Clustering of Heterogeneously Typed Data with Soft Computing - A case Study

Angel Kuri-Morales[1], Daniel Trejo-Baños[2] and Luis Enrique Cortes-Berrueco[2]
[1] *Instituto Tecnológico Autónomo de México, Río Hondo No. 1 México D.F. México*
[2] *Universidad Nacional Autónoma de México, Apartado Postal 70-600, Ciudad Universitaria,México D.F., México*
akuri@itam.mx, { l.cortes,d.trejo}@ uxmcc2.iimas.unam.mx

ABSTRACT
The problem of finding clusters in arbitrary sets of data has been attempted using different approaches. In most cases, the use of metrics in order to determine the adequateness of the said clusters is assumed. That is, the criteria yielding a measure of quality of the clusters depends on the distance between the elements of each cluster. Typically, one considers a cluster to be adequately characterized if the elements within a cluster are close to one another while, simultaneously, they appear to be far from those of different clusters. This intuitive approach fails if the variables of the elements of a cluster are not amenable to distance measurements, i.e., if the vectors of such elements cannot be quantified. This case arises frequently in real world applications where several variables (if not most of them) correspond to categories. The usual tendency is to assign arbitrary numbers to every category: to encode the categories. This, however, may result in spurious patterns: relationships between the variables which are not really there at the offset. It is evident that there is no truly valid assignment which may ensure a universally valid numerical value to this kind of variables. But there is a strategy which guarantees that the encoding will, in general, not bias the results. In this paper we explore such strategy. We discuss the theoretical foundations of our approach and prove that this is the best strategy in terms of the statistical behavior of the sampled data. We also show that, when applied to a complex real world problem, it allows us to generalize soft computing methods to find the number and characteristics of a set of clusters. We contrast the characteristics of the clusters gotten from the automated method with those of the experts.

Keywords. Clustering, Categorical variables, Soft computing, Data mining.

1 INTRODUCTION

1.1 Clustering

Clustering can be considered the most important unsupervised learning problem. As every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. In this particular case it is of relevance because we attempt to characterize sets of arbitrary data trying not to start from preconceived measures of what makes a set of characteristics relevant. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”.

A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

When the similarity criterion is distance two or more objects belong to the same cluster if they are “close” according to a given distance. This is called distance-based clustering. Another kind of clustering is conceptual clustering where two or more objects belong to the same cluster if this one defines a concept common to all those objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures [1,2,7,9]. Our contention is that conceptual clustering leads to biased criteria which have lead to the unsuccessful generalization properties of the models proposed in the past.

1.2 The Need to Encode

In recent years there has been an increasing interest to analyze categorical data in a data warehouse context where data sets are rather large and may have a high number of categorical dimensions [4,6,8,15]. However, many traditional techniques associated to the exploration of data sets assume the attributes have continuous data (covariance, density functions, PCA, etc.). In order to use these techniques, the categorical attributes have to be discarded, although they are potentially loaded with valuable information. With our technique the categorical attributes are encoded into numeric values in such a way that spurious correlations are avoided and the data can be handled as if it were numeric.

In [5] the authors propose a framework designed for categorical data analysis that allows the exploration of this kind of data with techniques that are only applicable to continuous data sets. By means of what the authors call “separability statistics", e.g. matching values with instances in a reference data set, they map any collection of categorical instances to a multidimensional continuous space. This way, instances similar to a reference data set, that could be the original dataset itself, will occupy the same region occupied by instances from the reference dataset and instances that are different will tend to occupy other regions. This mapping enables visualizing the categorical data using techniques that are applicable to continuous data. Their framework can be used in the context of several data mining tasks such as outlier detection, clustering and classification. In [3], the authors show how the choice of a similarity measure affects performance. By contrast, our encoding technique maps the categorical data to a numerical domain. The mapping is done avoiding the transmission of spurious correlations to the corresponding encoded numerical data. Once the data is numerically encoded, techniques applicable to continuous data can be used.

Following a different approach, in [11] the authors propose a distance named “distance hierarchy", based on concept hierarchies [10] extended with weights, in order to measure the distance between categorical values. This type of measure allows the use of data mining techniques based on distances, e.g. clustering techniques, when dealing with mixed data, numerical and categorical. With our technique, by encoding categorical data into numeric values, we can use then the traditional distance computa-

tions avoiding the need to figure out different ways to compute distances. Another approach is followed in [13]. The authors propose a measure in order to quantify dissimilarity of objects by using distribution information of data correlated to each categorical value. They propose a method to uncover intrinsic relationship of values by using a dissimilarity measure referred to as Domain Value Dissimilarity (DVD). This measure is independent of any specific algorithm so that it can be applied to clustering algorithms that require a distance measure for objects. In [14] the authors present a process for quantification (i.e. quantifying the categorical variables - assigning order and distance to the categories) of categorical variables in mixed data sets, using Multiple Correspondence Analysis, a technique which may be seen as the counterpart of principal component analysis for categorical data. An interactive environment is provided, in which the user is able to control and influence the quantification process and analyze the result using parallel coordinates as a visual interface. For other possible clustering methods the reader is referred to [12,16,17,18,24].

