5-2002

An iterative initial-points refinement algorithm for categorical data clustering

Ying Sun

Qiuming Zhu
University of Nebraska at Omaha, qzhu@unomaha.edu

Zhengxin Chen
University of Nebraska at Omaha, zchen@unomaha.edu

An iterative initial-points refinement algorithm for categorical data clustering

Ying Sun, Qiuming Zhu *, Zhengain Chen

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

Abstract

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

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

1. Introduction

Partitioning a set of objects in a data collection of multiple attributes into homogeneous groups (clusters) of certain intra-relations is a fundamental operation in data mining. The most distinct characteristic of clustering operation in data mining is that the data sets often contain both numeric and categorical (symbolic) attribute values. This requires the clustering algorithms to be capable of dealing with the complexity of the intra- and inter-relations of the data sets expressed in different

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

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

types of the attributes, no matter numeric or categorical (Michalski et al., 1998).

Among the clustering algorithms that have been developed, the *k*-means algorithm is the most popular one (Jain and Dubes, 1988). Many other clustering algorithms were derived from it, such as the fuzzy *k*-means algorithm, the ISODATA, the *k*-modes algorithm (Huang, 1998), etc. The *k*-means algorithm is well known for its efficiency in clustering large data sets (MacQueen, 1967; Anderberg, 1973). However, the original *k*-means algorithm works only on numeric data because it aims at minimizing a cost function that is numerically measured. This prohibits the *k*-means algorithm from being directly used in applications where categorical data are involved, such as the data mining applications.

Work on clustering data with categorical attributes has been done by several researchers. Ralambondrainy (1995) presented an approach by using the *k*-means algorithm to cluster categorical data. The approach is to convert multiple category attributes into binary attributes (using 0 and 1 to represent either a category absent or present) and to treat the binary attributes as numeric in the *k*-means algorithm. The main drawback of the approach is that the cluster means, given by values between 0 and 1, often do not indicate the exact characteristics of the clusters. Gower and Diday (1991) used a similarity coefficient and other dissimilarity measures to process data with categorical attributes. However, the quadratic computational cost makes them unacceptable for clustering large data sets.

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

The *k*-modes algorithm (Huang, 1997) extends the *k*-means paradigm to cluster categorical data by using (1) a simple matching dissimilarity measure for categorical objects, (2) modes instead of means for clusters, and (3) a frequency-based method to update modes in the *k*-means fashion to minimize the clustering cost function of clustering. Because the *k*-modes algorithm uses the same clustering process as *k*-means, it preserves the efficiency of the *k*-means algorithm, which is highly desirable for data mining. A similar work that aims to cluster large data sets is the CLARA (abbreviated from Clustering LARge Application) algorithm (Kaufman and Rousseeuw, 1990). CLARA is a combination of a sampling procedure and the clustering program Partitioning Around Medoids (PAM). Given a set of objects X and the number of clusters k, PAM clusters X by finding k medoids (representative objects of clusters) that can minimize the average dissimilarity of objects to their closest medoids. Ng and Han (1994) have analyzed that the computational complexity of PAM in a single iteration is $O(k(n-k)^2)$ where n is the number of objects in X. Obviously, PAM is not efficient when clustering large data sets. That makes CLARA inefficient in clustering large data sets.

As for the traditional clustering algorithms, most of the above-mentioned categorical data clustering algorithms, including the *k*-modes algorithm, require a random selection or setting up of initial data points in addition to a known or estimated number of clusters (also called starting conditions), before iteratively mapping the data records to separate clusters. This leads to the problem that the clustering results are often dependent on the selection of the initial points regardless of what measurement metric is used for the similarity (distance) evaluation operation. That is, the clustering solution is very much sensitive to the initial-point choices. An inappropriate setting up of initial points would lead to some unacceptable clustering results. For example, a large percent of data samples might be crowded into one or a few clusters with other clusters having only a few scarce samples, leaving users questioning its reality. Moreover, the clustering results often cannot be repetitively generated,

causing problems in the validation of the clustering results.

