Pattern Recognition Letters

Pattern Recognition Letters 24 (2003) 2367–2376

Clustering and its validation in a symbolic framework

Kalyani Mali [a], Sushmita Mitra [b,*]

[a] *Department of Computer Science, Kalyani University, Kalyani 741 235, India*
[b] *Machine Intelligence Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700 108, India*

Abstract

Clustering of symbolic data, using different validity indices, is proposed for determining the optimal number of meaningful clusters. Symbolic objects include linguistic, nominal, boolean, and interval-type of features, along with quantitative attributes. Clustering in this domain involves the use of symbolic dissimilarity between the objects. The novelty of the method lies in transforming the different clustering validity indices, like Normalized Modified Hubert's statistic, Davies–Bouldin index and Dunn's index, from the quantitative domain to the symbolic framework. The effectiveness of symbolic clustering is demonstrated on several real life benchmark data sets.

Keywords: Hierarchical clustering; Symbolic clustering; Symbolic data; Validity index

1. Introduction

Clustering is a useful technique for the discovery of some knowledge from a data set. It maps a data item into one of several clusters, where clusters are natural groupings of data items based on similarity metrics or probability density models. Clustering of data is broadly based on two approaches, viz., hierarchical and partitive (Jain and Dubes, 1988). Hierarchical methods can again be categorized as *agglomerative* and *divisive* algorithms, corresponding to bottom-up and top-down strategies, to build a hierarchical clustering tree (*dendogram*). The optimal number of clusters is usually determined based on a validation index. There exist several clustering algorithms and validation indices in literature (Milligan and Cooper, 1985; Jain and Dubes, 1988; Bezdek and Pal, 1998), that conventionally deal with numerical/quantitative data. Some other general methods are available (Ben-Hur et al., 2002; Fridlyand and Dudoit, 2001; Lange et al., 2002; Tibshirani et al., 2001) that use stability to determine roughly the highest number of clusters, such that the clustering is stable.

Symbolic/categorical clustering refers to the clustering of symbolic/categorical data. This is important from the point of view of data mining, where one has to mine for information from a set of symbolic data. Symbolic patterns are defined by attributes that can be quantitative (numeric or

intervals) as well as qualitative. The similarity and dissimilarity measures between symbolic data are determined based on their position, span and content (Chidananda Gowda and Diday, 1991; Chidananda Gowda and Ravi, 1995; Mali and Mitra, 2002). Validation of symbolic clusters is an issue mostly untouched in literature.

Different indices (Jain and Dubes, 1988; Bezdek and Pal, 1998), like Normalized Modified Hubert's statistic, Davies–Bouldin index and Dunn's index that were originally developed for the quantitative domain, are modified in this paper to work in the symbolic framework using dissimilarity measure, expressed in terms of symbolic distance. Their performance is then compared with that of the cluster indicator (Chidananda Gowda and Ravi, 1995) and conceptual clustering (Biswas et al., 1998).

Conceptual clustering (Wilson and Martinez, 1997; Biswas et al., 1998), from the Machine Learning community, is also applicable to a mixture of numeric, ordinal and symbolic data. Here the focus is on interpretability/meaningfulness of generated patterns. The algorithm preserves cohesiveness within clusters while maintaining clear distinctness between clusters. Nonparametric probabilistic measures are used to determine the groupings.

In the present article we concentrate on studying the effectiveness of symbolic clustering, using different validity indices in terms of symbolic dissimilarity, on real life data. Symbolic benchmark data sets like *Microcomputer*, *Zoo*, *Auto Import* and *Mushroom* (Blake and Merz, 1998) are used for the purpose. An optimal number of meaningful clusters are obtained in each case. Comparative study is made with conceptual clustering and the results are quantitatively evaluated.

Section 2 describes the clustering algorithm in terms of symbolic dissimilarity. The different validity indices, modified in the symbolic framework, are presented in Section 3. The implementation on the benchmark data sets is provided in Section 4. Section 5 concludes the article.

2. Symbolic clustering

In this section we describe symbolic data, the different dissimilarity measures (expressed as the dissimilarity between them) and an agglomerative clustering algorithm.

2.1. Symbolic patterns

Symbolic data are defined as the logical conjunction of events linking values and variables. The following are two examples of events: $e_1 = [\text{color} = \{\text{white}, \text{blue}\}]$, $e_2 = [\text{height} = [1.5\text{–}2.0]]$. Here e_1 indicates that the variable color takes a value either white or blue, while e_2 indicates that the variable height takes a value between 1.5 and 2.0. For simplicity, we can drop the variable name and only take the value of that feature variable. Two symbolic data A and B are written as the logical conjunction of feature values A_k and B_k as $A = A_1 \wedge \cdots \wedge A_n$ and $B = B_1 \wedge \cdots \wedge B_n$.

