1 Approximate Reasoning Schemes: Classifiers for Computing With Words

Andrzej Skowron, Marcin Szczuka

Institute of Mathematics
Warsaw University
Banacha 2, 02-097 Warsaw, Poland

Abstract. In the paper we discuss classifiers relevant to approximate reasoning. The approach is based on rough-fuzzy hybridization. We discuss its possible applications to computing with words.

1.1 Introduction

We propose to use classifiers for rough-fuzzy concepts (see, [8]) as a tool in searching for approximate reasoning rules, called productions. From such productions approximate reasoning schemes can be derived. They are the basic constructions in rough-neuro computing [4] based on rough mereological approach [6]. The approach can be treated as a way to Computing with Words (see, e.g., [11], [12]). The proposed approach splits approximate reasoning into the following stages. In the first stage classifiers for relevant concepts should be induced using the existing statistical and other methods (see, e.g., [2], [3], [7]). Next, productions are extracted from data. It is important to note that it is possible to develop productions in such a way that they are using only linguistic names. Using the productions approximate reasoning schemes can be derived. On the top level solely linguistic names names appear in the reasoning scheme.

1.2 Information Granule Systems and Parameterized Approximation Spaces

In this section, we present a basic notion for our approach, i.e., information granule system. An information granule system is a tuple

$$S = (G, R, Sem) \tag{1.1}$$

where

1. G is a finite set of parameterized constructs (e.g., formulas) called information granules;
2. R is a finite (parameterized) relational structure;
3. Sem is a semantics of G in R.

We assume that with any information granule system there are associated:

1. H a finite set of granule inclusion degrees with a partial order relation $<$ which defines on H a structure used to compare the inclusion degrees; we assume that H consists of the lowest degree 0 and the largest degree 1;
2. $\nu_p \subseteq G \times G$ a binary relation *to be a part to a degree at least* p between information granules from G, called *rough inclusion*. (Instead of $\nu_p(g, g')$ we also write $\nu(g, g') \geq p$.)

Components of an information granules system are parameterized. It means that we deal with parameterized formulas and a parameterized relational system. The parameters are tuned to make it possible to construct finally relevant information granules, i.e., granules satisfying specification or/and some optimization criteria. Parameterized formulas can consist of parameterized sub-formulas. The value set of parameters labelling a sub-formula is defining a set of formulas. By tuning parameters in optimization process or/and information granule construction a relevant subset of parameters is extracted and used for construction of the target information granule.

There are two kinds of computations on information granules. These are computations on information granule systems and computations on information granules in such systems, respectively. The first ones are aiming at construction of a relevant information granule systems defining parameterized approximation spaces for concept approximations used on different levels of target information granule constructions and the goal of the second ones is to construct information granules over such information granule systems to obtain target information granules, e.g., satisfying a given specification (at least to a satisfactory degree).

Examples of complex granules are tolerance granules created by means of similarity (tolerance) relation between elementary granules, decision rules, sets of decision rules, sets of decision rules with guards, information systems or decision tables (see, e.g., [8]). The most interesting class of information granules are information granules approximating concepts specified in natural language by means of experimental data tables and background knowledge.

One can consider as an example of the set H of granule inclusion degrees the set of binary sequences of a fixed length with the relation ν to be a part defined by the lexicographical order. This degree structure can be used to measure the inclusion degree between granule sequences or to measure the matching degree between granules representing classified objects and granules describing the left hand sides of decision rules in simple classifiers. However, one can consider more complex degree granules by taking as degree of inclusion of granule g_1 in granule g_2 the granule being a collection of common parts of these two granules g_1 and g_2.

New information granules can be defined by means of operations performed on already constructed information granules. Examples of such operations are set theoretical operations (defined by propositional connectives). However, there are other operations widely used in machine learning or pattern recognition [3] for construction of classifiers. These are the $Match$ and $Conflict\_res$ operations. We will discuss such operations in the following section. It is worthwhile mentioning yet another important class of operations, namely, operations defined by data tables called decision tables [8]. From these decision tables, decision rules specifying operations can be induced. More complex operations on information granules are so called transducers [1]. They have been introduced to use background knowledge (not necessarily in the form of data tables) in construction of new granules. One can consider theories or their clusters as information granules. Reasoning schemes in natural language define the most important class of operations on information granules to be investigated. One of the basic problems for such operations and schemes of reasoning is how to approximate them by available information granules, e.g., constructed from sensor measurements.

