Strengths of Fuzzy Techniques in Data Science

Bernadette Bouchon-Meunier

Strengths of Fuzzy Techniques in Data Science

Bernadette Bouchon-Meunier[a]

[a]*Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6, UMR7606, 4 place Jussieu, 75005 Paris, France*

Abstract

We show that many existing fuzzy methods for machine learning and data mining contribute to providing solutions to data science challenges, even though statistical approaches are often presented as major tools to cope with big data and modern user expectations of their exploitation. The multiple capacities of fuzzy and related knowledge representation methods make them inescapable to deal with various types of uncertainty inherent in all kinds of data.

Keywords: Data science, fuzzy technique, uncertainty, fuzzy knowledge representation, linguistic summary, similarity

1. Introduction

Data science is progressively replacing data mining in the realm of big data analysis, at the crossroad of statistics and computer science. In the latter, machine learning has been one of the main components of data mining for several decades, together with statistics and databases. The modern massive amounts of data have clearly requested more advanced methods than in the past, in terms of efficiency, scalability, visualisation, and also with regard to their capacity to cope with flows of data, huge time series or heterogeneous types of data. To extract information from big data is nevertheless not sufficient to satisfy the final user expectations, more and more demanding not only rough information but also understandable and easily manageable knowledge. Criteria such as data quality, information veracity and relevance of information have always been important but they are now playing a crucial role in the decision support process pertaining to data science.

The acronym VUCA (volatility, uncertainty, complexity, ambiguity), commonly used in strategic management, can also characterise the system to which data science is applied. It seems clear that it is not sufficient to consider the only data, and the technical and final users must also be put in the loop. The sources of data should also be regarded, as well as their interactions in some cases, as their mutual effects may influence information quality. It is therefore necessary to have a systemic approach of data science, taking into account globally sources, data and users, and managing their characteristics to come to an effective knowledge able to support decisions. This is why it is important to

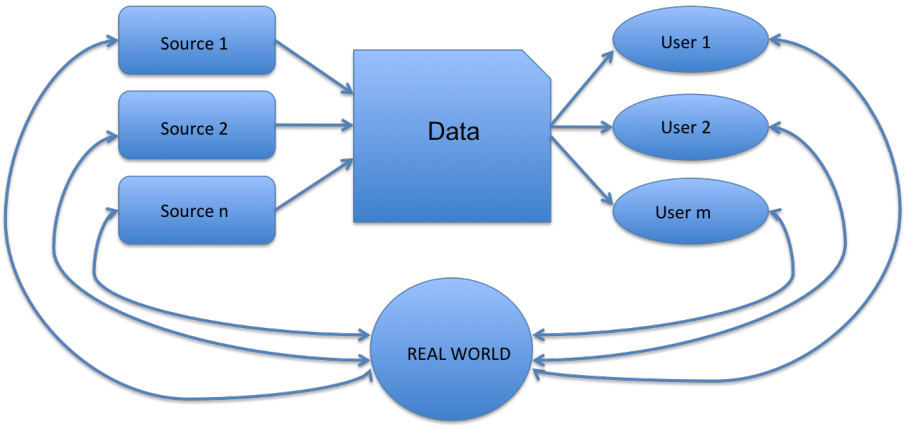

Figure 1: Description of the system

address the four characteristics we mentioned. The first characteristic is the *volatility* of information and it corresponds to the variability of the context, inherent in any evolving world, but made more significant in a digital environment in which data are produced and evolve constantly and quickly, for instance on the web, on social networks or when they are generated continuously by sensors. *Uncertainty* is the second characteristic and it refers to the handling of data subject to a doubt on their validity or being linked with forecasting or estimation, for instance in risk assessment under specific hypotheses. The third characteristic is the *complexity* of the real world about which data are available, only known through perceptions, measurements and knowledge representation, natural language being the most common. In addition, the complexity of human beings involved in the system must not be underestimated. The last characteristic of the considered system is the *ambiguity* of information which can result from the use of natural language, from conflicting sources or from incomplete information.

