International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 4 (2012) pp. 400–410
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Fuzzy Clustering based Methodology for Multidimensional Data Analysis in Computational Forensic Domain

Kilian Stoffel, Paul Cotofrei and Dong Han

Information management institute, University of Neuchâtel,
Pierre-à-Mazel 7, CH-2000 Neuchâtel, Switzerland
{*kilian.stoffel, paul.cotofrei, dong.han*}@*unine.ch*

Abstract: **As interdisciplinary domain requiring advanced and innovative methodologies, the computational forensics domain is characterized by data being, simultaneously, large scaled and uncertain, multidimensional and approximate. Forensic domain experts, trained to discover hidden pattern from crime data, are limited in their analysis without the assistance of a computational intelligence approach. In this paper, a methodology and an automatic procedure, based on fuzzy set theory and designed to infer precise and intuitive expert-system-like rules from original forensic data, is described. The main steps of the methodology are detailed, as well as the experiments conducted on forensic data sets - both simulated data and real data, representing robberies and residential burglaries.**

Keywords: fuzzy inference system, fuzzy clustering, forensic data, computational intelligence

I. Introduction

Defined as data collected on a crime scene, during operational surveillance or by the information services, forensic data is considered as an important resource of information, recorded to analyze certain criminal phenomenon in order to define preventive or repressive measures. Usually kept in different formats, such as texts, tables and lists, the primary objective of this data is to identify the phenomena as accurate as possible. The traditional approach of utilizing this data is to deliver it to domain experts, in order to be analyzed using the expertise and basic knowledge of specialists. Usually these experts are very experienced with different cases for summarizing their principles on top of the phenomena. But their capacity is limited, due to fatigue, misjudgment and slow response, when the cognitive complexity increases to a certain degree. As a result, it is fairly difficult to carry out the analytical tasks without the assistance of automatic and intelligent computational methods [1]. In addition, common solutions to other domains are difficult to be used onto forensic data due to its unique features. This affirmation is strongly supported by two characteristics of forensic data:

- *Dimensionality*: A crime event is represented by a huge number of attributes (or dimensions, including temporal and spatial ones) in order to capture the accuracy of the criminal phenomena. Moreover, the volume of crime to be recorded is huge (in a specific crime analysis unit covering a region of around 700'000 inhabitants, about 100 events are to be scrutinized daily).

- *Vagueness*: The pieces of evidence (e.g. smudged fingermark, fragmentary ear mark, disguised handwriting or unobtrusive paint scratch) are hidden in a mostly chaotic environment. The traces identified in a crime scene will never be identical to known specimen in a reference base, even if traces are caused by an identical source.

Therefore, forensic data is both large scaled and uncertain, multidimensional and approximate. It is hence valuable to find the analytical framework particularly adapted for the forensic domain. According to the analysis above, the following requirements are to be considered:

- Primary data is supposed to be kept and represented with a tradeoff between completeness and facility of using it. The data recorded should reflect the reality whilst being not too complicated for emerging methodologies.

- Advanced methodologies supporting theoretical and applied forensic related processes are needed to assist the analysis. These methodologies may be new models, algorithms or frameworks.

- The computational framework should be quite intuitive to be understood by forensic domain experts. The framework output must allow the experts to use it together with their domain knowledge to directly supply conclusive actions.

In recent years, mathematical, statistical and computational science methods have found extensive applications in developing new procedures for crime investigation, prosecution and the law enforcement. Computational Forensic (CF) is an emerging interdisciplinary research domain. It concerns the investigation of forensic problems using computational methods, with the primary goal of discovery and advancement of forensic knowledge [2]. CF works towards in-depth understanding of a forensic discipline, evaluation of a particular scientific method basis, and a systematic approach to forensic sciences by applying techniques of computer science, applied mathematics and statistics. The contribution where the expected impact of CF is potentially far reaching covers two important research aspects:

- Formal modelling of various forensic processes/procedures followed during crime investigation;
- The development of an intelligent framework in which sophisticated computational methods, driven by forensic processes, analyse data and extract potentially useful information.

Clearly in line with the second theme, the goal of the paper is to propose a methodology and an automatic procedure for inferring accurate and easily understandable expert-system-like rules from forensic data. Due to the imprecise nature of data and the uncertainties and conjectures which characterize the inference structures in this domain, the proposed analytical framework must be based on a *soft computing* technique. These methods (which include fuzzy logic, neurocomputing, genetic algorithms, probabilistic reasoning, etc.) are computational methods tolerant to sub-optimality, impreciseness (vagueness) and partial truth and giving quick, simple and sufficiently good solutions. Each of them plays an important role for data analysis, but from different perspectives. Among the list above, the paper concentrates on the fuzzy set theory, by analyzing the applicability of different fuzzy methods to improve the effectiveness and the quality of the data analysis phase for crime investigation. The test bed of this paper is the data from forensic case data representing robberies and residential burglaries in the region of Lausanne, Switzerland. This data is stored in a relational database formulated via a collaboration of forensic experts and computer scientists taking the consideration of real cases.

The paper is organized in the following way. The next section will review the state of the art for the literature relevant to our topic, including data mining and fuzzy logic theory. The sections III and IV present the basic concepts of fuzzy set theory on which our approach is constructed as well as a detailed description of the proposed methodology. The characteristics of the forensic data set on which the test bed is constructed are depicted in section V followed, in the next section, by a detailed analyse of the outputs of experiments, conducted on simulated data and on real data. Finally, the last section is the conclusive statement.

II. State of the Art

In the late nineties, the significance of computational methods for forensic investigation service was revealed by the research community (especially in the domains of fingerprint identification [3] and DNA analysis [4]), and the number of academic publication relevant to this domain increased gradually in the recent years, especially resulted by the emergence of the concept of computational forensic. According to [5], the computational methods allows the forensic practitioner to analyse and identify traces in an objective and reproducible manner, to standardise investigative procedures, to search large volumes of data efficiently, to assist in the interpretation of results and their argumentation, to reveal previously unknown patterns, to derive new rules, and to contribute to the generation of new knowledge.

Computational methods find their place in the forensic sciences in three ways [2]:

- They provide tools for the investigator to overcome limitations of human cognitive or perceptive abilities.
- They allow for the analysis of large amounts of data and the extraction of useful knowledge.
- They can ultimately be used to represent expert knowledge and provide automatic or semi-automatic reasoning and inference capabilities.

The inherent heterogeneity and fuzziness of the data naturally calls for data mining techniques in order to discover hidden knowledge and for fuzzy logic based reasoning to apply automatic or semi-automatic inference.