2 UNBIASED ENCODING OF CATHEGORICAL VARIABLES

We now introduce an alternative which allows the generalization of numerical algorithms to encompass categorical variables. Our concern is that such encoding:

a) Does not induce spurious patterns

b) Preserves legal patters, i.e. those present in the original data.

By "spurious" patterns we mean those which may arise by the artificial distance induced by our encoding. On the other hand, we do not wish to filter out those patterns which are present in the categories. If there is an association pattern in the original data, we want to preserve this association and, furthermore, we wish to preserve it in the same way as it presents itself in the original data. The basic idea is simple: "Find the encoding which best preserves a measure of similarity between all numerical and categorical variables".

In order to do this we start by selecting Pearson's correlation as a measure of linear dependence between two variables. Higher order dependencies will be hopefully found by the clustering algorithms. This is one of several possible alternatives. The interested reader may see [25,26]. Its advantage is that it offers a simple way to detect simple linear relations between two variables. Its calculation yields "r", Pearson's correlation, as follows:

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}} \tag{1}$$

Where variables X and Y are analyzed to search their correlation, i.e. the way in which one of the variables changes (linearly) with relation to the other. The values of "r" in (1) satisfy $-1 \le r \le +1$. What we shall do is search for a code for categorical variable A such that the correlation calculated from such encoding does not yield a

significant difference with any of the possible encodings of all other categorical or numerical variables.

2.1 Exploring the correlations

To exemplify let us assume that our data consists of only 10 variables. In this case there are 5,000 objects (or 10-dimensional vectors) in the data base. A partial view is shown in figure 1. Notice that two of the variables (V006 and V010) are categorical, whereas the rest are numerical.

A	B	C	D	E	F	G	H	I	J
V001	V002	V003	V004	V005	V006	V007	V008	V009	V010
0.00009070	0.00000001	1.00003023	0.00952325	-2.00000004	NEW HAMPSHIRE	0.41161556	0.20441124	-1.82342981	B
0.00025888	0.00000007	1.00008630	0.01608987	-2.00000034	ARKANSAS	-2.64698978	1.81799816	-0.82730115	C
0.00029245	0.00000009	1.00009748	0.01710103	-2.00000043	MONTANA	0.02173054	0.58433958	-1.92997804	E
0.00037022	0.00000014	1.00012341	0.01924119	-2.00000069	NEW HAMPSHIRE	-2.06263802	1.52399200	-1.53649553	C
0.00050723	0.00000026	1.00016908	0.02252174	-2.00000129	NEW YORK	0.41995754	0.22398177	-1.81276765	C
0.00070027	0.00000049	1.00023342	0.02646256	-2.00000245	CALIFORNIA	0.32585443	0.41401481	-1.80981725	C
0.00193430	0.00000374	1.00064477	0.04398064	-2.00001870	ARKANSAS	0.12610259	0.53361534	-1.88908012	D
0.00203758	0.00000415	1.00067919	0.04513957	-2.00002075	OREGON	0.41331723	0.20784753	-1.82143947	C
0.00212120	0.00000450	1.00070707	0.04605651	-2.00002249	ALASKA	0.37132492	0.37425889	-1.79485253	A
0.00231738	0.00000537	1.00077246	0.04813918	-2.00002684	ARKANSAS	-0.98345777	1.02154751	-2.09077872	A
0.00236024	0.00000557	1.00078675	0.04858231	-2.00002784	ALASKA	-2.03182318	1.50889034	-1.56527039	D
0.00238070	0.00000567	1.00079357	0.04879237	-2.00002833	MICHIGAN	-1.32471362	1.17449911	-2.01038178	A
0.00244130	0.00000596	1.00081377	0.04940947	-2.00002979	NEW MEXICO	-0.17984430	0.67499378	-2.00024213	F
0.00251393	0.00000632	1.00083798	0.05013913	-2.00003158	NEW HAMPSHIRE	0.24048391	0.05827254	-1.94235602	E
0.00268335	0.00000720	1.00089445	0.05180106	-2.00003598	NEW YORK	-1.00370495	1.03047794	-2.08821147	E
0.00287404	0.00000826	1.00095801	0.05361005	-2.00004128	NEW YORK	0.16833514	0.02837119	-1.97170118	D

Fig.1. Mixed type data.

