Proceedings Paper:

Muda, M.Z. and Panoutsos, G. orcid.org/0000-0002-7395-8418 (2021) An evolving feature weighting framework for granular fuzzy logic models. In: Jansen, T., Jensen, R., Mac Parthaláin, N. and Lin, C.-M., (eds.) Advances in Computational Intelligence Systems (UKCI 2021). UKCI 2021: 20th UK Workshop on Computational Intelligence, 08-10 Sep 2021, Aberystwyth, UK. Advances in Intelligent Systems and Computing, 1409 . Springer , pp. 3-14. ISBN 9783030870935

https://doi.org/10.1007/978-3-030-87094-2_1

An Evolving Feature Weighting Framework for Granular Fuzzy Logic Models

Muhammad Zaiyad Muda and George Panoutsos

Department of Automatic Control and Systems Engineering
The University of Sheffield, Sheffield, UK

Abstract. Discovering and extracting knowledge from large databases are key elements in granular computing (GrC). The knowledge extracted, in the form of information granules can be used to build rule-based systems such as Fuzzy Logic inference systems. Algorithms for iterative data granulation in the literature treat all variables equally and neglects the difference in variable importance, as a potential mechanism to influence the data clustering process. In this paper, an iterative data granulation algorithm with feature weighting called W-GrC is proposed. By hypothesising that the variables or features used during the data granulation process can have different importance to how data granulation evolves, the weight of each feature's influence is estimated based on the information granules on a given instance; this is updated in each iteration. The feature weights are estimated based on the sum of within granule variances. The proposed method is validated through various UCI classification problems:- Iris, Wine and Glass datasets. Result shows that for certain range of feature weight parameter, the new algorithm outperforms the conventional iterative granulation in terms of classification accuracy. We also give attention to the interpretability-accuracy trade-off in Fuzzy Logic-based systems and we show that W-GrC produces higher classification performance - without significant deterioration in terms of its interpretability (Nauck's index).

Keywords: Granular Computing, Iterative Data Granulation, Fuzzy Logic, Feature Weights, Feature Relevance.

1 Introduction

One of the key steps in building data driven Fuzzy Logic (FL) models is the process of extracting knowledge from data [1]. Granular Computing (GrC) and iterative data granulation algorithms are an effective approach to extract knowledge from data within the context of human-centric systems [2-3]. Among the most widely used techniques for this process are fuzzy c-means (FCM) and hierarchical clustering.

In recent years, iterative data granulation algorithms, also known as granular clustering proposed in [3-4] have become a proven alternative in data mining and developing FL rule-bases. The main idea of this algorithm is to merge two most compatible information granules iteratively until sufficient data compression is achieved [3]. The compatibility measure can simply be distance based (such as in hierarchical clustering algorithms) or potentially involve more complex formulations that combine granular density, cardinality, overlap etc.

GrC algorithms have a similarity with agglomerative hierarchical clustering in terms of its 'find and merge' strategy. However, one main distinction between these algorithms is that in GrC, every granule consist of sub-granules originating directly from the actual data. This contributes to strong connection between the raw data and the information granules. Moreover, the compatibility measure in GrC is very useful tool in monitoring the similarity between granules; this can be linked to a numerical criterion to terminate the granulation in order to avoid merging of low compatibility granules [4].

So far GrC algorithms treat all features equally during the data granulation process. This is not desirable especially when dealing with data consisting of high number of features [5]. In many cases, some features are not as crucial as other features in the development of the FL model [1], while other features may have an importance that changes throughout the granulation process. This leads to the idea of continuously measuring and assigning appropriate weight for each feature throughout the data granulation (as in adaptive feature weighting for classical clustering algorithms).

Even though the feature weight concept has been introduced elsewhere [4], most of the works regarding this algorithm such as [6] and [7] use fixed weight for each feature. Investigations in feature weighting for GrC are scarce; for example in [8] a Fast Correlation-Based Filter which is based on symmetrical uncertainty to determine the most relevant features of a welding process. However, this is a preprocessing step (acting as a filter method) where the feature weights are determined in advance, and their values are constant throughout the evolution of the granulation process.