A. *Data mining*

Data mining is a core technology in the specific field of digital forensics [6], dealing primarily with the investigation of material present on digital devices. An extensive survey of data mining techniques in this particular field has been conducted in [7]. Data mining can be used in both crime **investigation** and **detection**, both aspects of crime analysis. Techniques that have been applied successfully include classification via cluster analysis [8] and discriminant analysis [9], association rule mining [10, 11], content retrieval incorporating text mining [12] and N-gram-based text categorization techniques [13] - particularly in investigations pertaining to system logs, email or instant messaging evidence. Other specific data mining such as anomaly detection [14, 15, 16], data visualisation and various profiling techniques have been applied to intrusion detection, image analysis (both 2D and 3D) and criminal profiling [17, 18, 19, 20, 21, 22]. Furthermore an adaptive neural network is used to solve the forensic problem of image matching (personal identification [23] or postmortem identification using dental radiographs [24]), to discover existing trends in criminal activities [25] or to predict fraud litigation for assisting accountants [26].

B. *Fuzzy Theory*

Fuzzy methods (including fuzzy sets, fuzzy logic, fuzzy inference systems, fuzzy clustering) and hybrid fuzzy methods (genetic fuzzy clustering, fuzzy neural networks, etc..) play an important role in learning complex data structures and patterns, and classifying them to make intelligent decisions. Fuzzy clustering is used in [27] to detect the explanation of criminal activities for crime hot-spot areas and their spatial trends. Compared with two hard-clustering approaches (median and k-means clustering problem), the empirical results suggest that a fuzzy clustering approach is better equipped to handle crime spatial outliers. An approach based on fuzzy logic and expert system for network forensics that can analyze computer crimes in network environment and make digital evidences automatically is proposed in [28]. Experimental results show that the system can classify most kinds of attack types (91.5% correct classification rate on average) and provide analyzable and comprehensible information for forensic experts. A pseudo outer-product based fuzzy neural network (POPFNN) is trained to detect similarity between two fingerprints and decide whether they belong to the same person [29]. The characteristics of POPFNN, such as the learning, generalization, and high computational abilities, make fingerprint verification particularly powerful when verifying authentic fingerprints subjected to external conditions and recognizing spurious ones. A two stage fuzzy decision classifier, using reference fuzzy set information, is used in [30] to create a text-independent Automatic Speaker Identification. Finally, a framework of intelligent decision-support model based on a fuzzy self-organizing map (FSOM) network to detect and analyze crime trend patterns from temporal crime activity data is proposed in [31]. The resultant model can support police managers in assessing more appropriate law enforcement strategies, as well as improving the use of police duty deployment for crime prevention.

III. Fuzzy set theory

The literature review outlined in the previous section is a strong support of the approach considered in this paper, i.e. the integration of data mining and fuzzy theory is able to set up the framework for computational forensic and to produce meaningful results. Consequently, the proposed methodology for inferring expert-system-like rules from forensic data will be based on fuzzy set theory.

A. Fuzzy logic

The term Fuzzy Set was originally proposed by Lotfi A. Zadeh [32]. It differs from classical notion of set in that it provides the gradual assessment of the membership function, which is ranged within the interval [0, 1]. This function represents the degree of the statement (such as whether the temperature is high or low) in a fuzzy way.

On top of the theory of fuzzy set, fuzzy logic extends the case of multi-valued logic. It assigns a degree of truth - a value varying between absolutely true and absolutely false - to each proposition. Fuzzy logic (together with neurocomputing and genetic algorithms) is one of the techniques of soft computing, i.e. computational methods tolerant to suboptimality, impreciseness (vagueness) and partial truth and giving quick, simple and sufficiently good solutions. The guiding principle of these methods is perfectly adapted to the way in which reasoning and deduction have to be performed in forensic science (for searching hidden traces in a mostly chaotic environment, traces never identical with known specimens in a reference base), i.e. on the basis of partial knowledge, approximations, uncertainties and conjectures [2].

Among the general statements about fuzzy logic, we may enumerate [33] the flexibility, the tolerance of imprecise data, the capacity to model nonlinear functions of arbitrary complexity (matching any set of input-output data), the capacity to be built on the top of the experience of experts, and the facility of use (due to its basis built on natural language).

From another perspective, the utilisation of fuzzy theory fits well our demand of the forensic domain as the forensic data rarely presents in an absolute way. It reflects, instead, some flexible and probabilistic principles recorded based on a series of phenomena. We are thus able to investigate these phenomena by expressing and analysing the data with the assistance of fuzzy theory [34]. The output result will be more intuitive and meaningful for domain experts to understand and make use of.

B. Fuzzy Inference Systems

Describing generally vague concepts (as tall people, hot weather, morning hours, etc.), fuzzy sets have associated a membership function (denoted $\mu(x)$) which maps an input value to its appropriate membership value. A membership function may be any arbitrary function with values in $[0, 1]$, but in practice basic functions are used, as piece-wise linear functions, Gaussian distribution function, the sigmoid curve, quadratic and cubic polynomial curves. In a mathematical notation, a fuzzy set is the set of pairs $A = \{(x, \mu(x)\}$. The set of elements that have a non-zero membership is called the *support* of the fuzzy set.

The fuzzy logical reasoning is a superset of standard Boolean logic, i.e. the truth functions of connectives have to behave classically on the extremal values $0, 1$. For conjunction, a family of functions satisfying this condition is the set of binary T-norm operators [35] (*min* is a classical exemple), whereas for disjunction is the set of binary T-conorm operators (*max* is a classical exemple). Several parameterized T-norms and dual T-conorms have been proposed in the literature, such as those of Yager[36], Dubois and Prade[37] and Sugeno [38].

A fuzzy rule *if-then* has the form *If x is A Then y is B*, where A and B are fuzzy sets. Interpreting an *if-then* rule involves two distinct parts. Firstly, the premise of the rule is evaluated, which involves *fuzzifying* the input (i.e. calculate the membership value) and - if the premise have multiple parts - applying any necessary fuzzy operators. Secondly, the result is applied to the consequent (operation known as implication) using an *implication* function, which modifies the output fuzzy set to the degree specified by the antecedent. The modification is usually realized by truncation, using the *min* function, or by scaling, using the *prod* function, but other theoretical approaches have been proposed [39, 40].

The fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. The systems using fuzzy inference have been applied in different domains, as automatic control, data classification, decision analysis, expert systems or computer vision. In the literature two types of fuzzy inference systems (FIS), differing by the way the output is determined, are the most known: Mamdani-type and Sugeno-type. The Mamdani's fuzzy inference method [41] expects the output membership functions to be fuzzy sets. For Sugeno-type systems [42], the output membership function is a singleton, which simplifies the defuzzification process. In general, Sugeno-type systems can be used to model any inference system in which the output membership functions are either linear or constant.