We define the *i-th instance* of a categorical variable V_X as one possible value of variable X. For example, if variable V006 takes 28 different names, one instance is "NEW HAMPSHIRE", another instance is "ARKANSAS" and so on. We denote the number of variables in the data as V. Further, we denote with r_{ik} Pearson's correlation between variables i and k. We would like to a) Find the mean μ of the correlation's probability distribution for all categorical variables by analyzing all possible combinations of codes assignable to the categorical variables (in this example V006 and V010) plus the original (numerical) values of all non-categorical variables. b) Select the codes for the categorical variables which yield the closest value to μ. The rationale is that the absolute typical value of μ is the one devoid of spurious patterns *and* the one preserving the legal patterns. In the algorithm to be discussed next the following notation applies:

```
N    ← number of elements in the data
V    ← number of categorical variables
V[i] ← the i-th variable
Ni   ← number of instances of V[i]
```

$\bar{r}_j$ ← the mean of the j-th sample

```
s    ← sample size of a mean
```

$\mu_{\bar{r}}$ ← mean of the correlation's distribution of means

$\sigma_{\bar{r}}$ ← standard deviation of the correlation's distribution of means

Algorithm A1.
Optimal Code Assignment for Categorical Variables

```
01  for i=1 to V
02    j ← 0
03    do while r̄_j is not distributed normally
04      for k=1 to S
05        Assign a code for variable V[i]
06        Store this code
07        ℓ ← integer random number (1≤ ℓ ≤ V;  ℓ ≠i)
08        if variable V[ℓ] is categorical
09          Assign a code for variable V[ℓ]
10        endif
11        r_k = [N∑XY − ∑X∑Y] / √{[N∑X² − (∑X)²][N∑Y² − (∑Y)²]}
12      endfor
13      Calculate r̄_j ← = (1/S) Σ_{k=1}^{S} r_k
14      j ← j+1
15    enddo
16    μ = μ_r̄ ; the mean of the correlation's distribution
17    σ = √SS · σ_r̄ ; the std. dev. of the correlation's distribution
18    Select the code for V[i] which yields the r_k closest to μ
19  endfor
```

For simplicity, in the formula of line (11), X stands for variable V[i] and Y stands for variable V[ℓ]. Of course it is impossible to consider all codes, let alone all possible combinations of such codes. Therefore, in algorithm A1 we set a more modest goal and adopt the convention that to *Assign a Code* [as in lines (05) and (09)] means that we restrict ourselves to the combinations of integers between 1 and Ni (recall that *Ni* is the number different values of variable *i* in the data). Still, there are Ni! possible ways to assign a code to categorical variable *i* and Ni! x Nj! possible encodings of two categorical variables *i* and *j*. An exhaustive search is, in general, out of the question. Instead, we take advantage of the fact that, regardless of the way a random variable distributes (here the value of the random encoding of variables *i* and *j* results in correlation r_{ij} which is a random variable itself) the *means* of sufficiently large samples very closely approach a normal distribution. Furthermore, the mean value of a sample of means $\mu_{\bar{r}}$ and its standard deviation $\sigma_{\bar{r}}$ are related to the mean μ and standard

deviation σ of the original distribution by $\mu = \mu_{\bar{r}}$ and $\sigma = \sqrt{SS} \cdot \sigma_{\bar{r}}$. What a *sufficiently large* sample means is a matter of convention and here we made S=25 which is a reasonable choice. Therefore, the loop between lines (03) and (15) is guaranteed to end. In our implementation we split the area under the normal curve in deciles and then used a χ^2 goodness-of-fit test with p=0.05 to determine that normality has been achieved. This approach is directed to avoid arbitrary assumptions regarding the correlation's distribution and, therefore, not selecting a sample size to establish the reliability of our results. Rather, the algorithm determines at what point the proper value of μ has been reached. Furthermore, from Chebyshev's theorem, we know that

$$P(\mu - k\sigma \le X \le \mu + k\sigma) \ge 1 - \frac{1}{k^2} \tag{2}$$

If we make k=3 and assume a symmetrical distribution, the probability of being within three σ's of the mean is roughly 0.95. We ran our algorithm for the data of the example and show in figure 5 the values that were obtained.

Parameters of the Population

Variable #	Mu_r	Sigma_r	Mu	Sigma	Minimum R @95%	Maximum R @95%
6	-0.0028739	0.0040739	-0.0028739	0.0191084	-0.0601990	0.0544512
10	0.0006013	0.0027945	0.0006013	0.0131073	-0.0387207	0.0399232

Fig. 2. Values of categorical encoding for variables 6 and 10.

