Iranian Journal of Fuzzy Systems **Vol. 4, No. 1,** (2007) pp. 1-19

DISTRIBUTED AND COLLABORATIVE FUZZY MODELING

W. PEDRYCZ

ABSTRACT. In this study, we introduce and study a concept of *distributed* fuzzy modeling. Fuzzy modeling encountered so far is predominantly of a *centralized* nature by being focused on the use of a single data set. In contrast to this style of modeling, the proposed paradigm of distributed and collaborative modeling is concerned with distributed models which are constructed in a highly collaborative fashion. In a nutshell, distributed models reconcile and aggregate findings of the individual fuzzy models produced on a basis of local data sets. The individual models are formed in a highly synergistic, collaborative manner. Given the fact that fuzzy models are inherently granular constructs that dwell upon collections of information granules – fuzzy sets, this observation implies a certain general development process. There are two fundamental design issues of this style of modeling, namely (a) a formation of information granules carried out on a basis of locally available data and their collaborative refinement, and (b) construction of local models with the use of properly established collaborative linkages. We discuss the underlying general concepts and then elaborate on their detailed development. Information granulation is realized in terms of fuzzy clustering. Local models emerge in the form of rule-based systems. The paper elaborates on a number of mechanisms of collaboration offering two general categories of so-called horizontal and vertical clustering. The study also addresses an issue of collaboration in cases when such interaction involves information granules formed at different levels of specificity (granularity). It is shown how various algorithms of collaboration lead to the emergence of fuzzy models involving information granules of higher type such as e.g., type-2 fuzzy sets.

1. Introduction

Fuzzy modeling and fuzzy models have assumed a highly visible and crucial position in the overall research on fuzzy sets. We have witnessed a wealth of conceptual and algorithmic developments along with a plethora of applications and case studies. Over years we have noted a rapid progress in the area and a growing sophistication of the concepts and assessment mechanisms guiding the development of fuzzy models. It becomes apparent that fuzzy models are predominantly constructed in the framework of Computational Intelligence. Soft Computing or Computational Intelligence (CI), as these two notions seem to be used quite interchangeably, has brought together a wealth of information technologies of Granular Computing (and fuzzy sets, in particular), neurocomputing and biologically

Invited paper: Received in March 2007

Key words and phrases: Computational Intelligence, C^3 paradigm, Distributed processing, Fuzzy clustering, Fuzzy models.

inspired processing by forming a highly cohesive environment. The synergy being a CI cornerstone is intensively exploited in fuzzy modeling. The conceptual framework of the model is supported by the use of fuzzy sets. The substantial numeric optimization arises in the framework of neurocomputing and fuzzy neurocomputing, in particular. Evolutionary optimization and biologically-inspired computing support structural developments of fuzzy models. We have also moved a long way in articulating a suite of fundamental requirements of fuzzy modeling. In addition to the commonly encountered criterion of accuracy of fuzzy models, some other important aspects of interpretability, stability and human-centricity are taken into consideration. The phenomenon of human-centricity manifests quite vividly in several ways in fuzzy modeling. First, the results are presented at some suitable level of abstraction secured by the use of information granules. Likewise the semantics of the information granules that are used to organize findings about data is conveyed in the language of fuzzy sets whose interpretation is quite intuitive. In this sense, we envision that the available mechanisms of presentation of results to the end-user are quite effective. Second, the communication with the human at the entry point when the data sets become analyzed is not that well developed. Domain knowledge available there is crucial to the build up of models (say, fuzzy models) and the establishment of their transparency and readability. It is worth stressing that the transparency and accuracy are the two dominant requirements of fuzzy models we are interested in satisfying to the highest possible extent. The effective two-way communication is a key to the success of CI constructs, especially if we are concerned with the ways how all computing becomes navigated. For instance, the mechanisms of relevance feedback that become more visible in various interactive systems hinge upon the well-established and effective human-centric schemes of processing in which we effectively accept user hints and directives and release results in a highly comprehensible format. A succinct view at fuzzy modeling is portrayed in Figure 1.

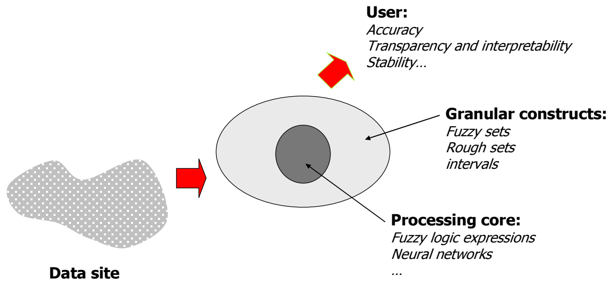


FIGURE 1. A general view of fuzzy modeling highlighting the essential functional aspects of fuzzy models

The current trends of distributed processing and distributed modeling are highly visible as we often encounter spatially and temporally distant sources of data. A joint processing of them could be highly beneficial yet there could be a number of factors preventing this from happening. For instance, there could be evident issues of privacy and security that do not allow us to share data. There might be some compelling technical constraints that make transfer of substantial amounts of data infeasible or impractical, Figure 2. Establishing a web of collaborative links is of paramount relevance to the effective realization of this modeling.