In the context of an universe comprising fuzzy sets and a number of weighted fuzzy rules *if-then*, a fuzzy inference process comprises five phases: (i) fuzzification of the input variables (those appearing in the antecedent part of the rules), (ii) application of the fuzzy operator (AND or OR) in the antecedent (if necessary), (iii) implication from the antecedent to the consequent, (iv) aggregation of the consequents across the rules, and (v) defuzzification.

If the system starts with more than one rule, the fuzzy sets representing the output of rules implying the same variable are combined (aggregated) into a single fuzzy set. This operation is applied during the fourth phase of the inference process. The output of the aggregation phase is one fuzzy set for each output variable. The fifth phase allows to obtain a single numerical value, by applying the defuzzification process on the output fuzzy sets. Among the most popular defuzzification methods we may enumerate the centroid calculation (for Mamdani-type systems) and the the weighted average (for Sugeno-type systems).

The basic model for a fuzzy inference system considers that membership functions, representing the characteristics input, are predetermined by the user. In the situation where these characteristics can't be "guessed" only by looking at the data, a neuro-adaptive learning technique may be used to *learn* information about a data set, by choosing the parameters so as to tailor the membership functions to the input/output data. The final system is called an *adaptive neuro-fuzzy inference system* because it uses a network-type structure similar to that of a neural network for the learning purpose. The fuzzy modeling approach comprises the classical system identification steps: hypothesizing a parameterized model structure, collecting input/output data in an appropriate form, training the FIS model according to a chosen error criterion and validating the model.

C. *Fuzzy Clustering*

Another essential element in our approach is Fuzzy Clustering. The aim of a cluster analysis is to partition a given set of data or objects into clusters (subsets, groups, classes), such that the data that belong to the same cluster should be as similar as possible, and the data that belong to different clusters should be as different as possible. The Fuzzy C-Means Clustering (FCM) is an unsupervised goal oriented clustering algorithm, introduced by Dunn [43] and generalized by Bezdek [44].

The term fuzzy is used here to refer to the way how the analysis of clusters is done: an item x_k can be assigned to several clusters c_i, through the membership functions μ_i. The goal function of the FCM is defined by

$$J(U,C)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}d_{ik}^{2} \qquad (1)$$

where $u_{ik} = \mu_i(x_k)$, $m \in (1,\infty)$ defines the degree of fuzziness and $d_{ik}^2 = (x_k - \overline{c}_i)^T(x_k - \overline{c}_i)$ is the squared distance (usually the euclidian distance) between the item x_k and the center $\overline{c}_i$ of the cluster c_i. The clusters' centers are stored in the matrix C whereas the matrix U contains the corresponding values of u_{ik}.

The optimization problem can be described as *minimize* $J(U,C)$ under the constraints

(i) $(\forall k)\sum_{i=1}^{c}u_{ik}=1$

(ii) $(\forall i)\sum_{k=1}^{n}u_{ik}>0$.

In order to solve this optimization problem the Lagrange method can be used.

$$J(U,C,\lambda)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}d_{ik}^{2}-\sum_{k=1}^{n}\left(\lambda_k\left(\sum_{i=1}^{c}u_{ik}-1\right)\right) \qquad (2)$$

The derivative for u, c and λ have to be calculated. This leads to the following solution

$$\overline{c}_i=\frac{\sum_{k=1}^{n}u_{ik}^{m}x_k}{\sum_{k=1}^{n}u_{ik}^{m}};\, u_{ik}=1/\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}} \qquad (3)$$

The FCM algorithm consists of the following steps:

1. initialize the matrix U^0
2. calculate the matrix $C^s = [c_i]$
3. calculate the matrix $U^s = [u_{ik}]$
4. if $||U^s - U^{s-1}|| < \epsilon$ then stop, else goto step2

The FCM algorithm is efficient, straightforward, and easy to implement, but it is sensitive to initialization (due to random selection of initial center points) and so is easily trapped in local optima. Moreover, the number of clusters to be generated for the given data set needs to be specified *a priori*. To overcome these problems, different solutions were proposed, such as a PSO-based fuzzy clustering algorithm [45], a genetic fuzzy k-Modes algorithm [46] or a hybrid fuzzy clustering algorithm based on FCM and rough sets theory [47].

IV. Fuzzy Clustering based Methodology

As we have stated in the introduction section, the main objective of this paper is to build up a methodology which constructs the expert-system-like if-then rules automatically from the forensic data. There are two main constraints for the rules. On the one hand, the rules extracted are supposed to be as precise as possible to reflect the phenomena in reality. On the other hand, they should be intuitive, understandable and applicable by the experts specialized in the forensic domain (not necessarily in expert systems or other information domains). We aim to find a tradeoff for these two factors above to provide a meaningful methodology as well as its experimental results.

The methodology we are proposing here, based on fuzzy clustering, is one of the many used for inferring membership functions for fuzzy variables from raw data. The overall procedure consists of three main steps:

1. clustering the raw data
2. extract the membership functions from the data
3. create the fuzzy inference system

We will give now a short description of each step.

A. *Clustering the Raw Data*

In order to provide meaningful membership functions for the fuzzy variables representing interesting dimensions, a clustering algorithm is applied during the first step of the methodology. In respect to the forensic data, these dimensions will correspond to attributes (or groups of attributes) from the forensic database (detailed in section V). The clustering procedure is based on a Fuzzy C-Mean algorithm, which has the important advantage - in respect to the general framework - to produce a membership value for each data point, defining its degree of membership to each cluster. Nevertheless the same algorithm has some negative points in the sense that it is not invariant to linear transformations and it is also sensible to the initialization of the cluster centers.

The FCM outcome which is important for the next step is the matrix U: the element u_{ik} defines the degree of membership of the element x_k to the cluster c_i. In order to get this output, the algorithm must receive as input the list of elements that have to be clustered (each element x_k being a multidimensional data point, represented as a tuple of values for each dimension), as well as the number of clusters to generate. For the experiments conducted in this paper the number of clusters was provided by the user, but it would also be possible to use automatic techniques, e.g. differential clustering to generate the number of clusters.

