
Michalski, R., Bratko, I., Kubat, M., 1998. *Machine Learning and Data Mining: Methods and Applications*. Wiley, New York. 551  
552  
553  
Michalski, R., Stepp, R., 1983. Automated construction of classifications: conceptual clustering versus numerical taxonomy. *IEEE Trans. Pattern Anal. Machine Intell.* 5 (4), 396–410. 554  
555  
556  
557  
Ng, R., Han, J., 1994. Efficient and effective clustering methods for spatial data mining. In: *Proc. 20th VLDB Conf.*, Santiago, Chile, pp. 144–155. 558  
559  
560  
Ralambondrainy, H., 1995. A conceptual version of the  $k$ -means algorithm. *Pattern Recognition Letters* 16, 1147–1157. 561  
562  
563