Towards the definition of a Data Mining process based in Fuzzy Sets for Cooperative Metaheuristic systems

J.M. Cadenas, M.C. Garrido, L.D. Hernández, E. Muñoz
Depto. Ingeniería de la Información y las Comunicaciones
Facultad de Informática. Universidad de Murcia. Spain.
jcadenas@dif.um.es, mgarrido@dif.um.es, ldaniel@dif.um.es, enrique.m.b@alu.um.es

Abstract

The aim of this paper is to apply all the phases of the process of the extraction of knowledge from data to the problem of describing the coordinator of a cooperative metaheuristic system applied to combinatory optimization. Once the system has been described, the most important aspects are analysed in each of the phases of the process, and the result is a set of fuzzy rules to describe the functioning of the system coordinator.

Keywords: Data Mining, Soft-Computing, Fuzzy Sets, Cooperative Metaheuristic Systems.

1 Data Mining and Soft Computing

Data mining is described in [4] as the process of extracting new, useful, comprehensible knowledge from data stored in different formats. In other words, the fundamental task of data mining is to find intelligible models, starting from the data. If the process is to work it should be automatic or semi-automatic and the models discovered should help to take safer decisions which will benefit the system.

There are, therefore, two challenges for data mining: in the first place, it must work with large volumes of data and the problems this entails (noise, missing data, vagueness, etc.); secondly, suitable techniques must be used to analyse the data and to extract new, useful information. In many cases the usefulness of the mined knowledge is closely related to the comprehensibility of the inferred model, so it is important that the information obtained be more comprehensible (through rules, etc.), [7].

In general terms, within the knowledge extraction process by data mining, we can identify the following phases, [4, 7, 16]:

- Preparation of data.
- Data Mining.
- Evaluation and use of models.

In the preparation of data, the data pertaining to our problem must be gathered and the selection, cleaning and transformation of the data is carried out. It is in this phase, then, that the strategy to follow for imperfect data is decided upon. As imperfection will, in general, be present in the data, we will operate within the paradigm of Soft-Computing.

The task to be performed is chosen during the data mining phase, along with the technique to be used. One of the most important aspects in determining the choice of the right technique is the type of model we wish to construct. This model should be able to help us explain the data.

Finally, during the evaluation and use of models phase, the models are evaluated and analysed by experts and, if necessary, we return to previous phases and perform a new iteration. This phase enables us to assess how the different techniques work with our problem and to select that which best adapts to it.

In this paper we apply all the phases described above to the problem of describing the coordinator of a cooperative metaheuristic system applied to combinatory optimization. Section 2 outlines the problem. Section 3 defines a knowledge extraction process based on fuzzy sets for a cooperative metaheuristic system and details all of the phases. Finally, we present our conclusions and propose future research on the actual deployment of the system.

2 Cooperative Metaheuristic Systems

Metaheuristics provide effective strategies for finding approximate solutions to optimization problems, although the execution times associated with the exploration of the search space may be high. Difficulties also exist since the problem is new and it is, therefore, practically impossible to determine which is the best metaheuristic to solve it. Even if we do know of a good metaheuristic, it may be the case that the type of instance we need to solve is complex for the metaheuristic. Furthermore, no practical methodology exists to find the best set of parameters to solve a particular instance. Thus, difficulties exist in finding metaheuristics in the literature to solve optimization problems possessing general characteristics that are independent of the problem, robust, easy to implement and which return acceptable results within a reasonable period of time.

All this leads us to consider the use of other metaheuristics within a single coordinated schema, i.e. a cooperative parallel strategy for solving combinatory optimization problems. Robustness will be obtained by using different combinations of metaheuristics which intercooperate to give high quality solutions for different types of instances. This cooperative execution can be carried out, among other ways, using a coordinating agent to control and modify the behaviour of the agents. The coordinator will therefore have two fundamental tasks: to gather information on the performance of each of the metaheuristics and to send orders to them to modify their search behaviour, [5].

We are going to work with this type of cooperative system in which the problem arises of describing the coordinator in such a way that it can control and modify the behaviour of the metaheuristics efficiently and lessen the problems outlined earlier regarding the behaviour of the metaheuristics at an individual level.

The problem we have outlined is the one which we seek to solve in this study through the use of data mining, and which we describe in the following section.

3 A Fuzzy Set based Data Mining Process for Cooperative Metaheuristic Systems

In this section we carry out the preliminary study of the process we wish to perform in order to obtain the description of the coordinator for the system outlined above. This description will be made using a set of fuzzy rules obtained through a data mining process.

Figure 1 illustrates the complete data mining process aimed at obtaining the set of rules for modelling the behaviour of the system coordinator.