The dissimilarity between two symbolic data A and B is defined as (Chidananda Gowda and Diday, 1991; Chidananda Gowda and Ravi, 1995)

$$D(A,B) = \sum_{i=1}^{n} D(A_i,B_i), \tag{1}$$

where

$$D(A_i,B_i) = D_p(A_i,B_i) + D_s(A_i,B_i) + D_c(A_i,B_i)$$

with D_p, D_s and D_c (normalized to [0,1]) indicating the components due to position, span and content, respectively.

We have

$$D_p(A_i,B_i) = \cos\left[\left(1 - \frac{|\text{lower limit of } A_i - \text{lower limit of } B_i|}{\text{length of maximum interval along feature } i}\right) * 90\right], \tag{2}$$

where the denominator indicates the difference between the highest and lowest values of the ith

feature over all the objects. This measure holds for quantitative attributes only. The remaining two measures are defined for both quantitative and qualitative attributes.

$$D_s(A_i,B_i)=\cos\left[\frac{|\text{length of } A_i|+|\text{length of } B_i|}{2*\text{ span length of } A_i \text{ and } B_i}*90\right], \tag{3}$$

where span length denotes the length of the minimum interval containing both A_i and B_i for quantitative values. The length of a qualitative feature value is the number of its elements, and the span length of two such feature values is the number of elements in their union.

$$D_c(A_i,B_i)=\cos\left[\frac{\text{length of intersection of } A_i \text{ and } B_i}{\text{span length of } A_i \text{ and } B_i}*90\right]. \tag{4}$$

2.2. Agglomerative clustering

Agglomerative algorithms typically involve a repetition of the steps (i) assignment of pattern vectors X to its own cluster, (ii) all intercluster distance computation, and (iii) merging two clusters which are closest to each other, until there is only one cluster left. Like typical hierarchical methods, the partitioning at any stage depends on the previously found clusters.

Let us now define two different measures for within cluster and between cluster scatters in the symbolic framework. Let $\{X_1,\ldots,X_{c_k}\}$ be a set of symbolic data lying in a cluster U_k. Then the *average scatter* within the cluster U_k is expressed as

$$S_a(U_k)=\frac{\sum_{i,i'} D(X_i,X_{i'})}{|c_k|(|c_k|-1)}, \tag{5}$$

where X_i, $X_{i'}\in U_k$, $i\neq i'$, $|c_k|$ is the number of samples in cluster U_k, and D indicates the symbolic dissimilarity of Eq. (1). The *between cluster scatter* is defined as

$$d_a(U_k,U_l)=\frac{\sum_{i,j} D(X_i,X_j)}{|c_k||c_l|}, \tag{6}$$

where $X_i\in U_k$, $X_j\in U_l$, such that $k\neq l$. We have used S_a and d_a in our computations, in terms of the symbolic dissimilarity D of Eq. (1).

2.3. Symbolic clustering algorithm

Agglomerative symbolic clustering tends to favor the merging of singleton clusters, or of small clusters with large ones, as compared to the merging of medium sized clusters. The algorithm is as follows.

(i) Let $\{X_1,X_2,\ldots,X_N\}$ be a set of N symbolic data forming the original data set. Let the initial number of clusters be N, with each cluster having a weight (number of objects) of 1. Therefore $X_i\in U_i$, $i=1,\ldots,N$.

(ii) Compute the weighted dissimilarities (Chidananda Gowda and Diday, 1991) between all pairs of clusters in the data set as

$$d_{aw}(U_i,U_j)=d_a(U_i,U_j)\left(\frac{|c_i|.|c_j|}{|c_i|+|c_j|}\right)^{0.5}, \tag{7}$$

where $|c_i|$, $|c_j|$ are the cluster weights of U_i, U_j respectively, and $d_a(U_i,U_j)$ is the average between cluster scatter/dissimilarity given by Eq. (6). Note that the weighting term on the r.h.s. of Eq. (7) yields a value $\sqrt{50}$ for $|c_i|=|c_j|=100$, while it results in $\sqrt{0.5}$ for $|c_i|=|c_j|=1$ (singleton clusters). Hence the dissimilarity is highlighted for larger clusters. However for $|c_i|=100$, $|c_j|=1$, we have $100/101\simeq 1$, naturally implying a higher dissimilarity than that for $|c_i|=|c_j|=1$. So there is a greater probability of merging a pair of smaller clusters as compared to larger clusters.