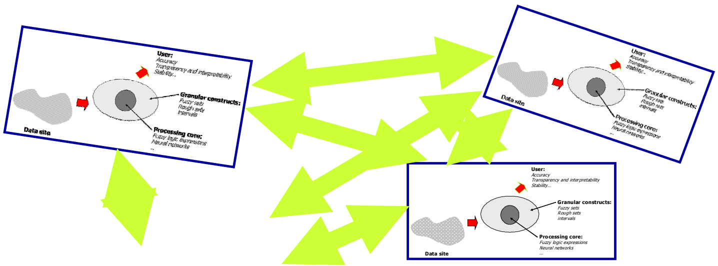


FIGURE 2. Distributed fuzzy modeling: building a web of collaborative linkages

Owing to the constraints of privacy implying that data cannot be shared, information granulation becomes a critical feature of distributed modeling. All possible communication pursuits leading to various facets of collaboration cannot be realized through a direct exchange of numeric data. In this sense, clustering methods that are pivotal to the design of information granules are an integral part of distributed fuzzy modeling. To accommodate the modeling needs, we offer enhancements of clustering techniques. Several fundamental concepts such as clustering with partial supervision and proximity knowledge hints are of interest with this regard. In this study, we propose and investigate a C^3 paradigm which offers a systematic operational framework for building constructs of fuzzy models in a coherent and well-orchestrated manner.

To stress the point of the granular nature of constructs of distributed fuzzy modeling, we refer to Figure 3 alluding to the rule-based modeling with local regression models (in sense of Takagi-Sugeno) assuming the following form

$$\text{-if } x \text{ is } A_i \text{ then } y = f(x, a_i) \tag{1}$$

where A_i is a certain information granule defined in the input space (condition part) and a_i denotes a vector of parameters of the local model; in particular we can envision the model to be linear, that is $y = a_i^T x$.

When it comes to collaborative development of models, individual data sites support the construction of such models at a local level whereas at the same time we envision collaborative pursuits in the reconciliation of the information granules (A_i, A_j, …) and the parameters of the local models.

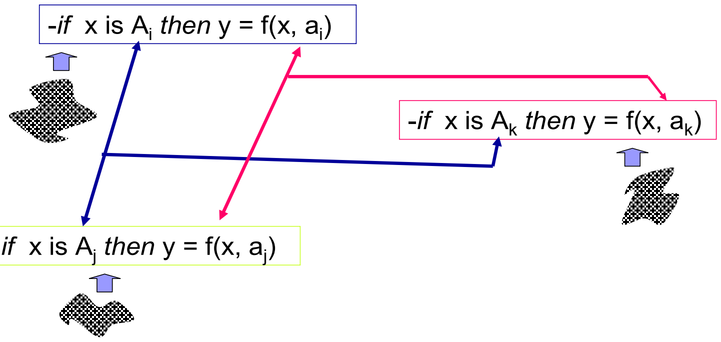

-*if* x is A_i *then* y = f(x, a_i)
-*if* x is A_k *then* y = f(x, a_k)
-*if* x is A_j *then* y = f(x, a_j)

FIGURE 3. The collaborative development of rule-based models with local regression models. Note

The presentation in this study is organized as follows. We start with the concept of the C^3 paradigm (Section 2). The role of the fuzzy clustering and the Fuzzy C-Means (FCM) as a suitable algorithmic vehicle of information granulation is briefly outlined in Section 3. The essence of distributed and collaborative schemes of CI is discussed in Section 4. Along this line, we next discuss the underlying principle of collaborative clustering (Section 5). Further algorithmic investigations concern hierarchies of clusters and their coordinated development, Section 6. A collaborative development of CI rule-based models in the Takagi-Sugeno sense (Section 7) presents a way of reconciling the parametwers of the locally developed fuzzy models (regression models). Higher-type constructs of CI emerging in the collaborative framework are discussed in Section 8. Finally, some closing remarks are covered in Section 9.

2. The C^3 Paradigm of the Distributed Environment of Fuzzy Modeling

Individual data sites can engage in some interaction between themselves when exchanging their findings and supporting each other in the buildup of the findings that are common or similar to all of them. The reconciliation mechanisms can be structured at three distinct conceptual and algorithmic levels. We distinguish here between communication, collaboration, and consensus building, hence an abbreviation of C^3 which captures the essence of these three mechanisms considered together. Let us elaborate on each of them in more detail by stressing what the underlying concepts are and what conceptual differences are profoundly visible.

Communication – in this mode of interaction there is an exchange of granular findings between the data sites. Information granules formed at one data site are communicated to others. The mode of this interaction is passive. While the granular results are made available (and could be eventually expressed in terms of the results available at the given data site), there are no provisions to adjust such local findings. Each data site fully adheres to its own structure of information granules.

Collaboration – here we encounter an active mode of interaction. The data sites exchange their findings but afterwards it can act upon it given the differences between the results coming from outside. It may adjust their local findings by reformulating the task of forming the information granules. Now they are based upon the locally available data but at the same time they take into consideration the particulars of the information granules supplied by other data sites. The intensity of collaboration itself can be established by forming a certain augmented objective function which incorporates data as well as quantifies the differences between local findings.

Consensus formation – it is similar to collaboration in the sense each data site receives findings from others and can act upon those yet their format could be somewhat incompatible with the format being used locally. This requires mechanisms of forming consensus that are elevated at the higher level of abstraction. For instance, when dealing with incompatible partition matrices conveying the structure (albeit being quite distinct as far as their granularity is concerned), we may need to effectively exchange the findings by looking at induced proximity matrices.

