Hacettepe Journal of Mathematics and Statistics
Volume 41 (3) (2012), 397 – 405

A NEW CLUSTERING SCHEME FOR CRISP DATA BASED ON A MEMBERSHIP FUNCTION AND OWA OPERATOR

Murat Alper Basaran*†, Alparslan A. Basaran‡, Biagio Simonetti§ and Antonio Lucadamo§

Received 12 : 01 : 2012 : Accepted 28 : 08 : 2012

Abstract

Clustering is a very important tool which is applied in several areas, ranging from pattern recognition and marketing to chemistry. A majority of the clustering algorithms classify observations based on distance measures. According to the literature, if the units of measurement of the variables are different, then the result of the clustering is said to be unreliable. Even sometimes, distance based clustering shows contradictory results when measurement units are closely related. Therefore, a new clustering scheme is proposed in this paper based on combining the membership function and OWA operator when classic clustering seems to have failed. For this purpose, a real data set from chemistry with ten variables are used to exemplify the new clustering scheme.

Keywords: Fuzzy membership function, Fuzzy set, OWA operator, Cluster analysis.

2000 AMS Classification: 03 E 72, 91 C 20.

*Akdeniz University, Faculty of Engineering at Alanya, Management Engineering Department, Alanya, 07425 Turkey. E-mail: muratalper@yahoo.com

†Corresponding Author.

‡Hacettepe University, Faculty of Economics and Administrative Science, Department of Public Finance, Ankara, 06800 Turkey. E-mail: aab@hacettepe.edu.tr

§University of Sannio, Department of Economical, Juridical and Social System Studies, Benevento, 82100 Italy. E-mail: (B. Simonetti) simonetti@unisannino.it (A. Lucadamo) alucadamo@unisannio.it

1. Introduction

Clustering is a widely used tool which is applied to problems ranging from pattern recognition and marketing to chemistry. Its aim is twofold. The first is to group observations using distance based methods, the second to prepare the data for further statistical analysis. Whichever the motivation, observations on the available data set need to be grouped as correctly as possible. The majority of clustering algorithms classify observations using distance functions which involve the values of the variables obtained for each observation. However, one of the drawbacks mentioned in the literature is that distance based clustering sometimes fail when the units of measurement are different. In order to overcome this problem, a new clustering scheme is proposed combining the notion of membership function and *OWA* operator. An experimental and a crisp data set consisting of 114 observations having ten variables is used to exemplify the new clustering scheme.

The organization of the paper is as follows: Section 2 gives references related to fuzzy set theory and explains briefly *OWA* operators. Section 3 explains the data set which is fundamental to the development of the new clustering scheme and includes the application of the new clustering scheme. Section 4 gives the proposed new method as an algorithm. The last section is the conclusion. In the appendix, the whole data set is given.

2. Preliminaries

Fuzzy Set Theory (FST) was introduced in the 60 's as a fundamental mathematical tool which enables one to deal with vague information in natural language. Afterwards it began to attract several researchers from a wide range of disciplines such as decision making, engineering and so on. FST was first introduced by [6]. Its potential to deal with vagueness or imprecision helps researchers employ it in several areas both in terms of application and methodology. Making computations with words also become a new tool to model data when its characteristic is verbal [7]. Therefore, (FST) has been witnessing rapid growth in several areas in both science and social disciplines [2]. The fundamental knowledge can be found in textbook [8]. A more detailed rigorous coverage can be found in [4]. Also, interested readers can benefit from the book [3].

One of the statistical methods using FST is clustering algorithms. The notion of membership understood in terms of the classical sense has deficiencies. Therefore, clustering algorithms combined with FST has provided gradual membership which helps determine some observations with partial membership. Fuzzy clustering methods have been proposed recently in order to deal with partial membership. On the other hand, employing distance based approaches have been reported as giving unreliable results since different units of measurement of variables is another drawback that is encountered in the classic case.

When the data set is large and data fusion is necessary, the OWA operator is a tool proposed by [5]. The OWA operator is an aggregation operator providing a parameterized family of aggregation operators between the minimum and the maximum. Its definition is given below:

2.1. Definition. An OWA operator is a function denoted by $F : R^n \to R$ which has an associated weighting vector W of dimension n such that w_i is in $[0, 1]$ and $w_1 + \cdots + w_n = 1$. Then, $F(a_1, \ldots, a_n) = w_1 \cdot b_1 + \cdots + w_n \cdot b_n$, where b_i is the i^{th} largest of the numbers a_i.