B. Extraction of Membership functions

Obviously, the construction of a membership function for a fuzzy set depends on the preselected type (shape) of such a function. For illustration purposes we will focus only on the symmetric Gauss membership functions. However, the same general procedure as outlined here can be applied to all kinds of membership functions. The main idea consists of inferring the parameters of the membership function from the output values obtained in the previous step. The Gauss membership function has the following form: $g(x, \sigma, c) = e^{\frac{-(x-c)^2}{2\sigma^2}}$. This function depends on two parameters c and σ, where c corresponds to the centers of the clusters found during the FCM clustering and σ is calculated using the forensic raw data and the U matrix obtained in step one.

Concretely, for each cluster c_i and each fuzzy variable j (data dimension) the following formula can be applied.

$$\sigma(i,j) = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{-(x_{kj}-\bar{c}_i)^2}{2\log(u_{ik})}} \qquad (4)$$

where N is the number of data points, x_{kj} is the value of data point k for the dimension (database attribute) j, $\bar{c}_i$ is the center of cluster c_i and u_{ik} is the degree of membership of point k in cluster c_i. However this function would go to infinity if one data point corresponds exactly to the center of the cluster. But these points are not adding or removing anything to the spread of the Gaussian curve and can therefore be removed, for our purposes. This concludes the second step; now a number of $\#cluster \times \#dimension$ membership functions are defined and the corresponding fuzzy sets will be used to construct the rules of the inference system.

C. Creating the Fuzzy Inference System

The rules to be included in the fuzzy inference system should fulfill the constraint of being easily understandable by a domain expert, e.g. a member of the police corps or a forensic specialist. As fuzzy rules, both components - the antecedent part and the consequent part - involve fuzzy sets. If, for example, the following fuzzy sets would be defined:

(A) *late evening*, over the universe represented by the slot 6.00p.m - 12.00p.m.

(B) *rural*, over the universe represented by geographical coordinates, and

(C) *high*, over the universe represented by the unit interval $[0, 1]$

then a meaningful and understandable fuzzy rule could be (expressed as a natural language sentence):

```
IF time is "late evening" and place is
"rural" THEN possibility of a robbery is
"high"
```

and expressed as a fuzzy logical formula as

$$(x_1 \in A) \wedge (x_2 \in B) \Rightarrow (y \in C)$$

For each cluster generated during the first step of the methodology, a certain number of fuzzy sets - equal with the number of dimensions - are defined by the the corresponding membership functions extracted during the second step. Therefore, all the elements for constructing rules are available. The first decision to be taken is to choose which dimensions will be used in the premise part of the rule and which will be used in the consequent part. During the experiments described in the paper, the fuzzy rules have been limited to those implying a single dimension in the antecedent part, but our approach also allows for conjunctions of dimensions in the premise (as in the previous example). Once the list is created, the fuzzy rules of the given form are defined for each cluster. The set of all rules will then constitute the fuzzy inference system. The last decision is to select the type of fuzzy inference system that will be used. This choice depends on the further usage of the fuzzy system. The experiments in the paper are based on a Mamdani-type system, however, in the practical application we prefer the Sugeno-type system, especially for further automatic improvements of the system.

After finishing these steps a complete system is available. It can be used in the process of investigation to perform tasks such as prediction, characterization, or validation. All of the steps needed in our methodology may be easily implemented in most of the data analysis environments (some of them having even comparable procedures predefined).

V. Multidimensional Data Set

The three main categories of data/information related to activities that are dealing with police data and information are the following:

1. Crime Scene Data: all the data collected at a crime scene.

2. Risk Assessment Data: data collected during operational surveillance or collected by the information services, in order to be able to analyse certain criminal phenomenon. This data and information is used for defining preventive or repressive measures.

3. Jurisdiction Data: data and information collected for use in justice. There are two groups in this category:

(a) data/information related with one particular crime or a series of crimes committed by the same author

(b) data/information that, once the author is known, will be used to construct proofs that can be used in court of justice

These three categories are following very different types of logic, all related to the divergent goals to achieve using this data and information. Generally speaking, the processes followed to acquire this knowledge and the logic applied in the different cases is very poorly documented and the links between them are unknown. The methodology proposed in this paper is focusing on the first and the second category of data, for which the current approaches leads to suboptimal results as it is illustrated by crime analysis approach - mainly based on environmental criminology which typically tends to neglect crime scene data and computational methods. Some Bayesian approaches have emerged in the third category but they seem not to be of general interest and usability for the first two categories. This is why we think that the results of this paper could make a significant scientific contribution.

The forensic database we are working with in our study contains all the collected data about events representing robberies and residential burglaries in the canton of Vaud, Switzerland. This high-dimensional database contains information about events, identities, vehicles, tests, relations, etc., characterized by about 70 attributes. We conducted tests on a wide range of objects, but for space constraints we decided in this paper to focus on the ”event” part of the database. This part is typical enough to present our main results and does not need long presentations. Each event is identified as a point having three dimensions (a dimension being represented by an attribute, a derived attribute or a group of attributes):

- Temporal dimension: characterized by starting/ending date and time of occurrence.
- Spatial dimension: characterized by the geographical x and y coordinates, based on the Swiss reference system CH1903 and by the address coordinates (county, city, zip, etc..).
- Typology dimension: characterized by the event type, offender type, address type (apartment, residential house, commercial store, etc..) and *modus operandi* type (from roof, alarm disabled, blowlamp, etc..).

VI. Experiments

In order to evaluate the capability of the proposed methodology to deliver meaningful rules (both precise and understandable), we decided to conduct experiments using two types of data. The first one is a simulated datasets, strictly respecting the structure of the original forensic database, however containing some hidden internal structures. This first experiment is intended to test if the methodology is able to unveil these structures and, at the same time, to check that the system is not discovering non-existing structures. The second datasets is represented by the original forensic database. Even if the rules generated during this last experiment revealed, in our opinion, interesting connections between the different dimensions of events, their interestingness and usability can be confirmed only by forensic domain experts.

A. Simulation

As explained above, an event database containing temporal, spatial and typological information was simulated according to the following requirements for the three dimensions:

- Temporal dimension: five years worth of data, divided in year, month, day (including weekday).
- Spatial dimension: $x - y$-coordinates of the region under investigation. Uniformly distributed events, cities and rural regions, and highways were simulated. The cities were simulated as independent, aligned in respect to the x-coordinate (longitude) and in respect to the y-coordinate (latitude).
- Typological dimension: simplified numerical value. In this basic structure of the database hidden rules were integrated, e.g. the typological value is probabilistically higher for rural regions, the probability of having events occurring on a weekday are higher than on weekends, etc. The goal is to test if the system is able to unveil them. Some of these examples are presented as follows.