3. **Fuzzy C-Means (FCM) as an Example of the CI Algorithm of Data Analysis and Information Granulation**

To make a consistent exposure of the overall material and assure linkages with the ensuing optimization developments, we confine ourselves to one of the objective function based fuzzy clustering. More specifically, we consider a Fuzzy C-Means (FCM) (Bezdek, 1981) governed by the following objective function

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} \| \mathbf{x}_k - \mathbf{v}_i \|^2 \qquad (2)$$

where x_k denotes an multidimensional data point (pattern) , v_i is an i-th prototype and $U=[u_{ik}]$, i=1, 2, …, c; k=1, 2,…,N is a partition matrix. $||.||$ denotes a certain distance function and “m” denotes a fuzzification coefficient; m>1.0. The minimization of (1) is realized with respect to the partition matrix and the prototypes. The optimization scheme and all specific features of the minimization of Q are well reported in the literature, refer for instance to Abonyi and Szeifert (2003) and Pedrycz (1985) and Pedrycz and Gomide (2007). What is of interest to us here is an observation that fuzzy clustering is inherently associated with the granularity of information. In a nutshell fuzzy clustering leads to the abstraction of data into a format of information granules. Two essential and somewhat orthogonal dimensions of the granulation process are envisioned: (a) numeric realization of the granulation through a collection of the prototypes, and (b) a collection of information granules – fuzzy sets represented by successive rows of the partition matrix. Interestingly enough, there is a direct correspondence between these two representations. Given a collection of prototypes we can determine the entries of the partition matrix. And vice versa, a given partition matrix along with the data gives rise to the prototypes. The interpretability of the results of the FCM is its significant and highly valuable feature of the algorithm. As a collection of fuzzy sets (described by the corresponding rows of the generated partition matrix) offer a holistic view at the structure of data, this feature of the FCM emphasizes its linkages with the main thrusts of Computational Intelligence.

There are three issues of paramount relevance when casting fuzzy clustering in the CI framework

Knowledge-based orientation. A heavy and visible reliance on numeric data is an evident feature of fuzzy clustering as it could be seen today. There are, however, other important factors one has to take into account when discovering the structure in data. Various sources of knowledge are available from experts, data analysts, users and they come in various formats. The fundamental challenge concerns efficient ways of their incorporation into the clustering schemes, both as a concept and the algorithmic enhancement. This is not a straightforward task given the fact that clustering has to reconcile numeric aspects (data) and knowledge component (human factors). In essence, the knowledge-based orientation of clustering is in line of human-centricity of Computational Intelligence and the development of interaction schemes.

Distributed character of processing. This challenge has emerged because of the inherently distributed nature of data. Those tend to be distributed at individual locations (say, sensor networks) and this poses an interesting quest as to the distributed clustering. The centralized mode that is predominant today in fuzzy clustering requires a careful revision. The clustering techniques available nowadays that predominantly revolve around a single, huge and centrally available dataset do require a careful re-visiting and reformulation.

Interaction mechanisms. All of those aspects are associated in one way or another with the distributed nature of data sets. Given the distributed character of data, it is also very likely that they cannot be shared because of the privacy and security restrictions. On the other hand, some collaboration and interaction would be highly desirable given the fact that the structure in some datasets could be quite similar and sharing the knowledge about the discovery of clusters within one dataset with other sites could be beneficial. How to facilitate collaboration and consensus building in data analysis while respecting security requirements becomes an evident challenge.

Each of these challenges comes with a suite of their own quite specific problems that do require a very careful attention both at the conceptual as well as algorithmic level. We have highlighted the list of challenges and in the remainder of this study present some of the possible formulations of the associated problems and look at their solutions. It is needless to say that our proposal points at some direction that deems to be of relevance however does not pretend to offer a complete solution to the problem. Some algorithmic pursuits are also presented as an illustration of some possibilities emerging there.

4. Distributed and Collaborative Schemes of CI and Fuzzy Modeling

Quite commonly we encounter situations where databases are distributed rather than centralized (Da Silva et al., 2005; Park and Kargupta, 2003; Tsoumakas et al., 2004). There are different outlets of the same company and each of them operates independently and collects data about customers populating their independent databases. The data are not available to others. In banking, each branch may run its own database and such databases could be geographically remote from each other. In health institutions, there could be separate datasets with a very limited communication between the individual institutions. In sensor networks (which become quite popular given the nature of various initiatives such as intelligent houses, information highway, etc.), we encounter local databases that operate independently from each other and are inherently distributed. They are also subject to numerous technical constraints (e.g., a fairly limited communication bandwidth, limited power supply, etc) which significantly reduce possible interaction between the datasets. Under these circumstances, the "standard" data mining activities are faced now new challenges that need to be addressed. It becomes apparent that processing all data in a centralized manner cannot be exercised. On the other hand, data mining of each of the individual databases could benefit from availability of findings coming from others. The technical constraints and privacy issues dictate a certain level of interaction. There are two general modes of interaction that is collaborative clustering and consensus clustering both of which are aimed at the data mining realized in the distributed environment. The main difference lies in the level of interaction. The collaborative clustering is positioned at the more active side where the structures are revealed in a more collective manner through some ongoing interaction. The consensus driven clustering is focused on the reconciliation of the findings while there is no active involvement at the stage of constructing clusters.

Collaborative clustering

Given the distributed character of data residing at separate databases, we are ultimately faced with the need for some collaborative activities of data mining. With the distributed character of available data come various issues of privacy, security, limited communication capabilities that have to be carefully investigated. We show that the notion of information granularity that is at heart of fuzzy sets plays a pivotal role in this setting.

