Applying a Fuzzy Decision Tree Approach to Soil Classification

Mariana V. Ribeiro[1], Luiz Manoel S. Cunha[2], Heloisa A. Camargo[1], and Luiz Henrique A. Rodrigues[2]

[1] Computing Department of the Federal University of São Carlos
São Carlos, São Paulo, Brazil
{mariana.ribeiro,heloisa}@dc.ufscar.br
http://cig.dc.ufscar.br

[2] College of Agricultural Engineering of the State University of Campinas
Campinas, São Paulo, Brazil
lique@feagri.unicamp.br, luiz.cunha@embrapa.br
http://www.feagri.unicamp.br

Abstract. As one of the most important factors that interfere in peoples life, the soil is characterized by quantitative and qualitative features which describe not only the soil itself, but also the environment, the weather and the vegetation around it. Different types of soil can be identified by means of these features. A good soil classification is very important to get a better use of the soil. Soil classification, when performed manually by experts, is not a simple task, as long as the experts opinions may vary considerably. Besides, different types of soil cannot be defined deterministically. With the objective of exploring an alternative approach towards solving this problem, we investigated in this paper the application of an automatic procedure to generate a soil classifier from data, using a fuzzy decision tree induction algorithm. In order to compare the results obtained by means of the fuzzy decision tree classifier, we used two well known methods for classifiers generation: the classic decision tree induction algorithm C4.5 and the fuzzy rules induction algorithm named FURIA.

Keywords: fuzzy rule based systems, decision tree, fuzzy decision tree, classification, soil classification, soil classification system.

1 Introduction

Due to its use to food cultivation, the soil is one of the most important factors that interfere in people's life, since good food requires good soil. To take advantage of all it's best qualities, not only in food branch, it is very important to know the characteristics of the soil present in each site [20]. Motivated by this, some different classes of soils have been created according to their characteristics. The soil characteristics are related to quantitative and qualitative features that describe the soil, the environment, the weather and the vegetation around them. By knowing its main characteristics, the class that the soil belongs to is

A. Laurent et al. (Eds.): IPMU 2014, Part I, CCIS 442, pp. 87–96, 2014.

also known and then, it is possible to make the best use of it. Although this classification is very useful, it involves very subjective criteria and as it is usually done by experts, it depends very much on the experts opinion. Aiming to support the experts task and reduce the influence of subjectiveness in the classification process, some classification systems can be constructed automatically from data.

Fuzzy decision trees combine the advantages of decision trees, such as the embedded feature selection and low computational cost, with the ability of processing uncertainty and imprecision of fuzzy systems. Some fuzzy decision trees algorithms have been proposed in the literature [2–7]. In this work, we use the fuzzy decision tree induction algorithm (FuzzyDT) described in [2], which is an algorithm based on the well known C4.5 algorithm, to generate fuzzy rules. FuzzyDT starts with the fuzzyfication of the continuous features before inducing the fuzzy decision tree. This algorithm has shown good results in a previous work, when it was applied to a real-world problem, the prediction and control of the coffee rust disease in Brazilian crops [8]. In the work presented here, we investigate the generation of a classification system to deal with the problem of soil classification using FuzzyDT. We also compare the results with the ones obtained by the classic C4.5 algorithm [9] and FURIA algorithm, proposed in [10]. We evaluated them by comparing their accuracy, measured by the correct classification rate, and interpretability, measured by the format and the number of rules generated by each algorithm.

The paper is organized as follows: in the next section, we describe briefly the soil classification problem. In section 3, a short description of decision trees and the main concepts of C4.5 are presented and we describe a general view of the fuzzy classification systems and of the FuzzyDT and FURIA algorithms. The experiments and analyses are presented in section 4. The final conclusions are discussed in section 5.

2 Soil Classification

The soil is very important to the human beings and was defined in [1] as "a collection of solid, liquid and gas parts, which could be three-dimensional, moving, formed by minerals and organic materials that occupy most of the surface of the continental extensions of our planet". In Brazil, the soil classification is governed by the Brazilian System of Soil Classification (SiBCS) [1], a hierarchical and multi categorical system, which is open to improvements and expansions.

According to Oliveira [11], researcher and member of the Executive Committee of the Brazilian System of Soil Classification, classifying the soil is very important because it allows:

a) to understand the relation between the individuals
b) to recall the properties of the classified objects
c) to predict the individuals' behavior
d) to improve the use of the soil in a place

e) to estimate the productivity of a stand
f) to provide research themes
g) to explore data from research or observations
h) to facilitate the communication

Actually, the soil classification is extremely important when used to identify the occurrence of different soil in the environment, as in the soil maps.