(iii) Determine the mutual pair (clusters) having the lowest average weighted dissimilarity d_{aw}min by Eq. (7). Form a composite cluster U_k by merging the individuals of this pair, such that $|c_k|=|c_i|+|c_j|$. Reduce the number of clusters by 1.

(iv) Repeat steps (ii) to (iii) until the number of clusters equals 1.

(v) Compute cluster validity index V_t by Eqs. (13)–(16). Determine the stage t_0, with clusters $c=c_0$, for $c=2,\ldots,\sqrt{N}$, at which V_t is optimum. This indicates the optimal number of clusters.

2.4. Conceptual clustering

Conceptual clustering (Wilson and Martinez, 1997; Biswas et al., 1998) is based on *Category Utility* CU of a cluster U_k defined as

$$\mathrm{CU}_k = P(U_k) * \left(\sum_i \sum_j P(A_i = V_{ij}|U_k)^2 - \sum_i \sum_j P(A_i = V_{ij})^2 \right), \quad (8)$$

where $P(U_k)$ is the a priori probability of cluster U_k, $P(A_i = V_{ij})$ is the probability of feature A_i taking on value V_{ij} and $P(A_i = V_{ij}|U_k)$ is the conditional probability of $A_i = V_{ij}$ in cluster U_k. Here i indicates the distinct number of features, while j denotes the distinct values that feature i can attain. This represents an increase in the number of feature values that can be correctly guessed for cluster $U_k(P(A_i = V_{ij}|U_k)^2)$, over the expected number of correct guesses, given that no class information is available $[P(A_i = V_{ij})^2]$. The partition score/utility of a partition structure made up of l clusters is defined as the average CU over the l clusters

$$\frac{\sum_{k=1}^{l} \mathrm{CU}_k}{l}. \quad (9)$$

The objective is to generate maximally cohesive clusters (intraclass similarity) while achieving maximum separability (interclass dissimilarity) among the clusters in a partition. Probabilistic measures, to evaluate the goodness of the partitioning, are expressed as (Biswas et al., 1998)

$$M_{dk} = \sum_{i,j \in \{A_i\}_d} (P(A_i = V_{ij}|U_k)^2 - P(A_i = V_{ij})^2) \quad (10)$$

and

$$\mathrm{Var}(U(k), U(l)) = \frac{1}{n} \sum_i^n \sum_j (P(A_i = V_{ij}|U_k) - P(A_i = V_{ij}|U_l))^2. \quad (11)$$

Here M_{dk} is the increase in predictability of an object d to cluster k. *Cohesion* of a partition structure is measured as the sum of the M_{dk} values of all objects in the data set. $\mathrm{Var}(U(k), U(l))$ is the variance of distribution match between clusters k and l in a given partition. *Distinctness* of a partition is taken as the average variance between clusters in that partition.

3. Cluster validity indices

To select the best among different partitioning, each of these can be evaluated using some validity index. The procedure is repeated for $c = 2, \ldots, \sqrt{N}$ number of clusters, where N is the size of the data set. Several validation methods have been proposed in literature (Bezdek and Pal, 1998) for quantitative data. These include Normalized Modified Hubert's statistic, Davies–Bouldin index and Dunn's index. In this section we modify these expressions in the symbolic framework. We use the average scatter S_a (Eq. (5)) within a cluster and d_a (Eq. (6)) between clusters, in the computations. This is in contrast to using the distance from cluster means/centroids in Hubert's and Davies–Bouldin, or the diameter in Dunn's, as in the quantitative domain (Jain and Dubes, 1988; Bezdek and Pal, 1998). Their performance is compared with that of the cluster indicator (Chidananda Gowda and Diday, 1991) and conceptual clustering (Biswas et al., 1998). The modified expressions are provided below.