The intrinsic problem of initial-point selection in clustering algorithms and the computation cost of the categorical data clustering call for an approach that provides a better organized initial setting for improving the performance of clustering processes. Hopefully, the improved initial-point sets would let the clustering algorithm converge with the global optimal or close to the optimal solution more accurately and repetitively. That is, the selection of initial data points fits more appropriately and consistently with the nature and underlying distributions of the data sample sets.

In this paper we present an experiment on applying the iterative refinement algorithm to the setting of the initial points so as to map the categorical data sets to clustering results that have better consistency rates. This paper is organized as follows. Section 2 discusses the basics of the k-modes algorithms (Huang, 1997). Section 3 describes Bradley and Fayyad's initial-points refinement algorithm and its principle (cf. Bradley and Fayyad, 1998). Section 4 presents our experimental results in applying the initial-points refinement to the k-modes algorithm for clustering categorical data samples. Section 5 concludes the presentation.

2. The k-modes algorithm for categorical data clustering

Let $\boldsymbol{S} = \{X_1, X_2, \ldots, X_n\}$ denote a set of n data objects, and $X_i = [X_{i1}, X_{i2}, \ldots, X_{id}]$, $i = 1, 2, \ldots, n$, be an object represented by d attribute values. Let k be a positive integer. The objective of k-means clustering is to find a partition that divides object set $\boldsymbol{S}$ into k disjoint regions that meet certain criteria and constraints. For a given n and k, the number of possible partitions is definite but could be extremely large. A common way of solving it is to choose a clustering criterion that guides the search for an approximate solution. The most common criterion has been the minimization of the total distances of the data points to their cluster centers. Formulated as a mathematical programming problem $P(W, Q)$, the k-means clustering algorithm has been traditionally expressed as the following (Hartigan, 1975; Bobrowski and Bezdek, 1991):

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

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

$$1 \leqslant i \leqslant n, \ 1 \leqslant l \leqslant k,$$

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

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

1. *Sub-problem* P_1: Fix $Q = \hat{Q}$ and solve the reduced problem $P(W, \hat{Q})$.
2. *Sub-problem* P_2: Fix $W = \hat{W}$ and solve the reduced problem $P(\hat{W}, Q)$.

Sub-problem P_1 is solved by

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

and sub-problem P_2 is solved by

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

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

1. Choose an initial Q^0 and solve $P(W, Q^0)$ to obtain W^0. Set $t = 0$.
2. Let $\hat{W} = W^t$ and solve $P(\hat{W}, Q)$ to obtain Q^{t+1}. If $P(\hat{W}, Q^t) = P(\hat{W}, Q^{t+1})$, output $\hat{W}, Q^t$ and stop; otherwise, go to 3.
3. Let $\hat{Q} = Q^{t+1}$ and solve $P(W, \hat{Q})$ to obtain W^{t+1}. If $P(W^t, \hat{Q}) = P(W^{t+1}, \hat{Q})$, output $W^t, \hat{Q}$ and stop; otherwise, let $t = t + 1$ and go to 2.

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

In principle the formulation of problem P in the above is also valid for categorical and mixed-type data objects. The reason that the k-means algorithm cannot cluster categorical objects is its dissimilarity measure used to solve problem P_2. These

barriers can be removed by making the following modifications:

1. Using a simple matching dissimilarity measure for categorical objects.
2. Replacing means of clusters by modes.
3. Using a frequency-based method to find the modes to solve problem P_2.

Let X, Y be two categorical objects described by m categorical attributes. The dissimilarity measure between X and Y can be defined by the total mismatches of the corresponding attribute categories of the two objects. This measure is often referred to as simple matching (Kaufman and Rousseeuw, 1990). Formally, we have

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

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

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

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

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

$$f_r(A_j = q_j|X) \geqslant f_r(A_j = c_{k,j}|X) \quad \text{for } q_j \neq c_{k,j},$$
$$\text{for all } j = 1, \ldots, m.$$

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

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

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

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

That is, to minimize the cost function, the k-modes algorithm proceeds by: (1) using the simple matching dissimilarity measure to solve the problem 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 clustering

As mentioned above, both k-means and k-modes algorithms draw an initial estimation (approximation model) of the clusters represented by Q^0 from a randomly selected subset of the data points in the constraint object space. The algorithms then subsequently extend the model asymptotically 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 (Ahrens 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 clusters of the larger (or the whole) data set.

