Thematic Image Segmentation by a Concept Formation Algorithm

J.J. Korczak[*] **D. Blamont**[†] **& A. Ketterlin**[‡]
Laboratoire des Sciences de l'Image, d'Informatique et de Télédétection, URA 1871, CNRS
Université Louis Pasteur
7, rue René Descartes, 67084 Strasbourg Cedex, France

Abstract

Unsupervised empirical machine learning algorithms aim at discovering useful concepts in a stream of unclassified data. Since image segmentation is a particular instance of the problem addressed by these methods, one of these algorithms has been employed to automatically segment remote-sensing images. The region under study is Nepalese Himalayas. Because of important variations in altitude, effects of lighting conditions are multiplied, and the image becomes a very complex object. The behavior of the clustering algorithm is studied on such data. Because of the hierarchical organization of the resulting classes, the segmentation produced may be interpreted in a variety of thematic mappings, depending on the desired level of detail. Experimental results prove the influence of lighting conditions, but also demonstrate very good accuracy on sectors of the image where lighting is almost homogenous.

1 Introduction

Research on Machine Learning is a sub-field of Artificial Intelligence which investigates the various ways to automatically acquire efficient and reusable knowledge. Beside strong reasoning abilities, any system needs a wide body of knowledge: machine learning techniques are aimed to resolve the knowledge acquisition problem. Among these techniques, inductive learning techniques provide algorithms to get the most out of data. Unsupervised techniques ("learning from observations") provide ways of learning without the help of conceptual knowledge about the domain under study.

Cartography by remote sensing is a way to produce knowledge about the area under study. Since the speech does not account for the complexity of a geographical system, a map is the usual way of representing simultaneously many aspects of the system. However, the automatic analysis of remote sensing images does not only produce an instantaneous map, it also raises questions about the region, by identifying and characterizing the objects of the study. Since this identification is much more "objective" than the one built by a geographer, it also often suggests new directions of investigation[5].

In this paper, thematic mapping is viewed as an unsupervised empirical learning problem, where the system, presented with a remote sensing image, must derive coherent and useful concepts. The first section describes a concept formation algorithm. Next, this algorithm is applied to a remote sensing image of Nepalese Himalayas, and the results analysed from a thematic point of view. Finally, some conclusions are drawn about these results, and the use of unsupervised methods for cartography by remote sensing.

[*]e-mail: jjk@dpt-info.u-strasbg.fr
[†]e-mail: denis@lune.u-strasbg.fr
[‡]e-mail: alain@dpt-info.u-strasbg.fr

2 A Concept Formation Algorithm

Concept formation has been shown to be one of the most basic capabilities in intelligent behavior. Research on conceptual clustering investigates the principles that govern the process of building concept hierarchies from observed data[6,10]. Given a set of objects, conceptual clustering algorithms determine — intensionaly defined — clusters (or concepts) that cover these objects, along with a hierarchical organization of these clusters. In the context of image analysis, this process is called hierarchical segmentation.

The Classit algorithm[3,6] addresses this task, with the additional property of working incrementally, i.e. it accepts instances one at a time and incorporates them in its memory structure. Incremental conceptual clustering will be referred to as *concept formation*[4]. To maintain its memory-structure, Classit performs a hill-climbing search through a space of potential hierarchies. Bi-directional operators are provided to reduce local optima sensitivity.

As all learning algorithms, Classit aims to improve some kind of performance task. Because of its unsupervised nature, the primary performance task is to predict missing values for partially described instances. This implies that hierarchies produced by the algorithm can be evaluated in terms of their predictive ability.

2.1 Representation of Knowledge

Classit works with a propositionnal formalism, i.e. instances are represented as a set of attribute-value pairs. Attributes may be of nominal type (i.e. they have a finite, discrete set of potential values) or of continuous type (which will be of interest here). Multiple values for the same attribute are not allowed, but some attribute values of an instance description may be missing.

The representation of a concept must sum up the class (the set of instances) it covers. Thus, it must contain a set of attribute-distribution pairs, along with its probability of occurrence according to its parent concept. A distribution may take two forms according to the kind of attribute it corresponds to. In the case of a nominal attribute, all possibles values are stored together with their frequency of occurrence. In the case of a continuous attribute, values are supposed to be normally distributed, and thus only the mean value and standard deviation must be kept.