3.1. Hubert's statistic

Let X_i be the ith object and $L(i) = k$ if $X_i \in U_k$. Modified Hubert's Γ statistic (Bezdek and Pal, 1998) for a particular cluster structure is expressed in terms of symbolic dissimilarity as

$$\Gamma = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} D(X_i, X_j) d_a(U_{L(i)}, U_{L(j)}). \quad (12)$$

If X_i and X_j lie in two different clusters, d_a is computed using Eq. (6). However, when they belong to the same cluster, $d_a = 0$. From this, we get Normalized Modified Hubert's statistic $\hat{\Gamma}$ as

$$\hat{\Gamma} = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{(D(X_i, X_j) - \overline{D})(d_a(U_{L(i)}, U_{L(j)}) - \overline{d_a})}{s_D s_{d_a}}. \quad (13)$$

Here

$$\overline{D} = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} D(X_i, X_j),$$

$$\overline{d_a} = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_a(U_{L(i)}, U_{L(j)}),$$

$$s_D^2 = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} D^2(X_i, X_j) - \overline{D}^2$$

and

$$s_{d_a}^2 = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_a^2(U_{L(i)}, U_{L(j)}) - \overline{d_a}^2,$$

where $M = (N(N-1))/2$ is the total number of terms under the double summation. Note that $M = N^2$ if the matrix under summation is not symmetric. The terms s_D and s_{d_a} are the standard deviations of the entries of the matrices D and d_a respectively, while s_D^2 and $s_{d_a}^2$ are the corresponding variances (Bezdek and Pal, 1998). The optimal partitioning occurs at $c = c_0$ for which $\Delta(\Delta\widehat{\Gamma})$ is minimum. This corresponds to a sharp change in slope (also called *knee*) of the piece-wise linear graph for Normalized modified Hubert's statistic, in case of well-separated clusters.

3.2. Davies–Bouldin index

The Davies–Bouldin index (Bezdek and Pal, 1998) is a function of the ratio of the sum of within-cluster scatter to between cluster separation. The best clustering, for $c = c_0$, minimizes

$$\frac{1}{c}\sum_{k=1}^{c}\max_{l\neq k}\left\{\frac{S_a(U_k) + S_a(U_l)}{d_a(U_k, U_l)}\right\}, \tag{14}$$

for $1 \leqslant k,\ l \leqslant c$. Here the within-cluster scatter is minimized and the between cluster separation is maximized. The index is expressed in the symbolic framework.

3.3. Dunn's index

Like Davies–Bouldin index, Dunn's index (Bezdek and Pal, 1998) is designed to identify sets of clusters that are compact and well separated. We maximize

$$\min_k\left\{\min_{l\neq k}\left\{\frac{d_a(U_k, U_l)}{\max_j S_a(U_j)}\right\}\right\}, \tag{15}$$

for $1 \leqslant j,k,\ l \leqslant c$. Here also we maximize the symbolic intercluster separation, while minimizing symbolic intracluster distances.

3.4. Cluster indicator

The cluster indicator value at the tth iteration is defined as

$$\mathrm{CI}_t = \frac{R_t}{R_{t+1}}, \tag{16}$$

where

$$R_t = \frac{\min_{k\neq l} d_a^t(U_k, U_l)}{\min_{k'\neq l'} d_a^{t+1}(U_{k'}, U_{l'}) + \min_{k''\neq l''} d_a^{t-1}(U_{k''}, U_{l''})} \tag{17}$$

This is maximized over the iterations. Note that initially, at $t = 0$, there are N singletons. At $t = 1$, the pair of closest clusters are merged, resulting in $N - 1$ clusters. Therefore, at the tth iteration we have $N - t$ clusters.

4. Results

The symbolic clustering algorithm was applied to four sets of symbolic data (including some quantitative features), viz., the *Microcomputer* data (Michalski and Stepp, 1983) and the benchmark data sets, viz., *Zoo*, *Auto Import* and *Mushroom* (Blake and Merz, 1998). Since the objective was to perform unsupervised classification, hence any class information was eliminated from the data. The results are compared with that of conceptual clustering, and the measures *Cohesiveness* and *Distinctness* of Eqs. (10) and (11) are used for this purpose.

Tables 2–5 enumerate the clustering results for the four data sets. The validity indices used, the number of elements in each cluster (in parentheses), and the individual elements according to their sequential order of entry in the corresponding cluster (while optimizing the different validity

indices of Section 3) are provided. The quantitative evaluation indices *Cohesiveness* and *Distinctness*, for the resulting partitions, are also provided. Plots for Normalized Hubert's statistic are provided in Figs. 1 and 2.

4.1. Microcomputer data

Table 1 shows the *Microcomputer* data (Michalski and Stepp, 1983) used. It consists of 12 patterns with two symbolic attributes, viz., display and MP, two interval-type attributes, viz., ROM size and keys, and one numeric attribute, i.e., RAM size.

