Abstract

This paper reflects the trends of the past years based on the diffusion of various traditional approaches and methods when tackling new problems. Two components of the computational intelligence (CI) are applied, rough and fuzzy sets theory. These components permit one to operate with uncertainty data. The current knowledge in the investigated field is summarized and briefly explained. It also deals with uncertainty in an information system and the two approaches, the fuzzy sets (FSs) and rough sets theory (RST), for operating it. The proposal and implementation of a rough-fuzzy classifier (RFC) is modified. RFC uses the rules generated by RSTbox. The databases IRIS and WINE were chosen for verification. The classification results were compared with the results of other classification methods applied on these databases. Finally, we summarized the presented problems. Based on the above stated facts it can be claimed that the proposed modified algorithm, RSTbox and RFC model are functional. The model is relatively successful (compared to other approaches), and by using it two classification databases can be carried out. This model is proposed in MATLAB.

Keywords: Classification; fuzzy inference system; rough-fuzzy approach; rules generation; UCI data repository

1. Introduction

In this section we will review the basic concepts and definitions of FSs \(^1^3\), rough sets \(^4^6\), and fuzzy-rough approach \(^7^1^0\) in relation to the process classification of systems. Systems can be usually described and defined
by humans. This section analyses the past years trend in the diffusion of various traditional methods and approaches when tackling new problems.

CI methods are used in creating of system models. Areas of CI (FSs, neural networks, genetic algorithms, rough sets etc.) belong to a fast developing field in the applied research. It is composed of several theories and approaches which despite being different from one another, have two common denominators. They are the non-symbolic representation of pieces of knowledge and ‘bottom-up’ architecture, in which the structures and paradigms appear from an unordered beginning.

These two denominators have been successfully used in various uncertainty information processing systems. The RST, attributed by prof. Pawlak, is based on the research in the logical properties of information systems, and the uncertainty in information systems which are expressed by a boundary region. Every investigated object is related to a specific piece of information. The objects which are characterized by the same pieces of information are mutually indistinguishable from the point of view of the accessible pieces of information. This is expressed in RST by the indiscernibility relations. The theory of FSs, attributed by prof. Zadeh, is a known approach to uncertainty. In this theory an element belongs to a set according to the membership function values (membership degree), i.e. to a closed interval. Theory of FSs is an expansion of the traditional sets theory in which an element either is or is not a set member. If we attempt to describe and model a particular reality problem we encounter a certain discrepancy. On one hand there is the accuracy of mathematical methods by which a specific problem is described, and on the other hand there is a very complicated reality demanding a range of simplifications and the consequent inaccuracy, infidelity of the model arising from them.

RST and FSs are applied in classifier modeling. The goal of the paper is a synthesis and analysis of the original RFC model with using of the RST tool (henceforth called RSTbox).

2. Classification based on rough and fuzzy sets

The role of a classification is to classify objects, events, and real-life situations into classes. Each of the reviewed objects is unique, original, and its classification means a certain degree of generalization. Let's define a system for the particular objects, i.e. input and output variables, elements (objects), and their mutual relations. Defining and collecting the data of input/output variables cannot be generalized, even though this stage influences the classification result. An application of classification methods based on CI represents an effective tool for the realization of a classification model.

On the basis of achieved classification results, it seems to be effective and up-to-date to tackle the classification problem using a hybrid approach combining rough sets and FSs. Both of them belong to the field of the CI research.

The application of the classification methods based on CI represents an effective tool for the classification model implementation. For example, we can speak about probabilistic rough classifiers, fuzzy classifiers etc. The probabilistic rough classifier combines all positive aspects of rule induction systems with the flexibility of statistical techniques for classification. Two natural approaches to classifier design are: to ask experts how they solve the problem and try to encapsulate the knowledge in a fuzzy-base classifier; to collect input-output data (i.e. a labeled data set) and extract the classifier parameters from the data. The first model represents transparent approach (is interpretable in the domain context) and the second based on data, may or may not be interpretable. Fuzzy classifier models are deemed to be able to integrate both approaches: human and data sources.

On the basis of specialized literature we can define some known interesting approaches to rough fuzzy hybridization.