Globally, the obtained results were very satisfactory. Following the proposed methodology, the system was able to unveil all the hidden structures integrated into the generated datasets. Also, while being more difficult to assess, the accuracy of the results was very positive. The datasets were generated using probabilistic algorithms which introduce errors. The errors produced by the proposed methodology had also comparable statistical distribution as the errors introduced by the generation procedure. This is the base on which we may conclude that the accuracy of the approach is satisfactory. Furthermore it has to be underlined that the fuzzy inference system constructed in the three steps outlined above is not intended to be in its final shape. These systems can further be improved using methods not mentioned here.

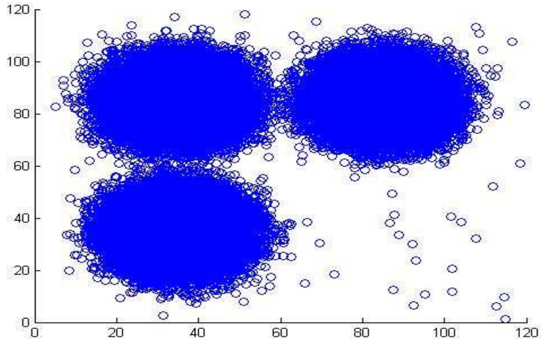

Figure. 1: Spatial distribution of the events.

In the following, two examples are presented as an illustration of the kind of results obtained in general. In both examples the same spatial distribution of events was used. It simulates three cities, two of them being aligned in respect to the x-axis and two in respect to the y-axis (see Figure 1). The goal of these examples is to predict the typology from the x and y values. *Remark*: the number of clusters (input parameter for FCM algorithm) was set to three. Therefore,

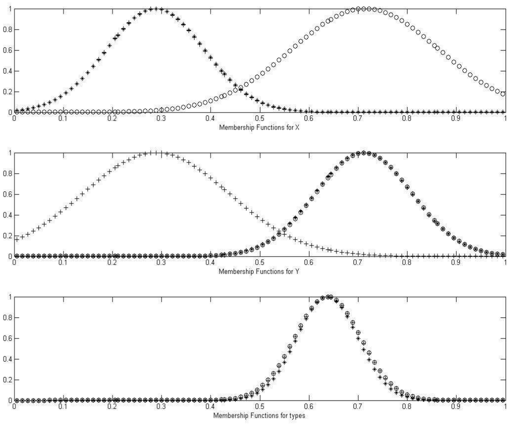

Figure. 2: Three cities with similar typologies of events.

the number of membership functions along each dimension is equal to three.

In the first example, a situation where the typology attribute has the same value for all the events occurred in the three cities was simulated. The density of events is higher in the center and then degrades as we move away from the center. In Figure 2 we can see in the first graph that the two left membership functions along the spatial dimension X are overlapping, as the two cities on the left share similar x-values. In the second graph one can see the same phenomenon in respect to the y-values for the upper cities. For the output variable (typology) we even have all three membership functions that are overlapping, which is correct as the three cities have similar values.

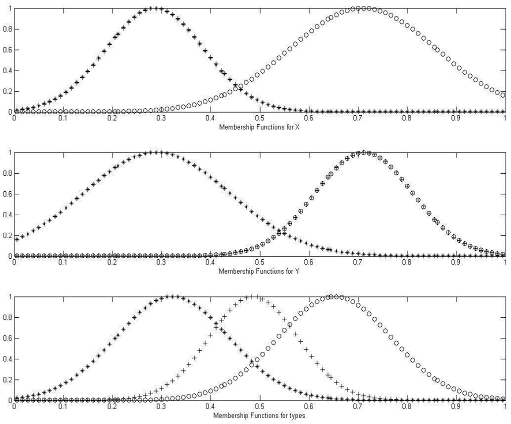

Figure. 3: Three cities with different typologies of events.

In the second example (see Figure 3) it was simulated a situation where the typology (output variable) for the events in the three cities has different values. Now we can see that the three membership functions of the output variable illustrate this fact.

It can be observed that, in the first example, only two membership functions for spatial dimension X, two for spatial dimension Y and one for the typology dimension would have been sufficient, and in the second example the same remark for X and Y dimension, but three membership functions for the topology. The current approach does not automatically adapt the number of membership functions, however they can be manually adapted.

Even though the results are very encouraging, there is one important drawback to mention. Following the proposed methodology, the number of membership functions per fuzzy variable (i.e. database attribute) is always equal to the number of clusters created (even if some membership functions may be identical). Although this approach allows to fix the number of clusters by the user, only one value is possible. For the type of forensic data involved in our experiments it is difficult to fix the number of clusters that would produce meaningful membership functions for the temporal and the spatial dimensions. This does not reduce the quality of the numerical results of the fuzzy inference system, however, it might be difficult to assign a clear semantics to these fuzzy sets (and implicitly, to fuzzy rules) which is important in systems designed for domain experts.

B. Residential Burglaries Cases

The main objective of the experiment based on simulated data was to demonstrate, by using some simple but persuasive examples, that the proposed methodology is appropriate, from different perspectives, for multidimensional and fuzzy forensic data. But the logic of a simulation process for artificial data is ”from rules to data”, contrary to the usual applications, requiring a rule discovery process ”from data to rules”. Therefore, to evaluate the real capability of the methodology it is necessary to conduct experiments on the real data.

In a first phase, we intend to focus on the data itself, aiming to find some simple rules from unsophisticated analysis of the data set. Figure 4 displays the geographical distribution of the cases using MapPoint application [48]. The visualisation of the distribution facilitates an intuitive understanding of data and reveals some principles which might be useful for further analysis phases:

(i) The outline of the graph follows exactly the geographical shape of the region Vaud. This, to some extent, reveals the completeness of our data.

(ii) The density of the cases in each region varies consistently with different socio-economics factors leading to the differences between the regions. Among these factors we may enumerate the intensity of economic activity or the size of population. For example, the region around the cities and littoral parts present more cases than other regions since there are more habitants and commerce in these regions, coherent with our direct impression. The result, in terms of longitude and latitude, of the four clusters in Figure 5 reflects exactly the four main centres of Vaud in Figure 4 - Nyon(denoted in black colour in these figures), Lausanne(red), Yverdon(green) and Montreux(blue).

(iii) There are still some data points in this figure which look interesting for our methodology and for domain experts, asking a deeper and detailed investigation. For instance, some noise points (one in the middle of Leman lake!) and some particular regions with a non-expected high rate of cases are found in the figure.