Nowadays, the SiBCS is constituted by six categorical levels:

1) Order
2) Suborder
3) Large group
4) Subgroup
5) Family
6) Series

So far, the 5th and 6th levels are not organized yet. The attributes which were used in the organization of each level are soils characteristics identified in the research or inferred from other attributes or previous knowledge from soil science. In each categorical level, a set of classes is defined by one or more rules. In this work, we approach specifically the classes Brown Latosol and Brown Nitosol. Brown (redish yellow colors) is a suborder of Latosol and Nitosol orders. Evaluating the soil as Brown Latosol and Brown Nitosol is a crucial problem to the research community because, with the development of the soil science, the understanding of the main diagnostic attributes is under discussion. Diagnostic attributes are characteristics or properties that are used to divide the soil by classification system's levels. Some issues arise in cases where it is difficult to distinguish the soil's characteristics or when it presents conceptual overlap, which hampers the characterization, separation and classification of the soils [12]. The suborder of brown soils has some peculiarities, which demands new investigations that provide a better differentiation among them.

The soil classification task performed by experts started with pedological studies, a practical activity, where over a hundred characteristics' data were collected. These characteristics are defined by quantitative and qualitative data which describe the soil, the environment, the weather and the vegetation around and are used to soil classification. Furthermore, some data were obtained from laboratory analyses done on the collected data, and some other derived from the previous ones.These features were all added to the database in order to complete the set of features that will be used in the classification. Then, these data are discussed by experts which classify the samples based on a predefined pattern of each class and the current soil classification system.

3 Classic and Fuzzy Classification Systems

Nowadays, it is very common to deal with a lot of data which are often available on open sources. However, analyzing these data and extracting useful information from them is not an easy task for humans.

In order to solve this problem, some methods of inductive learning have been developed. Among the most used inductive learning methods are the algorithms that generate classifiers. They consist in, given a set of examples, each one described by a set of attributes and a class (or label), learning from these examples and representing the extracted knowledge in a model that is capable of classifying new examples of unknown classes.

A fuzzy system is a system that includes at least one linguistic variable, whose values are linguistic terms represented by fuzzy sets [18].

A very popular and useful type of fuzzy systems are the rule-based fuzzy systems (RBFS), which have a knowledge base, formed by a fuzzy data base and a rule base and an inference mechanism, which processes the rules in the rule base using a reasoning method.

Generally speaking, a classification problem is the problem of assigning a given input data to one of a set of pre-determined set of classes. Rule-based fuzzy classification systems (RBFCS) are a type of RBFS which deals with fuzzy classification problems. After the rules have been constructed, they can be used to classify new instances by applying an inference mechanism such as the ones proposed in [17]. The rules of a RBFCS with n attributes and m classes have the form:

$$\textbf{IF } X_1 \textbf{ is } A_1 \textbf{ AND } X_2 \textbf{ is } A_2 \ldots \textbf{ AND } X_n \textbf{ is } A_n \textbf{ THEN Class is } C_j$$

Where X_i represents the attributes of the set of examples, A_i are the attribute values represented by linguistic terms and C_j is one of the classes in the set of classes $\{C_1, C_2, \ldots C_m\}$.

In the following we describe briefly the learning algorithms used in this work, namely the classic C4.5 algorithm, the FuzzyDT algorithm and the FURIA algorithm.

3.1 C4.5 Algorithm

C4.5 is one of the most popular algorithms of decision trees induction. It was proposed by Quinlan [9] and uses entropy and information gain measures to find the most informative attributes for each new split.

The information gain of an attribute is defined as the information that is provided to classification by splitting a set of examples, based on that attribute. It corresponds to its entropy reduction. Higher information gains implies more homogeneous subsets in term of class after splitting. According to Shannon [19], the entropy of a set S containing k possible classes is defined as:

$$E(S) = -\sum_{j=1}^{k} \frac{freq\,(C_j, S)}{|S|} \cdot log_2 \left(\frac{freq\,(C_j, S)}{|S|} \right)$$

Where $freq(C_j, S)$ represents the number of examples in S that belongs to class C_j and $|S|$ is the number of examples in S.

The entropy shows the average amount of information necessary to classify an example in S.