In the program corresponding to figure 2, Mu_r and Sigma_r denote the mean and standard deviation of the distribution of means; Mu and Sigma denote the corresponding parameters for the distribution of the correlations and the titles "Minimum R @95%" and "Maximum R@95%" denote the smallest and largest values at ±3 σ's from the mean. In this case, the typical correlation is close to zero, denoting no first order patterns in the data. With probability 0.95 the typical correlation for variable 6 lies in an interval of size 0.1147 while the corresponding value for variable 10 lies in an interval of size 0.0786. Three other issues remain to be clarified.

1) To *Assign a code* to V[i] means that we generate a sequence of numbers between 1 and Ni and then randomly assign a one of these numbers to every different instance of V[i].

2) To *Store the code* [as in line (06)] means NOT that we store the assigned code (for this would imply storing a large set of sequences). Rather, we store the value of the calculated correlation along with the root of the pseudo random number generator from which the assignment was derived.

3) Thereafter, selecting the best code (i.e. the one yielding a correlation whose value is closest to μ) as in line (18) is a simple matter of recovering the root of the pseudo random number generator and regenerating the original random sequence from it.

3 CASE STUDY: PROFILE OF THE CAUSES OF DEATH OF A POPULATION

In order to illustrate our method we analyzed a data base corresponding to the life span and cause of death of 50,000 individuals between the years of 1900 and 2007. The confidentiality of the data has been preserved by changing the locations and regions involved. Otherwise data are a faithful replica of the original.

3.1 The data base

This experiment allowed us to compare the interpretation of the human experts with the one resulting from our analysis. The database contains 50,000 tuples consisting of 11 fields: BirthYear, LivingIn, DeathPlace, DeathYear, DeathMonth, DeathCause, Region, Sex, AgeGroup, AilmentGroup and InterestGroup. A very brief view of 8 of the 11 variables is shown in figure 3.

BirthYear	LivingIn	DeathPlace	DeathYear	DeathMonth	DeathCause	Region	Sex
2005	MONTANA	MONTANA	2005	3	Not identified	NORTHWEST	F
1932	WASHINGTON	WASHINGTON	1997	5	Stroke	NORTHWEST	M
2006	NEVADA	NEVADA	2006	2	Malign lung neoplasia	SOUTHWEST	F
2005	FLORIDA	FLORIDA	2005	5	Stroke	SOUTHEAST	M
1959	TEXAS	TEXAS	1988	12	Stroke	SOUTHWEST	M
1946	WYOMING	WYOMING	2003	5	Stroke	NORTHWEST	F
1942	WASHINGTON	WASHINGTON	1997	12		NORTHWEST	F
1906	TEXAS	TEXAS	1984	9	Stroke	SOUTHWEST	F
1901	GEORGIA	GEORGIA	1992	9	No Identificado	SOUTHWEST	F
1943	NEW MEXICO	NEW MEXICO	1997	12	Malign breast neoplasia	SOUTHWEST	F

Fig. 3. Partial view of the data base.

The last variable (InterestGroup) corresponds to interest groups identified by human healthcare experts in this particular case. This field corresponds to a heuristic clustering of the data and will be used for the final comparative analysis of resulting clusters. It will not be included either in the data processing nor the data mining activities. Therefore, our working data base has 10 dimensions.

The first thing to notice is that there are no numeric variables. BirthYear, DeathYear and DeathMonth are dates (clearly, they represent the date of birth, year and month of death respectively). "Region" represents the place where the death took place. DeathCause and AilmentGroup are the cause of death and the illness group to which the cause of death belongs.

3.2 Preprocessing the Information

In order to process the information contained in the data base we followed the next methodology:

- At the offset we applied algorithm A1 and, once the coding process was finished we got a set of 10 codes, each code with a number of symbols corresponding to the cardinality of the domain of the variable.
- Each column of the data base is encoded.

- We get the correlation between every pair of variables. If the correlation between two columns is large only one of them is retained.
- We assume no prior knowledge of the number of the clusters and, therefore, resorted to the Fuzzy C-Means algorithm and the elbow criterion to determine it [see 19, 20]. For a sample of K objects divided in c classes (where μ_{ik} is the membership of an object k to class i) we determine the partition coefficient (pc) and the partition entropy (pe) from formulas (3) and (4) respectively [see 21, 22, 23].

$$pc = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{c} \mu_{ik} \tag{3}$$

$$pe = -\frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{c} \mu_{ik} ln(\mu_{ik}) \tag{4}$$

3.3 Processing the Information

We process the information with two unsupervised learning techniques: Fuzzy c-means and Kohonen's SOM.

There is only one difference in the pre-process phase. For the Kohonen's SOM case a filtering of the data set was conducted. It was found that in several tuples the death date precedes birth date resulting in an inconsistent representation of reality. The data set was scanned and all the cases presenting the error were deleted. As a result of this action the original set was reduced from 500,000 tuples to 485,289.