In this paper, we propose a new GrC algorithm that assigns and updates the feature weights based on the importance of the input features throughout the evolution of the iterative data granulation. With this approach we enable the more important features to have higher influence in the data granulation than the less important features, for a given iteration. Furthermore, instead of assigning the weight in the preprocessing step, the feature weighting in this research is embedded in the granulation process itself. This allows the feature weights to be adjusted according to the information granules that have been formed. The hypothesis here is that as information granules merge, and patterns develop, the importance of particular features to the evolution of such granules may change. While this approach is new in GrC, it has already been proven to be effective in other data mining and clustering algorithms.

Feature weighting has been applied in many clustering algorithms to overcome the problem of feature selection. Wu et al. introduced a new weighted fuzzy c-means algorithm taking into account the between-cluster separability [9]. The iterative formulas of the feature weights and membership degrees are obtained by maximizing the in-cluster compactness and the between-cluster separability. In another research, Huang et al. proposed W-*k*-means, the weighted version of k-means that outperformed the standard k-means in recovering clusters in data [5]. They also demonstrated that eliminating the irrelevant features based on the feature weights may enhance the clustering results. In the area of hierarchical clustering, Amorim implemented the feature-weighting scheme in an improved version of Ward, called Ward_p [10]. He demonstrated the effectiveness of Ward_p over the conventional Ward in particular in datasets comprising noisy data.

2 Background: The GrC-Fuzzy Logic model

The general framework for GrC-Fuzzy Logic modelling consists of two main steps, which are knowledge discovery and followed by the formation of a Fuzzy Logic rule-base. In the knowledge discovery step, granular computing and the process of iterative data granulation mimic the cognitive human abstraction in grouping entities with similar features (i.e. geometrical properties, cardinality, density, etc.) [6]. The knowledge discovered in the form of information granules defines the structure of the FL rule-base, specifically the parameters of the FL membership functions.

2.1 Knowledge discovery

The process of iterative data granulation starts with finding the pair of granules with the highest compatibility measure. Next, the granules are merged together in a new information granule that consists of original granules [7]. These steps are repeated until a satisfactory data abstraction level is accomplished. The compatibility measure of two granules A and B is defined as:

$$C(A,B) = Distance_{MAX} - Distance_{A,B}.\exp(-\alpha \times R) \quad (1)$$

where

$$Density\ R = \frac{C_{A,B}/Cardinality_{MAX}}{L_{A,B}/\ Length_{MAX}} \quad (2)$$

$Distance_{MAX}$ is defined as the sum of maximum distance in each dimension d:

$$Distance_{MAX} = \sum_{v=1}^{d}(distance_v\ ) \quad (3)$$

$Distance_{A,B}$ is the average multidimensional distance between granules A and B weighted by feature weight w_v :

$$Distance_{A,B} = \frac{\sum_{v=1}^{d} w_v(D_1 - D_2)}{\mathrm{d}} \quad (4)$$

in which

$$D_1 = \max(max_{Av}, max_{Bv}) \quad (5)$$

$$D_2 = \min\ (min_{Av}, min_{Bv}) \quad (6)$$

max_{Av}: maximum value in granule A for dimension v, min_{Av}: minimum value in granule A for dimension v, α: weights the requirement between distance and density, $Cardinality_{MAX}$: the total number of instances in the data set, $Length_{MAX}$: the maximum possible length of a granule in the data set, $C_{A,B}$: the cardinality when granule A merge with granule B, and $L_{A,B}$: length of the granule A and B, defined as:

$$L_{A,B} = \sum_{i=v}^{d}(max_{Xv} - min_{Xv}) \quad (7)$$

Typically, the feature weight parameter w_v in equation (4) used in most previous works is set to 1 for all dimensions (i.e. feature weighting is not used), or used at a fixed pre-determined value for each feature. The computation and adaptive adjustment of this parameter is the focus in this paper.

Fig.1 illustrates the evolution of a data granulation process for a two-dimensional synthetic data set with 150 instances. It starts with the initial raw data where every data instance is considered as one granule-point. These granules are then merged iteratively causing the number of granules to be reduced until the final information granules are established (in a predetermined manner, or using some termination criterion).