In an information granule system, the relation ν_p to be a part to a degree at least p has a special role. It satisfies some additional natural axioms and additionally some axioms of mereology [6]. It can be shown that the rough mereological approach built on the basis of the relation to be a part to a degree generalizes the rough set and fuzzy set approaches. Moreover, such relations can be used to define other basic concepts like closeness of information granules, their semantics, indiscernibility and discernibility of objects, information granule approximation and approximation spaces, perception structure of information granules as well as the notion of ontology approximation. One can observe that the relation to be a part to a degree can be used to define operations on information granules corresponding to generalization of already defined information granules. For details the reader is referred to [4].

Let us finally note that new information granule systems can be defined using already constructed information granule systems. This leads to a hierarchy of information granule systems.

1.3 Classifiers as Information Granules

An important class of information granules create classifiers. One can observe that sets of decision rules generated from a given decision table $DT = (U, A, d)$ (see, e.g., [8] can be interpreted as information granules. The classifier construction from DT can be described as follows:

1. First, one can construct granules G_j corresponding to each particular decision $j = 1, \ldots, r$ by taking a collection $\{g_{ij} : i = 1, \ldots, k_j\}$ of left hand sides of decision rules for a given decision.
2. Let E be a set of elementary granules (e.g., defined by conjunction of descriptors) over $IS = (U, A)$. We can now consider a granule denoted by
$$Match(e, G_1, \ldots, G_r)$$
for any $e \in E$ being a collection of coefficients ε_{ij} where $\varepsilon_{ij} = 1$ if the set of objects defined by e in IS is included in the meaning of g_{ij} in IS, i.e., $Sem_{IS}(e) \subseteq Sem_{IS}(g_{ij})$; and 0, otherwise. Hence, the coefficient ε_{ij} is equal to 1 if and only if the granule e matches in IS the granule g_{ij}.
3. Let us now denote by $Conflict\_res$ an operation (resolving conflict between decision rules recognizing elementary granules) defined on granules of the form $Match(e, G_1, \ldots, G_r)$ with values in the set of possible decisions $1, \ldots, r$. Hence,
$$Conflict\_res(Match(e, G_1, \ldots, G_r))$$
is equal to the decision predicted by the classifier
$$Conflict\_res(Match(\bullet, G_1, \ldots, G_r))$$
on the input granule e.

Hence, classifiers are special cases of information granules. Parameters to be tuned are voting strategies, matching strategies of objects against rules as well as other parameters like closeness of granules in the target granule.

The classifier construction is illustrated in Fig. 1.1 where three sets of decision rules are presented for the decision values $1, 2, 3$, respectively. Hence, we have $r = 3$. In figure to omit too many indices we write α_i instead of g_{i1}, β_i instead of g_{i2}, and γ_i instead of g_{i3}, respectively. Moreover, $\varepsilon_1, \varepsilon_2, \varepsilon_3$, denote $\varepsilon_{1,1}, \varepsilon_{2,1}, \varepsilon_{3,1}$; $\varepsilon_4, \varepsilon_5, \varepsilon_6, \varepsilon_7$ denote $\varepsilon_{1,2}, \varepsilon_{2,2}, \varepsilon_{3,2}, \varepsilon_{4,2}$; and $\varepsilon_8, \varepsilon_9$ denote $\varepsilon_{1,3}, \varepsilon_{2,3}$, respectively.

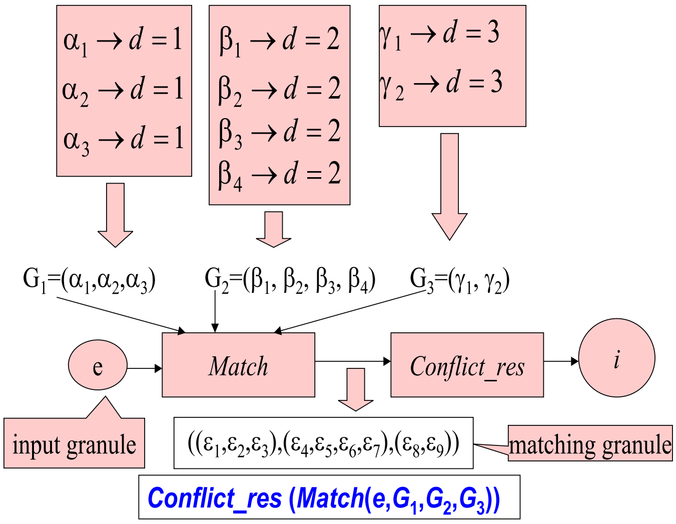

Fig. 1.1. Classifiers as Information Granules

The reader can now easily describe more complex classifiers by means of information granules. For example, one can consider soft instead of crisp inclusion between elementary information granules representing classified objects and the left hand sides of decision rules or soft matching between recognized objects and left hand sides of decision rules.

1.4 Approximation Spaces in Rough-Neuro Computing

In this section we would like to look more deeply on the structure of approximation spaces in the framework of information granule systems.