Privacy and security of computing versus levels of information granularity

While the direct access to the numeric data is not allowed because of the privacy constraints (Agarwal and Srikant, 2000; Claerhout and DeMoor, 2005; Clifton, 2000; Clifton et al., 1996, 2001; Coppi and d'Urso, 2003; Da Silva et al., 2005; Du and Zhan, 2002; Evfimievski et al., 2004; Johnsten et al., 2002; Kargupta et al., 2003; Lindell and Pinkas, 2000; Merugu and Ghosh, 2005; Verykios et al., 2004; Wang and Jafari, 2005; Wang et al., 2005) all possible interaction could be realized through some interaction occurring at the higher level of abstraction delivered by information granules. In objective function based fuzzy clustering, there are two important facets of information granulation conveyed by (a) partition matrices and (b) prototypes. Partition matrices are, in essence, a collection of fuzzy sets which reflect the nature of the data. They do not reveal detailed numeric information. In this sense, there is no breach of privacy and partition matrices could be communicated not revealing details about individual data points. Likewise prototypes are reflective of the structure of data and form a summarization of data. Given a prototype, detailed numeric data are hidden behind them and cannot be reconstructed back to the original form of the individual data points. In either case, no numeric data are directly made available.

The level of information granularity (Zadeh, 2005) is linked with the level of detail and in this sense when changing the level of granularity possible leakage of privacy could occur. For instance, in limit when the number of clusters becomes equal to the number of data points, each prototype is just the data point and not privacy is retained. Obviously, this scenario is quite unrealistic as the structure (the number of clusters) is kept quite condensed when contrasted with all data. The schematic view of privacy offered through information granulation resulting within the process of clustering is illustrated in Figure 4. We note here that the granular constructs (either prototypes or partition matrices) build some granular interfaces.

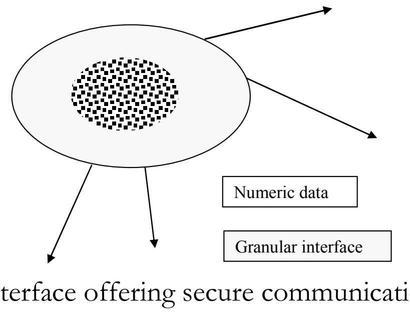


FIGURE 4. Granular interface offering secure communication and formed by the results of the fuzzy clustering (partition matrices and prototypes).

5. **The Underlying Principle of Collaborative Clustering**

When dealing with distributed databases we are often interested in a collaborative style of discovery of relationships (Pedrycz, 2002, 2005) that could be common to all of the databases. There are a lot of scenarios where such collaborative pursuits could be deemed highly beneficial. We could envision a situation where the databases are located in quite remote locations and given some privacy requirements as well as possible technical constraints we are not allowed to collect (transfer) all data into a single location and run any centralized algorithm of data mining, say clustering. On the other hand, at the level of each database each administrator/analyst involved in its collection, maintenance and other activities could easily appreciate the need for some joint activities of data mining. Schematically, we can envision the overall situation as schematically visualized in Figure 5.

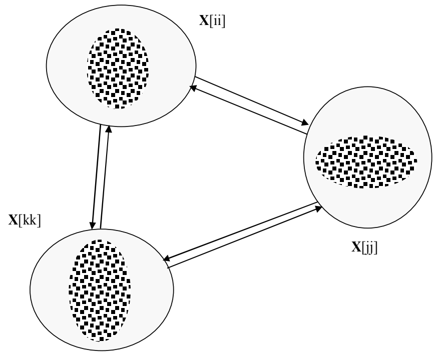


FIGURE 5. A scheme of collaborative clustering involving several datasets and interacting at the level of granular interfaces

While the collaboration can assume a variety of detailed schemes, the two of them are the most essential. We refer to them as horizontal and vertical modes of collaboration or briefly horizontal and vertical clustering. More descriptively, given are "P" data sets X[1], X[2], .. X[p] where X[ii] stands for the ii-th dataset (we adhere to the consistent notation of using square brackets to identify a certain data set) in *horizontal* clustering we have the same objects that are described in *different* feature spaces. In other words, these could be the same collection of patients coming with their records built within each medical institution. The schematic illustration of this mode of clustering portrayed in Figure 5 underlines the fact that any possible collaboration occurs at the structural level viz. through the information granules (clusters) built over the data; the clusters are shown in the form of auxiliary interface layer surrounding the data. The net of directed links shows how the collaboration between different data sets takes place. The width of the links emphasizes the fact that an intensity of collaboration could be different depending upon the dataset being involved and the intension of the collaboration say, a willingness of some organization to accept findings from external sources).

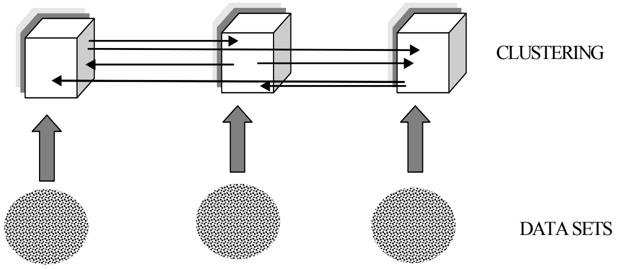


FIGURE 6. A general scheme of horizontal clustering; all communication is realized through some granular interface

The mode of *vertical* clustering, Figure 7, is complementary to the one already presented. Here the data sets are described in the same feature space but deal with *different* patterns. In other words, we consider that X[1], X[2], …, X[P] are defined in the same feature space while each of them consists of different patterns, dim(X[1]) = dim(X[2]) = … dim(X[P]) while X[ii] X[jj]. We can show the data sets as being stack on each other (hence the name of this clustering mode).