In both cases the categorical data was encoded to numbers, we obtained the correlation between the variables, Figure 4 presents the correlation matrix.

The largest absolute correlation does not exceed 0.3. Hence, there are no strongly correlated variables. It is important to notice that the highest correlations are consistent with reality: (1,6) Birth Place – Region of the country,(5,9) Pathology – Pathology Group.

1	-	-	-	-	-	-	-	-	-
-0.0080	1	-	-	-	-	-	-	-	-
0.0050	-0.3400	1	-	-	-	-	-	-	-
0.0300	0.0030	0.0000	1	-	-	-	-	-	-
0.0000	0.0010	-0.0040	0.0000	1	-	-	-	-	-
-0.0460	-0.0090	-0.0110	-0.0020	-0.0060	1	-	-	-	-
0.0020	0.2020	-0.2230	-0.0040	-0.0030	-0.0060	1	-	-	-
0.0040	0.0080	0.0090	-0.0060	0.0010	-0.0160	0.0270	1	-	-
-0.0110	0.0460	0.0150	0.0210	-0.0040	-0.0130	-0.0090	-0.0300	1	-
0.1400	-0.0600	-0.0150	-0.0180	0.0110	0.2810	-0.2700	-0.2200	-0.2580	1

1	-	-	-	-	-	-	-	-	-
0.0010	1	-	-	-	-	-	-	-	-
0.0030	-0.0270	1	-	-	-	-	-	-	-
-0.0170	-0.0200	-0.0200	1	-	-	-	-	-	-
0.0020	0.0010	0.0000	-0.0010	1	-	-	-	-	-
-0.0160	0.0020	-0.0130	-0.0330	0.0010	1	-	-	-	-
0.0030	0.2040	0.1830	0.0070	0.0050	-0.0120	1	-	-	-
-0.0010	0.0230	0.0170	-0.0010	-0.0010	-0.0200	0.0100	1	-	-
0.0030	0.0120	0.0040	0.0240	0.0010	-0.1230	0.0400	0.0360	1	-
-0.0010	0.0030	-0.0020	-0.0020	0.0020	0.2500	0.0360	0.0390	0.0330	1

Fig. 4. Correlation Matrix (up fuzzy z-means, down Kohonen's SOM)

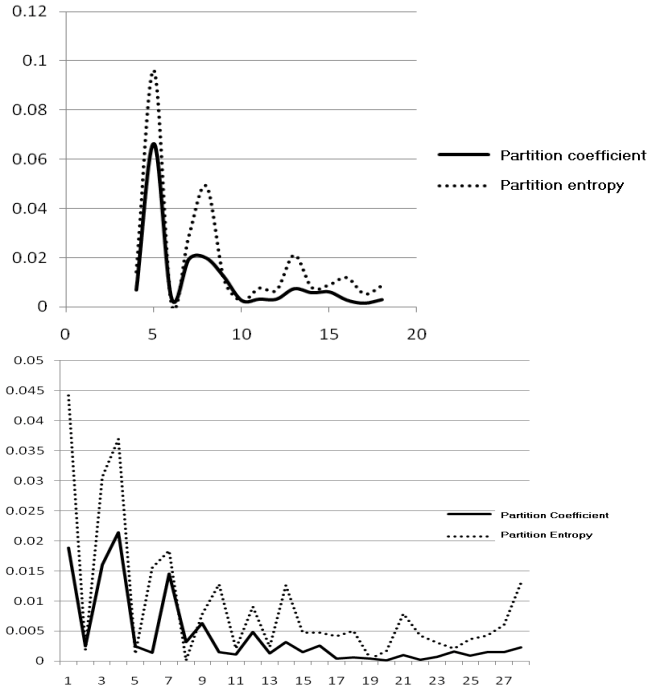

Fig. 5. Second differences graph (up fuzzy z-means, down Kohonen's SOM).

To determine the number of clusters we applied the fuzzy c-means algorithm to our coded sample. We experimented with 17 different possibilities (assuming from 2 to 18 clusters) for the fuzzy c-means case and with 30 different possibilities (from 2 to 31 clusters) for the Kohonen's SOM case. In figure 5 it is noticeable that the largest change occurs between 4 and 5 clusters for the first case and between 3 and 4 for the second case. In order to facilitate the forthcoming process we selected 4 clusters

(fuzzy c-means case) and for variety, we picked 3 clusters in the other case. This first approach may be refined as discussed in what follows.

Fuzzy c-means

Once the number of clusters is determined fuzzy c-means was applied to determine the cluster's centers. The result of the method, shown by the coordinates of the cluster centers is presented in figure 6. A brief graph showing the composition of the clusters centers can be seen in figure 10.