They are divided into two groups - supervised and unsupervised learning. Referring to the former, we can speak about supervised learning and information retrieval. In a fuzzy-rough ownership function that involves
the fuzzy uncertainty and the rough uncertainty is proposed. All training patterns influence the ownership function, and hence no decision is required as to the number of neighbors to consider, although there are other parameters that must be defined for its successful operation. This concept is further evolved in where “... classification efficiency of the conventional K-nearest neighbor algorithm is enhanced by exploiting fuzzy-rough uncertainty. The simplicity and nonparametric characteristics of the conventional K-nearest neighbor algorithm remain intact in the proposed algorithm. Unlike the conventional one, the proposed algorithm does not need to know the optimal value of \( K \). Moreover, the generated class confidence values, which are interpreted in terms of fuzzy-rough ownership values, do not necessarily sum up to one”.

In the classification task is divided into four stages. The first stage is the fuzzy-rough ownership calculation; in the second stage the training set is filtered. The further representative points are selected from the processed training set and fuzzy-rough ownership values updated based on mountain clustering during the third stage. In the fourth stage the test patterns are classified using the fuzzy-rough algorithm from.

In the second approach (unsupervised), we can discuss unsupervised learning and clustering. One of the first proposals of unsupervised learning, in the context of rough set theory, can be found in 1996. It describes the rough Kohonen’s neural network classifiers for the classification of complex objects. The axiomatic approach is taken in. In this text upper and lower approximations of a fuzzy subset, with respect to an indistinguishability operator, are studied and their relations with fuzzy rough sets are investigated. Both constructive and axiomatic approaches are used in. In the constructive approach, a pair of lower and upper generalized approximation operators are defined. In the axiomatic approach, various classes of fuzzy rough approximation operators are characterized by different sets of axioms. In the hybridization of rough sets model is investigated. Object-oriented rough set models are suggested for rule generation.

2.1. Design of Rough-Fuzzy Classifier Model

Our case deals with a hybrid RFC model. For a whole range of scientific papers dealing with the rule generation for analyzed data and a lot of various methods and procedures using CI see. This means that RST was used to define IF-THEN rules (conditioned rules) and FSs were applied as fuzzy controller with Mamdani inference. It is possible to speak about Mamdani’s fuzzy inference system (FIS).

The problem of classification in our model consists of three phases. The first phase is the pre-processing of real data that has been pre-processed and modified into a suitable format. Histograms were created for them, from which linguistic variables were derived. The whole data set was divided, in accordance to the ‘hold-out’ method, into training and testing sets. The second phase is the classification divided into RSTbox rules generation and FIS optimization, and the third phase is the output and class interpretation as we can see in Fig. 1. The learned knowledge is presented in the form of a set of decision rules that can easily be explained and understood by users. Rough sets approach is applied in RSTbox for the generation of minimal fuzzy rule base for FIS in \( \text{RFC}_1, \text{RFC}_2, \ldots, \text{RFC}_k \). These sets of \( \text{RFC}_i \) use various types of input membership functions. By using various membership functions shapes (symmetric, non-symmetric) and membership functions types (triangular, bell-shaped, gauss-shaped etc.) the outputs were compared. The shape of these membership functions is harmonized with the real data histograms, and particular rules stresses adjustments were made. The \( \text{RFC}_i \) can be described as multiple inputs and single output system where the inputs are the attributes of a real data set, and the output is the decision about the classification. Finally, the accuracy of \( \text{RFC}_i \) classification is used in order to choose the best fit. The models of the system were created and tested in MATLAB and Simulink and the results were collectively evaluated.

The goal of the experiments performed on the selected data is to verify the accuracy of the proposed RFC procedure (see Fig. 2), to reach the high testing data classification accuracy even in comparison with the algorithms hitherto known. The real data set represents the input of our procedure with the help of pre-processed data, we are able to use this information in the computation of histograms. On their basis FIS and
RSTbox, in which the pre-processed data are utilized, are subsequently modified. The classes are the outputs of the whole procedure.

It is inferred that a classifier is a unit (algorithm, model) executing a classification, a classifier input is a set of attributes, and a classifier output is a class allocation.

![Fig. 1. Model of rough-fuzzy classification](image)

The stated RFC model is based on the following assumptions:

- Let’s specify a set by attributes $A = a_1, a_2, ..., a_n$ and $A_r = v_r, 1, v_r, 2, ..., v_r, m$ is $n$-dimensional vector of attributes values where $r = 1, 2, ..., m$. 

![Fig. 2. Rough-fuzzy classifier model](image)
Let’s suppose the classification into $R$ classes be called $h_1, h_2, \ldots, h_R$. Let’s mark $N$-dimensional attributes space by $\Pi$.

Let’s mark $N$-dimensional attributes space by $\Pi$.

