1 INTRODUCTION

In this paper results of accidents classification using fuzzy cluster algorithms are reported. The object of clustering is to divide a given data set into homogeneous groups, that is all accidents in the same cluster share similar attributes and they do not share similar attributes with accidents in other clusters. The process was carried out using the fuzzy k-means method with extragrades. The basic idea is to classify accidents according to their patterns and causes into one or a combination of classes recognizing the complex interaction among accident factors and the uncertainties associated with accident data. The object of the study are the accidents happened along the highway A3 Naples-Salerno during the period August 1998 - March 2000. The characterizations of the accidental events have been encoded, to the purposes of the elaborations, through variables whose values are represented by theirs levels.

2 METHODOLOGY

The object of clustering is to divide a given data set into homogeneous groups: all points in the same cluster share similar attributes and they do not share similar attributes with points in other clusters. The separation of clusters and the meaning of similar ones are fuzzy notions.

The Fuzzy k-Means (F k-M) is the process of classification of the data that seeks a fuzzy k-partition of the set X of the sample points or, in other words, a family of set $A_i$ of fuzzy type ($i=1,\ldots,k$). As already stated, the n sample points can partially belong to more than a group assuming, in each of the fuzzy cluster, a different membership degree representable by any value included in the interval $[0,1]$. Continuous classes should provide better representations of atypical individuals than discontinuous classes. This is especially the case with atypical individuals located between clusters in property space (intragrades). Fuzzy k-Means method will indeed give intermediate memberships to intragrades. Atypical individuals outside the main body of data points (extragrades) are still not suitably represented by fuzzy k-means algorithm. The F k-M objective function $J(U)$ was modified to account for extragrades (de
Gruijter et al., 1998). This improvement makes the memberships directly depend upon the distances to the class centroids as:

\[ J_{MG} (M, G) = \alpha \sum_{i=1}^{n} \sum_{c=1}^{k} m_{ic}^\phi \cdot d_{ic}^2 + (1 - \alpha) \sum_{i=1}^{n} m_{ic}^\phi \cdot \sum_{c=1}^{k} d_{ic}^2 \]

where \( m^* \) denotes the membership to a fuzzy class of atypical individuals and \( \alpha \) is a parameter that determines the mean value of \( m^* \). The aim is to fix the atypical individuals in a special class to decrease these effects on classification. The members of this particular class are not concentrated in a fuzzy hypersphere around a defined class centre, as with regular classes. Atypical individuals are spread over regions of larger distances between an individual and the class centres. The solution for the membership:

\[ m_{ic} = \frac{d_{ic}^{-2/(\phi-1)}}{\sum_{j=1}^{k} d_{ij}^{-2/(\phi-1)} + \left( \frac{1 - \alpha}{\alpha} \cdot \sum_{j=1}^{k} d_{ij}^{-2} \right)^{-1/(\phi-1)}} \]

\[ m_{ic} = \frac{\left( \frac{1 - \alpha}{\alpha} \cdot \sum_{j=1}^{k} d_{ij}^{-2} \right)^{-1/(\phi-1)}}{\sum_{j=1}^{k} d_{ij}^{-2/(\phi-1)} + \left( \frac{1 - \alpha}{\alpha} \cdot \sum_{j=1}^{k} d_{ij}^{-2} \right)^{-1/(\phi-1)}} \]

\[ c_c = \frac{\sum_{i=1}^{n} \left\{ m_{ic}^\phi \cdot (1 - \alpha) \cdot d_{ic}^{-2} \cdot m_{ic}^\phi \right\} \cdot x_i}{\sum_{i=1}^{n} \left\{ m_{ic}^\phi \cdot (1 - \alpha) \cdot d_{ic}^{-2} \cdot m_{ic}^\phi \right\}} \]