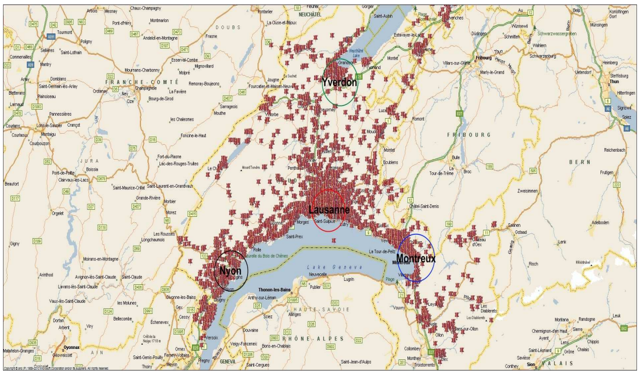

Figure. 4: Case distribution in the region of Vaud, Switzerland

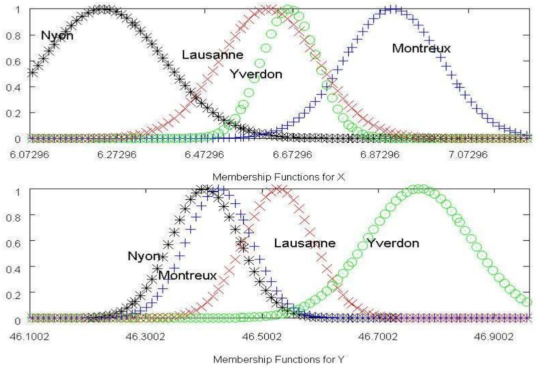

Figure. 5: Clusters of the case distribution

For the next step, since this data is multidimensional, with a number of attributes, our objective is to discover the rules and principles among these attributes. We are going to design some typical experiments to illustrate how the fuzzy clustering methodology works on our data set.

As already known from previous forensic studies, the frequency of reported cases varies obviously regarding different periods of the year. We would like to carry out an experiment, trying to discover the relationship between each month (or group of months) and its corresponding frequency of the occurred cases. Therefore the output fuzzy variable is given by the ”frequency” dimension, which is derived from typology dimension, whereas the input fuzzy variable is given by the temporal dimension ”month”. In order to set the initial number of clusters (the input parameter C of the methodology), the best approach is to use the concept of linguistic variable [49]. Such a variable takes as values words or sentences in a natural or artificial language. If ”frequency” is a linguistic variable, it may takes as values, e.g., the expressions $\{low, average, high\}$ (which implies three initial clusters) or {*very low, low, average, high, very high*} (implying five initial clusters). Consequently, the parameter C will be set according to the meaning a user intends to give to the values of output linguistic variable.

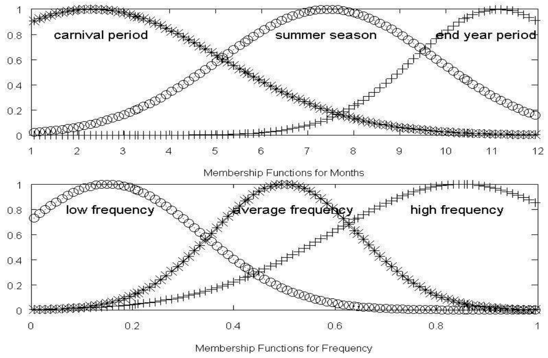

Figure. 6: The input and output fuzzy variables, ”months” and ”frequency”

Figure 6 depicts the three clusters along the temporal dimension ”months” (upper graph) and the typological dimension ”frequency” (bottom graph, normalized over the interval $[0, 1]$). The correspondence between the fuzzy sets from the bottom graph and the natural expressions {*low, average, high*} (values of linguistic variable ”relative frequency”) is straightforward. On the other hand, the meaning of the fuzzy sets from the upper graph, conceived as values for the linguistic variable ”year period” are not obviously: the left one may be identifies as *carnival period* (in Switzerland, most carnivals are held in February), the middle as *summer season* and the right one as *end year period*. Therefore, the fuzzy clustering methodology generated in this context a fuzzy inference system containing three fuzzy rules:

(A) IF month in "carnival period" THEN relative frequency of robberies is "average"

(B) IF month in "summer season" THEN relative frequency of robberies is "low"

(C) IF month in "end year period" THEN relative frequency of robberies is "high"

The rules seem interpretable. For example, economical and social activities are occurring more often during the months following the summer holidays. Taking the traffic as an example, the number of passengers taking public transportation increases from the end of September due to the start of autumn semester for schools at all the levels and for universities. There are thus more cases of pickpocketing. On the contrary, during summer time, the duration of the day time is much longer compared to other seasons, leading to a lower frequency of *reported* cases.

The following experiment is to find the rules relating the week days (input fuzzy variable) and the case frequency (output fuzzy variable). This experiment is interesting in that week day is an important periodical unit in our work and life. Plenty of plans are scheduled based on the week days. If it possible to find some positive results, it will be very meaningful for both domain experts and citizens to take some further actions as prevention.

The meaning of the three fuzzy sets over the temporal dimension ”week day”, depicted in the upper graph of Figure 7, are easily identified with natural language expressions *start of week, middle of week* and *weekend*. On the other hand, the meaning of the second fuzzy set over the ”frequency” dimension is not exactly *average*, like in the previous experiment, but rather *average to high*. Therefore, the following fuzzy rules are members of the fuzzy inference system:

(A) `IF day in "start of week" THEN relative frequency of cases is "low"`

(B) `IF day in "middle of week" THEN relative frequency of cases is "average to high"`

(C) `IF day in "end of week" THEN relative frequency of cases is "high"`

One of the explanations is that people spend more time on Saturday and Sunday with their families and friends for social and commercial activities, enlarging the possibility of burglary. On the other hand, their awareness is usually high at the beginning of the week, as the reason why the cases happened less frequently at this period.