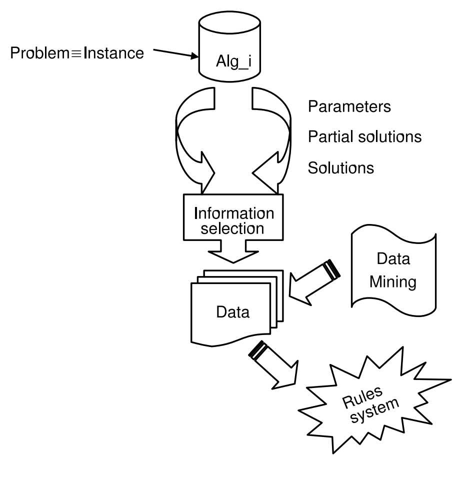

Figure 1: Data Mining process

The different phases below are represented in the figure:

- Preparation of data. As starting point we take a set of data extracted experimentally on solving sets of instances of a problem or problems using a set of metaheuristics.
- Data mining. This applies a set of inductive techniques which give a set of fuzzy rules as system model to the data obtained in the previous phase.
- Evaluation. Although this does not explicitly appear in the figure, it would evaluate the behaviour of the coordinator using the rules previously obtained and would repeat the whole process using the data and adapting them to the new knowledge about the problem.

Below we show each of the phases separately and we indicate all the necessary decisions to create a prototype process to solve the problem in question.

3.1 Preparation of Data

During the data preparation phase we seek to obtain a database of information which is of interest to us and one to which we can apply such data mining techniques as interest us. The first step is to gather information. Our sources of information come from the application of various metaheuristics to sets of instances of the problem or problems. As information of interest we can find:

- the specific parameters used in the execution of metaheuristics,
- a description of the specific instance of a problem,
- the final solution obtained and its cost,
- information obtained during the execution of an algorithm.

Figure 2 shows the complete process to obtain information.

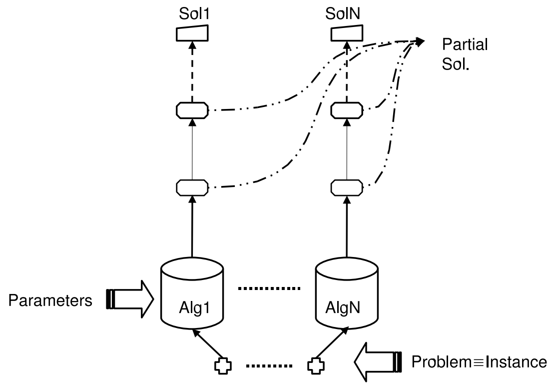

Figure 2: Massive extraction of information

Using this information we can create the initial structure from which an example for us will be, formed by a vector defined according to a series of characteristics. In order to work with these data we need to define a data structure that can support all types of information: heterogeneous, numeric and nominal characteristics which may present imperfect values (uncertainty and imprecision) as well as allowing for the case of a value of a characteristic of a specific example being unknown, [1].

Once the data are stored in the structure, we apply characteristic selection techniques to establish which are the most relevant for the aim sought. We tackle this data pre-processing stage using techniques like Bayesian networks [9], fuzzy integral [8] and heuristic methods based on "Fuzzy Pattern Matching" [3], as well as others.

When this step is applied to the set of data we obtain the data base to which we will apply the data mining techniques. Figure 1 shows how, from any initial information we have, after applying the step we have called information selection, created the set of data to which the data mining techniques can be applied.

3.2 Data mining techniques

The knowledge extracted from the data may be in the form of relations, patterns, rules or even a more concise description. These relations constitute the model of the analysed data. Many ways of representing models exist and each way determines the type of technique that can be used to infer the models.

There is a wealth of data mining techniques based on different theoretical proposals. However, most model construction techniques for inference tasks have paid relatively scant attention to sources of uncertainty and incomplete and imprecise data have been either discarded or ignored, both for learning processes and for later inference.

Since imperfect observations inevitably appear in realistic domains and situations, and since we do not know the starting data with which we will be working, we will focus on those techniques which currently allow greater and fuller treatment of imperfect information, i.e. those which include handling of observations with heterogeneous attributes (both numeric and nominal) and which may, moreover, present uncertainty and imprecision in both the learning phase and in inference.

Efforts have been made to incorporate uncertainty and imprecision into the learning and inference phases of well known data mining processes. Although not exhaustively, some techniques allow handling of examples with some unknown attribute, as in the case of techniques based on density estimation from sample patterns [16], regression/classification tree construction [6, 14], the multivariant normal classifier [12], mixture models [15], etc. Tree construction techniques and mixture models also allow observations with nominal attributes which present uncertain values from the probabilistic viewpoint, as well as continuous attributes expressed through crisp intervals (crisp imprecision) [14, 15]. Also incorporated is the handling of attributes expressed through fuzzy values in tree construction techniques (although only for classification) [10], as well as in neuronal networks [13] and in mixture models [2].