Table 2 provides the results of the symbolic clustering obtained, over the microcomputer data, using the different validity indices. Davies–Bouldin index, CI and Dunn's index generate four clusters. Cohesion and Distinctness for the two cases are computed as 25.8, 1.2 (four clusters) and 18.84,

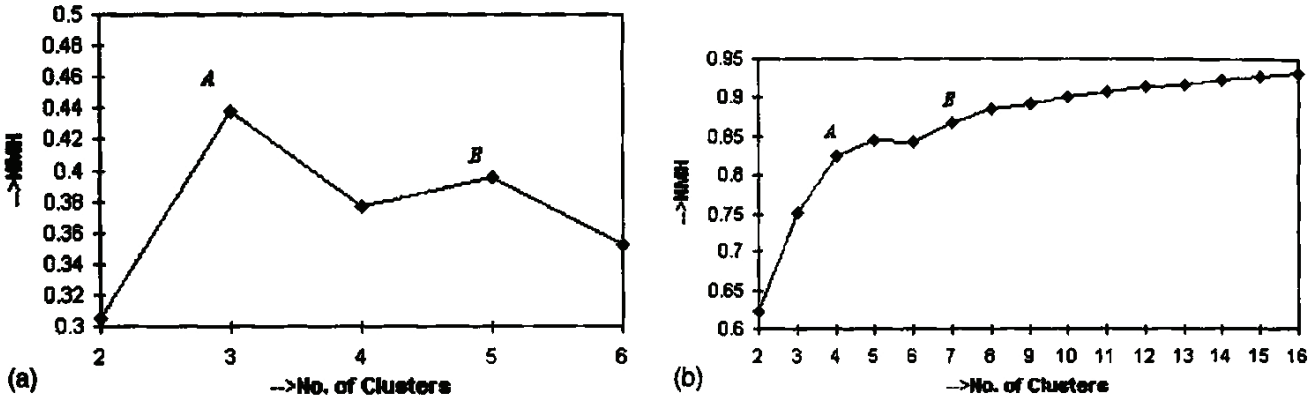

Fig. 1. Plot of Hubert's statistic for (a) *Microcomputer* and (b) *Zoo* data.

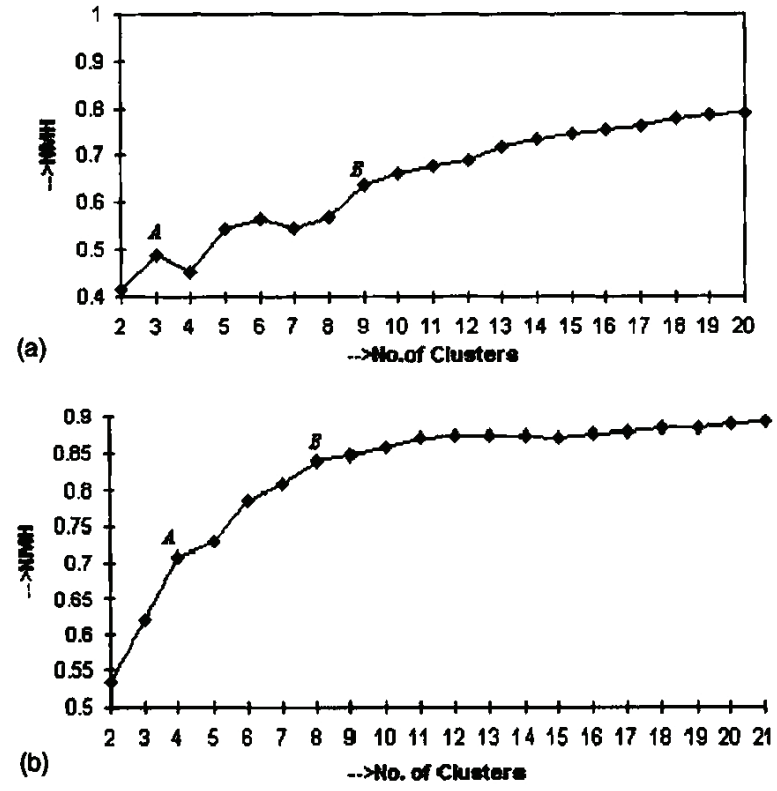

Fig. 2. Plot of Hubert's statistic for (a) *Auto Import* and (b) *Mushroom* data.