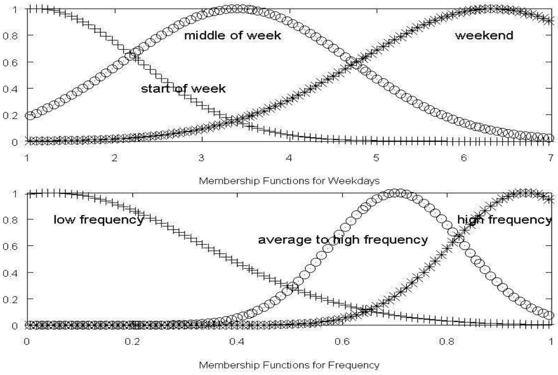

Figure. 7: The input and output fuzzy variables ”week day” and ”frequency”

The last experiment concernes the relation between the hour when a case occurred and the frequency of cases. The goal is to discover during which hours there are more records and also in which period it happens more rarely. Similarly to the previous two experiments, three clusters will be established. As shown in Figure 8, the upper graph depicts the three fuzzy sets over the temporal dimension ”daily hours”, identified as *early morning, afternoon* and *evening*. The bottom graph of Figure 8 displays the fuzzy sets along the dimension ”frequency”. Contrary to the previous experiences, it is not possible to identify these sets with the values {*low, average, high*}; in fact, two of the sets, almost identical, may be identified with the value *average* and the last set with the value *high*. This particular situation proves that the rule fixing the initial number of clusters as the number of expected values for the output linguistic variable is not always appropriate; a second option is to set this parameter according to the expected values for the input variable.
The resulted fuzzy inference system will include finally only two rules:

(A) `IF hour in "early morning" OR in "evening" THEN relative frequency of cases is "average"`

(B) `IF hour in "afternoon" THEN relative frequency of cases is "high"`

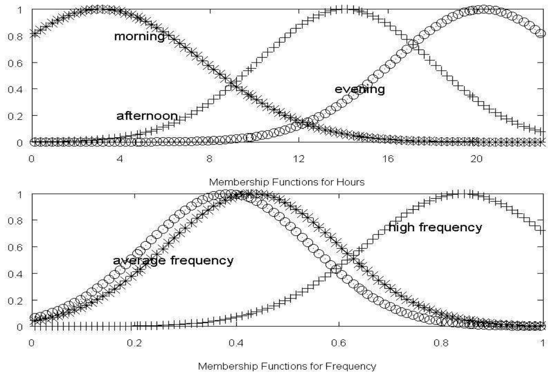

Figure. 8: The input and output fuzzy variables ”daily hour” and ”frequency”

A possible explanation of the peak frequency during the afternoon is that few people are at home and more social activities are taken during this period of the day. But, as we already emphasized, the validation of the interestingness, usability and accuracy of the fuzzy rules generated by this methodology can not be performed without the collaboration of a forensic domain expert.

VII. Conclusion and Future Work

In this paper we described a cluster based methodology to automatically extract expert-system-like if-then rules from forensic databases. The development framework is related to a new interdisciplinary research domain, the computational forensic domain. Based on the analysis emphasizing the main characteristic of forensic data - large scaled and uncertain, multidimensional and approximate - we decided to consider both data mining and fuzzy set theory as our research fundamentals. The proposed methodology, consisting of three main phases (fuzzy clustering data, membership function extraction and fuzzy rules generation) is proven to be easily implementable in most data analysis environments. A series of experiments, using both simulated and real data set, showed satisfactory results: the generated fuzzy rules captured all the hidden structures from simulated data and also known forensic dependencies from real robberies and residential burglaries data. The accuracy of the inferred rules was clearly higher than the minimum level required to make them usable in a practical setting. However, the tests have also shown a drawback that should not be neglected: the fact that it is not always obviously to find an intuitive semantics for some of the fuzzy sets (even though they are producing high quality rules). This fact complicates the communication with the domain experts, which are supposed to validate the meaning of the generated rules. A possible solution to this drawback is to extend the methodology, in the future, through the incorporation of forensic domain knowledge capable to drive the inference rules process.

Acknowledgements

The authors acknowledge support from Swiss National Science Foundation (SNSF), grant number CR21I2-135236/1. The authors also like to thank Prof. Olivier Ribaux, Institute of Forensic Science, University of Lausanne, and Mr. Jacques-François Pradervand, Le Chef de la Police de sûreté du canton de Vaud, Switzerland, for supporting this project and authorizing the access to the extensive forensic datasets.

References

[1] O. Ribaux, S. J. Walsh, and P. Margot, "The Contribution of Forensic Science to Crime Analysis and Investigation: Forensic Intelligence," *Forensic Science International*, vol. 156, pp. 171–181, 2006.

[2] K. Franke and S. Srihari, "Computational forensic: An overview," in *Computational Forensics*, 2008.

[3] A. K. Jain, L. Hong, and R. Bolle, "On-line fingerprint verification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 302 – 314, 1997.

[4] I. Evett, L. Foreman, J. Lambert, , and A. Emes, "Using a tree diagram to interpret a mixed dna profile," *Journal of Forensic Sciences*, vol. 43, pp. 472– 476, 1998.

[5] K. Franke and S. Srihari, "Computational forensics: Towards hybrid-intelligent crime investigation," in *Third International Symposium on Information Assurance and Security*, 2007.

[6] I. Ricci, "Forza - digital forensics investigation framework that incorporate legal issues," *Digital Investigation*, vol. 3, pp. 29–36, 2006.

[7] N. L. Beebe and J. G. Clark, *Research Advances in Digital Forensics*. Springer, 2005, ch. Dealing with Terabyte Datasets in Digital Investigations, pp. 3–16.

[8] de Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," *AGM SIGMOD Record*, vol. 30, pp. 55–64, 2001.

[9] M. Carney and M. Rogers, "The trojan made me do it: A first step in statistical based computer forensics event reconstruction," *Digita! Evidence*, vol. 2, no. 4, pp. 11– , 2004.

[10] T. Abraham and O. de Vel, "Investigative profiling with computer forensic log data and association rules," *Proceedings of the IEEE International Conference on Data Mining*, pp. 11–18, 2002.

[11] T. Abraham, R. Kling, and O. de Vel, "Investigative profile analysis with computer forensic log data using attribute generalization," in *Proceedings of the Fifteenth Australian Joint Conference on Artificial Intelligence*, 2002.

[12] M. Shannon, "Forensics relative strength scoring: Ascii and entropy scoring," *Digital Evidence*, vol. 2, pp. 1– 19, 2004.

[13] W. Cavnar and J. Trenkle, "N-gram-based text categorization," *Proceedings of the Thin! Annual Symposium on Document Analysis and Infomation Retrieval*, pp. 161–175, 1994.

[14] D. Barbara, J. Couto, S. Jajodia, and N. Wu, "ADAM: A testbed for exploring the use of data mining in intrusion detection," *ACM SIGMOD Record*, vol. 30, no. 4, pp. 15–24, 2001.

[15] S. Mukkamala and A. Sung, "Identifying significant features for network forensic analysis using artificial intelligence techniques," *Digita! Evidence*, vol. l, no. 4, pp. 1–17, 2003.

[16] S. Stolfo, W. Lee, P. Chan, W. Fan, and E. Eskin, "Data mining based intrusion detectors: An overview ofthe columbia ids project," *ACM SIGMOD Record*, vol. 30, no. 4, pp. 5–14, 2001.