After splitting S into n subsets $S_i (i = 1, ...n)$ by a node test with attribute X (which is the attribute that provided the highest information gain) the information gain InfGain is given by S's entropy reduction [19]:

$$InfGain(X) = E(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} E(S_i)$$

Once the tree is induced, C4.5 performs a post pruning method, which is based on the estimation of the real error of the model, according to its apparent error, aiming to generalize the final model and avoid overfiting.

3.2 FuzzyDT

FuzzyDT is an algorithm to induce fuzzy decision trees based on the classic C4.5.The first steps are the definition of fuzzy partitions in the continuous attributes domains and the fuzzification of the attribute values. After that, the tree induction process is applied to generate the fuzzy rules. Algorithm 3.2.1 presents the main steps of FuzzyDT [2].

Algorithm 3.2.1. The FUZZYDT algorithm [2]

1. Define the fuzzy data base, i.e., the fuzzy granulation for the domains of the continuous features;
2. Replace the continuous attributes of the training set using the linguistic labels of the fuzzy sets with highest compatibility with the input values;
3. Calculate the entropy and information gain of each feature to split the training set and define the test nodes of the tree until all features are used or all training examples are classified with the same class label;
4. Apply a pruning process.

3.3 Fuzzy Unordered Rule Induction Algorithm: FURIA

FURIA, the Fuzzy Unordered Rule Induction Algorithm, was proposed in [10] as a modification and extension of the famous RIPPER algorithm.

The algorithm considers that a rule covers an example $x = (x_1, x_2, ..., x_n)$ if, and only if, the value of the attribute x_i satisfy all predicates of the rules antecedent. Then, it orders the training examples according to the relative frequency of classes, from the least to the most frequent class. So, it learns rules for all classes, except for the last, which is the most frequent one. Once a rule is created, the examples covered by it are removed from the set of training examples.

The algorithm proceeds with the next class until there are no more examples in the training set or the last created rule is too much complex, according to a predefined measure.

Finally, RIPPER builds a default rule to the last class, which is the most frequent one. Intuitively, creating a default rule could be questionable, since it can privilege the most frequent class. One of the changes to this algorithm, that originated FURIA, is concerned with this default rule.

The main difference between FURIA and RIPPER is that FURIA infers fuzzy rules instead of crisp rules. Moreover, it does not order the training examples to infer the rules. Consequently, FURIA does not build a default rule, using a one-vs-rest decomposition to infer unordered rules.

When using an unordered rule set without default rule to classify a new instance, two problems can occur: First, a conflict may occur when the instance is equally well covered by rules from different classes. Second, it may happen that the instance is not covered by any rule. The first problem is rather unlikely to occur and, in case it still does, it is resolved by calculating the support of the rules and classifying the new instance as the class that occurs in the consequent of the rule which has higher support value. The second one is not so simple to resolve. For this, in [10], Cohen proposes a rule stretching method. The idea is to modify the rules in a local way so as to make them applicable to the instance that is been classified. It is done by replacing the rules by their minimum generalizations for the given instance. As proposed by [10], a generalization or stretching of a rule is obtained by deleting one or more of its antecedents, and it is minimal if it does not delete more antecedents than necessary to cover the instance. Thus, the minimal generalization of a rule is simply obtained by deleting all antecedents that are not satisfied by the instance.

Once all minimal generalizations are derived, FURIA re-evaluates each rule by its Laplace accuracy on the training data and then classify the instance by the rule with the highest evaluation.

4 Experiments

In this section we present the experiments developed, aiming to determine which of the three methods cited above (FuzzyDT, C4.5 or FURIA) gives better results for the soil classification problem.

The tests were performed using a real data set which instances were extracted from Brazilian System of Soil Information [11] and from researches on soil profiling, assigned by the Brazilian Institute of Geography and Statistics (IBGE) from Santa Catarina and by the Center of Agroveterinary Science of State University of Santa Catarina (CAV-UDESC).

To obtain the data set, a filter was applied to extract observations which follow the characteristics below:

- Altitude: upper than or equals to 600 meters;
- Soil's classes: Brown Latosol, Red Latosol , Red-yellow Latosol, Yellow Latosol, Brown Nitosol, Haplic Nitosol, Humic Cambisol, Haplic Cambisol;

- Location: from Paraná, Santa Catarina and Rio Grande do Sul states;
- Profile's sub-horizon B with largest amount of data.