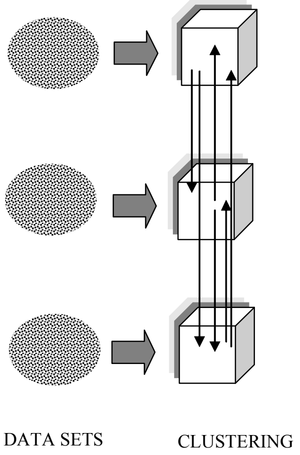


FIGURE 7. A general scheme of vertical clustering; note a "stack" of data sets communicating through some layer of granular communication

Collaboration happens through some mechanisms of interaction. While the algorithmic details are presented in the subsequent section, it is instructive to underline the nature of the possible collaboration.

- in horizontal clustering we deal with the same patterns and different feature spaces. The communication platform one can establish is through the partition matrix. As we have the same objects, this type of collaboration makes sense. The confidentiality of data has not been breached: we do not operate on individual patterns but the resulting information granules (fuzzy relations, that is partition matrices). As this number is far lower than the number of data, the low granularity of these constructs moves us quite far from the original data.
- in vertical clustering we are concerned with different patterns but the same feature space. Hence the communication at the level of the prototypes (which are high level representatives of the data) becomes feasible. Again, because of the aggregate nature of the prototypes, the confidentiality requirement has been satisfied.

There is also a number of hybrid models of collaboration where we encounter data sets with possible links of vertical and horizontal collaboration.

The collaborative clustering exhibits two important features:

- The databases are distributed and there is no sharing of their content in terms of the individual records. This restriction is caused by some privacy and security concerns. The communication between the databases can be realized at the higher level of abstraction (which prevents us from any sharing of the detailed numeric data).
- Given the existing communication mechanisms, the clustering realized for the individual datasets takes into account the results about the structures of other datasets and *actively* engages them in the determination of the clusters; hence the term of collaborative clustering

Depending upon the nature of the data located at each database and their mutual characteristics, we distinguish between two main fundamental modes of clustering such as horizontal and vertical clustering.

6. **Clusters of Clusters – a Hierarchical and Coordinated Development of Information Granules**

In case the number of clusters at each data site is different and the development of information granules has been realized at various levels of granularity, the interaction mechanism can be realized by clustering prototypes produced at the individual data sites, assessing them at the higher level and returning the findings down to the individual data site. The crux of this concept is illustrated in Figure 8.

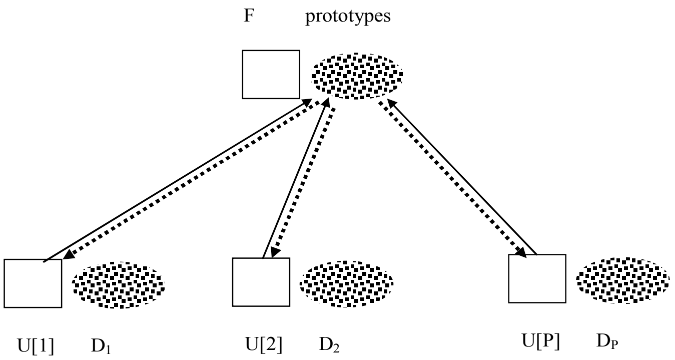


FIGURE 8. Fuzzy clustering of prototypes and their impact on the clustering realized at the individual data sites

At the algorithmic end, the clustering of the prototypes into "c" clusters results in some partition matrix F whose number of rows is equal to the number of all prototypes. Denote by f_i the maximal value of the membership in the i-th column of F. Then original objective function at the ii-th data site is modified by the values of f_i's where the corresponding distances are weighted by f_i

$$Q[ii]= \sum_{i=1}^{c[ii]} \sum_{ik=1}^{N[ii]} u_{ik}^{m} \, || \mathbf{x}_k - \mathbf{v}_i[ii] \, ||^2 \, f_i \tag{3}$$

ii=1, 2,…,P. The clustering pertaining to each site is completed by minimizing the above objective function; the results are again used at the higher level and returned to the lower level of processing; this interaction is also highlighted in Figure 5.

7. The Collaborative Development of CI Models

Information granules (say, fuzzy sets or fuzzy relations) form a conceptual and algorithmic backbone of granular constructs. They arise as the building blocks of fuzzy or neurofuzzy models. The general category of Takagi-Sugeno models is inherently associated with the information granules formed in the multivariable space that afterwards constitute conditions of the resulting rule-based system. The conclusion part is typically a certain linear or nonlinear functions typically produced as some regression constructs. Alluding to the framework developed so far, at each data site there is a certain rule-based model of the form

$$- \quad \text{if x is } A_i\,[ii] \ \text{ then } y = f_i\,(a_i\,[ii], ii) \tag{4}$$

i=1, 2,…, c[ii] and as before we are dealing with "P" data sites, ii=1, 2,…, P. The parameters of the conclusion part of the relationship occurring there are denoted by $a_i[ii]$.

When we were dealing with the collaborative pursuits, we have investigated some mechanisms of the highly orchestrated developments of the information granules which are positioned in the condition part of the model. The conclusion parts of the rules are to be subject to some collaborative developments. Assuming (which is quite typical) that the conclusion parts are multivariable linear relationships (dependencies), we intend to reconcile the existing regression models. To focus attention, let us consider regression models standing in the corresponding rules that are positioned at the individual data sites. Refer to Figure 9 portraying the essence of the anticipated interaction.