One practical problem of a simple sub-sampling approach is that severely sub-sampling the data will naturally bias the samples to representatives "near" the tails (or edges) of the data sets, while it makes nonsense to sub-sample a sufficiently large subset that is close to the actual data set. Fig. 1 shows a data set drawn from a mixture of two Gaussian models (clusters) in 2-D with means at $[1.5, 1.5]$ and $[4, 4]$, respectively. A small sub-sample set is shown in Fig. 2, which was expected to provide information on the models of the joint probability density function of the original data set of Fig. 1. Each of the points on the Fig. 2 may be thought of as a "guest" at the possible location of a model in the underlying distribution. It is seen that the sub-sample set exhibits certain "expected" behavior of the original data set. Worthy of note here is that the sub-sampling points are fairly spread out over the distribution region of the original data set. This observation indicates that the solutions obtained by clustering over a small sub-sample may not provide good initial estimates of the true means, or centroids, in the data. The simple sub-sampling method often produces noisy estimates due to single small sub-samples, especially in skewed distributions and high dimensions.

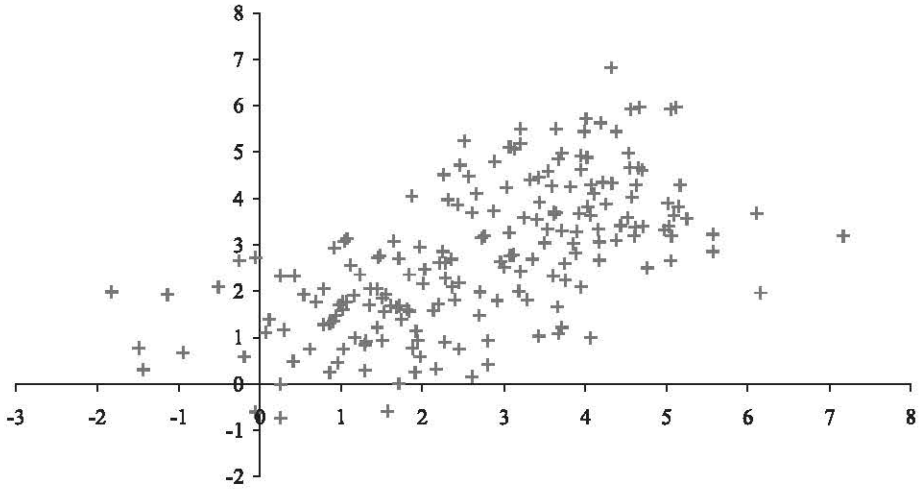

Fig. 1. Samples of Gaussian distribution in 2-D.

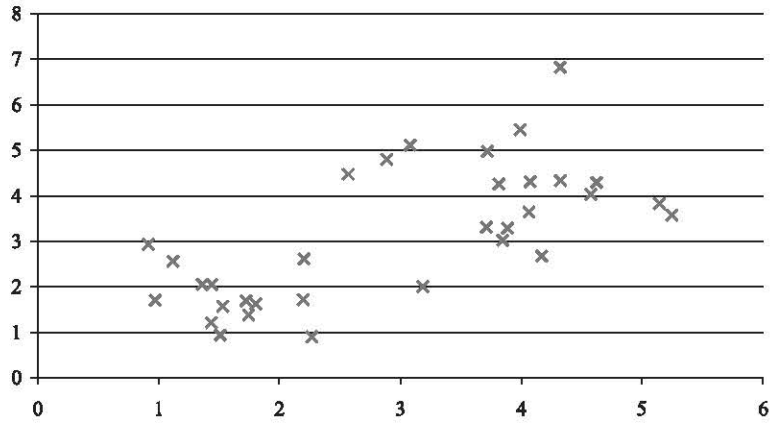

Fig. 2. Subset sampling of the Gaussian distribution of Fig. 1.