As can be seen in figure 7, the values for BirthYear and DeathCause are the ones that change the most within the cluster centers. An intuitive explanation is that the date of birth (and consequently the age) has had direct influence on the cause of death. The next step was a recall of the data. We grouped the tuples in one of the four classes, the one for which the tuple has the largest membership value. Now we achieve the classification of tuples on four crisp clusters. The clusters may then be analyzed individually.

C	Birth Year	Living In	Death Place	Death Year	Death Month	Death-Cause	Region	Sex	Age-Group	Ailment Group
1	19.038	15.828	17.624	16.493	6.446	62.989	2.960	0.498	10.461	5.181
2	59.085	15.730	17.685	15.223	6.432	68.0876	2.970	0.507	10.464	5.611
3	58.874	15.980	17.355	15.576	6.427	28.632	2.959	0.465	10.671	3.860
4	106.692	15.646	17.613	17.211	6.453	64.647	3.026	0.492	10.566	5.317

Fig. 6. Clusters centers

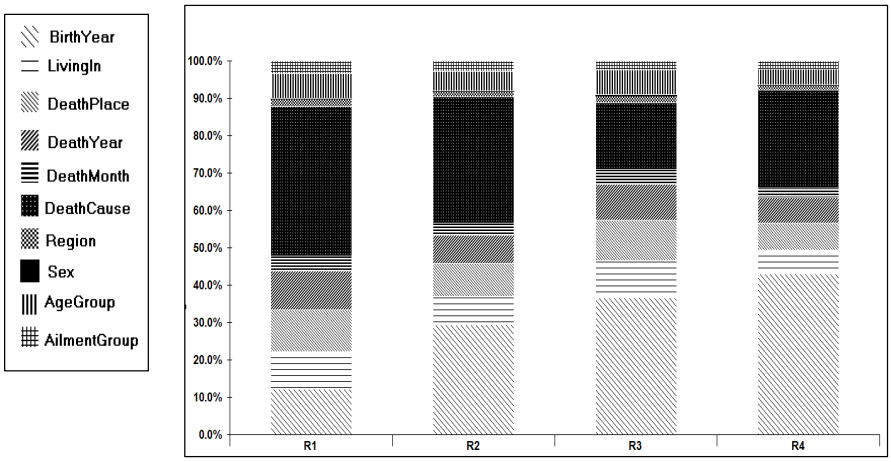

Fig. 7. Composition of the clusters

Limitations of space allow us to present, only, limited examples. In figure 8 we show the results for cluster 2. From this analysis various interesting facts were ob-

served. The values of the means tend to be very close between all clusters in all variables except BirthYear and DeathCause. Cluster 2 has a mean for BirthYear close to that of cluster 3, but the mean of DeathCause is very different. Some very brief and simple observations follow.

Cluster 2	Birth Year	Living In	Death Place	Death Year	Death Month	Death Cause	Region	Sex	Age Group	Ailment Group
Mean	58.91	15.80	17.70	15.49	6.42	72.77	3.02	0.51	10.47	5.46
Mode	52	25	20	4	7	68	3	1	11	1
Variance	146.53	97.48	73.77	112.02	13.54	201.69	2.68	0.25	5.92	16.73
S.Deviation	12.10	9.87	8.59	10.58	3.68	14.20	1.64	0.50	2.43	4.09
Range	52.00	31.00	32.00	34.00	12.00	67.00	5.00	1.00	14.00	12.00
Skewness	0.83	0.49	-1.68	1.77	-1.19	4.83	-2.98	-0.31	-10.48	1.39
Kurtosis	1.44 E+06	1.03 E+06	1.46 E+06	1.25 E+06	1.30 E+06	1.95 E+06	1.53 E+06	659795 .45	4.75 E+06	1.14 E+06

Fig. 8. Basic Statistics of cluster number two.

In cluster 1, for instance, the mode of BirthYear is 4, whose decoded value is the year 2006. The mode for DeathYear is 15 (decoded value 2008) and DeathCause corresponds to missing data. In cluster 2 the mode for BirthYear is 52, (1999), the mode for DeathCause is 68 (Diabetes type2). In cluster 3 the mode for BirthYear is 58 (2007 when decoded). For DeathCause the mode is 28 which correspondes to heart stroke. In cluster 4, the value of the mode for BirthYear is 4 (which corresponds to the year of 1900).

Kohonen's SOM

For this case we attempted to interpret the results according to the values of the mean, we rounded the said values for BirthYear and DeathCause and obtained the following decoded values:

- For cluster 1 the decoded values of the mean for BirthYear and DeathCause correspond to "1960" and cancer.
- In cluster 2 the values are "1919" and Pneumonia
- In cluster 3 the values are "1923" and Heart stroke

Interestingly, this approach seems to return more meaningful results than the mode based approach, by noting that people in different age groups die of different causes.