It is always possible to cope with these characteristics in data science by the only use of statistics and statistical machine learning. But are we sure that we don't lose substantial information and that we choose the most appropriate way to provide knowledge to the users? Can we consider alternative solutions or at least can we reinforce the existing ones by complementary approaches when appropriate? Such are the questions we would like to try to answer, looking at the existing methods proposed in data science.

It is interesting to compare these four characteristics of the whole system involved in data science to the commonly used Four V's introduced by IBM (volume, velocity, variety, veracity) to characterise the efficiency of solution proposed in Big Data, *volume*, *velocity* and *variety* corresponding to the capacity to manage huge amounts of data with a swiftness of the solution adapted to the volatility of data we mentioned earlier and taking into account heterogeneous

data. We can remark that volume, velocity and variety are parts of the complexity of the system. Veracity of data is related to the concept of uncertainty described above.

A proven means to deal with ambiguity, uncertainty, complexity and incompleteness in a system is to use a knowledge representation based on fuzzy modelling. In [1], the question of the need of fuzzy logic in machine learning is asked. If we extend this question to data science, we must ask if fuzzy logic is useful at the various levels of the process: in the representation of objects involved in the system, in the technique used to mine data regarding objects, in the presentation of results to the users, in the decision process resulting from the data analysis. We propose to see the methods already proposed at these levels for machine learning, data mining or data science. Our purpose is not to prepare a survey on fuzzy approaches to data science, which should have a considerable extension going far beyond the size of this article, given the variety of works existing on this topic, but to point out the diversity of tools available to cope with imperfect information.

In this paper, we propose to analyse the capacity of fuzzy set modelling to provide solutions to cope with these characteristics of the system, in what concerns knowledge representation in the first section and in data analysis techniques in the second one. Our purpose is not to provide an exhaustive state of the art of works on these two domains, which would require a complete book, but to draw the attention of the user to solutions which can cooperate with statistical or symbolic methods in order to solve the mentioned problems.

2. Knowledge representation

We must first remark that fuzzy sets, at the root of fuzzy modelling, are nothing else than a means to represent knowledge, as are natural numbers, percentages, words or images. There is obviously no fuzzy object in the real world, as there is no crisp object, and it is only our perception of the real world, our information or knowledge about it and the purpose of our task which can lead to a fuzzy or crisp representation. For instance, can a forest be regarded as a crisp object? Sure, it has a name and it is well identified by crisp boundaries on a map. On the other hand, can a forest be regarded as a fuzzy object? Of course, as to define the limit of the forest on the earth depends on the compatibility between the cadastral plans and the requested level of precision and it is difficult to claim that a bush at the limit of the forest is inside or outside. We can draw a precise map of the forest because an approximation is done and the scale of the map does not enable us to see significant difference between the possible limits of the forest. If we now consider an artefact such as a spot detected on a mammogram, expert analyses show that it does not have precise limits [2] and it is better represented as a fuzzy object, whose attributes are automatically evaluated by means of fuzzy values.

The existence of fuzzy objects is one reason to justify the use of fuzzy modelling in data science. We can always decide to ignore the fuzziness of an object

but we must note that some utilisations of the objects may require a crisp representation of them while some others take advantage of preserving a flexible description of the object. Another reason is the existence of non standard methods in the framework of fuzzy modelling, enriching the toolbox of data scientists. Fuzzy knowledge representation is multiple and, even though the use of fuzzy sets to represent approximate values or imprecisely defined objects is its most common aspect, we must not ignore associated methods to represent data, information and knowledge. First of all, there exist many knowledge representation methods classic in artificial intelligence which have been extended to or replaced by fuzzy ones in specific environments. It is the case of ontologies, description logic or causal networks for instance. In addition, related methods based on possibility and necessity measures correspond to the representation of uncertainty rather than imprecision associated with the available information. We should also mention other methods in the fuzzy modelling family such as rough sets, intuitionistic fuzzy sets, or type-2 fuzzy sets that have their specificity and propose to manage more complex aspects of imprecision and uncertainty. Another important knowledge representation method in the fuzzy framework corresponds to linguistic summaries of time series, based of fuzzy description of variables and fuzzy quantifiers. Last but not least, fuzzy modelling includes similarity measures, be they used to compare fuzzy or crisp objects.