Table 1
Microcomputer data

Microcomputer	Display	RAM (K)	ROM (K)	MP	Keys
Apple II	Color TV	48	10	6502	52
Atari 800	Color TV	48	10	6502	57–63
Commodore VIC 20	Color TV	32	11–16	6502A	64–73
Exidi Sorcerer	B&W TV	48	4	Z80	57–63
Zenith H8	Built-in	64	1	8080A	64–73
Zenith H89	Built-in	64	8	Z80	64–73
HP-85	Built-in	32	80	HP	92
Horizon	Terminal	64	8	Z80	57–63
Ohio Sc. Challenger	B&W TV	32	10	6502	53–56
Ohio Sc. II Series	B&W TV	48	10	6502C	53–56
TRS-80 I	B&W TV	48	12	Z80	53–56
TRS-80 III	Built-in	48	14	Z80	64–73

Table 2
Symbolic clustering and evaluation of *Microcomputer* data

Validity index	Cluster no.	Microcomputer make
Davies–Bouldin,	1 (5)	Apple II, Atari 800, Commodore VIC 20, Ohio Sc. Challenger, Ohio Sc. II Series
CI and Dunn's	2 (3)	Exidi Sorcerer, TRS-80 I, Horizon
	3 (3)	Zenith H8, Zenith H89, TRS-80 III
	4 (1)	HP-85
Normalized Modified Hubert's	1 (5)	Apple II, Atari 800, Commodore VIC 20 Ohio Sc. Challenger, Ohio Sc. II Series
and Conceptual	2 (3)	Exidi Sorcerer, TRS-80 I, Horizon
	3 (4)	Zenith H8, Zenith H89, TRS-80 III, HP-85

0.85 (three clusters) respectively. It is observed that the partitions are meaningful on the basis of the features *display* and MP. Normalized modified Hubert's statistic and conceptual clustering provide three clusters by combining elements from partitions 3 and 4, based on feature *display*. The plot of Hubert's statistic for the data is provided in Fig. 1(a). It is observed that a sharp *knee* occurs for three clusters. By comparing the Cohesion and Distinctness measures it is found that better partitioning is obtained with four clusters.

4.2. Zoo data

The *Zoo* data (Blake and Merz, 1998) consists of 100 instances of animals with 17 features and 7 output classes. The name of the animal constitutes the first attribute. There are 15 boolean features corresponding to the presence of hair, feathers, eggs, milk, backbone, fins, tail; and whether airborne, aquatic, predator, toothed, breathes, venomous, domestic, catsize. The character attribute corresponds to the number of legs lying in the set $\{0, 2, 4, 5, 6, 8\}$.

The clustering algorithm provided four clusters for the *Zoo* data with validity indices CI, Normalized Modified Hubert's, Davies–Bouldin and Dunn's, while conceptual clustering generated two clusters (merging clusters 2–4 into cluster 2 here). The first case is enumerated in Table 3. Cohesion and Distinctness for the two cases are computed as 458.43, 0.58 (four clusters) and 132.49, 0.50 (two clusters) respectively. It is observed that the resulting partitions in the first case (proposed symbolic clustering) are semantically meaningful, and very similar to those obtained by Kohonen's self-organizing feature map (Alahakoon et al., 2000). Conceptual clustering generates just two partitions, based on the attribute 'presence of milk'. On the other hand, the proposed method achieves finer partitioning of cluster 2 depending on the other attributes. For example, clusters 2 and 3 are

Table 3
Symbolic clustering and evaluation of *Zoo* data

Index	Cluster no.	Animals
CI, Normalized Modified Hubert's, Davies–Bouldin and Dunn's	1 (41)	Aardvark, bear, girl, boar, cheetah, leopard, lion, raccoon, wolf, lynx, mongoose, polecat, puma, mink, platypus, dolphin, porpoise, seal, sealion, antelope, buffalo, deer, elephant, giraffe, oryx, gorilla, wallaby, calf, goat, pony, reindeer, pussycat, cavy, hamster, fruitbat, vampire, squirrel, hare, vole, mole, opossum
	2 (21)	Bass, catfish, piranha, chub, herring, carp, haddock, seahorse, sole, dogfish, pike, tuna, stingray, frog, toad, newt, tuatara, pitviper, slowworm, scorpion, seasnake
	3 (21)	Chicken, dove, parakeet, lark, pheasant, sparrow, wren, flamingo, ostrich, tortoise, crow, hawk, vulture, kiwi, rhea, penguin, duck, swan, gull, skimmer, skua
	4 (17)	Clam, seawasp, crab, starfish, crayfish, lobster, octopus, flea, termite, slug, worm, gnat, ladybird, housefly, moth, honeybee, wasp

distinguishable based on features 'presence of feathers', 'whether airborne' and 'whether aquatic'. Plot of Hubert's statistic for the data, in Fig. 1(b), shows a sharp knee four clusters. The cohesion and Distinctness measures indicate better partitioning for four clusters.