A class indicator $d \in \{h_1, h_2, \ldots, h_n \}$ is assigned to every $A_r \in \Pi$. Function $d = f(A_r)$ is the rule defining this assignment.

### 2.2. Experimental comparison of proposed and other classifiers

The most important goal in designing a classifier is to achieve the highest possible classification accuracy or the lowest possible error rate. We are referring to the accuracy measure and error measure. The classification accuracy is the ratio of correctly classified objects to the total amount of objects $x$ in a set, expressed in percent (here denoted $P_x$). The parameter of the total classification error of a classifier model, obtained as the difference $100 - P_x$, is also frequently used. The resubstitution error is the next well known numeric parameter which is obtained as the ratio of correct classified objects to the total amount of training data objects in a set, and are expressed as a percent.

The methods used for the classifier accuracy evaluation according to are:

- Testing of the whole training data
- K-fold validation
- Leave-one-out
- Testing of the testing data by ‘holdout’ method
- Bootstrap

In our experiments we used testing on the whole training data, which is based on using one data set for both training and testing. This method is applicable, however, it bears the highest threat of overfitting, decreasing the testifying parameter abilities, and it is affected by the resubstitution error. The method is called “optimistic” here.

The “holdout” method was used, too. It means an accidental data divided into two independent sets, training and testing. The usual division proportion is 2/3 to 1/3 up to 4/5 to 1/5. The training set serves for the model (classifier) creation and derivation, and the testing set for the classification accuracy determination. This method gives a more pessimistic $P_x$. Once more the goal of the selected data experiments is to verify the legitimacy of the proposed RFC procedure (see Fig. 3), to reach a high testing data classification accuracy even in comparison with the algorithms hitherto known. The method is called “pesimistic” here.

For the first part of the experiments IRIS-called data were used. This data is often cited and may perhaps be the best known database to be found in the pattern recognition area. The database contains 150 records of size measurements for iris flowers. The length and width of sepal and petals were measured. Three kinds of iris were investigated - setosa, virginica and versicolor, where each iris plant refers to a class. One class is linearly separable from the other two, the latter being not linearly separable from each other. The second group of experiments was carried out with WINE-called database (wine recognition data). This data came into existence as the results of a chemical analysis three distinct Italian-region-grown wines. The data contains chemical elements values from 178 samples. The analysis determined the quantities of 13 constituents found in each of the three types of wines. All attributes are continuous.

The experiments run according to Fig. 1 and Fig. 2. Firstly, the data was pre-processed and converted in a suitable format. Consequently, the histograms for the preprocessed data have been calculated (see Fig. 3).

To proceed in “holdout”, the IRIS data was divided into training (120 objects) and testing (30 objects). The second part proceeded concurrently with the whole data set. A 30- and 150-object set was used for testing. The systems created in this way were then tested in Simulink-created models and the results collectively evaluated. We can see the example of the notation of FIS type Mamdani with non-symmetric triangular membership functions in Fig. 4. The membership function (non-symmetric triangular) for the petal-width (PW) parameter is presented in Fig. 5.
Fig. 3. Histogram for IRIS data, PW parameter

Fig. 4. Part of Mamdani’s FIS algorithm

[System]
Name="trimfbyexpert"
Type="mamdani"
...

[Rules]
1 2 1 1, 1 (0.085271) : 1
1 3 1 1, 1 (0.24806) : 1
1 1 1 1, 1 (0.00775) : 1
1 1 2 2, 2 (0.03876) : 1
1 1 2 3, 3 (0.007752) : 1
2 3 3 3, 3 (0.0386) : 1
2 3 1 1, 1 (0.03876) : 1
2 1 2 2, 2 (0.093) : 1
2 2 3 3, 3 (0.0155) : 1
2 3 2 3, 3 (0.007752) : 1
3 3 2 2, 2 (0.10978) : 1
3 2 2 3, 3 (0.0155) : 1
3 1 2 2, 2 (0.023256) : 1
3 2 2 2, 3 (0.0155) : 1
3 1 2 3, 3 (0.007752) : 1
1 3 1 2, 1 (0.0235) : 1
The outputs are demonstrated in the following tables (Table 1, 2 and 3). The same procedure was used for the WINE data in this case. The data set was divided into training (138 objects) and testing (40 objects). The resulting classification accuracy denoted $P_x$ is the ratio of correctly classified objects to the total amount of objects $x$ in a set, expressed as a %, as we can see in Table 1.