2.1. Fuzzy sets and possibility degrees

Speaking of fuzzy modelling to cope with information ambiguity, it is immediate to refer to the representation of linguistic terms by fuzzy sets as an interface between numerical and symbolic data taking their imprecision into account, such as "big" or the representation of approximate numerical values such as "approximately 120", through a membership function lying on the universe of discourse and taking values in $[0, 1]$, with a core corresponding to membership degrees equal to 1 associated with elements of the universe belonging absolutely to the fuzzy set, and a support out of which the elements of the universe do not belong at all to the fuzzy set. Many solutions exist to define membership functions, from psychometric ones to automatic ones by means of machine learning methods. Such fuzzy sets are used in the more elaborate fuzzy models described in the next three subsections.

We should nevertheless not forget the option to represent subjective uncertainty by means of possibility distributions associated with fuzzy or crisp sets. Possibility degrees correspond to the consideration of a doubt on the validity of a piece of information and the dual necessity degrees represent the certainty on such a piece of information. They have for instance been used in the evaluation of data quality to deal with the uncertainty in the system and veracity of available data we mentioned in the first section [3] [4].

2.2. Rule bases

Even though E. Hüllermeier [1] claims that the interpretability of fuzzy models is a myth, the expressiveness of fuzzy models is certainly one of their most interesting qualities. Fuzzy rules such as "$if\ V1\ is\ A1\ and\ V2\ is\ A2\ and ...\ then\ W\ is\ B$"

have long been considered as the most common fuzzy knowledge representation tool because it was considered as an easy way to elicit knowledge from experts. They are extensively used in decision-making support, more than in data science where they mainly appear in the interpretation of decision trees. Many criteria have been proposed to evaluate their interpretability [5] and, more generally their appropriateness to establish an interface between the system and the user, on the basis of compactness, completeness and consistency of a collection of rules, as well as coverage, normality and distinguishability of fuzzy modalities used in the rules [6]. It is well recognized that a too complex system of fuzzy rules makes it lose its interpretability capacity, and a trade off must be found between understandability of the system and accuracy of the provided information.

2.3. Linguistic summaries

The concept of interpretability itself is difficult to define, depending on the domain and the category of users. However, among other interesting fuzzy models, we would like to focus on linguistic summaries [7][8], that combine the understandability of simplified natural language and the capacities of automatic learning and quality checking, the quality being understood in various senses. Their purpose is to sum up information contained in large volumes of data into simple sentences and the interpretability is at the core of the process [9]. The most generally used sentences, called protoforms, are of the form "$Q\,B\,x's\,are\,A$", where Q is a fuzzy quantifier representing a linguistic quantifier such as "most" or "a majority of", or, in the case of time series, a temporal indication such as "often" or "regularly", B is a fuzzy qualifier of elements x of the dataset to be summarised, sometimes omitted, and A is a fuzzy description of these elements called a summariser. Examples of such protoforms are "Most of the cold days are windy" or "Approximately every day, the amount of CO2 is high".

Fuzzy linguistic summaries can be compared to other methods to extract information from large datasets such as temporal series. Since their main quality is their expressiveness, it looks pertinent to compare them with linguistic summaries obtained by means of natural language generation. Even if the latter is naturally semantically richer, the information provided by fuzzy linguistic summaries has the advantage of not requiring any expert knowledge as it is generally the case for natural language generation-based summaries. It is also made of simple sentences, the form of which depends on the needs of the user, in adequacy with the context. A degree of truth is calculated from the dataset for each protoform. Either the user is directly provided with a collection of protoforms as a summary of the dataset or he/she uses queries to obtain information regarding summarizers and qualifiers of interest for him/her [10]. In a general environment, the number of sentences generated by a list of quantifiers, qualifiers and summarisers may be big and the most interesting ones can be selected automatically on the basis of their level with regard to a chosen criterion, for instance the degree of focus, specificity or informativeness [11]. In the case of queries, various interactive solutions have been proposed to enable