Actually, the literature related to OWA operators is wide and covers several aspects ranging from determining weights optimally to applying it to many frameworks such as fuzzy and decision making problems. However, in this paper, we use it just as an information fusion technique.

3. Data analysis

Instead of conducting long and tedious experiments, theoretical chemistry or computational chemistry, as a new emerging field aimed at calculating the values of variables using quantum theory with the help of computers, tries to determine, for example, which group of molecules can be used to develop a medicine or to be effective as corrosion inhibitors using statistical or mathematical models [1]. This helps construct models which are used instead of experiments. For example, if the chemist knows that one of the molecules is effective for the investigation in advance, then he does not need to conduct the same experiment for the other molecules. What he does is to use some statistical analysis such as clustering in order to find some other molecules which have similar characteristics. Therefore, he does not conduct same long expensive experiments for the other molecules.

In this paper, a crisp data set consisting of 114 observations, which are molecules in our case, having ten variables is used. Those 114 molecules are used as corrosion inhibitors in chemistry. The ten variables, called descriptors, are quantum chemical descriptors which are calculated by computers. Those variables or descriptors are called and abbreviated as the activation value (X_1) , Ehomo (X_2), Elumo (X_3), polarizity (X_4), hardness (X_5), softness (X_6), chemical potential (X_7), electro negativity (X_8), dipole moment (X_9), and SEZPE (X_{10}).

First a classic clustering algorithm, namely Hierarchical Clustering, is employed on the data set in order to classify similar molecules in order to understand which group of molecules can be used as corrosion inhibitors, since we do not know in advance how many clusters exist. However, as mentioned before, this fails to generate reliable clusters due to one of the drawbacks mentioned in the literature, which is that the measurement units of the variables are different. Hence, a new scheme combining the concept of membership function and a OWA operator is applied to data set to group the molecules.

An empirical approach is adapted to construct the membership functions for each variable. Some descriptive statistics such as mean, skewness and kurtosis are obtained using SPSS version 17.0. Based on those characteristics, the symmetric features around the mean suggest using symmetric triangular fuzzy membership functions for each variable. Also, five linguistic values such as very low, low, average, high and very high are used to split the data set for linguistic values. For instance, the membership functions for the variable called activation value (X_1) are constructed based on those linguistic values and their corresponding membership functions are defined as follows:

$$\mu_{\mathrm{VL}}(x) = \begin{cases} 0, & x < -3.9, \\ 1 - |\frac{-3.9-x}{0.7}|, & -3.9 \le x \le -3.2 \end{cases} \tag{3.1}$$

$$\mu_{\mathrm{L}}(x) = \begin{cases} |\frac{-3.25-x}{0.35}|, & -3.25 \le x \le -2.9, \\ 1 - |\frac{-2.9-x}{0.35}|, & -2.9 \le x \le -2.55 \end{cases} \tag{3.2}$$

$$\mu_{\mathrm{AV}}(x) = \begin{cases} |\frac{-2.6-x}{0.35}|, & -2.6 \le x \le -2.25, \\ 1 - |\frac{-2.25-x}{0.35}|, & -2.25 \le x \le -1.9 \end{cases} \tag{3.3}$$

$$\mu_{\mathrm{H}}(x) = \begin{cases} |\frac{-1.95-x}{0.35}|, & -1.95 \leq x \leq -1.6, \\ 1 - |\frac{-1.6-x}{0.35}|, & -1.6 \leq x \leq -1.25 \end{cases} \tag{3.4}$$

$$\mu_{\mathrm{VH}}(x) = \begin{cases} 0, & x > -1.95 \\ |\frac{-1.3-x}{0.6}|, & -1.6 \leq x \leq -1.25 \end{cases} \tag{3.5}$$

Similar constructions for the rest of the other 9 variables can be given. Then, the crisp values of the activation value (X_1) are plugged into the corresponding membership functions. Those calculations lead to membership grades which will be then be used as weights for information fusion. Then, those weights and the crisp data are used in order to generate a single datum as a representative of the fusion process for each molecule. Based on those representative data, a classic clustering algorithm, which is hierarchical clustering, is conducted. The results, showing better accuracy rates, are obtained when cross-validation is conducted, that is, while the new scheme has 0.89 cross-validation rate, classic hierarchical clustering has 0.67. The new scheme based on membership functions and an OWA operator as a data fusion method when compared with the classic cluster, which is a hierarchical cluster, resulted in more clusters. Also, it detected a group of molecules which are located between clusters. A sample calculation will be provided in detail in the next paragraph. Also, its algorithm will be presented in the next section.