Instance descriptions with mixed types of attributes (some nominal along with others that are continuous) are allowed. Although not theoretically stated, mixed distributions are manageable by the evaluation function used by Classit.

2.2 Learning

2.2.1 Hill-climbing through a Space of Hierarchies

The basic Classit algorithm performs a hill-climbing search through a space of concept hierarchies[9]. In the context of this study, the original one-level instance structuring mechanism[6] is discarded from the algorithm. In the present work Classit will be considered simply as an extension of Cobweb[3] which deals with continuous attributes.

Classit incorporates any new instance in a hierarchical organization of concepts built from earlier observations. For each incoming instance, starting at the root of the tree, Classit descends the instance through the tree. At each level, a set of operators is evaluated. According to the results of that test, the algorithm recurses one level deeper in the tree, until a leaf is reached, or certain stopping criterion fulfilled.

2.2.2 Learning Operators

At each level in the tree, several transformations (called *learning operators*) are tentatively applied to the current partitionning. Classit evaluates the partitionning resulting from an integration of the instance in one of the sub-classes. The creation of a new class restricted to the instance is also examined. These are standard classification operators in the maintenance of hierarchical organization. In addition, and in order to give the algorithm the ability to recover from hill-climbing related problems (such as local optima and ordering effects), CLASSIT also evaluates two other operators.

The first additional operator considers merging two sub-classes. Merging is only considered for the two sub-classes that best host the incoming instance (i.e. the two sub-classes with the highest scores after integration has been evaluated).

The second additional operator considers splitting one sub-class, and replacing it with its own sub-classes. Again, splitting will only be considered for the best host for the instance.

2.3 EVALUATION FUNCTION

In order to select the operator to apply, CLASSIT needs a way to evaluate the quality of the partitionings it generates. To accomplish this, it uses a measure called *category utility.* It is expressed in terms of the intra-class similarity and inter-class dissimilarity of a partitioning. The former is expressed by the conditional probability $P(A_i = V_{ij} \mid C_k)$ (i.e. the probability for an object to take value V_{ij} on attribute A_i given its membership to C_k), the latter by $P(C_k \mid A_i = V_{ij})$. The overall probability of one individual value $P(A_i = V_{ij})$ is used as a measure of the frequency of occurrence for attribute values, and is independent from any class-membership.

An overall measure of a partitioning's quality may be stated as a sum over all classes, all attributes and all values of a simple product of the above characteristics

$$\sum_k \sum_i \sum_j P(A_i = V_{ij}) \cdot P(A_i = V_{ij} \mid C_k) \cdot P(C_k \mid A_i = V_{ij}) \tag{1}$$

Such a quantity allows us to compare different partitionings of the same set of instances, selecting the one that has the highest value for that measure. It can be transformed to obtain:

$$\sum_k P(C_k) \sum_i \sum_j P(A_i = V_{ij} \mid C_k)^2$$

This simple transformation has much more importance than it would first appear to have. The expression in equation (1) was based on measurable properties of individual attribute-values, while the transformed expression has a strong class-oriented flavor. The computational feasability of the second expression replaces the considerable complexity of equation (1). Moreover, the second formulation expresses a weighting of some measurable characteristic of individual classes of the partitioning. In fact, the innermost double summation can be interpreted as the number of attribute values that can be correctly guessed for a member of the class. It is then a measure of the quality of a class, knowing that better classes are those which can predict more attribute values.

The *category utility* is then formulated as the increase of this measure obtained by partitioning a class C_0 in K sub-classes $\{C_1, \ldots, C_K\}$.

$$\frac{1}{K}\left[\sum_{k=1}^{K} P(C_k) \sum_{i=1}^{I} \sum_{j=1}^{J(i)} P(A_i = V_{ij} \mid C_k)^2 \quad - \quad \sum_{i=1}^{I} \sum_{j=1}^{J(i)} P(A_i = V_{ij} \mid C_0)^2\right] \tag{2}$$

This measure is used at each level in the tree. For one class with K sub-classes, the algorithm computes at most $(K + 3)$ scores: one for each possible application of the *integration* operator, one for the *creation* operator, and one for each restructuring operators (when applicable).