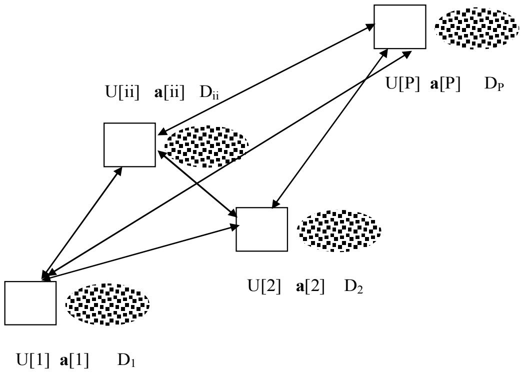


FIGURE 9. A collaborative scheme of the development of the fuzzy models; shown is a case of a single rule in which we modify the parameters of the regression model occurring in this setting

Two schemes of collaborative pursuits are envisioned here which are referred to as the
There are two fundamental modes of interaction (collaboration):

(a) centralized mode. In this mode, we consider one data, say D_i , set up to take the lead, for which we are going to reconcile the findings (its local model) with the modeling results available at all remaining datasets D_1, D_2, …, D_{i-1}, D_{i+1}, …D_P.

(b) distributed mode. Here we allow all data sites interact between each other and the resulting local models are shared. Each data site affects each other when optimizing their parameters.

Let us briefly elaborate on these two schemes by emphasizing the associated functionality and the optimization setup.

The centralized mode of interaction. Here, only one site, that is D_i revises the values of the parameters of its model. It is done through the minimization of the following objective function

$$\sum_{k=1}^{N[i]} \|\mathbf{a}_j^T[i]\mathbf{x}_k - y_k\|^2 \; u_{jk}[i] + \beta \sum_{\substack{jj=1 \\ jj\neq i}}^{N[i]} \sum_{k=1}^{N[i]} \|\mathbf{a}_j^T[i]\mathbf{x}_k - \mathbf{a}_j^T[jj]\mathbf{x}_k\|^2 \; u_{jk}[i] \qquad (5)$$

This concerns the optimization of the j-th rule in the i-th data site, that is

$$\text{-if } \mathbf{x} \text{ is } A_j[i] \text{ then } y = \mathbf{a}^T{}_j[i]\,\mathbf{x} \qquad (6)$$

where $A_j[i]$ is the linked with the j-th row of the partition matrix standing there, U[i].
Its first component optimizes the location of the regression line with respect to the data in D_i. The second one tries to reconcile the differences between this model and the others coming from data sites and made available to the i-th data site. Here a[jj] is the vector of the parameters of the linear regression model of the jj-th data site. As we have assumed correspondence between the rules, the summation is taken over the data that are in context of this given rule (as stressed by the notation of the weights produced by the corresponding row of the partition matrix). The positive weight β sets up some balance between the two components of the objective function.

The distributed mode of interaction. In this mode of interaction, each data site interacts with all remaining when reconciling the differences between the models and building the optimal findings.

The overall scheme of interaction can be outlined as follows. We assume that the level of interaction quantified by α is provided in advance.

Step 0. For each data site derive an optimal regression model standing in the conclusion parts of the corresponding rules of the fuzzy model. These locally optimal models are described by the vectors of the parameters a[1], a[2] ,…, a[P]. Evaluate the overall quality of the models by determining the value of some global performance index V.

iterate
Step 1. All data sites establish communication between themselves and exchange the regression models. Each model realizes the optimization where the original vector is updated through the minimization of a certain performance index. Consider the i-th data site. The performance index Q to be minimized at this site is expressed in the following form

$$Q = \sum_{\mathbf{x}_k \in X_i} (\mathbf{a}^T\mathbf{x}_k - y_k)^2 + \alpha \sum_{\substack{j=1 \\ j\neq i}}^{P} \sum_{\mathbf{x}_k \in X_i} (\mathbf{a}^T\mathbf{x}_k - \mathbf{a}^T[j]\mathbf{x}_k)^2 \qquad (7)$$

The result of the optimization is denoted by a˜[i]. More specifically, a˜[i] = arg $_{\mathbf{a}}$Min Q
This step is repeated for each data site, i=1,2,…,P.

Compute the value of the global performance index V for the updated vectors of the parameters of the models. If the current value of V is lower than the previous one, them update the parameters by accepting the new computed values, a[i] is replaced by a~[i] and repeat the optimization of (xx)

If no further improvement of V is reported, we stop the iterations.

end of iteration

8. **Higher-type of Constructs of Distributed and Collaborative Fuzzy Modeling**

It becomes quite remarkable that when engaging in some collaborative activities, the resulting constructs start to get closer to each other (in some sense predetermined by the essence of the assumed collaboration) yet they could be collectively described with the aid of constructs that are of higher type than the original ones. Given a collection of numeric entities, we end up with a granular construct such as e.g., fuzzy numbers. In the case of fuzzy sets, we come up with fuzzy sets of type-2 (viz. fuzzy sets whose membership degrees are rather fuzzy sets defined in [0,1] rather than single numeric values of the membership grades). Similarly we envision an aggregation of numeric values leading to a single fuzzy set. In distributed fuzzy modeling (because of distributivity) this phenomenon of aggregating more specific data becomes quite apparent. Two scenarios cast in the setting of vertical clustering are outlined. The first one focuses on the a way which we form a fuzzy set of prototypes while in the second we show how type-2 fuzzy sets of membership are formed with the use of induced partition matrices. Before we move on with the detailed discussion, we present a way in which numeric data could be aggregated into a single information granule.