In order to exemplify what we have done, just one observation, molecule 27, is picked. Then calculations are conducted to generate two crisp data instead one since it is a fact that each membership function has two sides, namely increasing and decreasing parts. Therefore, when data fusion is realized, the values related to the increasing parts and to the decreasing parts are grouped and combined by taking account of this fact. In any fuzzy data analysis, roughly speaking, two different approaches are taken which are either to obtain a crisp value as a representative value or to obtain a fuzzy number. However, when a crisp value is used, it is said that just one single value cannot represent the fuzziness. In a similar a way, when a fuzzy number is obtained, there is an ongoing problem of wider fuzziness resulting from the fuzzy arithmetic. In our case, neither way is adapted. In the fusion process, the values on the decreasing part of and on the increasing part of the membership functions are treated separately. For example, when the case 27, namely molecule 27, is picked as said before, the values of the variables abbreviated as (X_1), (X_2), (X_3), (X_4),(X_5), (X_6), (X_7), (X_8), (X_9), and (X_{10} correspond to 4 coming from the increasing parts of the membership functions and 6 coming from the decreasing parts of the membership functions.

Therefore, the increasing parts constitute the crisp value which represents the left series of data and the decreasing parts represent the right side of the data. In the fusion process the weights which are membership grades of the data are normalized since the total weight is equal to 1. The increasing part is tabulated as follows:

Table 1. Increasing Parts of Membership Functions for Molecule 27

"M27"	X_2	X_3	X_6	X_7
"W"	0.0	0.63	0.11	0.16
"NW"	0.0	0.7	0.12	0.18
"Val"	-5.6	-1.04	1.14	-3.32

As said before, the values and weights coming from the increasing parts of the membership functions of variables are grouped together in order to obtain a single value which represent the left side of the membership functions for molecule 27. It is noted that, corresponding weights are normalized. Then normalized weights are multiplied by the

real values of the variables to obtain the single value. Also it is noted that the constant 1/4 is multiplied by the single value since just four out of ten consists of the increasing parts of the membership functions:

$$[0.0 \cdot (-5.6) + 0.7 \cdot (-1.04) + 0.12 \cdot (1.14) + 0.18 \cdot (-3.32)]/4 = -0.30,$$

and for the decreasing part,

Table 2. Decreasing Parts of Membership Functions for Molecule 27

"M27"	X_1	X_4	X_5	X_8	X_9	X_{10}
"W"	0.57	0.12	0.46	0.87	0.70	0.59
"NW"	0.17	0.04	0.14	0.26	0.21	0.18
"Val"	-2.08	336.822	2.28	50.26	6.0711	-69111

and

$$[0.17 \cdot (-2.80) + 0.04 \cdot (336.822) + 0.14 \cdot (2.28) + 0.26 \cdot (50.26) + 0.21 \cdot (6.0711) + 0.18 \cdot (-69111)]/6 = -2068.72.$$

Similar arguments are valid for the decreasing parts of the membership functions.

In both tables above, M27, W, NW and Val stand for Molecules 27, Weights, Normalized Weights and Values, respectively. Also, it is noted that the same procedure is realized for the rest of the molecules. Therefore, two sets of data are obtained which represent the so-called left and right sides respectively. Then, the classic clustering algorithm, namely hierarchical clustering, is employed in order to obtain the number of clusters and the molecules, which in this case can be found in the intersection of the clusters.

4. Proposed Method

The explanations and single computation given in the previous section exemplify how the new scheme for clustering is realized. In this section, we give it as an algorithm in order to follow the procedure easily.