2.2 Formation of Fuzzy Logic rule-base

Taking a Gaussian Fuzzy Logic membership function (MF) as an example, the MF depends on two parameters σ and c which represent the width and the centre of a fuzzy set [11]. The standard deviation and median of data in each information granule can be used to determine the σ and c, respectively. Each information granule characterises one fuzzy rule [12]. For example, five information granules in Fig. 1 will lead

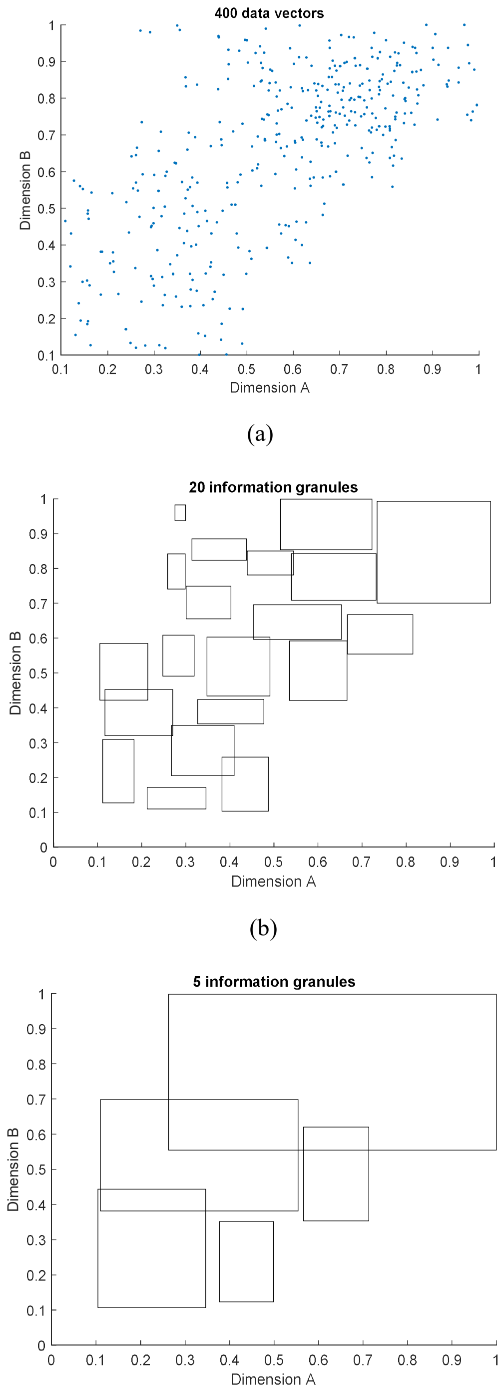


Fig. 1. Data granulation process from (a) 400 data vectors to (b) 20 information granules and (c) 5 information granules

to the formation of five fuzzy rules. Fig. 2 shows the overview of GrC-Fuzzy Logic modelling framework.

By determining the parameters σ and c across each input dimension individually in a multi-input single-output (MISO) system, the rules based on Mamdani fuzzy inference system (FIS) can be written as follows:

