Mining and Visualizing Uncertain Data Objects and Named Data Networking Traffics by Fuzzy Self-Organizing Map

Amin Karami[1,2] and Manel Guerrero-Zapata[1,2]

[1] Computer Architecture Department (DAC), Universitat Politècnica de Catalunya
(UPC), Campus Nord, C. Jordi Girona 1-3. 08034 Barcelona, Spain
[2] amin@ac.upc.edu and guerrero@ac.upc.edu

Abstract. Uncertainty is widely spread in real-world data. Uncertain data -in computer science- is typically found in the area of sensor networks where the sensors sense the environment with certain error. Mining and visualizing uncertain data is one of the new challenges that face uncertain databases. This paper presents a new intelligent hybrid algorithm that applies fuzzy set theory into the context of the Self-Organizing Map to mine and visualize uncertain objects. The algorithm is tested in some benchmark problems and the uncertain traffics in Named Data Networking (NDN). Experimental results indicate that the proposed algorithm is precise and effective in terms of the applied performance criteria.

1 Introduction

Uncertainty is a frequent issue in data analysis. The various factors that lead to data uncertainty include: approximate measurement, data sampling fault, transmission error or latency, data integration with noise, data acquisition by device error, and so on [1] [2]. These factors produce vague and imprecise data. Visualizing uncertain data is one of the new challenges in the uncertain databases [3]. Among the many visualization techniques, the Self-Organizing Map (SOM) [4] is widely and successfully applied due to its good result. SOM is a very popular unsupervised learning algorithm based on the classical set theory. An important application of SOM is discovering the topological relationship among multidimensional input vectors and mapping them to a low dimensional output which is easy for further analysis by experts [5] [6]. The process of SOM training requires a certain and an unambiguous input data either belongs or not belong to a weight vector (cluster), where the membership evaluation is boolean. In contrast, uncertain and vague input vectors are not either entirely belong or not belong to a weight vector. A data may be considered vague and imprecise where some things are not either entirely true nor entirely false and where the some things are somehow ambiguous. For instance, fuzzy location in the right side of Fig. 1 is a way to represent the item of vague information: the object is *approximately* at position (4, 3), in which the grey levels indicate membership values with white representing 0 and black representing 1. In contrast, the left side of Fig. 1 shows the exact position of a certain data where the membership evaluation of centers (weights) is boolean. There has been a lot of research in the application of Fuzzy sets theory to model vague and uncertain information [7]. The Fuzzy set (FS) theory introduced by Zadeh [8] is a more flexible approach than classical set theory, where objects belong to sets (clusters) with certain degree of membership ranging [0..1]. This makes FS theory suitable for representing and visualizing uncertain data [9]. Therefore, a combination of SOM and FS is able to illustrate dependencies in the uncertain data sets in a very intuitive manner. SOM is indeed originally intended as a classification method, not a visualization method so there are a few additions to apply SOM for visualization. Li et al. [3] proposed a mining and visualizing algorithm for uncertain data, called USOM which combines fuzzy distance function and SOM. In this paper, we employ the FS theory in the context of SOM algorithm to mine and visualize the uncertain objects in the uncertain databases. Experimental results over four classic benchmark problems and a new network architecture as Named Data Networking (NDN) show that the proposed method outperforms standalone SOM and USOM [3] in terms of the applied performance metrics. The remainder of the paper is organized as follows: Section 2 presents self-organizing map. Section 3 presents our contribution. Section 4 evaluates the new approach experimentally. Section 5 is the conclusion and future work.

2 Self-Organizing Map (SOM)

SOM (also known as Kohonen SOM) is a very popular algorithm based on competitive and unsupervised learning [4]. The SOM projects and represents higher dimensional data in a lower dimension, typically 2-D, while preserving the relationships among the input data. The main process of SOM is generally introduced in three main phases: competition, cooperation and adaptation which are described in detail in [4].