It was said earlier that in continuous domains, random variables are assumed to be normal. In such cases, we have:

$$\begin{aligned}\sum_{j=1}^{J(i)} P(A_i = V_{ij} \mid C_k)^2 &= \int_{-\infty}^{+\infty} \left(\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \right)^2 dx \\ &= \frac{1}{\sqrt{4\pi}\sigma_i}\end{aligned}$$

where $\mathcal{N}(\mu_i, \sigma_i^2)$ is the assumed probability distribution of the values of attribute A_i in C_k. The parameters μ_i and σ_i are estimated from the set of objects covered by C_k. This approximation leads to the following expression of the evaluation function:

$$\frac{1}{K\sqrt{4\pi}} \sum_{k=1}^{K} P(C_k) \sum_{i=1}^{I} \left(\frac{1}{\sigma_{ik}} - \frac{1}{\sigma_{i0}} \right) \tag{3}$$

It has to be noted that the normality assumption has been experimentally verified and seems to be valid on the data used for the experiments. More precisely, the hypothesis that data are distributed according to a normal law was not rejected, and thus the standard deviation has been used as a good approximation. Indeed, in situations were such a model could not be assumed, the term $\sum_{j=1}^{J(i)} P(A_i = V_{ij} \mid C_k)^2$ could easily be computed from an histogram.

2.4 Performance

2.4.1 Prediction

Prediction in this context may be understood as the task of making assumptions upon an unknown value of an attribute given the class to which the instance belongs to. Due to the predictive nature of learnt concepts, the task of prediction is often considered as the primitive performance task for computed hierarchies.

To predict an attribute value for an instance, one must first classify this instance through the tree (see below), to find the most specific concept which covers this instance. Once identified as a member of a class, the corresponding concept gives a distribution for the unknown value, according to which the most frequently occuring value will be assigned to the missing value.

2.4.2 Recognition or Classification

Recognition is the task of assigning a class label to an incoming instance. The way to classify an instance does not differ from the way to learn it, except for the fact that the *split* and *merge* operators are not evaluated. This corresponds to the assumption that the classification (or recognition) task does not lead to any modification of the knowledge used to perform it. Thus, at each level in the tree, the *integration* and *creation* operators are evaluated using *category utility*. If an integration is found to be the best solution, the recognition task is applied recursively to the tree starting at the node having the best score. If creation of a new class is found to be the best alternative, the current node is returned, but no new sub-class is created. In the latter case, the assumption is that the best fitting class has been found.

In the case of image analysis, there is no consensus about objective evaluation strategies for computer-built classifications. The common practice is to let a human expert give his feeling about the resulting segmentation.

2.5 Time and Space Complexity

Evaluating time complexity of the Classit algorithm is equivalent to quantify the number of partitionning evaluations performed. If the cost of evaluation of a partitionning is taken as a unity, it is easy to derive that learning (i.e. incorporating) the $(n+1)^{th}$ object requires

$$(B+3)\log_B(n)$$

computations of *category utility*, where B is the average *branching factor* of the hierarchy. This makes the whole cost of incorporating N objects equal to $O(N\log(N))$.

To evaluate the amount of space used by the resulting hierarchy, one must notice that individual objects form the leaves of the tree. The space complexity of incorporating N' (distinct) objects is thus the number of nodes of a tree having N' leaves, which is

$$\frac{BN'-1}{B-1}$$

where B is the branching factor of the tree. Recall that N' is the number of distinct objects, and not the sample size. This expresses the fact that some redundant data (i.e. the same pixel appearing at several places) does not increase the size of the hierarchy. This fact also entails corrections in the time complexity, which were ignored in the previous paragraph.

3 The Radiometry of Nepalese Himalayas

3.1 Methodological Notes

The main goal of the work exposed in this paper is to apply the algorithm described in the previous section to a remote sensing sattellite image. The region under investigation is the Nepalese Himalayas. Three main motivations have to be stressed.