8.1. From Numeric Data to Information Granules. Consider a finite collection of numeric data b_1, b_2, …, b_N and a certain numeric representative such as e.g., a prototype or the membership value denoted by "m". In the realization of the induced granular construct we engage two intuitively appealing criteria, that is (a) first, we expect that the resulting fuzzy set should highly "reflect" (match) the available numeric entities, and (b) second, the fuzzy set should be kept specific enough so that it comes with a well-defined semantics.

We construct a membership function separately for its rising and declining sections formed around "m". Assuming the simplest scenario when using the linear type of membership functions, the essence of the optimization is such that we rotate the linear section of the membership function around "m" so that the criteria presented above are satisfied. Refer to Figure 7 outlining the essence of the construct.

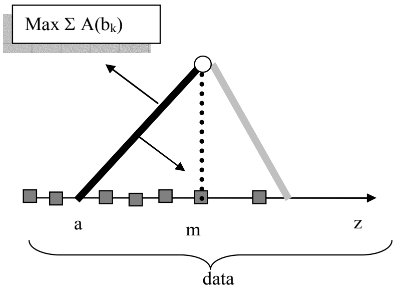


FIGURE 10. Optimization of the linearly increasing section of the membership function of A. The point of rotation of the linear segment of the membership function is marked by an empty circle

The two requirements guiding the design of the fuzzy set are and transformed into the corresponding optimization problem as outlined as follows:

(a) maximize the experimental evidence of the fuzzy set; this implies that we tend to "cover" as many numeric data as possible, viz. the coverage has to be made as high as possible. Graphically, in the optimization of this requirement, we rotate the linear segment up (clockwise) as illustrated in Figure 7. Formally, the sum of the membership grades $A(b_k)$ $\sum_k A(b_k)$ (where A is the linear membership function to be optimized and b_k is located to the left to "m") has to be maximized.

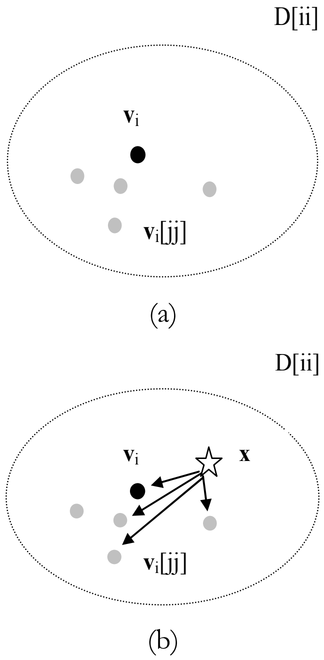
(b) Simultaneously, we would like to make the fuzzy set as specific as possible so that is comes with some well defined semantics. This requirement is met by making the support of A as small as possible, that is $\min_a |m - a|$

To accommodate the two conflicting requirements, we combine (a) – (b) in the form of the ratio that is maximized with respect to the unknown parameter of the linear section of the membership function

$$\text{Max}_a \frac{\sum_k A(b_k)}{|m-a|} \tag{8}$$

The linearly decreasing portion of the membership function is optimized in the same manner. The overall optimization returns the parameters of the fuzzy number in the form of the lower and upper bound (a and b, respectively).

8.2. The Aggregation of Results of Collaborative Clustering. When dealing with vertical clustering, the collaboration mechanism establishes an essential correspondence between the clusters (their prototypes). This correspondence is crucial to the formation of fuzzy sets of prototypes. Consider the i-th prototype at data site by v_i. The corresponding prototypes coming from remaining data sites are denoted by $v_i[1]$, $v_i[2]$, …., $v_i[P]$, Figure 11(a).



(a)

(b)

FIGURE 11. Emergence of granular constructs through the process of collaboration: (a) a fuzzy set of prototypes, and (b) fuzzy set of type-2 of membership grade of x to cluster "i"

The aggregation scheme presented in the previous section deals with scalar numeric entities and its usage to the prototypes requires that we consider each variable separately. Given the dimensionality of the data to be "n", the formation of the fuzzy set over the j-th coordinate of the data uses the experimental evidence of the form $v_{ij}[1]$, $v_{ij}[2]$, …., $v_{ij}[P]$ by splitting it into the sets of values lower than v_{ij} (that serves the role of "m") and higher than v_{ij}. The resulting fuzzy sets formed over the corresponding variables and denoted as V_{i1}, V_{i2}, …, V_{in} are put together by taking their Cartesian product

$$\mathbf{V}_i = V_{i1} \times V_{i2} \times \ldots \times V_{in} \tag{9}$$

which is regarded as a granular prototype V_i of the i-th cluster where its granular nature is reflective of the reconciliation of structural findings across various data sets.

Another facet of granular information is illustrated in Figure 11 (b). Here we encounter a situation of computing membership grades for a given pattern x. Given the prototypes {v$_i$}, i=1, 2, …, c obtained as a result of collaboration, the expression for the membership grade in the i-th cluster reads

$$u_i(\mathbf{x}) = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|\mathbf{x}-\mathbf{v}_i\|}{\|\mathbf{x}-\mathbf{v}_j\|}\right)^{\frac{2}{m-1}}} \qquad (12)$$

where || .|| is the same distance as being used in the clustering process. The same applies to the value of the fuzzification coefficient. The i-th prototype at other data sites are available here. Using them, the computations using (12) leads to the corresponding membership degrees u$_i$[1], u$_i$[2],…, u$_i$[P]. Those considered altogether with u$_i$ being computed by means of (xx) are then used to determine a membership function of type-2 fuzzy set as outlined in Section 8.1.