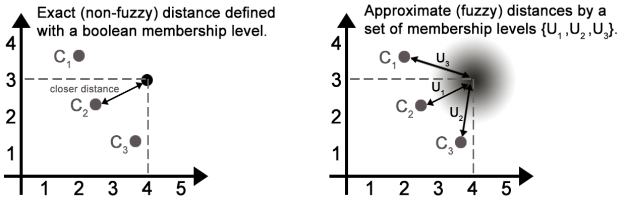

Exact (non-fuzzy) distance defined with a boolean membership level.
closer distance
C1
C2
C3
1 2 3 4 5
1 2 3 4
Approximate (fuzzy) distances by a set of membership levels {U1 ,U2 ,U3}.
U1
U2
U3

Fig. 1: An example of exact (non-fuzzy) and approximate (fuzzy) distances in a 2-D space for a certain and vague data.

3 The Proposed Method

The procedure of the proposed method, application of fuzzy set theory in the context of SOM for mining and visualizing uncertainties is as follows. A diagram of the proposed method is shown in Fig. 2.

1. Fuzzy competition: in *hard competition*, the input vector is divided into distinct weights (clusters), where each input element belongs to exactly one weight. In *fuzzy competition*, input vector can belong to more than one weight, and associated with each element by a set of membership levels. Fuzzy c-means (FCM) [10] method allows one piece of input data to belong to two or more clusters (weights). The standard function is:

$$U_x = \frac{1}{\sum_j \left(\frac{d(weight_k,x)}{d(weight_j,x)}\right)^{\frac{2}{m-1}}} \tag{1}$$

   Where, U_x is the membership value of each input vector x to all weights, $j = 1, 2, ..., w$, and m is the level of cluster fuzziness which is commonly set to 2. By the fuzzy competition all the neurons are wining neurons with the membership degree ranging [0..1].
2. Fuzzy cooperation: in fuzzy cooperation, all wining neurons cooperate with their neighboring neurons in terms of the membership degree by Eq. 2. For the size of the neighborhood, we employed the Gaussian function that shrinks on each iteration until eventually the neighborhood is just the BMU itself.

$$h(j,i) = U_{xi} \times exp(\frac{-{d_{j,i}}^2}{2\sigma^2}) \quad i, j = 1, 2, .., n; \;\; i \neq j \tag{2}$$

   Where, i is the number of the wining neurons including all the neurons with different membership degrees, j is the number of the cooperating neighbor neurons. U_{xi} is the membership value of input vector x from i^{th} wining neuron. $h(j,i)$ is the topological area centered around the wining neuron i and the cooperating neuron j. The size σ of the neighborhood needs to decrease with time. A popular time dependence is an exponential decay by:

$$\sigma(t) = \sigma_0 exp(\frac{-t}{\lambda}) \tag{3}$$

   Where, $\sigma(t)$ is the width of the lattice at time t, σ_0 is the width of the lattice at time t_0, and λ is the time constant.
3. Fuzzy adaption: the adaption phase is the weight update by:

$$w_j = w_j + U_j \times (\eta h(j,i) \times (x - w_j)) \quad i, j = 1, 2, .., n; \;\; i \neq j \tag{4}$$

   Where, U_j is the membership value of input x from neuron j.

These three phases are repeated, until the maximum number of iterations is reached or the changes become smaller than a predefined threshold.

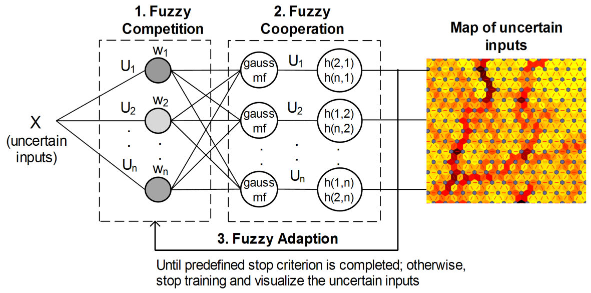


Fig. 2: The proposed method for mining and visualizing uncertainties.