4.3. Auto Import data

The *Auto Import* data (Blake and Merz, 1998) uses 193 instances with 24 features. There are 14 continuous/quantitative (wheel-base, length, width, height, curb-weight, engine-size, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg, price) and 10 nominal (make, fuel-type, aspiration, number-of-doors, body-style, drive-wheels, engine-location, engine-type, number-of-cylinders, fuel-system) attributes. Each quantitative attribute is discretized to five quantiles (Davies and Yoder, 1937). [1]

The clustering results are provided in Table 4. Cohesion and Distinctness for the two cases are computed as 530.67, 0.39 (three clusters) and 309.50, 0.33 (two clusters) respectively. It is observed from the results that cluster 1 contains cars with high price (>12,500), high curb-weight (>2600), high engine-size (>127), high compression-ratio (>105), low highway-mpg (<28), low city-mpg (<24) and high horsepower (>105). On the other hand, cluster 2 consists of cars having low price (<9000), low curb-weight (<2400), low engine-size (<98), low compression-ratio (<84), high highway-mpg (>32), high city-mpg (>27) and low horsepower (<84). While cluster 1 contains cars with fuel-system *mfi*, *mpfi*, *spdi*, *spfi* (*bbl*) and number-of-cylinders $\geqslant 6$, cluster 2 constitutes those with fuel-system 1, 2 (*bbl*) and number-of-cylinders $\leqslant 4$. Cluster 3 can be distinguished from cluster 1 on the basis of high height (>55).

Normalized modified Hubert's statistic and Dunn's index generate three partitions. Davies–Bouldin index and conceptual clustering lead to two clusters, merging partitions 1 and 3 in the process. CI does not provide meaningful partitions. Plot of Hubert's statistic for the data set is provided in Fig. 2(a). The Cohesion and Distinctness measures show a better partitioning for three clusters.

4.4. Mushroom data

The *Mushroom* data (Blake and Merz, 1998) consists of 8124 instances with 21 nominal attributes and two output categories (poisonous (p)/edible (e)). The input attributes correspond to cap-shape, cap-surface, cap-color, whether bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, veil-type, veil-color, ring-number, ring-type, spore-print-color, population-type and habitat.

Table 5 provides the clustering results. Plot of Hubert's statistic for the data set is provided in

[1] Quantiles are the values of a variate which divide the total frequency into a number of equal parts.

Table 4
Symbolic clustering and evaluation of *Auto Import* data

Index	Cluster no.	Car make
Normalized Modified Hubert's and Dunn's	1 (56)	Alfa-romeo, toyota, bmw, isuzu, mazda, dodge, plymouth, mitsubishi, mercury, porsche, nissan, jaguar, mercedes-benz
Conceptual	2 (105)	Audi, volkswagen, nissan, toyota, dodge, plymouth, mitsubishi, mazda, subaru, chevrolet, honda
	3 (32)	Mercedes-benz, peugeot, saab, volvo
Davies–Bouldin and Conceptual	1 (88)	Alfa-romeo, toyota, bmw, isuzu, mazda, dodge, plymouth, mitsubishi, mercury, porsche, nissan, jaguar, mercedes-benz, peugeot, saab, volvo
	2 (105)	Audi, volkswagen, nissan, toyota, dodge, plymouth, mitsubishi, mazda, subaru, chevrolet, honda

Table 5
Symbolic clustering and evaluation of *Mushroom* data

Index	# Clusters	Cohesion	Distinctness
Davies–Bouldin	10	3132.8	0.64
Norm. Modified Hubert's and conceptual	4	2079.6	0.57
Dunn's	2	924.0	0.38

Fig. 2(b). It is observed that there is a sharp knee for four clusters. In case of Davies–Bouldin index we obtained four purely edible groups, three purely poisonous groups and three mixed groups of mushrooms. The edible groups are clustered depending on whether bruises exist, and different gill-spacing, stalk-shapes, ring-number and habitat. The poisonous groups are clustered depending on different gill-size, gill-color, stalk-shape, ring-type and spore-print-color. The partitioning generated by Normalized Hubert's statistic and conceptual clustering results in two purely poisonous groups and and two mixed groups. Dunn's index generates one purely poisonous and one mixed group. According to the Cohesion and Distinctness measures, Davies–Bouldin index generates the best clustering.