Fuzzy k-means algorithm with extragrades recognizes that some objects might not fit well in any of the groups that are formed and puts these objects in an additional miscellaneous outlier group, the extragrades. The 'distance' measure should be chosen with care. Euclidean should not be used because different attributes have widely varying average values and standard deviations, since large numbers in one attribute will prevail over smaller numbers in another. With the diagonal metric, the input data are transformed before use. Choosing diagonal distance produces the transformation of the data set to one, where all attributes have equal variance. Correlations between variables are taken into account. The fuzzy k-means with extragrades algorithm is not sensitive to direction of distance. A useful analogy is to think of the algorithms as defining hyperspheres in multi-dimensional space to cover the data points. The performance of the k-means with extragrades algorithm depends on the choice of the 'distance' measure. Separations, measured by diagonal distance, might be better than Euclidean distance as they compensate for marked departures from the
hyperspherical shape. Better separations might be given when more attributes are used, because measurements of many properties are likely to provide more distinctions and thus greater dissociation. The diagonal norm is induced by the $p \times p$ diagonal matrix $A_D$, whose elements are $(1/\sigma_i)^2$ on the principal diagonal and 0 on the others positions: $\sigma$ is the standard deviation of a measured variable. Thus the diagonal norm is

$$(d_{ij})^2 = (x_i - c_j)' A_D (x - c_j)$$

Like the Euclidean, the diagonal norm is insensitive to statistically dependent variables but would compensate for distortions in the assumed spherical shape, caused by disparities in variances among the measured variables (Dell’Acqua, 2000).

### 3 ELABORATIONS

To demonstrate the applicability of the fuzzy cluster algorithms to the accidents classifications the aforementioned techniques were run on a sample data set (Minasny et al., 2000). The elaborations have been realized using a sample set related to the 20 months of observations. The sample set is composed therefore of a matrix with 836 accidents and 9 variables. The variables that appeared most useful in describing accidents were selected. The variables included a variety of accident, road and traffic characteristics (see Table 1).

To determine the number of clusters, it is necessary starting from the indications of the Separate distance ($S$) $c=2$; after this, through the application of the Fuzzy k-means with extragrades and the determination of the matrix $U$ ($2x836$) on the space $M_c$ of the possible partitions, a partition of the initial matrix in two classes has been reached. The elements of the matrix $U$ are the membership degrees of every accident to every considered class. The objective function value (OFV) decreases monotonically with increasing number of classes and increasing value of fuzzy exponent.

The value $dJ/d\phi$ is the derivative of the objective function value over the fuzzy exponent $f = -[(dJ/df) \sqrt{\text{class}}]$ to obtain an optimal value for $f$. The best value of “$\phi$” for a given class is at maximum of the curve $-[(dJ/d\phi) \sqrt{\text{class}}]$ versus class. Fuzziness Performance Index (FPI) estimates the degree of fuzziness generated by a specified number of classes. Modified Partition Entropy (MPE) estimates the degree of disorganization created by a specified number of classes.

The optimum number of classes was established on the basis of minimising these two measures. Separate distance measures the total middle compactness and the separation of a fuzzy k-partition.

The low value of $S$ points out to partition in which all the groups are very compact and separated. The best number of groups $k$ is, therefore, the one in correspondence of which the function $S$ has a minimum (see Fig. 1, 2).
One of the most important concepts of fuzzy sets is the concept of an $\alpha$-cut and its variant, a strong $\alpha$-cut. Given a fuzzy set $A$ defined on $X$ and any number $\alpha\in[0,1]$, the $\alpha$-cut, $^\alpha A$, and the strong $\alpha$-cut, $^\alpha+ A$, are the crisp sets: $^\alpha A = \{x \mid A(x) \geq \alpha\}$ and $^\alpha+ A = \{x \mid A(x) > \alpha\}$. That is, the $\alpha$-cut (or the strong $\alpha$-cut) of a fuzzy set $A$ is the crisp set $^\alpha A$ (or the crisp set $^\alpha+ A$) that contains all the elements of the universal set $X$ whose membership grades in $A$ are the greater than or equal to (or only the greater than) the specified value of $\alpha$. As an example, the following is complete characterization of strong $\alpha$-cuts ($\alpha = 0.6-0.7-0.9$) for the fuzzy sets class $A$, class $B$ and extragrades given in Fig. 3.