9. **Conclusions**

In this study, we have introduced a concept of distributed and collaborative fuzzy modeling. Given the inherently distributed nature of data and existing constraints of privacy, security and technical constraints, there is a genuine need to develop effective schemes of interaction. We have established the concept of the C^3 interaction by distinguishing between categories of mechanisms of communication, collaboration and consensus formation. The essence of design of the fuzzy models in the distributed format leads to the reconciliation of the parameters of the individual models and carrying out optimization by taking into account the local data and the parameters of other models. Interestingly, the collaborative developments lead to a diversity of constructs and this in turn brings forward the ideas information granules of higher type (such as type-2 fuzzy sets).

References

[1] R. Agarwal and R. Srikant, *Privacy-preserving data mining.*, Proc. of the ACM SIGMOD Conference on Management of Data, ACM Press, New York, May (2000), 439–450.

[2] A. M. Bensaid, L. O. Hall, J. C. Bezdek and L. P. Clarke. *Partially supervised clustering for image segmentation,* Pattern Recognition*,* **29(5)** (1996), 859-871.

[3] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, NY. (1981)

[4] C. Clifton and D. Marks, *Security and privacy implications of data mining,* Workshop on Data Mining and Knowledge Discovery, Montreal, Canada, (1996), 15–19.

[5] J. C. Da Silva, C. Giannella, R. Bhargava, H. Kargupta and M. Klusch, *Distributed data mining and agents*, Engineering Applications of Artificial Intelligence, **18 (7)** (2005), 791-807.

[6] W. Du and Z. Zhan, *Building decision tree classifier on private data*, Clifton, C., Estivill-Castro, V. (Eds.), IEEE ICDM Workshop on Privacy, Security and Data Mining,

Conferences in Research and Practice in Information Technology, Vol. **14**, Maebashi City, Japan, ACS, (2002), 1–8.

[7] T. Johnsten and V. V. Raghavan, *A methodology for hiding knowledge in databases,* Clifton, C., Estivill-Castro, C. (Eds.), IEEE ICDM Workshop on Privacy, Security and Data Mining, Conferences in Research and Practice in Information Technology, Vol. **14**. Maebashi City, Japan, ACS, (2002), 9–17.

[8] H. Kargupta, L. Kun, S. Datta, J. Ryan and K. Sivakumar, *Homeland security and privacy sensitive data mining from multi-party distributed resources,* Proc. 12th IEEE International Conference on Fuzzy Systems, *FUZZ '03,* .Volume **2**, May (2003), 25-28, Vol. **2** (2003), 1257 – 1260.

[9] S. Merugu, and J. Ghosh, A privacy-sensitive approach to distributed clustering, *Pattern Recognition Letters,* **26 (4)** (2005), 399-410.

[10] B. Park and H. Kargupta, *Distributed data mining: algorithms, systems, and applications*, In: Ye, N. (Ed.), The Handbook of Data Mining. Lawrence Erlbaum Associates, New York, (2003), 341–358.

[11] W. Pedrycz, *Algorithms of fuzzy clustering with partial supervision*, Pattern Recognition Letters, **3** (1985), 13 - 20.

[12] W. Pedrycz, and J. Waletzky, *Fuzzy clustering with partial supervision,* IEEE Trans. on Systems, Man and Cybernetics, **5** (1997), 787-795.

[13] W. Pedrycz and J. Waletzky, *Neural network front-ends in unsupervised learning,* IEEE Trans. on Neural Networks, **8** (1997), 390-401.

[14] W. Pedrycz, V. Loia and S. Senatore, *P-FCM: A proximity-based clustering,* Fuzzy Sets & Systems, 148, (2004), 21-41.

[15] W. Pedrycz, *Collaborative fuzzy clustering*, Pattern Recognition Letters, **23(14)**(2002), 1675-1686.

[16] W. Pedrycz, *Knowledge-Based Clustering: From Data to Information Granules*, J. Wiley, New York (2005).

[17] W. Pedrycz and F. Gomide, *Fuzzy Systems Engineering: Toward Human-Centric Computing*, J. Wiley, NJ Hoboken, (2007).

[18] A. Strehl and J. Ghosh, *Cluster ensembles—a knowledge reuse framework for combining multiple partitions*, Journal of Machine Learning Research, **3**, (2002), 583–617.

[19] H. Timm, F. Klawonn and R. Kruse, *An extension of partially supervised fuzzy cluster analysis***,** Proc. Annual Meeting of the North American Fuzzy Information Processing Society, *NAFIPS,* (2002), 63 –68.

[20] G. Tsoumakas, L. Angelis and I. Vlahavas, *Clustering classifiers for knowledge discovery from physically distributed databases,* Data & Knowledge Engineering, **49(3)** (2004), 223-242.

[21] V. S. Verykios, et al. State of the art in privacy preserving data mining, SIGMOID Record, **33(1)** (2004), 50-57.

[22] L. A. Zadeh, *Toward a generalized theory of uncertainty (GTU) – an outline*, Information Sciences, **172(1-2)** (2005), 1-40.

Witold Pedrycz, Department of Electrical & Computer Engineering, University of Alberta, Edmonton T6R 2G7 Canada and Systems Research Institute of the Polish Academy of Science, Warsaw, Poland

E-mail address: pedrycz@ee.ualberta.ca