Step 1. Determine linguistic values for each variable.
Step 2. Construct membership function based on the data set.
Step 3. Calculate membership grades for each variable.
Step 4. Normalize weights.
Step 5. Calculate crisp values for increasing and decreasing parts by using an OWA operator.
Step 6. Run the classic clustering algorithm.

5. Conclusion

In this study, one of the drawbacks often encountered in classic clustering problems is remedied by combining the concept of membership function and an OWA operator. In order to realize this scheme, the linguistic variables and their corresponding values are determined. Then membership functions are constructed based on the data. Using a crisp data set with membership functions generates membership grades which are weights. Since the sum of the weights must be equal to one, the weights need to be normalized.

In fuzzy data analysis, there exist two main approaches. Either a crisp value is chosen as a representative of the fuzzy set, which is criticized because a loss of information has taken place, or the fuzzy set operations lead to wider fuzzy results which does not make sense. For this purpose, neither just one crisp value nor a wider fuzzy set is allowed.

Two crisp data sets are generated using the increasing part of and decreasing part of the membership functions, respectively. Therefore, it is possible to observe the molecules found in the intersection of the clusters. Also, the new scheme lead to a better cross-validation rate, which is 0.89, then that of the classic clustering method has. As a result, some of the molecules are found located in the intersection of the clusters so that some new molecules emerge as potential candidates. This information cannot be obtained with the classic clustering algorithms.

6. Appendix

Table 3. Data for 114 Molecules

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
-2.43	-5.58	-1.07	352.568	2.26	1.13	-3.33	49.86	4.5258	-71862.28
-1.00	-5.97	-1.12	356.078	2.43	1.21	-3.55	60.95	6.2301	-73907.64
-1.18	-6.19	-1.13	357.336	2.53	1.27	-3.66	67.78	6.9155	-75954.05
-2.26	-5.82	-1.03	344.991	2.40	1.20	-3.43	56.19	4.1902	-62096.13
-1.70	-5.24	-1.03	341.398	2.11	1.05	-3.14	41.38	3.2718	-62532.72
-0.70	-6.02	-1.49	348.567	2.27	1.13	-3.76	63.87	5.5206	-64110.71
-1.65	-6.16	-1.12	344.868	2.52	1.26	-3.64	66.78	5.6503	-64548.04
-2.00	-6.13	-1.11	356.611	2.51	1.26	-3.62	65.78	5.1155	-65617.01
-1.54	-5.73	-1.04	354.548	2.35	1.17	-3.39	53.74	4.4475	-65616.98
-1.54	-5.66	-0.99	355.71	2.34	1.17	-3.33	51.63	2.4025	-65617.12
-2.08	-5.34	-1.02	378.147	2.16	1.08	-3.18	43.69	2.2004	-78528.15
-2.00	-5.85	-1.05	348.942	2.40	1.20	-3.45	57.13	3.8412	-64142.62
-2.41	-5.63	-1.07	341.029	2.28	1.14	-3.35	51.17	4.3958	-63073.84
-2.38	-6.01	-1.53	351.078	2.24	1.12	-3.77	63.67	4.5324	-66157.62
-2.48	-5.72	-1.01	358.52	2.36	1.18	-3.37	53.33	7.1825	-67122.09
-1.95	-5.62	-1.04	350.531	2.29	1.15	-3.33	50.79	4.4662	-66054.01
-2.40	-5.32	-1.07	334.383	2.13	1.06	-3.20	43.38	3.2485	-61463.62
-2.51	-5.89	-1.04	304.69	2.43	1.21	-3.47	58.23	4.1668	-57852.26
-2.60	-5.41	-1.02	322.636	2.20	1.10	-3.22	45.38	4.1132	-59989.97
-1.48	-5.71	-1.04	328.025	2.34	1.17	-3.38	53.19	4.0894	-59990.08
-2.85	-5.03	-1.08	328.141	1.98	0.99	-3.06	36.87	3.1814	-60426.44
-3.38	-5.23	-1.06	336.69	2.09	1.04	-3.15	41.25	3.8198	-61495.46
-2.43	-5.37	-1.01	264.659	2.18	1.09	-3.19	44.37	4.5701	-53704.8
-2.60	-5.62	-1.02	310.383	2.30	1.15	-3.32	50.70	4.412	-57981.16
-2.60	-5.72	-1.04	339.002	2.34	1.17	-3.38	53.47	3.7273	-61059.2
-1.85	-5.42	-0.99	341.87	2.22	1.11	-3.21	45.51	5.0886	-63105.44
-2.08	-5.60	-1.04	336.862	2.28	1.14	-3.32	50.26	6.0711	-69161.0