In general, we can state that there are a number of limitations on imperfection allowed in the initial observations. Current techniques, which perform a more complete handling of imperfect information are classification trees and mixture models.

Elsewhere, when we seek to obtain a model for the important dependencies between the known and the unknown attributes in order to perform inference tasks, most techniques lead to models which capture partial dependencies (usually for one attribute with respect to the others). This occurs, for example, when using techniques based on tree construction, neuronal networks, multivariant normal classifier, etc., where a model must be learnt for each attribute to be inferred. It is more interesting to obtain models which capture global dependencies (among all the attributes) since these allow inference of the values of any subset of attributes from any other subset of attributes of known values, and using the same model. This is the case when using techniques based on Bayesian networks, mixture models, etc.

Thus, given the characteristics described above and the fact that these impose fewer limits, we propose to tackle the data mining phase to describe the cooperative system with a mixture model, known as EMFGN [2], with Bayesian networks and a tree construction technique.

The use of EMFGN and Bayesian techniques, which model the global dependence of the attributes used to describe the system, means that we can determine the dependence of any attribute, or any group of attributes, from any other subset of attributes. Thus, at any given moment, we can modify the trajectory of an algorithm, given the values which another algorithm that is achieving better results is working with at the same moment.

Since we wish to modify the system by using a set of fuzzy rules, these will be obtained through two of the techniques selected (tree construction and mixture models). We will use the techniques based on Bayesian networks to add knowledge to those rules.

These rules will be those used to describe the behaviour of the cooperative system and will be of the type IF < antecedent> THEN <consequent>. We will propose, among others, rules to model the following behaviour:

- Rules which will indicate to us which algorithm should be reinitiated for a solution which seems to be the best at the present moment.
- Rules which will indicate to us to what extent the behaviour of an algorithm should be modified with respect to its search for the solution.
- Rules which will indicate to us which algorithm should be reinitiated with different parameters.
- Rules which will indicate to us when we should stop, either because the available resources have been exceeded or because a solution which is satisfactory for the user has been found.
- Rules which will indicate to us which rule should be selected in those cases when more than one of the system rules can be applied.

3.3 Evaluation of the model obtained

In this phase we seek to check the efficiency of the set of fuzzy rules for modelling the coordinator, and also that the latter is performing the task efficiently with respect to the computational cost and the goodness of the solutions obtained. We selected as pilot problem the p-median one and we make a set of experiments for instances of different difficulties which are known and measurable. Taking these instances as starting point, and once the coordinator has been modelled with the set of rules obtained in the previous phase, we must ascertain if its behaviour is suitable or decide whether it needs to be adapted (system feedback) and improved.

In order to measure the behaviour of the system (with respect to the costs of the solutions obtained) we will use several measures, such as the mean quadratic error and the relative quadratic error. To perform the experiments we will use crossed and bootstrap validation methods. We will also use the information provided by the system regarding the rules which have been enabled, which metaheuristics have been modified in their execution, which metaheuristics do not seem to present suitable behaviour with respect to the others and to the solution they give, etc.

Once all the tests have been made, the evaluation methods have been calculated and the necessary information on the executions has been obtained, we can tackle the problem of how to modify (feedback) the system. We can return to the process of knowledge extraction (remove/add characteristics, remove/add information corresponding to a metaheuristic, change of data mining method, etc.) in order to improve the behaviour of the coordinator by using this information. Figure 3 show a schema the feedback process.

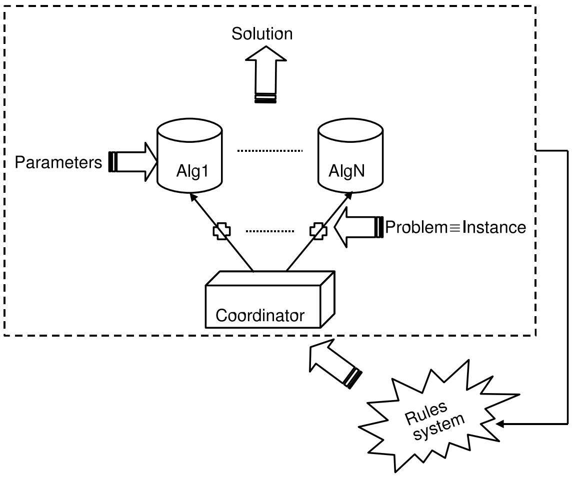

Figure 3: System feedback

An algorithmic approach to this feedback process which could improve or adapt the set of fuzzy rules which model the coordinator could be the following:

- If the error obtained from the evaluation is good, then the system of rules modelling the coordinator seems suitable. STOP

- Otherwise,
  - If the additional information indicates that an algorithm exists which does not provide anything to the system, or is redundant, then it is removed from the system. Any information provided by that algorithm should be deleted from the data base.
  - If the additional information indicates that the behaviour of an algorithm within the coordinated system has worsened with respect to its individual performance, then remove/add characteristics relative to the algorithm.
  - If the additional information indicates that the behaviour of an algorithm within the coordinated system has worsened with respect to its individual performance on a set of instances, then remove/add descriptive characteristics of those instances.
  - In any other case, we should add more metaheuristics to the cooperative system, increase the volume of starting data or change the data mining technique for modelling the problem.
- Apply again all the phases of the knowledge extraction process.