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>LEVELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lighting conditions</td>
<td>day light ÷ night light</td>
</tr>
<tr>
<td>Accident time</td>
<td>non-rush hour ÷ rush hour</td>
</tr>
<tr>
<td>Traffic</td>
<td>Value</td>
</tr>
<tr>
<td>Accident location</td>
<td>Normal lane ÷ overtaking lane</td>
</tr>
<tr>
<td>Degree of curvature</td>
<td>Straight ÷ curve</td>
</tr>
<tr>
<td>Speed limit</td>
<td>Value</td>
</tr>
<tr>
<td>Pavement condition (CAT)</td>
<td>Value</td>
</tr>
<tr>
<td>Surface condition</td>
<td>Wet, Dry, ....</td>
</tr>
<tr>
<td>Vehicle type</td>
<td>Passenger, truck, ....</td>
</tr>
</tbody>
</table>

4 RESULTS

The elements of the matrix $U$ determined are the membership degrees (included among zero and one) of every accident to every considered cluster.

Introducing a value of threshold ($\alpha$-cut) for the membership degree, a number $c$ of fuzzy set is got, whose elements are the accidents of the sample set that introduce a membership degree to the considered cluster greater to the imposed threshold.

The result of the elaborations has been a number of 3 sets of the original sample data base, equal to the number of considered clusters, whose elements result as much similar among them as much dissimilar from those of the others, just for the way in which they have been built.

The object of the conclusive readings of the two classes has been both the accidents set of every single cluster and the subsets drawn by the imposition of the aforesaid threshold from the three sets, the results are the prototypes:

1. the accidents that, having the most elevated membership degree among those of the considered set, have resulted the most representative of the same set;
2. the three sets, have brought to the identification of the classes of accident, therefore to the qualification of that whole situations, that, have primarily caused the accidental event.

<table>
<thead>
<tr>
<th>CLASS</th>
<th>LIGHTING CONDITIONS</th>
<th>ACCIDENT TIME</th>
<th>TRAFFIC (VEIC/DA Y)</th>
<th>ACCIDENT LOCATION</th>
<th>DEGREE CURVE</th>
<th>SPEED LIMIT (KM/H)</th>
<th>CAT</th>
<th>SURFACE CONDITIONS</th>
<th>VEHICLE TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>night</td>
<td>non rush hour</td>
<td>40893</td>
<td>Normal lane</td>
<td>straight</td>
<td>90</td>
<td>45.8</td>
<td>dry</td>
<td>Passenger only</td>
</tr>
<tr>
<td>B</td>
<td>day</td>
<td>rush hour</td>
<td>40945</td>
<td>Overtaking</td>
<td>straight</td>
<td>90</td>
<td>50.5</td>
<td>dry</td>
<td>Passenger only</td>
</tr>
<tr>
<td>Extr.</td>
<td>day</td>
<td>rush hour</td>
<td>40207</td>
<td>Normal lane</td>
<td>curve</td>
<td>60</td>
<td>47.8</td>
<td>frozen</td>
<td>Articulated truck</td>
</tr>
</tbody>
</table>

Table 2: prototypes

The difference between classes is related prevalently to the levels assumed by the accidents in the variables: lighting condition, traffic intensity and accidents location. The accidents of class A happened prevalently at night with scarce traffic and in normal lane; the accidents of class B, happened in full daylight with normal traffic conditions and in overtaking lane. The accidents of the extragrades class happened prevalently in full daylight with normal traffic and in normal lane.

5 CONCLUSIONS

The work reported in this paper describes a method to classify accidents on an assessment of factors that contribute to them. The basic idea is to classify accidents according to their patterns into one of the c groups or into a combination of these groups. The method uses fuzzy pattern recognition techniques for the classification process. There are several fuzzy methods for classification analysis such as Fuzz k-Means with extragrades. This method suits the purpose of this study for the following reasons:

- the same accident can belong to more than one class with varying degrees;
- the abnormal accidents (extragrades) are partly collocated in a special class;
- the relationships between the variables associated with accidents are of an uncertain nature.

Fuzzy logic can handle the uncertainty due to the nature of accident data and because of these reasons fuzzy techniques are adopted for the classification process. Fuzzy k-Means with extragrades algorithm have allowed to realize an innovative classification of road
accidents that considered an ample and complex number of factors and that, therefore, furnished some indications on the same phenomenon. The results are only meant to illustrate, in a concise way, the usefulness of the proposed method in a real case. We are elaborating other studies with encouraging results.
REFERENCES