In general, one cannot guard against the possibility of having points from the tails appearing in the sub-samples. However, if we sub-sample the data set enough times, the tailing will disappear and the combination of the sub-samples will represent the actual data set in certain precision. Thus, in Bradley's random sampling refinement procedure, the k-means algorithm is first applied to a small percentage of the samples randomly selected under the assumption that the smaller initialization set has the same distribution as the full sample set (Bradley and Fayyad, 1998). The refinement algorithm is featured with an iterative, multiple subset sampling and refinement process to derive a proper initial-point set for the clustering algorithms. The k-means obtained from these random samples are then used as initial points for a full scale conducting of the k-means algorithm on the entire data set.

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

In the second step, where the refinement actually takes place, the CM is treated as the data set and clustered via the selected clustering algorithm again. The CM is clustered J times using CM_i, $i = 1, \ldots, J$ as the initial points. Each clustering solution over CM forms the sets FM_i, $i = 1, \ldots, J$. Let FMS be the union of sets FM_i, $FMS = \bigcup_{i=1}^{J} FM_i$. The FMS represents a candidate set of the refined initial points. The candidate initial-point sets FM_i, $i = 1, \ldots, J$, in FMS are further evaluated by a distortion measurement function *Distortion* (FM_i, CM) for the selection of the initial points to 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 parameters 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 $i = 1, \ldots, J$
  - 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 $i = 1, \ldots, J$
  - 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 numerically valued data sets. We present our testing and experimental results of the algorithm on the categorical data clustering in the following section.

4. Experimental results of initial-point refinement for categorical data clustering

We used the Michalski well-known soybean disease data set (Michalski and Stepp, 1983) to test the refined initial-point setting algorithm with the *k*-modes method. We compare the result with the method of picking the initial points randomly from the data set. We also compare our result obtained from applying *k*-modes clustering algorithms with initial-point refinement to those from Michalski's conceptual clustering algorithm.

Table 1
Multi-valued variables used to describe cases of soybean disease

Environment variables		Plant global variables
Time of occurrence	(7)	Severity
Plant stand	(3)	Seed treatment
Precipitation	(2)	Seed germination
Temperature	(4)	Plant height
Occurrence of hail	(3)	
Number of years crop repeated	(2)	
Damaged area	(3)	
Plant local variables		
Conditional of leaves	(2)	Condition of stem
Leaf spots	(2)	Presence of lodging
Leaf spots margin	(2)	Stem cankers
Leaf spot size	(5)	Canker lesion color
Shotholing/shreading	(2)	Fruiting bodies on stem
Leaf malformation	(2)	External decay of stem
Leaf mildew growth	(2)	Mycelium on stem
Condition of seed	(3)	Internal discoloration
Seed mold growth	(4)	Sclerotia internal/external
Seed discoloration	(2)	Condition of fruit pods
Seed size	(2)	Fruit spots
Seed shriveling	(2)	Condition of roots

Table 2
Summary of 20 results set

Accuracy	Cases with refinement	Cases with no refinement
0.98	14	5
0.94		2
0.89		2
0.77		3
0.70	5	
0.68		5
0.66	1	3

Table 3
k-modes algorithm results with accuracy of 0.98 for soybean disease data set. (Each row represents a cluster mode.)

Cluster	Mode																																			
D1	5	0	2	1	1	1	1	1	1	1	1	0	2	2	0	0	0	1	1	3	1	1	0	0	0	4	0	0	0	0	0	0	0	0	0	0
D2	5	0	0	1	1	1	2	1	1	1	1	0	2	2	0	0	0	1	1	0	3	0	0	0	2	1	0	4	0	0	0	0	0	0	0	0
D3	2	1	2	0	1	1	1	1	1	1	1	0	2	2	0	0	0	1	1	1	1	0	1	1	0	0	3	4	0	0	0	0	0	0	0	1
D4	1	1	1	1	1	1	1	1	1	1	1	0	2	2	0	0	0	1	1	1	2	0	1	0	0	0	3	4	0	0	0	0	0	0	0	1

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 *k*-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	Irrelevent	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

The Michalski problem is to reconstruct a classification of selected soybean diseases. Given in the data set are 47 cases of soybean diseases each characterized by 35 multi-valued variables. These cases are drawn from four populations – each population representing one of the following soy-

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

bean diseases: D1 – Diaporthe stem canker, D3 – Rhizoctonia root rot, D2 – Charcoat rot, and D4 – Phytophthorat rot. Table 1 shows the 35 variables to categorize these diseases. Ideally, a clustering method should partition these given cases into four groups corresponding to the diseases.