While selecting the data, some attributes were ignored because they are not used neither to soils characterization nor classification.

Since there was only a small amount of observations from classes Red Latosol, Red-yellow Latosol, Yellow Latosol, Haplic Nitosol, Humic Cambisol and Haplic Cambisol, they were grouped into a single class which were named Other Latosols, Nitosols and Cambisols (OLNC).

After the preprocessing, the remaning data included instances from three possible classes:

(a) Brown Nitosol (BN)
(b) Brown Latosol (LB)
(c) Other Latosols, Nitosols e Cambisols (OLNC)

The characteristics of these soils were expressed by 25 attributes,described in Table 1 by means of the name, the type (discrete or continuous), the number of values in the case which the attribute is discrete and a brief descrition of each attribute.

The tests were carried out by using 10-fold Cross Validation. For C4.5 and FURIA algorithms, it was used the implementation of these algorithms available in the software WEKA [16] and for FuzzyDT, our own Java implementation. The parameters of the algorithms C4.5 and FURIA were maintained as the default ones and FuzzyDTs data fuzzyfication was done using partitions with three triangular fuzzy sets per attribute. The results, comprising the accuracy and number of rules generated by each method, are shown in Table 2. The rules format generated by each one of the algorithms are illustrated by the examples presented in Table 3.

As can be seen in Table 2, FuzzyDT obtains the best result in terms of accuracy, followed by FURIA and then C4.5. Concerning the number of rules, FuzzyDT generates the worse result, with a higher number than the other two methods. Although FURIA gives the lowest number of rules, the rules format do not favor comprehensibility of the system as a whole. While in the rules generated by C4.5 and Fuzzy DT it is possible to clearly identify both, the attribute which is been tested and its partition, with FURIA this recognition is not so simple to be done. This is mainly because the attribute values are represented by the parameters of trapezoidal membership function of its fuzzy sets. Besides that, analyzing the rule base constructed by FURIA, we realize that a different partition is generated for each attribute in each rule. This way, the fuzzy sets generated by the algorithm do not have a semantic meaning shared by all rules and the interpretability of the system is deteriorated.

Table 1. Attribute's characteristics

Attribute	**Discrete/Continuous**		**Description**
l_texture	Discrete	3	Level of texture of the soil.
l_structure	Discrete	3	Level of structure of the soil.
s_structure	Discrete	3	Structures size of the soil.
sh_structure	Discrete	5	Structures shape of the soil.
consistency_moist	Discrete	3	Level of consistency of the moist soil.
l_plasticity	Discrete	3	Level of plasticity of the soil.
l_tackiness	Discrete	3	Level of tackiness of the soil.
waxy	Discrete	2	Presence or absence of waxy and shiny appearance.
l_waxy	Discrete	4	Waxys level of the soil.
q_waxy	Discrete	4	Quantity of waxy of the soil.
l_distinctness	Discrete	4	Level of distinctness of the soil.
horizon_A	Discrete	4	Type of horizon A.
source_material	Discrete	7	Source material of the soil.
clay	Continuous		Clay content of the soil.
cxc_clay	Continuous		Cation exchange capacity of the clay.
fine_sand	Continuous		Fine sand content of the soil.
grit	Continuous		Grit content of the soil.
total_sand	Continuous		Total sand content of the soil.
sulfuric_attack_SiO_2	Continuous		Si by sulfuric acid attack expressed by SiO_2.
sulfuric_attack_Al_2O_3	Continuous		Al by sulfuric acid attack expressed by Al_2O_3.
carbon_nitrogen	Continuous		Carbon/Nitrogen.
Fe_2O_3_clay	Continuous		Fe_2O_3/Clay content.
Al_2O_3_clay	Continuous		Al_2O_3/Clay content.
SiO_2_clay	Continuous		SiO_2/Clay content.
Ki_clay	Continuous		Ki/Clay content.