4 Experimental Results

The proposed method, USOM and SOM were implemented by the MATLAB on an Intel Pentium 2.13 GHz CPU, 4 GB RAM running Windows 7 Ultimate.

4.1 Uncertain data modeling

To assess the accuracy and performance of the proposed method, four classic benchmark problems from the UCI machine learning repository [11] are applied. The selected data sets are Iris (4-D), Glass (9-D), Wine (13-D), and Zoo (17-D). In practice, uncertainties are usually modeled in the form of Gaussian distributions [2]. For some attributes in data sets, we add a Gaussian noise with a zero mean and the standard deviation with the normal distribution $[0, f]$, where, f is an integer parameter from the set of $\{1, 2, 3\}$ to define different uncertain levels.

4.2 Assessing the quality of visualizations

To assess the quality of the proposed method, several measures have been applied, including Quantization Error (QE), Topographic Error (TE), Trustworthiness of a Visualization, and Continuity of the Neighborhoods [12].

4.3 Visualization results

The experiments on each method were repeated 10 times independently. We evaluate the several SOM network structures on applied uncertain data sets which the optimal ones are Iris with 16x16 nodes, Glass with 16x16 nodes, Wine with 17x17 nodes, and Zoo with 15x15 nodes.

Table 1 shows that our proposed method outperforms SOM and USOM methods in terms of the Quantization Error (QE) and Topographic Error (TE). The proposed method seems to be more time consuming (with Exec.) than the other

Table 1: Performance improvements achieved by the proposed scheme

Data	SOM				USOM				Proposed Method			
			Time				Time				Time	
	QE	TE	Exe.	Inc.	QE	TE	Exe.	Inc.	QE	TE	Exe.	Inc.
Iris (16x16)	0.024	0.0404	9.6	5.45	0.023	0.034	11.34	4.16	0.02	0.0267	16.43	2.17
Glass (16x16)	0.066	0.0312	21.17	15.45	0.042	0.02	23.11	14.23	0.028	0.0174	26.82	10.1
Wine (17x17)	0.072	0.0381	18.87	12.05	0.06	0.022	19.12	10.16	0.049	0.0102	22.07	7.74
Zoo (15x15)	0.067	0.0215	15.54	11.23	0.046	0.016	18.51	10.36	0.039	0.0103	21.34	8.52

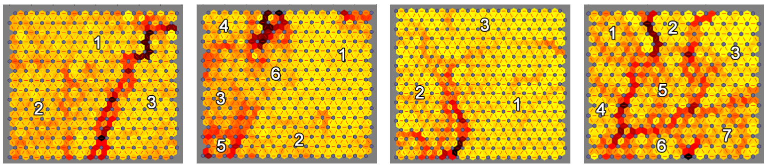
Table 2: The quality measurement by Trustworthiness

Data	SOM			USOM			Proposed Method		
	K=1	K=10	K=20	K=1	K=10	K=20	K=1	K=10	K=20
Iris (16x16)	0.937	0.94	0.93	0.95	0.962	0.968	0.962	0.968	0.974
Glass (16x16)	0.913	0.903	0.898	0.914	0.921	0.933	0.915	0.93	0.939
Wine (17x17)	0.917	0.921	0.904	0.924	0.941	0.953	0.925	0.951	0.962
Zoo (15x15)	0.957	0.96	0.96	0.962	0.963	0.968	0.963	0.97	0.978