Such information granule systems are satisfying some conditions related to their information granules, relational structure as well as semantics. These conditions are the following ones:

1. Semantics consists of two parts, namely relational structure R and its extension R^*.
2. Different types of information granules can be identified: (i) object granules (denoted by x), (ii) neighborhood granules (denoted by n with subscripts), (iii) pattern granules (denoted by pat), and (iv) decision class granules (denoted by c).
3. There are decision class granules $c_1, \cdots, c_r$ with semantics in R^* defined by a partition of object granules into r decision classes. However, only the restrictions of these collections to the object granules from R are given.
4. For any object granule x there is a uniquely defined neighborhood granule n_x.
5. For any class granule c there is constructed a collection granule $\{(pat, p) : \nu_p^R(pat, c)\}$ of pattern granules labelled by maximal degrees to which pat is included in c (in R).
6. For any neighborhood granule n_x there is distinguished a collection granule $\{(pat, p) : \nu_p^R(n_x, pat)\}$ of pattern granules labelled by maximal degrees to which n_x is at least included in pat (in R).
7. There is a class of $Classifier$ functions transforming collection granules (corresponding to a given object x) described in two previous steps into the powerset of $\{1, \cdots, r\}$. One can assume object granules to be the only arguments of $Classifier$ functions if other arguments are fixed.

The classification problem is to find a $Classifier$ function defining a partition of object granules in R^* as close as possible to the partition defined by decision classes.

Any such $Classifier$ defines the lower and the upper approximations of union of decision classes c_i over $i \in I$ where I is a non-empty subset of $\{1, \cdots, r\}$ by

$$\underline{Classifier}(\{c_i\}_{i \in I}) = \{x \in \bigcup_{i \in I} c_i : \emptyset \neq Classifier(x) \subseteq I\}$$

$$\overline{Classifier}(\{c_i\}_{i \in I}) = \{x \in U^* : Classifier(x) \cap I \neq \emptyset\}.$$

The positive region of $Classifier$ is defined by

$$POS\,(Classifier) = \underline{Classifier}(\{c_1\}) \cup \cdots \cup \underline{Classifier}(\{c_r\}).$$

The closeness of the partition defined by the constructed $Classifier$ and the partition in R^* defined by decision classes can be measured, e.g., using ratio of the positive region size of $Classifier$ to the size of the object universe. The quality of $Classifier$ can be defined taking, as usual, only into account objects from $U^* - U$:

$$quality(Classifier) = \frac{card(POS\,(Classifier) \cap (U^* - U))}{card((U^* - U))}.$$

One can see that approximation spaces have many parameters to be tuned in order to construct the approximation of high quality class granules.

One more interesting issue is the direct connection between descriptions using classifier-based granules and the characterization in terms of the Dempster-Shafer theory of evidence. This inter-connection derives from the relationships that exist between rough set theory and evidence theory as described in e.g. [9]. We may introduce belief and plausibility functions that characterize granules defined by classifiers in the following way (with previous notation):

$$Bel_{Classifier}(I) = \frac{|\{x \in U^* : Classifier(x) \subseteq I\}|}{|U^*|}$$
$$= \frac{|\{\underline{Classifier}(\{c_i\}_{i \in I})|}{|U^*|}$$
$$Pl_{Classifier}(I) = \frac{|\{x \in U^* : Classifier(x) \cap I \neq \emptyset\}|}{|U^*|}$$
$$= \frac{|\{\overline{Classifier}(\{c_i\}_{i \in I})|}{|U^*|}$$

1.5 Standards, Productions, and *AR*-schemes

AR-schemes have been proposed as schemes of approximate reasoning in rough neurocomputing (see, e.g., [4], [8]). The main idea is that the deviation of objects from some distinguished information granules, called standards or prototypes, can be controlled in appropriately tuned approximate reasoning. Several possible standard types can be chosen. Some of them are discussed in the literature (see, e.g., [4]). We propose to use standards defined by classifiers. Such standards correspond to lower approximations of decision classes or (definable parts of) boundary regions between them.

Rules for approximate reasoning, called productions, are extracted from data (for details see [8]). Any production has some premisses and conclusion. In the considered case each premiss and each conclusion consists of a triple *(classifier, standard, deviation)*. This idea in hybridization with rough-fuzzy information granules (see, e.g., [8]) seems to be especially interesting. The main reasons are:

- standards are values of classifiers defining approximations of cut differences and boundary regions between cuts [8],
- there is a natural linear order on such standards defined by classifiers.