5. Conclusions and discussion

Real life data is essentially not restricted to the numeric domain. Hence the need for symbolic clustering to efficiently handle data like linguistic, nominal, boolean, interval, shape, color, etc., arises. Partitioning of such data demands the use of symbolic measures for determining the similarity and dissimilarity between objects. In this article we have studied the effectiveness of symbolic clustering on several benchmark symbolic data, viz., *Microcomputer*, *Zoo*, *Auto Import* and *Mushroom*.

We have used intercluster and intracluster scatter in the symbolic framework. Different clustering validity indices have been modified to incorporate the symbolic computations using dissimilarity measures. The resultant optimal clusters are found to be stable for the different validity indices used, viz., Normalized Hubert's statistic, Davies–Bouldin Index, Dunn's index and cluster indicator. The generated clusters are also observed to be *naturally* meaningful for the symbolic data used. Comparative studies are made with conceptual clustering using quantitative indices Cohesion and Distinctness. The validity indices, expressed here in the symbolic framework, are generally found to provide better partitioning as compared to conceptual clustering.

It is observed from the plots of Normalized Modified Hubert's statistic that *Zoo* and *Mushroom* data, being purely symbolic, exhibit almost similar behavior with a single knee between A and B lying between cluster numbers four-seven and four-eight respectively. Here A corresponds to the

point where steepness of the curve begins to fall, while B indicates the point from where stability of the curve begins. On the other hand, *Auto Import* and *Microcomputer* data depict a different behavior perhaps because of the presence of a mixture of symbolic as well as numeric attributes. For example, there exists multiple knees between A and B in Figs. 1(a) and 2(a), representing cluster numbers three to five for *Microcomputer* data and three to nine, for *Auto Import* data. This will be an issue for future investigation.

Clustering has useful applications in data mining, pattern recognition, image segmentation, rule extraction and web mining. The importance of symbolic clustering in real world data is all the more evident, considering the availability of large volumes of mixed-media data that are distributed over the Internet.

The determination of the optimal number of clusters is an open problem. We have tried to attack this issue in the symbolic framework, using different clustering validity indices. In all cases, some objective function is required to be optimized. The number of meaningful clusters selected depends on the application domain. For example, if one desires plain clustering/segmentation then one should go for coarser granules. On the other hand, if the goal is data condensation for data mining then one should concentrate on finer granules/clusters as representative points.

References

Alahakoon, D., Halgamuge, S.K., Srinivasan, B., 2000. Dynamic self-organizing maps with controlled growth for knowledge discovery. IEEE Trans. Neural Networks 11 (3), 601–614.

Ben-Hur, A., Elisseeff, A., Guyon, I., 2002. A stability based method for discovering structure in clustered data. In: Altman, R.B., Dunker, A.K., Hunter, L., Lauderdale, K., Klein, T.E. (Eds.), Pacific Symposium on Biocomputing. World Scientific, pp. 6–17.

Bezdek, J.C., Pal, N.R., 1998. Some new indexes for cluster validity. IEEE Trans. Systems Man Cybernet.—Part-B 28, 301–315.

Biswas, G., Weinberg, J.B., Fisher, D.H., 1998. ITERATE: A conceptual clustering algorithm for data mining. IEEE Trans. Systems Man Cybernet.—Part C: Appl. Rev. 28, 219–230.

Blake, C.L., Merz, C.J., 1998. UCI repository of machine learning databases, 1998. University of California, Irvine, Dept. of Information and Computer Sciences. Available from <http://www.ics.uci.edu/~mlearn/MLRepository.html>.

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

Chidananda Gowda, K., Ravi, T.V., 1995. Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity. Pattern Recognition 28 (8), 1277–1282.

Davies, G.R., Yoder, D., 1937. Business Statistics. John Wiley & Sons, Inc., London.

Fridlyand, J., Dudoit, S., 2001. Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. Technical Report 600, UC Berkeley, September 2001.

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice Hall, NJ.

Lange, T., Braun, M.L., Roth, V., Buhmann, J.M., 2002. Stability-based model selection. In: Advances in Neural Information Processing Systems.

Mali, K., Mitra, S., 2002. Clustering of symbolic data and its validation. In: Pal, N.R., Sugeno, M. (Eds.), Advances in Soft Computing—AFSS 2002. In: Lecture Notes in Artificial Intelligence. Springer Verlag, Heidelberg, pp. 339–344.

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

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179.

Tibshirani, R., Walther, G., Botstein, D., Brown, P., 2001. Cluster validation by prediction strength. Technical report, Stanford.

Wilson, R., Martinez, T.R., 1997. Improved heterogeneous distance functions. J. Artificial Intelligence Res. 6, 1–34.