Table 3: The quality measurement by Continuity

Data	SOM			USOM			Proposed Method		
	K=1	K=10	K=20	K=1	K=10	K=20	K=1	K=10	K=20
Iris (16x16)	0.945	0.901	0.892	0.961	0.964	0.966	0.97	0.974	0.982
Glass (16x16)	0.911	0.898	0.873	0.914	0.916	0.92	0.92	0.931	0.937
Wine (17x17)	0.921	0.892	0.883	0.93	0.931	0.935	0.935	0.939	0.941
Zoo (15x15)	0.86	0.841	0.812	0.89	0.898	0.902	0.91	0.918	0.927

methods due to the application of fuzzy set theories in the context of the SOM, in which all the neurons are winner with different membership grading. However, the proposed method can find a better solution with less times of increment on computational time (with Inc.) than the other methods due to its fast convergence speed. The trustworthiness and continuity values for K={1, 10, 20} are shown in Tables 2 and 3, respectively. The trustworthiness and continuity measures show that the proposed method obtains the better results as compared to SOM and USOM. The results show that the proposed method with the application of fuzzy set theory in the context of the SOM yields high accuracy as compared to other methods without very much computational cost. Since our proposed method performs well as compared to SOM and USOM, we visualize uncertainties in the applied uncertain data sets. To facilitate the interpretation of results, we use the U-Matrix (unified distance matrix) where visualize the high-dimensional uncertain data into a 2-D space in Fig. 3. In this figure, the blue hexagons represent the neurons (weights). The darker colors in the regions between neurons represent larger distance, while the lighter colors represent smaller distances. Fig. 3(a) shows that the constructed 4-D uncertain Iris SOM network has been clearly clustered into three distinct groups. The Glass SOM



(a) (b) (c) (d)

Fig. 3: U-Matrix of the applied benchmark problems: 3(a) Iris 16x16 SOM, 3(b) Glass 16x16 SOM, 3(c) Wine 17x17 SOM, and 3(d) Zoo 15x15 SOM.

Table 4: NDN traffic generation

Type of traffic		Frequency	Pattern
Normal (526 records)		[100..500]	Exponential
Attack (211 records)	Cache pollution	[200..800]	Locality-disruption attacks uniformly
	DoS attacks	[400..1500]	Interest flooding attacks for non-existent and existent content uniformly and exponentially

network (Fig. 3(b)) has been apparently classified 9-D uncertain data objects into six distinct types of glass. Figs. 3(c) and 3(d) show the three and the seven distinct groups of 13-D and 17-D uncertain data from Wine and Zoo data sets, respectively. The results confirm that the proposed method performs well in mining and visualizing uncertain data into somewhat expected distinct groups.

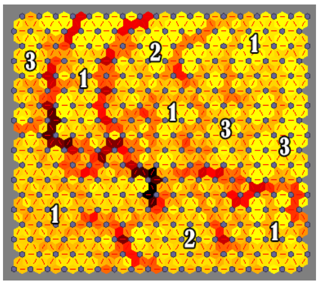


Fig. 4: U-Matrix of the NDN traffic. 1: normal, 2: DoS attack, 3: cache pollution attack.

4.4 Visualizing uncertain traffics in Named Data Networking

After evaluating the robustness and the accuracy of our proposed method with some benchmark problems, we apply the proposed method for visualizing uncertain traffics in Named Data Networking (NDN). NDN [13] is a promising network architecture being considered as a possible replacement to overcome the fundamental limitations of the current IP-based Internet. Traffic uncertainty refers to traffic volumes belong to more than one pattern, and associated with each pattern by a set of membership levels. Fuzzy approach can reduce the false positive rate with higher reliability in identifying the pattern of traffic volumes, due to any uncertain attack data may be similar to some normal patterns [14]. We conduct the same testbed configuration from papers [14] [15]. The employed features for traffic generation come from paper [14] as well as the ratio of (1) cache hit, (2) dropped Interest packet, (3) dropped data packets, (4) satisfied Interest packet, and (5) timed-out Interest packets in each 1 sec time interval. The structure of the traffic generated is shown in Table 4. We modeled uncertainties for some attributes in NDN traffic samples in the form of Gaussian distributions similar to Section 4.1. Fig. 4 maps the 11-D uncertain traffic samples to the 2-D space through our proposed method. This figure shows that the our proposed method performs somewhat well in mining and visualizing uncertainties into predefined distinct groups. Fig. 4 illustrates that there are some small groups of clustered data points with the lighter regions. These small clusters may contain some normal or attack data that try to be incorrectly placed in the neighboring regions, due to their uncertain nature. The results in Table 5 show that our proposed method offers the best performance and outperforms sufficiently other preexisting methods.