Table 2. Tests' results to C4.5, FURIA and FuzzyDT algorithms

C4.5		**FURIA**		**FuzzyDT**	
Accuracy	# of Rules	Accuracy	# of Rules	Accuracy	# of Rules
82.85	50	83.99	20	92.99	104

Table 3. Example of rules format generated by the three methods

Method	**Example of Rule**
C4.5	grit ≤ 120 **and** SiO_2_clay ≤ 0.088222: BN
FURIA	(source_material = Sao_Bento) **and** (sulfuric_attack_Al_2O_3 **in** $[-\infty, -\infty, 227, 232]$) $\geq$ classe=OLNC
FuzzyDT	**IF** source_material **IS** 4 **AND** grit **IS** low **THEN** CLASS **IS** 3

5 Conclusion

The best use of the different types of soil depends on a proper classification. This is not an easy task since it implies very subjective expert opinions. Aiming at solving this problem, we proposed and tested the use of a fuzzy decision tree approach, named FuzzyDT to build a fuzzy classification system which deals with the problem of soil classification. We compared the generated FuzzyDT with two other classification systems, obtained from the algorithms C4.5 and FURIA. Analysing the results, it is possible to observe that FuzzyDT reaches the highest accuracy but generates the highest number of rules. In spite of that,, its rules are interpretable, following the format of standard fuzzy rules. FURIA and C4.5 obtained very similar results with respect to accuracy while FURIA generates the lower number of rules. Nevertheless, it generates rules that are not interpretable, once for each rule, a different partition for each attribute is generated, which implies that the fuzzy sets generated by the algorithm are not interpretable.

In the future work we intend to investigate techniques to be applied on the set of fuzzy rules generated by FuzzyDT, to reduce the number of rules, while still preserving the good accuracy obtained.

References

1. Santos, H., et al.: Brazilian System of Soil Classification. Embrapa Solos, Rio de Janeiro, RJ, Brazil (2013) (in Portuguese)
2. Cintra, M.E., Monard, M.C., Camargo, H.A.: A Fuzzy Decision Tree Algorithm Based on C4.5. Mathware & Soft Computing 20, 56–62 (2013)
3. Chang, R.L.P., Pavlidis, T.: Fuzzy decision tree algorithms. IEEE Transactions on Systems Man and Cybernetics 7, 28–35 (1977)
4. Janikow, C.Z.: Fid4.1: an overview. In: Proceedings of NAFIPS 2004, pp. 877–881 (2004)
5. Kazunor, U., Motohide, S.: Fuzzy C4.5 for generating fuzzy decision trees and its improvement. Faji Shisutemu Shinpojiumu Koen Ronbunshu 15, 515–518 (2001) (in Japanese)
6. Olaru, C., Wehenkel, L.: A complete fuzzy decision tree technique. Fuzzy Sets and Systems 138(2), 221–254 (2003)
7. Tokumaru, M., Muranaka, N.: Kansei impression analysis using fuzzy c4.5 decision tree. In: International Conference on Kansei Engineering and Emotion Research (2010)
8. Cintra, M.E., Meira, C.A.A., Monard, M.C., Camargo, H.A., Rodrigues, L.H.A.: The use of fuzzy decision trees for coffee rust warning in Brazilian crop. In: 11th International Conference on Intelligent Systems Design and Applications, vol. 1, pp. 1347–1352 (2011)
9. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)
10. Huhn, J., Hullermeier, E.: Furia: An algorithm for unordered fuzzy rule induction. Data Mining and Knowledge Discovery (2009)

11. Oliveira, J.B.: Soil classification and its agricultural and nonagricultural use (2013) (in Portuguese), http://jararaca.ufsm.br/websites/dalmolin/download/textospl/classif.pdf (accessed in December 30, 2013)
12. Almeida, J.: Brief History of the Subtropical Bruno Soils in Brazil. VIII RCC Brazilian Meeting of Soil Correlation and Classification. Embrapa Solos (2008) (in Portuguese)
13. Cohen, W.W.: Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 115–123. Morgan Kaufmann (1995)
14. Furnkranz, J., Widmer, G.: Incremental Reduced Error Pruning. In: Proceedings of the 11th International Conference on Machine Learning (ML 1994), pp. 70–77. Morgan Kaufmann, New Brunswick (1994)
15. Furnkranz, J.: Separate-and-conquer rule learning. Artificial Intelligence Review 13(1), 3–54 (1999)
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
17. Cordón, O., del Jesús, M.J., Herrera, F.: A proposal on reasoning methods in fuzzy rule-based classification systems. Int. J. Approx. Reasoning 20(1), 21–45 (1999)
18. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
19. Shannon, C.: A mathematical theory of communication. Bell System Technical Journal 27 (1948)
20. Martinez, C.V.O., Souza, V.F.: Importance of the soils classification in the Brazilian system and the ability of land use of rural properties for their sustainable management. In: International Meeting of Cientific Production Cesumar (2009) (in Portuguese)