We run the program that implements the k-modes with iterative initial-point refinement algorithm 20 times and compared the results with that of non-refinement initializations. The results are evaluated using clustering an accuracy rate r defined as

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

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

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

Next we compare our cluster centroids (cluster modes) obtained from the k-modes algorithm to refined initialization with Michalski's results (Michalski and Stepp, 1983). Michalski used conceptual clustering algorithm to cluster the same soybean disease data set. We use the cases that have 0.98 accuracy in our algorithm to compare with Michalski's results. Table 3 listed the *modes* of the four clusters from k-modes algorithm.

Table 4 represents the complete complex for cluster D1 – Diaporthe stem canker. The first column is the name of the 25 attributes used to describe the characteristics of cluster D1. The second column contains the values for the 25 variables from Michalski's conceptual clustering algorithm. The third column represents values of variables used by an expert plant pathologist to describe the same disease for diagnosis. We reconstructed the values of all attributes for cluster D1 and listed them in the fourth column of the table. As we can see from the table that the description of the disease determined by k-modes algorithm contains all the symptoms of the disease specified by Michalski's conceptual clustering algorithm and by the plant pathologist. Table 5 shows the values of discriminate variables for each cluster derived from conceptual clustering algorithm. In Table 6 we show the same values derived from the k-modes with initial-point refinement

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

5. Conclusion

In this paper, an experiment on an iterative initial-point refinement process to k-modes clustering algorithm for clustering data set containing categorical (symbolically valued) values is presented. The procedure is motivated by the observation that sub-sampling can provide guidance regarding the location of the data modes governed by a joint probability density function assuming to have generated the data. The refinement algorithm operates over small sets of sub-samples of a given data set, hence requiring a small portion of the total memory needed to store the full data and making this approach very appealing for large-scale clustering problems. By initializing a general clustering estimation near the true modes, the true clusters are discovered more often in the repetitive applications of the program. However, more study is needed on the scalability and adaptiveness of the algorithms for much larger and complicated distributed data sets. On the other hand, the technique dealt with in this research is independent of the data set size in terms of algorithmic analysis of the technique presented. Therefore it can be expected that the algorithm is to perform equally well on other data sets in principle.

Acknowledgements

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

Anderberg, M., 1973. Cluster Analysis for Applications. Academic Press, New York.

Bobrowski, L., Bezdek, J., 1991. C-means clustering with the l_1 and l_∞ norms. IEEE Trans. Systems Man Cybernet. 21 (3), 545–554.

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.

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.

Fisher, D., 1987. Knowledge acquisition via incremental conceptual clustering. Machine Learning 2 (2), 139–172.

Gower, J., Diday, E., 1991. Symbolic clustering using a new dissimilarity measure. Pattern Recognition 24 (6), 567–578.

Hartigan, J., 1975. Clustering Algorithms. Wiley, New York.

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.

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.

Jain, A., Dubes, R., 1988. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ.

Kaufman, L., Rousseeuw, P., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Lebowitz, M., 1987. Experiments with incremental concept formation. Machine Learning 2 (2), 103–138.

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.

Michalski, R., Bratko, I., Kubat, M., 1998. Machine Learning and Data Mining: Methods and Applications. Wiley, New York.

Michalski, R., Stepp, R., 1983. Automated construction of classifications: conceptual clustering versus numerical taxonomy. IEEE Trans. Pattern Anal. Machine Intell. 5 (4), 396–410.

Ng, R., Han, J., 1994. Efficient and effective clustering methods for spatial data mining. In: Proc. 20th VLDB Conf., Santiago, Chile, pp. 144–155.

Ralambondrainy, H., 1995. A conceptual version of the k-means algorithm. Pattern Recognition Letters 16, 1147–1157.