the user to easily find appropriate answers to his queries [12]. The number of sentences can also be reduced by taking into account properties of inclusion between quantifiers or summarizers, for instance. Another consideration enabling to reduce the number of protoforms is the management of oppositions in order to ensure the consistency of the collection of protoforms proposed to the user [13], eliminating contradictions and exploiting duality and antonymy. Constraints on membership degrees can be taken into account [14] to analyse the coherence of fuzzy descriptions. In the particular case of the summarisation of temporal series, which has given rise to many methods in statistical learning, the diversity of sentences used in the summaries must be pointed out, going beyond the usual protoforms. Trends are often taken into consideration [15], as well as fuzzy temporal propositions [16], detection of local changes [17], to cite but a few examples.

We focus on the analysis of periodicity of time series, which can obviously be approximative or described imprecisely, for instance of the form "$Many\ x's\ are\ A\ most\ of\ the\ time$" [18]. To analyze the regularity of high and low values, the periodicity of such behaviors and their approximate duration can for instance be achieved through an efficient scalable and robust method [19] requesting neither any hypothesis on the data nor any tuning of parameters, automatically detecting groups of high and low values and providing simple natural language descriptions of the periodicity.

2.4. Fuzzy ontologies

Ontologies are an important knowledge representation tool used in many aspects of information or image retrieval and semantic web to manage concepts and their relationships in a structured environment. Description logic is an efficient way to construct ontologies in order to manage concepts, roles and individuals. If we assume that most concepts are imprecise and their relations as well, there is a clear need of fuzzy ontologies which have been extensively studied and applied. Fuzzy description logics have been proposed [20] to construct fuzzy ontologies in the case where concepts and relations are imprecise, in the framework of fuzzy logic. They can correspond to the idea of unclear boundaries of concepts or relationships, or imperfect knowledge about individuals, which goes far beyond taking into account synonymy or forms of words like plural or tense, as commonly managed by natural language processing, or even misprint correction. It is not the style of descriptions which is addressed but their content itself. A number of works [21] [22] have extended the Web Ontology Language (OWL) based on Description Logic to construct fuzzy ontologies. Among the most recent ones, fuzzyDL is an ontology reasoner supporting fuzzy logic reasoning tasks such as concept satisfiability, concept subsumption or entailment. [23]. An alternative to fuzzy description logic when the available knowledge is uncertain consists of possibilistic description logic [24][4], dealing with uncertain roles and individuals. It is based on possibilistic logic, managing gradual and subjective uncertainties and assigning confidence degrees to pieces of information. Fuzzy ontologies have been extensively used in medical applications, in ubiquitous learning, in sentiment analysis on social media sites or in information

retrieval and in particular semantic similarity, for instance. Possibilistic logics have been used in military intelligence services and for the semantic web.

2.5. Similarity measures

Similarity pertains to knowledge representation as it contributes to the construction of categories or classes representing the available knowledge. In addition, similarity can be viewed as a way to represent knowledge on relations between elements in a system observed in data science, for instance. It is a complex concept, much investigated in psychology from psychometrical and cognitive points of view. It is involved in categorization to reduce the amount of available information and cognitive categories have been pointed out to be fuzzy, for instance by E. Rissland [25]. who considers that many concepts have "grey areas of interpretation" with a core and a boundary region. Similarities are then useful to construct categories.