$$Rule\ 1:$$

$$IF\ (inputA = A_1\ and\ inputB = B_1\ and\ \ldots)$$

$$then\ (output = O_1)$$

$$Rule\ 2:$$

$$IF\ (inputA = A_2\ and\ inputB = B_2\ and\ \ldots)$$

$$then\ (output = O_2) \quad (8)$$

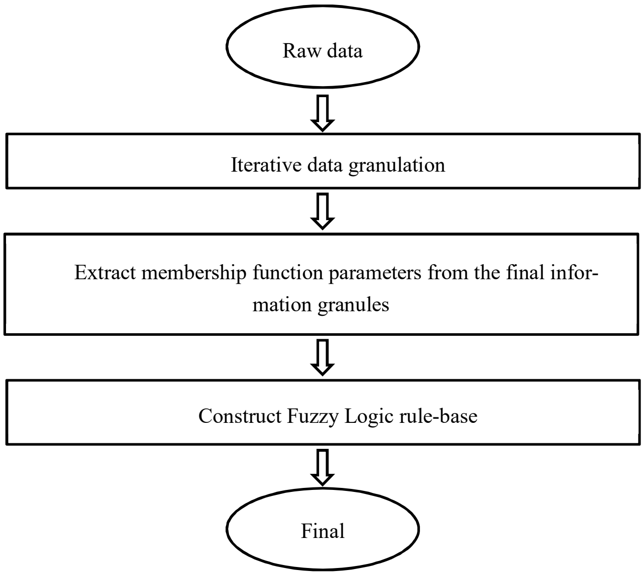


Fig. 2. The overview of GrC-Fuzzy Logic modelling framework

3 Proposed Methodology: Evolving Feature Weighting GrC

The Weighted K-Means (WK-Means) algorithm introduced by Huang et al. [5] minimises the following object function:

$$W(S, C, w) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v \in V} w_v^{\beta} d(y_{iv}, c_{kv}) \quad (9)$$

The Equation above is minimised by an iterative method, optimising (9) for S, C, and w, where $S = \{S_1, S_2, \ldots, S_k, \ldots, S_K\}$, $c_k \in C$ is the centroid for each granule k, y_i is an object in dataset Y, and β is the feature weighting parameter that balances the degree of effect between the weight and its contribution to the distance. There are two possibilities for the update of w_v, with S and C fixed, subject to $\beta > 1$:

$$w_v = \begin{cases} 0,\ if\ D_v = 0 \\ \dfrac{1}{\sum_{j=1}^{h} \left[\frac{D_v}{D_j}\right]^{\frac{1}{\beta - 1}}},\ if\ D_v \neq 0 \end{cases} \quad (10)$$

where h is the number of features where $D_v \neq 0$.

The parameter w_v (feature weight) in equation (4) has a fixed value, often pre-determined, in works related to GrC. In this paper, the weight for each feature v is defined and iteratively updated based on equation (10).

As shown in the equation, nonzero weight is only assigned to a feature where $D_v \neq 0$. $D_v = 0$ indicates that the *vth* feature consists of single value in each granule [5] and will be assigned zero weight. In this research, D_v is set as the sum of within granule variance:

$$D_v = \sum_{k=1}^{K} \frac{1}{N-1} \sum_{i=1}^{N} |y_{iv} - c_{kv}|^2 \qquad (11)$$

where N is the cardinality in the granule k.

The underlying principle here is to assign higher weights for features with lower within granule variance i.e. high variance in granules is set to be undesirable, hence penalised in the compatibility index. High variance would translate into high *sigma* (width) MFs. Hence, features that drive the creation of low variance granules, in any given iteration step, are promoted by the use of this adaptive weight and such features are considered here as more important for the evolution of the granulation process towards the development of FL rule-bases for classification problems.

3.1 Simulations and Empirical Results

Simulations were conducted on three datasets with regard to classification problems:– Iris, Wine and Glass (UCI Machine Learning Repository). All features are scaled to the interval of [0,1]. The ratio of training and testing data is set to 80:20. The range of feature weighting parameter β is selected between 1.5 and 10. The root mean square error (RMSE) and prediction accuracy % were calculated as the average of ten trials.

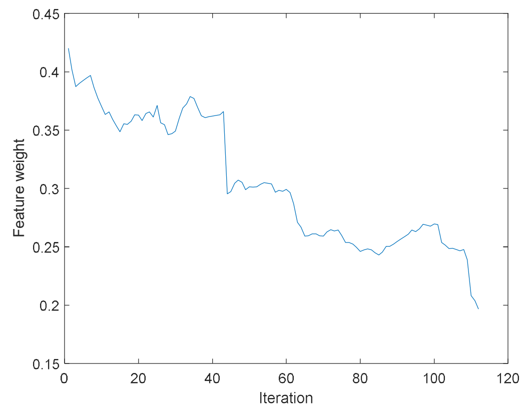
The Iris data consists of 150 instances with four input features. Next, the experiment is scaled up to datasets with higher feature dimensionality, which are Glass and Wine data with 10 and 13 input features, respectively**.** A bootstrapping method is applied to Glass data to balance the number of instances for each class. Due to this, the number of instances increases from 214 to 371**.** For comparison purposes, based on previous work [12], the number of granules selected for Iris and Wine is 5, while for Glass is 30 granules.