Table 1. The best results for IRIS and WINE data sets

| IRIS data set | WINE data set |
|---------------|---------------|
| trimfhyexpert-test$_{150}$ | trimfhyexpert-test$_{178}$ |
| “optimistic” | “optimistic” |
| $P_{\text{IRIS}}$ (%) | $P_{\text{WINE}}$ (%) |
| 95.31 | 96.60 |
| 93.33 | 95.00 |

The classification results have been compared with methods published in, as we can see in Table 2 and 3, too.

Table 2. Classification accuracy $P_{\text{IRIS}}$ for IRIS data

| Original RFC | Other classifications methods |
|--------------|-------------------------------|
| C5 | ID3 | EFUNN | Hong-Chen | PRISM |
| $P_{\text{IRIS}}$ (%) | 93.33 | 92.00 | 90.70 | 96.00 | 96.67 | 90.00 |

Table 3. Classification accuracy $P_{\text{WINE}}$ for WINE data
3. Conclusion

This paper dealt with the data classification. When describing real classification problem, it is possible to express their description by using a natural language. This description is uncertainty-loaded. To operate with uncertainty it is suitable to use RST and FSs theories, or possibly their combinations. For this reason the introduction section summarizes the basic ideas of the presented theories. The data IRIS and WINE from 39 were used for a classification problem testing. It is generally known and used as benchmark data. The experiments verified the proposed model for the data classification, and the results were compared with other available classification methods presented in 33, 38, and were applied to the same data.

The presented RFC turned out to appear suitable. The classification accuracy for IRIS data reached 93.33% (see in Table 1 and 2). The classification accuracy was 95% for WINE data (see in Table 1 and 3). On the basis of the above stated facts it can be claimed that the proposed RFC model is functional, relatively successful compared with other methods, and can be used to carry out various databases classification.

In the field of RFC it is feasible to proceed from the supervised learning technique to the combined approach by using pre-processed data in the first phase, e.g. a self-organization map in the future.

Acknowledgements

This work was supported by the project No. SGFES02 of the Ministry of Education, Youth and Sports of CR with title Research and Development Activities in the area of System Engineering and Informatics at the Faculty of Economics and Administration, University of Pardubice in 2014.

References

1. Nguyen HT., et al. First course in fuzzy and neural control, Chapman and Hall/CRC: Boca Raton, 2003.
2. Ross TJ., Fuzzy logic with engineering applications. Wiley: Hoboken New York, 2004.
3. Zadeh L.A., The roles of fuzzy logic and soft computing in the conception, design and deployment of intelligent systems. Software Agents and Soft Computing. 1997, 183-190.
4. Stefanowski J., Sowiski R., Rough set reasoning about uncertain data. Fundamenta Informaticae, 1996, 27:229-243.
5. Duntsch I., Gediga G., Rough set data analysis- a road to non-invasive knowledge discovery. Methodos: Angor, 2000.
6. Greco S., Matarazzo B., Slowinski R., The use of rough sets and fuzzy sets in MCDM. In: Multicriteria decision making: Advances in MCDM models, algorithms, theory, and applications. Kluwer Academic Publisher: Dordrecht, 1999.
7. Pal, SK., Soft data mining, computational theory of perceptions, and rough-fuzzy approach. Information Science, 2004, 163:5-12.
8. Shen Q., Chouchoulas A., A rough-fuzzy approach for generating classification rules. Pattern Recognition, 2002, 35:2245-2438.
9. Jirava P., Krupka J., Classification model based on rough and fuzzy sets theory. In: Proc. of the 6th WSEAS Int. Conf. on CINMACS97, 2007, pp. 199-203.
10. Dai J., Tian H., Fuzzy rough set model for set-valued data. Fuzzy Sets and Systems, 2013, 229, 54-68.
11. Bezdek JC., What is computational intelligence? Computational Intelligence: Imitating Live, (1994, 1-12.
12. Komorowski J., Pawlak Z., Polkowski L., Skowron A., Rough sets: a tutorial. In: Pal SK., Skowron A. (Eds.) Rough-fuzzy hybridization: a new trend in decision-making, Springer-Verlag: Singapore, 1998, pp. 3-98.
13. Pawlak Z., Rough sets. Int. J. of Information and Computer Sciences, 1982, 11:341-356.
14. Pawlak Z., A primer on rough sets: a new approach to drawing conclusions from data. Cardozo Law Review, 2001, 22, 5-6:1407-1415.
15. Zadeh LA., Fuzzy sets. Information and Control, 1965, 8:338-353.
Zadeh LA., Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans S.M.C.*, 1973, 3:28-44.