They have been used in data science in a restrictive approach, which could be used with more diversity and richness than it is, especially considering a fuzzy environment. In data science, similarity is often regarded as the dual of a distance, which requires a metrical space; another commonly used similarity measure is the cosine of the angle between two vectors, but such similarity measures neglect conceptual and perceptual aspects of similarities. Considering that two objects are similar clearly depends on the point of view: images of bats and squirrels can be regarded as similar with regard to the concept of mammals; images of bats and owls can be regarded as visually similar because they represent animals flying in the night. The concept of animal flying in the night itself is fuzzy, since squirrels partly belong to it because of the existence of flying squirrels for instance. The similarity between two elements clearly depends on the purpose of the analysis being performed. We must note that similarities can be symmetrical or not, according to Tversky's seminal work [26]. For instance, if one of the elements serves as a reference, appearing in a query or being the prototype or the representative of a category to which an unknown element is compared, then the similarity is not necessarily symmetrical. In the case when elements to compare are fuzzy, similarities take into account membership functions describing them. Classes of measures of similarity have been exhibited [27], including (non-symmetrical) satisfiability measures, (symmetrical) resemblance measures, inclusion measures involved in the comparison of categories, for instance. The richness of the available classes of similarity measures provides appropriate solutions to all utilisations of similarities related to fuzzy knowledge representation: to find relevant answers to database queries, taking into account the term fuzziness as well as a flexible matching between terms and fuzzy ontology-based similarity between terms [28], for missing data imputation [29]. An utilization of similarities of particular interest is the construction of prototypes of categories. Again on the basis of psychological studies [30], fuzzy prototypes can be defined as representatives of an imprecisely characterized class, the most similar to all members of the class and the most dissimilar to members of other classes [31].

3. Conclusion

We have pointed out various reasons to use fuzzy techniques to cope with all characteristics of data pertaining in data science, in particular volatility, uncertainty, complexity, ambiguity, incompleteness, heterogeneity. The major problems of data and information quality have not yet been enough tackled in data science, but it is clear that some already existing fuzzy and possibilistic methods are promising and should give rise to efficient solutions in the future.

References

[1] E. Hüllermeier, Does machine learning need fuzzy logic?, Fuzzy Sets and Systems 281 (2015) 292–299, special Issue Celebrating the 50th Anniversary of Fuzzy Sets.

[2] S. Bothorel, B. Bouchon Meunier, S. Muller, A fuzzy logic based approach for semiological analysis of microcalcifications in mammographic images, International Journal of Intelligent Systems 12 (11-12) (1997) 819–848.

[3] M.-J. Lesot, T. Delavallade, F. Pichon, H. Akdag, B. Bouchon-Meunier, P. Capet, Proposition of a semi-automatic possibilistic information scoring process, in: The 7th Conf. of the European Society for Fuzzy Logic and Technology (EUSFLAT-2011) and LFA-2011, Atlantis Press, 2011, pp. 949–956.

[4] O. Couchariere, M.-J. Lesot, B. Bouchon-Meunier, Consistency checking for extended description logics, in: International Workshop on Description Logics (DL 2008) CEUR vol. 353, Dresden, Germany, 2008.

[5] M. Gacto, R. Alcala, F. Herrera, Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures, Information Sciences 181-20 (2011) 4340–4360.

[6] J. Casillas, O. Cordon, F. Herrera, L. Magdalena, Interpretability improvements to find the balance interpretability-accuracy in fuzzy modeling: an overview, Springer Berlin Heidelberg, 2003, pp. 3–22.

[7] R. R. Yager, A new approach to the summarization of data, Information Sciences 28 (1) (1982) 69–86.

[8] J. Kacprzyk, R. R. Yager, Linguistic quantifiers and belief qualification in fuzzy multicriteria and multistage decision making, Control and Cybernetics 13 (3) (1984) 154–173.

[9] M.-J. Lesot, G. Moyse, B. Bouchon-Meunier, Interpretability of fuzzy linguistic summaries, Fuzzy Sets and Systems 292 (2016) 307–317, special Issue in Honor of Francesc Esteva on the Occasion of his 70th Birthday.

[10] J. Kacprzyk, S. Zadrozny, Linguistic database summaries and their protoforms: Towards natural language based knowledge discovery tools, Inf. Sci. Inf. Comput. Sci. 173 (4) (2005) 281–304.