The first characteristic of our system is that its learning phase is totally unsupervised. Only the raw data are given, in contrast with many approaches where a predefined classification has to be provided[1,2]. This pre-classification may in fact be a major problem. Terrains under study (particularly in Nepalese Himalayas) are not easily accessible, and little knowledge is available about their "true" nature. Moreover, and more generally, it may be hard for an expert to assign class labels to some pixels of the image, since (in the case of SPOT) the value of a pixel is an average of radiometry over a 20 meters side square. Instead, we prefer to adopt a "discovery" approach to classification where the expert has to evaluate the proposed segmentation. A second, more technical, aspect of the unsupervised nature of this algorithm is that it does not require a preliminary upper bound on the number of clusters, as it is often the case[7,8].

Secondly, the knowledge produced is hierarchically organized. This means that several levels of generality are represented, from the most general (which could be labelled `any-pixel` to the most specific (which usually represents one single point in measurement space). We do agree with the idea that efficient knowledge lie between these extremes. This seems also correspond to the discourse of a thematician, since legends are often hierarchical (even though it is not always explicit: for instance, it may appear in the choice of colors in a thematic map). The thematic knowledge may always be represented as a taxonomy. Our main interest is on how the hierarchy produced will fit the taxonomy of a thematician.

The third and last point is that our "exploratory" approach will be based on radiometric values only. It has often be argued that other descriptors (or *attributes*) help classification. These additionnal characteristics of individual pixels are of two types: some are *calculated* (for instance, vegetation indexes), other are purely *extrinsic* (for instance, slope value and slope orientation). Calculated attributes only rely on radiometric values, and thus, from an informational point of view, do *not* add anything to the raw data. However, they often increase classification quality by stressing tendancies of radiometry. Extrinsic attributes, on the other hand, form external knowledge: no relation exists between radiometric values and the values of such attributes. In that sense, they are usually used when radiometry is *unable* to distinguish between several thematic objects. In that paper, we take the subjective position to simply ignore such additionned information. It is a well-known fact that radiometry is insufficient to represent the universe of thematic discourse, and we are conscious of this fact. Indeed, our motivation is that radiometry has not been investigated in depth, and we think that a safe methodology would be to first understand "the logic of radiometry".

3.2 The Data

Experiments are conducted on a SPOT image of Nepalese Himalayas. This region has several characteristics that makes it a complex geographic object. The image was shooted at 09:30 *a.m.* Sun elevation at this time entails very high contrasts, and thus high radiometric discrepancies for similar soils under distinct lightings. Haze effects must also be mentionned, since they strongly affect shadowy versants. The three images used are shown on Figure 1.

Figure 1: Three Original Spot Images of Nepalese Himalayas

Lighting conditions and haze effects can be considered as external factors acting on the image. There are, however, other complexity factors that are inherent to the image. The studied area is very complex for two kinds of reasons. First, the multiplicity of landscape units: altitudinal amplitudes may attain 5000 meters on the same versants, leading to graded climates (from subtropical to alpine), as well as graded vegetations and landuses. This fact is even amplified by the various exposure levels to sun light and winds. The second complexifying factor is the size of the landscape units, which is due in part to the steepness of the slopes. It is needless to say that human intervention (through the management of the environment) also directly affects the structure of landscape units organization, but this influence appears much more on areas where the slopes remain gentle or where access is easy.

3.3 Resulting Segmentation

The experimental setting was as follows:

- 50000 pixels were randomly drawn from the original image (which has a size of 345×410): this is approximately a third of the available data.
- a hierarchy was built from these data. No *pruning* mechanism (i.e. any way to reduce the size —the depth— of the tree) was employed. The growing of the hierarchy could easily have been stopped by setting an upper limit on the value of the *category utility* measure. However, such an

C		
	μ	σ
V_1	24.67	6.778
V_2	16.55	7.527
V_3	36.78	16.315

C_0		
$P(C_0) = 0.157$		
V_1	16.52	0.845
V_2	7.77	0.707
V_3	10.38	2.678

C_1		
$P(C_1) = 0.194$		
V_1	19.67	1.152
V_2	10.77	0.902
V_3	25.15	7.881

C_2		
$P(C_2) = 0.649$		
V_1	28.14	5.886
V_2	20.40	6.587
V_3	46.64	9.101

Figure 2: The first level of the hierarchy

"accuity reducing" mechanism may affect the organization of higher levels of the hierarchy. Thus, in our experiments, the terminal nodes represent the ultimate degree of precision available in the measurement space.