3.2 Evolving feature weights

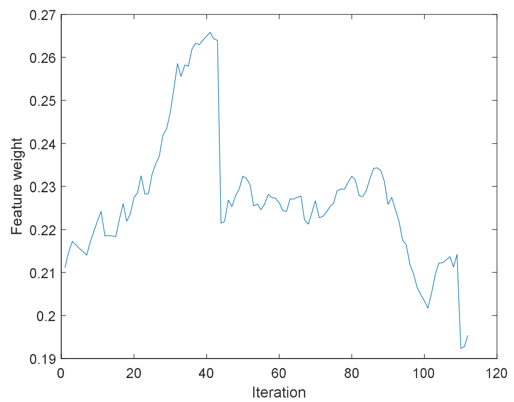
Fig. 3 shows how the feature weights evolve throughout the iterative granulation process, as an example for two features in the Iris dataset. The weights are plotted starting from the fourth iteration (out of 115 iterations), after which the feature weights are observed to be stable. This is due to the fact that the feature weights are assigned based on the within granule variance, while the merging process at the beginning only involves singleton granules (i.e. $D_v = 0$).

The feature weight average is computed and is compared with other measures such as mutual information and feature importance score as shown in Table 1. Mutual information gives information about the relevance between two random variables and normally being estimated between each feature and the given class labels [13]. The feature importance score ranks the features using a chi-square (χ^2) test [14]. The feature importance score is the negative log of chi-square tests' p-value [15].

This result shows that the feature weight ranking is consistent with the other two independent measures, confirming our hypothesis in capturing feature importance via the proposed method. All these three measures rank Petal width as the most important feature, followed by Petal length, Sepal length and Sepal width.



(a)



(b)

Fig. 3. Feature weights throughout the granulation for (a) Sepal length and (b) Sepal width

Table 1. Comparison of average feature weight in W-GrC with the feature importance score and mutual information

	Average weight (W-GrC)	Feature importance score	Mutual information
Sepal length	0.2721	41.7358	0.6415
Sepal width	0.2062	19.1551	0.3935
Petal length	0.3072	97.8866	1.2663
Petal width	0.3623	101.1028	1.3245

3.3 Empirical Results using Simulations

Table 2 summarises the performance of W-GrC with different values of β. The 'no feature weighting' row presents the results for the GrC without feature weighting, also known as conventional GrC. It is observed that with careful selection of β, the proposed W-GrC outperforms the standard GrC in terms of RMSE and accuracy. β needs to be treated as a hypermeter here, which will be identified in each case (problem specific).

For the Iris data, good results were obtained at $\beta \in \{3,4,5,6,7,8,10\}$. The highest accuracy was achieved when $\beta = 3$ and $\beta = 6$ with 96.33% of correct prediction as compared to 94% in the conventional GrC. For Wine data, improvement can be observed at β ranging from 3 to 6. Most experiments showed accuracies of above 90%, except for $\beta = 1.5$. This result is comparable to other literature results, however it is recognised that this specific case study may be too simple to stress test the proposed methodology (Glass and Wine data offer higher complexity and dimensionality).

In the case of the Glass dataset, we can see more clearly that higher values of β ($\beta \geq 3$) are more desirable to produce good result. The best performance is recorded at $\beta = 5$ with 71.86% accuracy in comparison with 62.79% in conventional GrC.

From Table 2, it can be observed that in general, W-GrC outperforms the conventional GrC. It achieves highest accuracy for all datasets, when an appropriate value of β is selected. This is because features that are more important for a given instance during the iterative granulation process are assigned with larger weights in forming the information granules. However, it is noted that the selection of β is important to obtain high classification accuracy. From the result, we suggest ($\beta \geq 3$) as the appropriate value of β, for this particular case study.

Results are benchmarked against other research such as [16] with 96.67% in Iris, [17] with 97.14% in Wine and [13] with 71.66% in Glass.