Table 3 (Continued)

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
-2.30	-5.83	-1.07	341.852	2.38	1.19	-3.45	56.66	3.7698	-72496.56
-1.40	-5.78	-1.06	328.535	2.36	1.18	-3.42	55.21	3.8025	-62690.54
-2.40	-5.50	-1.02	335.193	2.24	1.12	-3.26	47.61	4.5741	-61059.09
-1.81	-5.68	-1.03	349.797	2.33	1.16	-3.36	52.34	4.0733	-62128.26
-0.70	-5.66	-1.02	359.025	2.32	1.16	-3.34	51.76	4.3236	-63197.34
-3.01	-5.72	-0.94	297.135	2.39	1.20	-3.33	53.0	4.9219	-49940.9
-3.00	-5.62	-0.91	290.882	2.36	1.18	-3.27	50.21	3.7805	-48904.39
-3.00	-5.60	-0.90	324.874	2.35	1.18	-3.25	49.64	3.7663	-52111.95
-1.48	-5.55	-0.90	346.844	2.33	1.16	-3.23	48.36	5.074	-54249.91
-2.18	-5.58	-0.90	302.212	2.34	1.17	-3.24	49.13	5.177	-49973.2
-1.60	-5.58	-0.90	313.165	2.34	1.17	-3.24	49.13	4.0598	-51042.35
-1.30	-5.57	-0.90	324.405	2.34	1.17	-3.24	48.87	3.9961	-52111.44
-1.18	-5.57	-0.90	335.315	2.34	1.17	-3.24	48.87	3.9933	-53180.48
-2.00	-5.55	-0.89	358.717	2.33	1.17	-3.22	48.32	5.1672	-56258.4
-1.60	-5.56	-0.90	325.033	2.33	1.17	-3.23	48.62	4.0951	-52111.58
-1.00	-5.64	-0.91	364.02	2.37	1.18	-3.28	50.73	4.2786	-56258.5
-0.70	-5.57	-0.90	363.57	2.34	1.17	-3.24	48.87	3.9361	-56258.43
-3.00	-5.52	-0.89	295.46	2.32	1.16	-3.21	47.56	4.3064	-48363.47
-2.08	-5.50	-0.89	317.625	2.31	1.15	-3.20	47.06	4.2578	-50501.4
-2.00	-5.56	-0.90	354.608	2.33	1.17	-3.23	48.62	3.9399	-54648.7
-2.08	-5.55	-0.90	309.027	2.33	1.16	-3.23	48.36	4.9453	-48995.45
-2.00	-5.55	-0.90	371.787	2.33	1.16	-3.23	48.36	4.199	-55280.52
-3.00	-5.60	-1.14	365.18	2.23	1.12	-3.37	50.65	3.367	-84368.6
-1.54	-5.60	-1.14	365.179	2.23	1.12	-3.37	50.65	3.367	-84368.6
-2.00	-6.25	-1.20	369.946	2.53	1.26	-3.73	70.07	6.6169	-88460.36
-1.78	-6.01	-1.42	363.917	2.30	1.15	-3.72	63.35	3.72	-95142.86
-2.08	-6.27	-1.44	365.427	2.42	1.21	-3.86	71.78	5.4171	-97189.27

Table 3. (Continued)