[17] M. Chau, J. Xu, and I.-I. Chen, "Extracting meaningful entities from police narrative reports," *Proceedings ofthe National Conference for Digital Government Research*, pp. 271–275, 2002.

[18] H. Chen, W. Chung, Y. Qin, M. Chau, J. Xu, G. Wang, R. Zheng, and H. Atabakhsh, "Crime data mining: An overview and case studies," *Proceedings of the National Conference for Digital Government Research*, p. 4, 2003.

[19] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: A general framework and some examples," *IEEE Computer*, vol. 37, pp. 50–56, 2004.

[20] R. Usual, H. Atabaiamh, P. Ongvmuh, H. Gupta, and H. Chen, "Using coplink to analyze criminal justice data," *IEEE Computer*, vol. 35, pp. 30–37, March 2002.

[21] H. C. G. Wang and H. Atabakhsh, "Automatically detecting deceptive criminal identities," *Communications of the ACM*, vol. 47, pp. 71–76, 2004.

[22] J. Xu and H. Chen, "Fighting organized crimes: Using shortest-path algorithms to identify associations in criminal networks," *Decision Support Systems*, vol. 38, pp. 473 – 487, 2004.

[23] P. Sinha, "A symmetry perceiving adaptive neural network and facial image recognition," *Forensic Science International*, vol. 98, pp. 67–89, 1998.

[24] D. M. Nassar and H. H. Ammar, "A neural network system for matching dental radiographs," *Pattern Recognition*, vol. 40, pp. 65–79, 2007.

[25] K. Kaikhah and S. Doddameti, "Discovering trends in large datasets using neural networks," *Applied Intelligence*, vol. 24, pp. 51–60, 2006.

[26] H.-J. Chen, S.-Y. Huang, and C.-L. Kuo, "Using the artificial neural network to predict fraud litigation: Some empirical evidence from emerging markets," *Expert Systems with Applications*, vol. 36, pp. 478–1484, 2009.

[27] T. H. Grubesic, "On the application of fuzzy clustering for crime hot spot detection," *Journal of Quantitative Criminology*, vol. 22, pp. 77–105, 2006.

[28] N. Liao, S. Tian, and T. Wang, “Network forensics based on fuzzy logic and expert system,” *Comput. Commun.*, vol. 32, no. 17, pp. 1881–1892, 2009.

[29] C. Quek, K. B. Tan, and V. K. Sagar, “Pseudo-outer product based fuzzy neural network fingerprint verification system,” *Neural Networks*, vol. 14, pp. 305–323, 2001.

[30] P. Castellano and S. Sridharan, “A two stage fuzzy decision classifier for speaker identification,” *Speech Communication*, vol. 18, pp. 139–149, 1996.

[31] S.-T. Li, S.-C. Kuo, and F.-C. Tsai, “An intelligent decision-support model using fsom and rule extraction for crime prevention,” *Expert Systems with Applications*, vol. 37, pp. 7108–7119, 2010.

[32] L. A. Zadeh, “Fuzzy sets,” *Information and Control*, vol. 8, pp. 338 – 353, 1965.

[33] *MatLab R2009a*, MATHWORKS, 2009.

[34] D. Li, A. Laurent, and P. Poncelet, “Discovery of unexpected fuzzy recurrence behaviors in sequence databases,” *Int. Journal of Computer Information Systems and Industrial Management Applications*, vol. 2, pp. 279–288, 2010.

[35] E. Klement, R. Mesiar, and E. Pap, *Triangular norms*. Kluwer, 2000.

[36] R. Yager, “On a general class of fuzzy connectives,” *Fuzzy Sets and Systems*, vol. 4, pp. 235 – 242, 1980.

[37] D. Dubois and H. Prade, *Fuzzy Sets and Systems: Theory and Applications*. Academic Press, New York, 1980.

[38] M. Sugeno, “Fuzzy measures and fuzzy integrals: a survey,” *Fuzzy Automata and Decision Processes*, pp. 89 – 102, 1977.

[39] L. A. Zadeh, “Outline of a new approach to the analysis of complex systems and decision processes,” *IEEE Trans. Systems, Man & Cybernetics*, vol. 1, pp. 28 – 44, 1973.

[40] S. Fukami, M. Mizumoto, and K. Tanaka, “Some considerations of fuzzy conditional inference,” *Fuzzy Sets and Systems*, vol. 4, pp. 243 – 273, 1980.

[41] E. Mamdani and S. Assilian, “An experiment in linguistic synthesis with a fuzzy logic controller,” *International Journal of Man-Machine Studies*, vol. 7, pp. 1 – 13, 1975.

[42] M. Sugeno, *Industrial applications of fuzzy control*. Elsevier Science, 1985.

[43] J. Dunn, “A fuzzy relative of the isodata process and its use in detecting compact, well separated clusters,” *Journ. Cybern*, vol. 3, pp. 95 – 104, 1974.

[44] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.

[45] L. Li, X. Liu, and M. Xu, “A novel fuzzy clustering based on particle swarm optimization,” in *First IEEE International Symposium on Information Technologies and Applications in Education*, 2007, pp. 88 – 90.

[46] G. Gan, J. Wu, and Z. Yang, “A genetic fuzzy k-modes algorithm for clustering categorical data,” *Expert Systems with Applications*, vol. 36, pp. 1615 – 1620, 2009.

[47] M. P. and P. SK., “Rough set based generalized fuzzy c-means algorithm and quantitative indices,” *IEEE Trans Syst Man Cybern B Cybern.*, vol. 37, pp. 1529 – 1540, 2007.

[48] Microsoft, “Microsoft Mappoint Homepage,” http://www.microsoft.com/mappoint/en-us/home.aspx.

[49] L. A. Zadeh, “The concept of a linguistic variable and its application to approximate reasoning,” *Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975.

Author Biographies

Kilian Stoffel is a professor of computer science in the Information Management Institute at the University of Neuchâtel (Switzerland). His main research interests include knowledge representation, machine learning and data mining.

Paul Cotofrei is senior lecturer in Information Management Institute, University of Neuchâtel. He received a PhD in statistics from the University of Bucarest (Romania) and a PhD in computer science from the University of Neuchâtel (Switzerland). He is currently working on stochastic frameworks for temporal data mining, process-driven data mining and qualitative modelling.

Dong Han is PhD student in Information Management Institute, the University of Neuchâtel. He received his bachelor degree of computer science from the University of Science and Technology Beijing(China) and his master degree of computer science from Beihang University(China). His research interests include knowledge representation, ontology, and data mining.