Table 2. Average RMSE and % accuracy performance of W-GrC with various β values, testing (unseen) data, 10 runs per β value

	Iris		Wine		Glass	
	RMSE	Accuracy (%)	RMSE	Accuracy (%)	RMSE	Accuracy (%)
No feature weighting	0.1415	94	0.1173	92.3	0.2020	62.79
$\beta = 1.5$	0.1473	91.67	0.3101	66.67	0.4274	26.74
$\beta = 2.0$	0.1551	90.67	0.1238	91	0.3365	32.33
$\beta = 3.0$	0.1205	96.33	0.1082	94	0.2235	63.02
$\beta = 4.0$	0.1302	94.67	0.1123	92.67	0.2164	69.30
$\beta = 5.0$	0.1253	94.33	0.1033	95.67	0.2144	71.86
$\beta = 6.0$	0.1251	96.33	0.1067	93	0.2165	66.98
$\beta = 7.0$	0.1285	95.67	0.1230	92	0.1980	66.51
$\beta = 8.0$	0.1189	96	0.1342	90.33	0.2219	66.05
$\beta = 9.0$	0.1346	93.67	0.1186	91.67	0.2105	68.14
$\beta = 10.0$	0.1273	95	0.1212	91.33	0.2224	65.81

3.4 Interpretability index

In designing Fuzzy Logic systems (FLS), interpretability and accuracy are often conflicting objectives; one can be enhanced by sacrificing the other, a situation that is termed as interpretability-accuracy trade-off. For example the enhanced interpretability of Mamdani-based FLS, versus the enhanced predictive accuracy of TSK-based FLS. Interpretability, within the FLS context, can be defined as the trait of a model to enable human to understand a system's behavior by scrutinising its rule base [18]. In this study, we use the models developed using values of β that perform the best in terms of accuracy as in Table 2 to assess if the models' interpretability is affected by the enhanced predictive performance.

The impact of feature weighting on interpretability measure is investigated using Nauck's index. Nauck's index is a numerical index introduced by Nauck in order to assess the interpretability of FL rule-based classification systems [19-20]. It is computed as the product of three terms: complexity of FLS ($comp$), average normalized coverage of fuzzy partition ($\overline{cov}$) and average normalized partition index ($\overline{part}$) given by:

$$Nauck\ index = comp\ \times \overline{cov} \times \overline{part} \qquad (12)$$

(readers are referred to [19] and [20] for further details).

Table 3 summarises the comparison of the interpretability index for the proposed W-GrC and the conventional GrC. It is demonstrated that W-GrC is able to producing higher accuracy without a statistically significant deterioration in terms of model interpretability. The impact on interpretability index is minor, less than 2% on the Iris data, and even less for the Wine and Glass case studies. Note that the Nauck's index in Glass is comparatively to the other cases small due to the high number of rules (30 as opposed to 5 in Iris and Wine).

Table 3. Comparison of the interpretability index

	Nauck's index	
	W-GrC	GrC
Iris	0.3076	0.3129
Wine	0.0929	0.0928
Glass	7.02×10^{-4}	7.07×10^{-4}

Conclusion

In this paper, a new iterative data granulation algorithm is presented with evolving feature weighting to characterise the importance of data features and use such weights to drive the information granulation process. The weight for each feature is determined based on the sum of within granule variances from the granules that have been formed, at any given iteration. In each iteration, the importance of all features is evaluated to identify the most important features that contribute most to the computation of the granules' compatibility measure.

The resulting importance of features, estimated via averaging feature weights throughout the data granulation process, are compared with other methods such as chi-square test and mutual information; agreement in feature ranking is demonstrated.

Simulation results in UCI classification problems have shown that the proposed W-GrC algorithm outperforms the conventional GrC in terms of classification accuracy. Improvement can be seen, in more complex datasets such as Glass case study. The experiment results showed that the proposed GrC-Fuzzy-modelling framework is able to handle data with various dimensionality.

The interpretability of the resulting models is assessed, using Nauck's index, and no significant deterioration of predictive performance is observed despite the higher resulting % accuracy in the classification tasks. While this study shows positive preliminary results, a greater range of complexity in case studies can be investigated in the future, as well as performance can be assessed more extensively against non GrC-based methods.