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
-2.53	-5.96	-0.89	341.55	2.54	1.27	-3.43	59.47	6.6731	-83772.42
-2.77	-5.97	-0.93	386.169	2.52	1.26	-3.45	59.99	5.3769	-87483.17
-2.93	-5.98	-0.94	323.136	2.52	1.26	-3.46	60.34	5.7133	-81198.0
-2.00	-5.97	-0.96	365.408	2.51	1.25	-3.47	60.15	5.5811	-85441.81
-2.00	-5.95	-0.90	375.104	2.53	1.26	-3.43	59.24	5.2761	-86511.19
-1.90	-5.93	-1.41	386.487	2.26	1.13	-3.67	60.88	7.9139	-98920.23
-1.85	-6.02	-1.46	381.534	2.28	1.14	-3.74	63.78	4.0083	-98920.37
-1.60	-6.02	-1.45	382.46	2.29	1.14	-3.74	63.75	4.2236	-98920.38
-2.18	-6.37	-2.92	387.07	1.73	0.86	-4.65	74.44	3.6169	-94024.90
-1.78	-6.37	-3.03	389.522	1.67	0.84	-4.70	73.78	5.0577	-94024.90
-1.60	-5.99	-0.96	387.894	2.52	1.26	-3.48	60.74	5.4152	-89529.50
-2.90	-6.01	-1.24	365.364	2.39	1.19	-3.63	62.68	4.9086	-95142.93
-2.00	-5.99	-1.57	364.859	2.21	1.11	-3.78	63.15	5.6579	-86850.52
-2.53	-5.98	-0.84	364.112	2.57	1.29	-3.41	59.77	6.5673	-87319.58
-2.65	-6.10	-3.50	347.63	1.30	0.65	-4.80	59.90	5.8623	-59249.45
-1.54	-5.93	-1.06	344.366	2.44	1.22	-3.50	59.49	6.6629	-61401.17
-1.48	-6.10	-1.08	345.631	2.51	1.26	-3.59	64.70	8.2746	-63447.57
-3.30	-5.95	-1.01	353.682	2.47	1.24	-3.48	59.83	7.4605	-73907.58
-3.11	-6.14	-1.03	354.88	2.56	1.28	-3.59	65.67	9.4918	-75953.98
-2.43	-5.58	-1.07	352.543	2.26	1.13	-3.33	49.86	4.5108	-71862.28
-1.00	-5.97	-1.12	356.076	2.43	1.21	-3.55	60.95	5.4714	-73907.64
-1.18	-6.19	-1.13	357.344	2.53	1.27	-3.66	67.78	7.4275	-75954.05
-2.43	-5.97	-1.15	357.63	2.41	1.21	-3.56	61.09	6.4353	-73907.63
-2.76	-5.96	-1.12	345.191	2.42	1.21	-3.54	60.65	5.7477	-64101.61
-1.95	-6.54	-3.82	369.788	1.36	0.68	-5.18	72.98	5.6464	-86239.88
-2.04	-6.26	-1.22	368.788	2.52	1.26	-3.74	70.50	7.0983	-88460.47
-2.30	-5.78	-1.05	363.032	2.37	1.18	-3.42	55.16	8.1963	-66530.84

Acknowledgement

We gratefully acknowledge the help of A. Baykal in preparing the LaTeX 2_ε source file.

References

[1] Amim, M. A., Arida, H. A., Kandemirli, F., Saracoglu, M., Arslan, T. and Basaran, M. A. *Monitoring corrosion and corrosion control of iron in HCI by non-ionic surfactants of the TRITON-X series-Part III. Immersion time effects and theoretical studies*, Corrosion Science **53**, 1895–1909, 2011.

[2] Bellman, R. E. and Zadeh, L. A. *Decision-making in a fuzzy environment*, Manage. Sci. **17**, 141–164, 1970.

[3] Klir, G. J. and Folger, T. A. *Fuzzy Sets, Uncertainty and Information* (Prentice-Hall, Englewood Cliffs, NJ, 1988).

[4] Nguyen, H. T. and Walker, E. A. *A First Course in Fuzzy Logic (Third Edition)* (Chapman & Hall/ CRC, Boca Rotan, FL, 2006).

[5] Yager, R. R. *On ordered weighted averaging aggregation operators in mulitcriteria decision making*, IEEE Trans. Syst. Man. Cybern. **18**, 183–190, 1988.

[6] Zadeh, L. A. *A computational approach to fuzzy quantifiers in natural languages*, Comput. Math. Appl. **9**, 149–184, 1983.

[7] Zadeh, L. A. *Fuzzy logic – computing with words*, IEEE Trans. Fuzzy. Syst. **4**, 103–111, 1996.

[8] Zadeh, L. A. and Kacpryzk J. *Computing with words in information/intelligent 1* (Springer-Verlag, Heidelberg, 1999).