Dubois D., Prade H., *Fuzzy information engineering and soft computing: a guided tour of applications*. John Wiley Sons: New York, 1997.

Gorzalczany MB., Piasta Z., Neuro-fuzzy approach versus rough-set inspired methodology for intelligent decision support. *Information Sciences*, 1999, 120, 1-4:45-68.

Kunccheva LI., *Fuzzy classifier design*. Physica-Verlag: Heidelberg New York, 2000.

Lingras P., Jensen R., Survey of rough and fuzzy hybridization. In: *Fuzzy Systems Conference 2007*, 23-26 July, 2007, 1-6.

Cheng Y., The incremental method for fast computing the rough fuzzy approximations. *Data & Knowledge Engineering*, 2011, 70:84-100.

Zhou J., Pedrycz W., Miao D., Shadowed sets in the characterization of rough-fuzzy clustering. *Pattern Recognition*, 2011, 44:1738-1749.

Gorzalczany MB., Piasta Z., Neuro-fuzzy approach versus rough-set inspired methodology for intelligent decision support. *Information Sciences*, 1999, 120, 1-4:45-68.

Kunccheva LI., *Fuzzy classifier design*. Physica-Verlag: Heidelberg New York, 2000.

Lingras P., Jensen R., Survey of rough and fuzzy hybridization. In: *Fuzzy Systems Conference 2007*, 23-26 July, 2007, 1-6.

Cheng Y., The incremental method for fast computing the rough fuzzy approximations. *Data & Knowledge Engineering*, 2011, 70:84-100.

Zhou J., Pedrycz W., Miao D., Shadowed sets in the characterization of rough-fuzzy clustering. *Pattern Recognition*, 2011, 44:1738-1749.

Zhang Z., On interval type-2 rough fuzzy sets. *Knowledge-Based Systems*, 2012, 35:1-13.

Maji P., Pal SK., Rough set based generalized fuzzy c-means algorithm and quantitative indices. *IEEE Trans Syst Man Cybern B*, 2007, 37, 6: 1529-40.

Zhou L., Wu WZ., Zhang WX., On intuitionistic fuzzy rough sets and their topological structures. *International Journal of General Systems*, 2009, 38, 2:589-616.

Zhang Z., An interval-valued rough intuitionistic fuzzy set model. *International Journal of General Systems*, 2010, 39, 2:135-164.

Sarkar M., Fuzzy-rough nearest neighbor algorithms in classification. *Fuzzy Sets and Systems*, 2007, 158, 19:2134-2152.

Wang X., etc. Fuzzy-rough set based nearest neighbour clustering classification algorithm. *Lecture Notes in Computing Science*, 2005, 3613:370-373.

Lingras P., Unsupervised learning using rough kohonen neural network classifiers. In: *Proc. of the symposium on modelling, analysis and simulation, CESA96 IMACS Multiconference*, 1996, pp. 753-757.

Boixoixader D., Jacas J., Recasens J., Upper and lower approximations of fuzzy sets. *International Journal of General System*, 2000, 29:555-568.

Wu WZ., Mi JS., Zhang WX., Generalized fuzzy rough sets. *Information Sciences*, 2003, 151:263-282.

Jirava P., *Information system analysis based on rough sets*. PhD Thesis, University of Pardubice: Pardubice, 2007.

Guo G., etc. Similarity-based data reduction techniques. *Jour. of Research and Practice in Information Technology*. Australian Computer Society, 2005, 37, 2:211-232.

Kudo Y., Murai T., A method of generating decision rules in object oriented rough set models. In: *RSCTC06*, Kobe, 2006, pp. 338-347.

Sakai H., Nakata M., On possible rules and apriori algorithm in non-deterministic information systems. In: *RSCTC06*, Kobe, 2006, pp. 264-273.

Han J., Kamber M., *Data mining*. Elsevier: San Francisco, 2006.

Witten IH., Frank E., *Data mining: practical machine learning tools and techniques*. 2nd edition, Morgan Kaufmann: Amsterdam, 2005.

Stefanowski J., On rough set based approaches to induction of decision rules. *Rough Sets in Knowledge Discovery*, 1998, 1:500-529.

Asuncion A., Newman DJ., *UCI repository of machine learning databases and domain theories*, 1999

http://www.ics.uci.edu/MLRepository.html