SOMs results were, as expected, similar the ones gotten from fuzzy C-means. However, when working with SOMs it is possible to split the clusters into subdivisions by increasing the number of neurons.

3.4 Clusters proposed by human experts.

Finally we present the general statistics for the clusters proposed by human experts as defined in the last column of the database. In the experts' opinion, there are only three clusters (see figure 9).

Cluster 1	**Birth Year**	**Living In**	**Death Place**	**Death Year**	**Death Month**	**Death Cause**	**Region**	**Sex**	**Age Group**	**Ailment Group**
Mean	33.82	15.38	16.96	15.84	5.46	37.83	1.88	0.51	10.23	7.78
Variance	313.98	54.54	81.22	78.85	11.44	423.76	1.85	0.25	13.99	43.26
Mode	4	25	20	15	7	28	3	0	11	1
S.Deviation	37.23	9.87	8.71	9.97	3.68	26.59	1.60	0.50	2.35	4.73
Range	129.00	31.00	32.00	34.00	12.00	117.00	5.00	1.00	14.00	12.00
Skewness	4.25	4.39	-10.89	5.45	-9.23	3.19	-22.26	6.35	-94.55	28.92
Kurtosis	32E+06	26E+06	36E+06	32E+06	33E+06	34E+06	40E+06	17E+06	14E+06	28E+06
Cluster 2	**Birth Year**	**Living In**	**Death Place**	**Death Year**	**Death Month**	**Death Cause**	**Region**	**Sex**	**Age Group**	**Ailment Group**
Mean	59.28	16.27	17.72	16.35	6.38	71.69	2.93	0.62	10.60	6.61
Mode	4	25	20	15	7	69	3	1	11	7
Variance	1233.71	98.22	71.74	106.25	13.60	103.20	2.84	0.24	5.32	1.34
S.Deviation	35.12	9.91	8.47	10.31	3.69	10.16	1.68	0.49	2.31	1.16
Range	129.00	31.00	32.00	34.00	12.00	114.00	5.00	1.00	14.00	12.00
Skewness	0.59	0.04	-1.66	0.82	-1.09	13.88	-2.59	-2.42	-9.23	-12.16
Kurtosis	7.6E+06	5.5E+06	8.2E+06	6.8E+06	7.0E+06	3.7E+07	7.8E+06	4.4E+06	2.6E+07	3.4E+07
Cluster 3	**Birth Year**	**Living In**	**Death Place**	**Death Year**	**Death Month**	**Death Cause**	**Region**	**Sex**	**Age Group**	**Ailment Group**
Mean	63.08	16.59	17.95	16.83	6.48	72.58	2.71	0.54	9.76	10.97
Mode	52	25	20	18	7	69	0	1	11	11
Variance	1340.82	95.95	65.18	111.74	13.25	76.32	3.23	0.25	10.33	0.24
S.Deviation	36.62	9.80	8.07	10.57	3.64	8.74	1.80	0.50	3.21	0.49
Range	129.00	31.00	32.00	34.00	12.00	95.00	5.00	1.00	14.00	11.00
Skewness	5.72 E-02	-8.21 E-02	-4.35 E-01	-9.75 E-02	-2.34 E-01	1.73 E+00	-3.63 E-01	-1.78 E-01	-1.30 E+00	-1.79 E+01
Kurtosis	1340.82	95.95	65.18	111.74	13.25	76.32	3.23	0.25	10.33	0.24

Fig. 9. Statistical characteristics of the three clusters.

In this case we note that the value of the mean changes most for BirthYear. Cluster 1 has a very different value for the mean for DeathCause than the other two clusters. The decoded values of the mode for BirthYear and DeathCause are “2008” and Heart stroke, for cluster 2 “2008” and “Unknown”, and for cluster 2 “1990” and “Unknown”. Additionally we observe also significant changes in the mean for AilmentGroup. When decoding the values of the mode in each cluster we get that for cluster 1 the mode is Trombosis (in effect a heart condition), for cluster 2 it is Diabetes type 2 and for cluster 3 it is Diabetes type 1.

4 DISCUSSION AND PERSPECTIVES

We have shown that we are able to find meaningful results by applying numerically oriented non-supervised clustering algorithms to categorical data by properly encoding the instances of the categories. We were able to determine the number of clusters arising from the data encoded according to our algorithm and, furthermore, to interpret the clusters in a meaningful way. When comparing the clusters determined by our method to those of human experts we found some coincidences. However, some of our conclusions do not match those of the experts.