[11] J. Kacprzyk, A. Wilbik, Towards an efficient generation of linguistic summaries of time series using a degree of focus, in: Proceedings of the NAFIPS, 2009, pp. 1–6.

[12] J. Kacprzyk, S. Zadrożny, Fuzzy linguistic data summaries as a human consistent, user adaptable solution to data mining, Springer, 2005, pp. 321–340.

[13] G. Moyse, M. J. Lesot, B. Bouchon-Meunier, Oppositions in fuzzy linguistic summaries, in: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2015), 2015, pp. 1–8.

[14] M. Delgado, M. D. Ruiz, D. SÃ¡nchez, M. A. Vila, Fuzzy quantification: a state of the art, Fuzzy Sets and Systems 242 (2014) 1–30.

[15] J. Kacprzyk, A. Wilbik, S. Zadrozny, Linguistic summarization of time series using a fuzzy quantifier driven aggregation, Fuzzy Sets and Systems 159, 12 (2008) 1485–1499.

[16] P. Cariñena, A. Bugarín, M. Mucientes, S. Barro, A language for expressing fuzzy temporal rules., Mathware and Soft Computing 7 (2-3) (2000) 213–227.

[17] R. Castillo-Ortega, N. Mann, D. Sánchez, Linguistic local change comparison of time series, in: Fuzzy Systems (FUZZ), 2011 IEEE International Conference on, IEEE, 2011, pp. 2909–2915.

[18] R. J. Almeida, M.-J. Lesot, B. Bouchon-Meunier, U. Kaymak, G. Moyse, Linguistic summaries of categorical time series for septic shock patient data, in: 2013 IEEE International Conference on Fuzzy Systems, IEEE, 2013, pp. 1–8.

[19] G. Moyse, M.-J. Lesot, Linguistic summaries of locally periodic time series, Fuzzy Sets and Systems 285 (2016) 94–117.

[20] U. Straccia, Reasoning within fuzzy description logics, Journal of Artificial Intelligence Research 14 (2001) 137–166.

[21] S. Calegari, D. Ciucci, Fuzzy ontology, fuzzy description logics and fuzzy-owl, in: International Workshop on Fuzzy Logic and Applications, Springer, 2007, pp. 118–126.

[22] J. Liu, B. Zheng, L. Luo, J. Zhou, Y. Zhang, Z. Yu, Ontology representation and mapping of common fuzzy knowledge, Neurocomputing 215 (2016) 184–195.

[23] F. Bobillo, U. Straccia, The fuzzy ontology reasonerÂ fuzzydl, Knowledge-Based Systems 95 (2016) 12 – 34.

[24] G. Qi, Z. Pan, Q. Ji, Extending description logics with uncertainty reasoning in possibilistic logic, in: Proc. of the European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU 2007, 2007, pp. 828–839.

[25] E. L. Rissland, Ai and similarity, IEEE Intelligent Systems 21 (3) (2006) 39–49.

[26] A. Tversky, Features of similarity., Psychological review 84 (4) (1977) 327–352.

[27] B. Bouchon-Meunier, M. Rifqi, S. Bothorel, Towards general measures of comparison of objects, Fuzzy sets and systems 84 (2) (1996) 143–153.

[28] J. Liu, B.-J. Zheng, L.-M. Luo, J.-S. Zhou, Y. Zhang, Z.-T. Yu, Ontology representation and mapping of common fuzzy knowledge, Neurocomputing 215 (2016) 184–195.

[29] D. Li, J. Deogun, W. Spaulding, B. Shuart, Towards missing data imputation: a study of fuzzy k-means clustering method, in: Rough sets and current trends in computing, Springer, 2004, pp. 573–579.

[30] E. Rosch, Principles of categorization, in: E. Rosch, B. Lloyd (Eds.), Cognition and categorization, Lawrence Erlbaum, 1978, pp. 27–48.

[31] M.-J. Lesot, M. Rifqi, B. Bouchon-Meunier, Fuzzy prototypes: From a cognitive view to a machine learning principle, in: Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, Springer Berlin Heidelberg, 2008, pp. 431–452.