- the data (i.e. pixels) were presented as triples of radiometric values. No other information was provided to the system.

The results were the following:

- a hierarchy constituted of approximately 7000 nodes. One might be surprised by the high number of nodes, but in fact it corresponds to a 12 levels deep hierarchy. The first level segments are shown in Figure 2. The root node gives estimated mean and standard deviation over the whole image. Each first level node also gives the proportion of pixels covered.
- such a somewhat "abstract" definition of spectral classes can not be easily interpreted. The whole image (constituted of 141450 pixels) was then classified, i.e. driven through the tree. This process gave a visual representation for each node in the hierarchy.

3.4 Comments

Once the whole image was classified, the obtained segmentation was compared with available thematic maps, and, most importantly, given to the expert for validation. The following objects were recognized:

- the classes denoted by C_0 and C_1 cluster pixels from shadowy sectors, C_0 corresponding to deepest soils, C_1 to shallow soils.
- class C_2 corresponds to lighted versants, with the least dense vegetation, as well as cultivated areas on grazing light versants.
- sub-classes of C_2 (numbered $C_{2.0}$, $C_{2.1}$ and $C_{2.2}$) represent respectively
  - moderatly dense forests and clear forests ($C_{2.0}$)
  - cultivated areas, degraded forests, bushes and contact areas ($C_{2.1}$)
  - cleared ("deforested") areas ($C_{2.2}$)

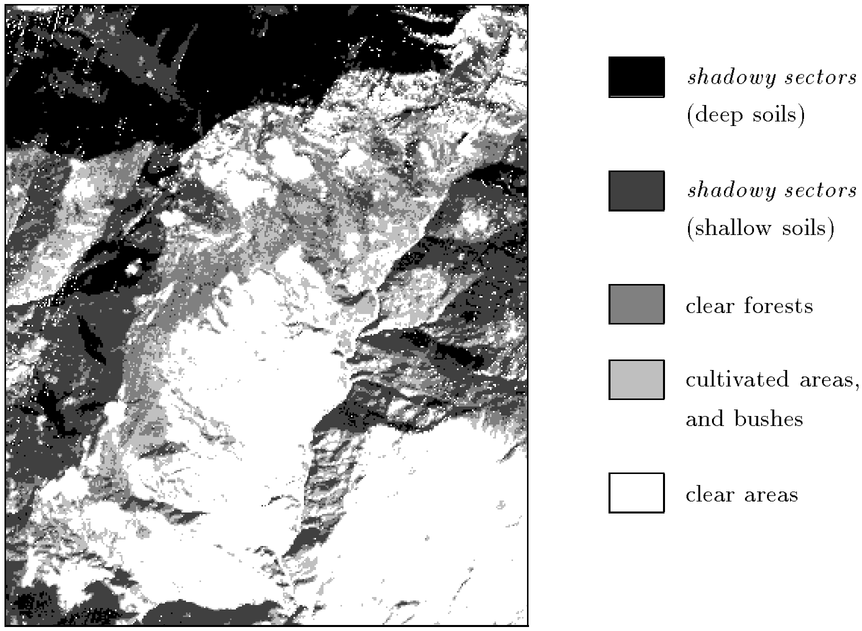

Figure 3: The first map, based on highest levels of the hierarchy

These remarks lead to a first map, built from only first and second level classes. This map is shown in Figure 3.

Let us now concentrate on the main region (i.e. the region corresponding to lighted versants, noted C_2 previously, and represented as the three clearest colors on the first map). It was said above that this class corresponds to differnet kinds of vegetation on lighted sectors. This class is further broken into three sub-classes, noted $C_{2.0}$ to $C_{2.2}$ (see the first map). Let us now examine these sub-classes:

- class $C_{2.0}$ (moderatly dense forests) further divide into two classes: one of them corresponds to continuous shrub cover, the other to clear forests.
- class $C_{2.1}$ (degraded forests) splits into bushes on one hand, and degraded forests on the other hand.
- class $C_{2.2}$ is divided into 5 sub-classes:
  - high-grade pastures (*type-I* grass),
  - wheat fields and *type-II* grass,
  - rice fields, stubbles and *type-III* grass,
  - eroded gullies, rice fields on shallow soils, fallow fields and valley bottoms (also with some clouds on SPOT image),
  - non-covering shrubs on pastures and contact areas.