To explain the meaning of productions let us consider the following example of a production with two premisses:

$$\textbf{if} \quad (C_1, stand_1, \varepsilon_1) \ \textbf{and} \quad (C_2, stand_2, \varepsilon_2) \ \textbf{then} \ (C, stand, \varepsilon)$$

In the production classifiers C_1, C_2, C are labelled by standards $stand_1$, $stand_2$, $stand$ and deviations ε_1, ε_2, ε. The deviation ε is showing the range in which (in the considered linear order) can the deviation move the standard *stand*. The intended meaning of such production is that if the deviation of input from standards $stand_1$, $stand_2$ are respectively at most ε_1, ε_2 then the conclusion deviates from *stand* to degree at most ε.

From production (extracted from data) *AR*-schemes can be derived (see, e.g., [8]).

One more important step that can be performed in order to bring this framework closer to the idea of pure computing with words is by substituting the degrees of closeness (deviations $\varepsilon, \varepsilon_1, \varepsilon_2$ in our case) by linguistic variables. What we want to make possible is the formulation of granule production in a purely linguistic way, for example:

if similarity between C_1 output and standard $stand_1$ is *high*
and similarity between C_2 output and standard $stand_2$ is *low*
then similarity between C output and standard $stand$ is *medium*

To achieve this task we have to define partitions for the ranges of deviation as the deviation is used to measure similarity between classifier and corresponding standards. Let us consider the deviation ε for the classifier C output and standard *stand*. It is quite natural to assume that the subsets of ε range are ordered linearly. Also, their layout should be fuzzy-like. We may e.g. take three such sets stating represented as $\{low, medium, high\}$. As these sets may (and in fact should) overlap, in turn we get more possible linguistic values e.g. $\{low, low\ or\ medium, medium, medium\ or\ high, high\}$.

The retrieval of proper sets for deviation ranges should be devised as an interactive data-driven process. By analysis of standards and classifiers and matching them against the training data we attempt to establish an initial layout for deviations. This layout (the choice and setting of subsets) is then verified and possibly modified in order to achieve high compliance with the underlying data sets. The choice of proper parameters for the sets of deviation ranges may be based on various known techniques in data analysis such as clustering, statistical analysis, density analysis etc.

Conclusion

We have proposed to use standards defined by classifiers. Such standards can next be used in the process of extracting of productions from data and for deriving AR-schemes. This is also a step towards implementation of the general idea of computing with words.

Acknowledgements

The research has been supported by the State Committee for Scientific Research of the Republic of Poland (KBN) research grant 8 T11C 025 19 and by the Wallenberg Foundation.

References

1. Doherty, P., Łukaszewicz, W., Skowron, A., Szałas, A.: *Combining rough and crisp knowledge in deductive databases*, (to appear in [4]).
2. Hastie, T., Tibshirani, R., Friedman, J. H.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer Verlag, Berlin 2001.
3. Mitchell, T.M.: *Machine Learning.* Mc Graw-Hill, Portland, 1997.
4. Pal, S.K., Polkowski, L., Skowron, A. (Eds.): *Rough-Neuro Computing: Techniques for Computing with Words.* Springer-Verlag, Berlin, 2002. (to appear).
5. Pawlak, Z.: *Rough Sets. Theoretical Aspects of Reasoning about Data.* Kluwer Academic Publishers, Dordrecht, 1991.
   Towards an

6. Polkowski, L., Skowron, A.: Rough mereology: a new paradigm for approximate reasoning. *International J. Approximate Reasoning* **15**(4), 1996, 333–365.
7. Skowron, A.: Rough sets in KDD. in: Z. Shi, B. Faltings, and M. Musem (Eds.), *16-th World Computer Congress (IFIP'2000): Proceedings of Conference on Intelligent Information Processing (IIP'2000)*, Publishing House of Electronic Industry, Beijing, 2000, 1–17.
8. Skowron, A.: Toward intelligent systems: Calculi of information granules. *Bulletin of the International Rough Set Society* **5**(1-2), 2001, 9–30.
9. Skowron, A., Grzymała–Busse, J.W.: From rough set theory to evidence theory. In: R.R. Yaeger, M. Fedrizzi, and J. Kacprzyk (eds.), *Advances in the Dempster Shafer Theory of Evidence*, John Wiley & Sons, Inc., New York (1994) 193–236 2001, 57–86.
10. L.A. Zadeh. Fuzzy sets. *Information and Control* **8**, 1965, pp. 333–353.
11. Zadeh, L.A.: Fuzzy logic = computing with words. *IEEE Trans. on Fuzzy Systems* **4**, 1996, 103–111.
    certainty in
12. Zadeh, L.A.: A new direction in AI: Toward a computational theory of perceptions. *AI Magazine* **22**(1), 2001, 73–84.
    1-2.