4 Conclusions and works by developing

Since it seems to be more efficient to use a coordinated system of metaheuristics with a centralized focus for solving complex problems, we have proposed here the process to be followed to apply the data mining paradigm in the definition of this system's coordinator.

In each phase of the process, we have presented the focus with which we intend to tackle the associated problematics. We have proposed the general work schema with which we will obtain the model of the coordinator's behaviour when faced with a specific problem.

Starting from this general schema, we propose, as studies to be developed, the construction and implanting of a prototype which will carry out each of the phases in a specific way and using as starting point the set of data obtained from a set of problems and their corresponding instances.

Acknowledgements

The authors thank the "Ministerio de Educación y Ciencia" of Spain and the "Fondo Europeo de Desarrollo Regional" (FEDER) for the support given to this work under the project TIN2005-08404-C04-02.

References

[1] MC. Garrido, J.M. Cadenas, A. Ruiz (1998). Una estructura de almacenamiento en memoria para datos imperfectos. *Documento Técnico Dpto. Informática y Sistemas*, TR DIS 3/98, pages 1-11, Universidad de Murcia, Spain, 1998.

[2] J.M. Cadenas, M.C. Garrido (2002), Imperfección Explícita versus adaptación de la información en MFGN extendido. In *Proceedings XI Congreso Español sobre Tecnologías y Lógica Fuzzy*, pages 547-552, León, Spain, 2002.

[3] J.M. Cadenas, M.C. Garrido, J.J. Hernández (2004). Improving fuzzy pattern matching techniques to deal with non discrimination ability features. In *Proceedings of the IEEE Internacional Conference on System, Man and Cybernetics*, Netherlands, 2004.

[4] P. Clark, R. Boswell (2000). Data mining. Practical machine learning tools and techniques with java implementations. *Morgan Kaufmann Publishers*, 2000.

[5] C.A. Cruz (2005). Estratégias coordinadas paralelas basadas en soft-computing para la solución de problemas

de optimización. *Tesis doctoral del Dpto. de Ciencias de la Computación e Inteligencia Artificial*, Universidad de Granada, 2005.

[6] R.O. Duda, P.E. Hart (1992). Pattern classification and scene analysis. *Wiley*, 1992.

[7] J. Hernández, M.J. Ramírez, C. Ferri (2004). Introducción a la Minería de Datos. *Pearson-Prentice Hall*, 2004.

[8] M. Grabisch (2000). Fuzzy integral for classification and feature extraction. In *M. Grabisch, et al., eds., Fuzzy Measures and Integrals: Theory and Applications*, pages 415-34. Physica-Verlag, New York, 2000.

[9] I. Inza, P. Larrañaga, B. Sierra (2002). Feature Subset Selection By Estimation of Distribution Algorithms. In *P. Larrañaga et al., eds, Genetic Algorithms and Evolutionary Computation*, chapter 13, pages 269-293, Kluwer Academic Publishers, 2002.

[10] C.Z. Janikow (1998). Fuzzy Decision Trees: Issues and Methods. *IEEE Transactions of Man, Systems, Cybernetics*, volume 28, pages 1-14, 1998.

[11] Y. Kodratoff, R. Michalski (1990). Machine Learning. *Morgan Kaufmann Publishers, Inc.*, volume 3, 1990.

[12] G.L. McLachlan, T. Krishnan (1997). The EM Algorithm and Extensions. *Wiley Series in Probability and Statistics*, 1997.

[13] S. Mitra, S.K. Pal (1995). Fuzzy Multi-Layer Perceptron, Inferencing and Rule Generation. *IEEE Transactions on Neural Networks*, volume 6, pages 51-63, 1995.

[14] J.R. Quinlan (1993). C4.5: Programs for machine learning. *The Morgan Kaufmann Series in Machine Learning*. California, 1993.

[15] A. Ruiz, P.E. López de Teruel, M.C. Garrido (1998). Probabilistic Inference from Arbitrary Uncertainty using Mixtures of Factorized Generalized Gaussians. *Journal of Artificial Intelligent Research*, volume 9, pages 167-217, 1998.

[16] I.H. Witten, E. Frank (2000). Data Mining. *Morgan Kaufmann Publishers*, San Francisco, 2000.