All these classes are represented on Figure 4. Other regions (represented dark on first map) are left blank.

The segmentation presented on Figure 4 shows a reasonably detailed map of the landscape. However the thematician could be interested in greater details on specific classes. As a reminder, the hierarchy

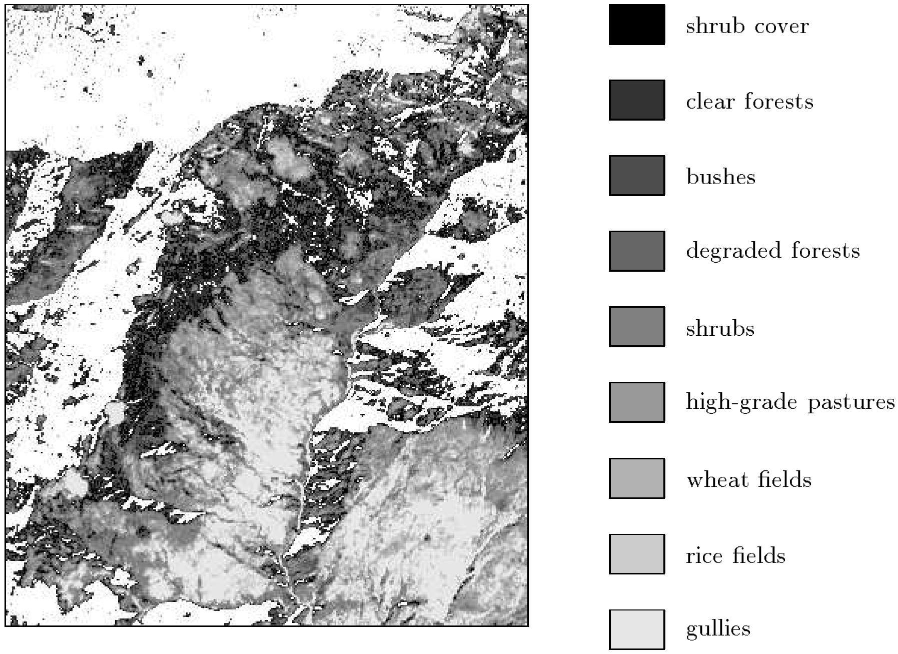

Figure 4: The second map, based on some lower levels of the hierarchy

starting at the node called $C_{2.2}$ (white area on the first map) still has more than 3700 nodes under it. Any segment of the second map could be further divided between 6 and 9 times each. However, "deep" classes (according to the hierarchy) represent less and less pixels, and are difficultly interpretable. Even such segmentations as the one depicted on the second map becomes hard to read. As an illustration, let us examine only one of the classes shown on Figure 4. It was said that the class labelled "shrubs" contained non-covering shrub covers and contact areas. Figure 5 shows a very simple map on which are shown:

- the class $C_{2.2}$ (shown in white on the first map) which corresponds to cleared forests; this class was labelled "cleared forests".
- the class labelled "shrubs" on the second map, which corresponds to non-covering shrubs and contact areas.

As one can see, the "shrubs" class is a precise characterization of limits between distinctly vegetated areas. Such a class is an important indicator when creating a thematic map.

4 Conclusion

As shown on the first map, the influence of lighting conditions is determinant to the accuracy of the resulting segmentation. This —well-known— fact usually leads to a first, preliminary, segmentation based on lighting information only. Each derived class of lighting is then further segmented on the base of radiometric values.