Rather than assuming that this is a limitation of our method, we would prefer to suggest that machine learning techniques such as the one described, yield a broader scope of interpretation because they are not marred by limitations of processing capabilities which are evident in any human attempt to encompass a large set of data.

At any rate, the proposed encoding does allow us to tackle complex problems without the limitations derived from the non-numerical characteristics of the data. Much work remains to be done, but we are confident that these are the first of a series of significant applications.

REFERENCES

1. A. Agresti. Categorical Data Analysis. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition, 2002.
2. D. Barbará, Y. Li, and J. Couto. Coolcat: an entropy-based algorithm for categorical clustering. In CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management, pages 582-589, New York, NY, USA, 2002. ACM.
3. Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. In SDM, pages 243-254, 2008.
4. Eugenio Cesario, Giuseppe Manco, and Riccardo Ortale. Top-down parameter-free clustering of high-dimensional categorical data. IEEE Trans. on Knowl. and Data Eng., 19(12):1607-1624, 2007.
5. Varun Chandola, Shyam Boriah, and Vipin Kumar. A framework for exploring categorical data. In SDM, pages 185-196, 2009.
6. Chia-Hui Chang and Zhi-Kai Ding. Categorical data visualization and clustering using subjective factors. Data Knowl. Eng., 53(3):243-262, 2005.
7. V. Ganti, J. Gehrke, and R. Ramakrishnan. Cactus---clustering categorical data using summaries. In KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 73-83, New York, NY, USA, 1999. ACM.
8. David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Clustering categorical data: an approach based on dynamical systems. The VLDB Journal, 8(3-4):222-236, 2000.
9. S. Guha, R. Rastogi, and K. Shim. ROCK : A robust clustering algorithm for categorical attributes. In ICDE Conference, pages 512-521, 1999.
10. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco, 1st edition, 2001.
11. Chung-Chian Hsu and Sheng-Hsuan Wang. An integrated framework for visualized and exploratory pattern discovery in mixed data. IEEE Trans. on Knowl. and Data Eng., 18(2):161-173, 2006.

12. Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2(3):283-304, 1998.
13. L. Jeonghoon, L. Yoon-Joon, and P. Minho. Clustering with domain value dissimilarity for categorical data. In ICDM '09: Proceedings of the 9th Industrial Conference on Advances in Data Mining. Applications and Theoretical Aspects, pages 310-324, Berlin, Heidelberg, 2009. Springer-Verlag.
14. S. Johansson, M. Jern, and J. Johansson. Interactive quantification of categorical variables in mixed data sets. In IV '08: Proceedings of the 2008 12th International Conference Information Visualisation, pages 3-10, Washington, DC, USA, 2008. IEEE Computer Society.
15. Mehmet Koyuturk, Ananth Grama, and Naren Ramakrishnan. Compression, clustering, and pattern discovery in very high-dimensional discrete-attribute data sets. IEEE Trans. on Knowl. and Data Eng., 17(4):447-461, 2005.
16. K. Wang, C. Xu, and B. Liu. Clustering transactions using large items. In ACM CIKM Conference, pages 483-490, 1999.17. H. Yan, K. Chen, and L. Liu. Efficiently clustering transactional data with weighted coverage density. In CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management, pages 367-376, New York, NY, USA, 2006. ACM.
17. Y. Yang, X. Guan, and J. You. Clope: a fast and effective clustering algorithm for transactional data. In KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 682-687, New York, NY, USA, 2002. ACM.
18. Haykin, Simon, Neural networks: A comprehensive foundation, MacMillan,1994
19. Halkidi, Maria, Batistakis, Yannis, and Vazirgiannis, Michalis, On Clustering Validation Techniques, J. Intell. Inf. Syst. 17(2-3), volume 17, 107–145, 2001
20. Jenssen, R., Hild, KE, Erdogmus, D., Principe, J.C., and Eltoft, T., Clustering using Renyi's entropy, Neural Networks, 2003. Proceedings of the International Joint Conference on, volume 1, 523–528, 2003
21. Lee, Y., and Choi, S., Minimum entropy, k-means, spectral clustering, Neural Networks, 2004. Proceedings IEEE International Joint Conference on, volume 1, 2005.
22. Shannon, C. E., and Weaver, W., The Mathematical Theory of Communication, Scientific American, July 1949
23. Vinh, N.X., Epps, J., and Bailey, J., Information theoretic measures for clustering's comparison: is a correction for chance necessary?, Proceedings of the 26th Annual International Conference on Machine Learning, 1073–1080, 2009
24. Kohonen Teuvo, Self-organizing maps, Springer-Verlag, New York, Inc., 1999.
25. http://udel.edu/~mcdonald/statspearman.html (08/26/2011)
26. http://www.mei.org.uk/files/pdf/Spearmanrcc.pdf (09/26/2011)