Table 5: Comparing results of visualizing NDN traffic samples

Criteria		Methods		
		SOM	USOM	Proposed Method
Quantization Error		0.042	0.029	0.0125
Topographic Error		0.074	0.053	0.031
Trustworthiness	K=1	0.91	0.95	0.968
	K=15	0.905	0.943	0.954
	K=30	0.877	0.925	0.942
Continuity	K=1	0.914	0.922	0.954
	K=15	0.893	0.931	0.941
	K=30	0.867	0.917	0.936

5 Conclusion

In this paper, we propose a new hybrid algorithm for mining and visualizing uncertain data. We investigate the implementation of fuzzy set theory in the design of SOM neural network in order to improve the accuracy of visualizing uncertain data bases. The experimental results over the uncertain benchmarking data sets and the uncertain traffics in Named Data Networking show that the proposed method is effective and precise in terms of the applied performance criteria. We plan to improve the proposed method for various uncertain models and big uncertain network traffic data in the future.

6 Acknowledgment

This work was partially supported by projects TIN2013-47272-C2-2 and SGR-2014-881.

References

1. Pavle Milošević, Bratislav Petrović, Dragan Radojević, and Darko Kovačević. A software tool for uncertainty modeling using interpolative boolean algebra. *Knowledge-Based Systems*, 62:1 – 10, 2014.
2. Jiaqi Ge, Yuni Xia, and Yicheng Tu. A discretization algorithm for uncertain data. In *Proceedings of the 21st International Conference on Database and Expert Systems Applications: Part II*, pages 485 – 499, 2010.
3. Le Li, Xiaohang Zhang, Zhiwen Yu, Zijian Feng, and Ruiping Wei. Usom: Mining and visualizing uncertain data based on self-organizing maps. In *International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 804 – 809, 2011.
4. T. Kohonen. *Self-Organizing Maps.* Springer, Berlin, Heidelberg, 1995.
5. Mohammed Khalilia and Mihail Popescu. Topology preservation in fuzzy self-organizing maps. *Advance Trends in Soft Computing*, 312:105 – 114, 2014.
6. Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586 – 600, 2000.
7. D. Herrero-Pérez, H. Martnez-Barberá, K. LeBlanc, and A. Saffiotti. Fuzzy uncertainty modeling for grid based localization of mobile robots. *International Journal of Approximate Reasoning*, 51(8):912 – 932, 2010.
8. Lotfi A. Zadeh. Fuzzy sets. *Information Control*, 8:338 – 353, 1965.
9. Feng Qi and A-Xing Zhu. Comparing three methods for modeling the uncertainty in knowledge discovery from area-class soil maps. *Computers & Geosciences*, 37(9):1425 – 1436, 2011.
10. J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algoritms.* Plenum Press, New York, 1981.
11. A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.
12. Gonzalo A. Ruz and Duc Truong Pham. Nbsom: The naive bayes self-organizing map. *Neural Comput. Appl.*, 21(6):1319 – 1330, 2012.
13. Diana K. Smetters James D. Thornton Michael F. Plass Nicholas H. Briggs Jacobson, Van and Rebecca L. Braynard. Networking named content. In *In Proceedings of the 5th international conference on Emerging networking experiments and technologies*, pages 1 – 12, 2009.
14. Amin Karami and Manel Guerrero-Zapata. A fuzzy anomaly detection system based on hybrid pso-kmeans algorithm in content-centric networks. *Neurocomputing*, 2014.
15. Amin Karami. Data clustering for anomaly detection in content-centric networks. *International Journal of Computer Applications*, 81(7):1 – 8, 2013.