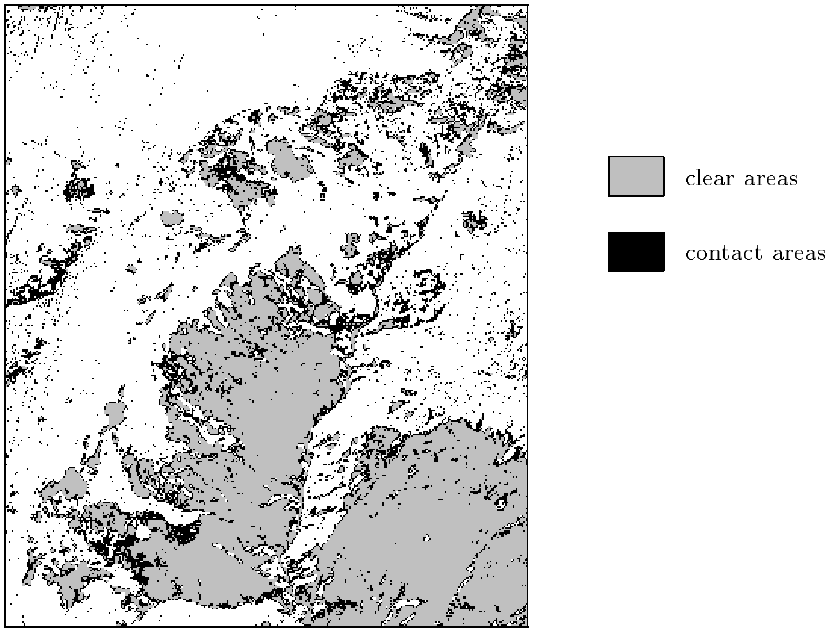

Figure 5: The third map, showing contact areas

As a result of the influence of lighting, the hierarchy produced by the algorithm described in this paper is strongly “biased”: the first map (in Figure 3) illustrates the fact that (even though not given as an attribute) lighting determines the first level segments. This is also noticed by an important decrease of variance between the root and the first level of the hierarchy: this is especially true for the third spectral band.

Once the influence of lighting conditions decreases (typically in deeper levels of the hierarchy, where lighting is almost homogenous), the concept formation algorithm gives remarkably accurate results. As our experiments suggest, the resulting segmentation is at least as good as the ones built by the best supervised methods. This, though counterintuitive (since additional knowledge should increase quality), means that efficient knowledge can be directly derived from the data. Due to the tremendous complexity of the landscape, expert provided knowledge can not be relied on.

From a thematic point of view, the strong influence of lighting conditions forbid us to say that the concept hierarchy built by our system “fits” the geographer’s thematic hierarchy. However, under homogenous lighting conditions, formed classes have an indeniable thematic meaning. In terms of the hierarchy produced, the coincidence between discovered classes and thematic classes starts after one or two levels have been traversed.

5 References

1. Blamont, D., & Mering, C. (1989). CIME: An Expert System for Remote Sensing in Montainous Areas. *International Conference on Expert Systems for Development*, Katmandu, March 1989.

2. Blamont, D., & Ganascia, J.G. (1990). Machine Learning and Knowledge Acquisition Applied to Cartography. *Cognitive and Linguistic Aspects of Geographic Space. NATO ASI.*

3. Fisher, D.H. (1987). Knowledge Acquisition Via Incremental Conceptual Clustering. *Machine Learning, 2*, pp. 139-172.

4. Fisher, D.H., Pazzani, M.J. & Langley, P. (1991). *Concept Formation: Knowledge and Experience in Unsupervised Learning*. Morgan Kaufmann.

5. Ganascia, J.G., Blamont, D., & Mering, C. (1990). A Symbolic of Geographic Space Through Remote Sensing. *Cognitive and Linguistic Aspects of Geographic Space. NATO ASI.*

6. Gennari, J.H., Langley, P., & Fisher, D.H. (1989). Models of Incremental Concept Formation. *Artificial Intelligence, 40*, pp. 11-61.

7. Korczak, J.J. (1991). A Computer-Assisted Method for the Creation of Spectral Classes from Remotely Sensed Images. *Actes des Journées Symboliques-Numériques*, Versailles, France. Nova Sc. Publ.

8. Korczak, J.J., & Rymarczyk, M. (1993). Application of Classical Clustering Methods to Digital Image Analysis. *Research Report*, Université Louis Pasteur, Strasbourg, France.

9. Langley, P., Gennari, J.H., & Iba, W. (1987). Hill-Climbing Theories of Learning. *Proceedings of the Fourth International Workshop on Machine Learning*, pp. 312-323.

10. Michalski, R.S., & Stepp, R.E. (1983). Learning from Observation: Conceptual Clustering. In R.S. Michalski, J.G. Carbonell, & T.M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach.* Los Altos, CA: Morgan Kaufmann.