Acknowledgement

This research is sponsored by Universiti Teknologi MARA, Ministry of Education, Malaysia and The University of Sheffield, UK.

References

1. Ahmad, S. S. S., Pedrycz, W.: Data and feature reduction in fuzzy modeling through Particle Swarm Optimization. Applied Computational Intelligence and Soft Computing (2012).
2. Hu, J., Li, T., Wang, H., Fujita, H.: Hierarchical cluster ensemble model based on knowledge granulation. Knowledge-Based Systems 91, 179-188 (2016).
3. Pedrycz, W., Bargiela, A.: Granular clustering: a granular signature of data. In IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics, vol. 32(2), pp. 212-224 (2002).
4. Panoutsos, G., Mahfouf, M.: A neural-fuzzy modelling framework based on granular computing: Concepts and applications. Fuzzy Sets and Systems 161(21), 2808-2830 (2010).

5. Huang, J. Z., Ng, M. K., Rong, H., Li, Z.: Automated variable weighting in k-means type clustering. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27(5), pp. 657-668 (2005).
6. Rubio-Solis, A., Panoutsos, G.: Granular computing neural-fuzzy modelling: A neutrosophic approach. Applied Soft Computing 13(9), 4010-4021 (2013).
7. Baraka, A., Panoutsos, G., Mahfouf, M., Cater, S.: A Shannon entropy-based conflict measure for enhancing granular computing-based information processing. In 2014 IEEE International Conference on Granular Computing (GrC), pp. 13-18 (2014).
8. Rubio-Solis, A., Panoutsos, G.: An ensemble data-driven fuzzy network for laser welding quality prediction. In 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1-6 (2017).
9. Wu, Z., Li., C., Zhou, F., Liu. L.: A new weighted fuzzy c-means clustering approach considering between-cluster separability. Journal of Intelligent & Fuzzy Systems 40(1), 1017-1024 (2021).
10. Amorim, R.: Feature relevance in Ward's hierarchical clustering using the Lp norm. Journal of Classification 32, 46-62 (2015).
11. Liadis, L., Skopianos, S., Tachos, S., Spartalis, S.H.: A fuzzy inference system using Gaussian distribution curves for forest fire risk estimation. In IFIP Advances in Information and Communication Technology, 339, pp. 376-386 (2010).
12. Muda, M. Z., Panoutsos, G.: An entropy-based uncertainty measure for developing granular models. In 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI), pp. 73-77 (2020).
13. Xing, H., Ha, M., Tian, D., Hu, B.: A novel support vector machine with its features weighted by mutual information. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 315-320 (2008).
14. Alkhodari, M., Islayem, D. K., Alskafi, F. A., Khandoker, A. H.: Predicting hypertensive patients with higher risk of developing vascular events using heart rate variability and machine learning. In IEEE Access, vol. 8, pp. 192727-192739 (2020).
15. Fernando, H., Marshall, J.: What lies beneath: Material classification for autonomous excavators using proprioceptive force sensing and machine learning. Automation in Construction 119 (2020).
16. Wang, Y., Cang, S.,Yu, H.: Mutual information inspired feature selection using kernel canonical correlation analysis. Expert Systems with Applications: X 4 (2019).
17. Dousthagh, M., Nazari, M., Mosavi, A., Shamshirband, S., Chronopoulos, A.T.: Feature weighting using a clustering approach. International Journal of Modeling and Optimization, 9(2), 67-71 (2019).
18. Shukla, P., Tripathi, S.: A Review on the Interpretability-Accuracy Trade-Off in Evolutionary Multi-Objective Fuzzy Systems (EMOFS). Information 3, 256-277 (2012).
19. Razak, T. R., Garibaldi, J., Wagner, C., Pourabdollah, A., Soria, D.: (2017). Interpretability indices for hierarchical fuzzy systems. In IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2017) (2017).
20. Nauck, D. D.: Measuring interpretability in rule-based classification systems. In The 12th IEEE International Conference on Fuzzy Systems, vol.1, pp. 196-201 (2003).