Transactions on Machine Learning
and Data Mining
Vol. 1, No. 2 (2008) 49 - 65
©ISSN:1865-6781 (Journal),
IBaI Publishing ISSN 1864-9734

ibai Publishing
www.ibai-publishing.org

Experiences Using Clustering and Generalizations for Knowledge Discovery in Melanomas Domain

A. Fornells[1], E. Armengol[2], E. Golobardes[1], S. Puig[3], and J. Malvehy[3]

[1] Grup de Recerca en Sistemes Intel·ligents
Enginyeria i Arquitectura La Salle, Universitat Ramon Llull
Quatre Camins 2, 08022 Barcelona (Spain)
email:{afornells,elisabet}@salle.url.edu
[2] IIIA - Artificial Intelligence Research Institute,
CSIC - Spanish Council for Scientific Research,
Campus UAB, 08193 Bellaterra, Catalonia (Spain)
email: eva@iiia.csic.es
[3] Melanoma Unit, Dermatology Department
IDIBAPS, U726 CIBERER, ISCIII
Hospital Clinic i Provincial de Barcelona (Spain)
email: {spuig,jmalvehy}@clinic.ub.es

Abstract. One of the main goals in prevention of cutaneous melanoma is early diagnosis and surgical excision. Dermatologists work in order to define the different skin lesion types based on dermatoscopic features to improve early detection. We propose a method called SOMEX with the aim of helping experts to improve the characterization of dermatoscopic melanoma types. SOMEX combines clustering and generalization to perform knowledge discovery. First, SOMEX uses Self-Organizing Maps to identify groups of similar melanoma. Second, SOMEX builds general descriptions of clusters applying the anti-unification concept. These descriptions can be interpreted as explanations of groups of melanomas. Experiments prove that explanations are very useful for experts to reconsider the characterization of melanoma classes.

Keywords: Melanoma, Skin Tumour, Dermoscopy, Medicine, Knowledge Discovery, Clustering, Self-Organizing Maps, Explanations.

1 Introduction

Early diagnosis and surgical excision are the main goals in the secondary prevention of cutaneous melanoma. Nowadays, the diagnosis of melanoma is based on the ABCD rule [1] which considers four clinical features commonly observed in this kind of tumour: asymmetry, border irregularity, colour variegation, and a diameter larger than 5 mm. Although most of melanomas are correctly diagnosed following this rule, a variable proportion of melanomas does not comply with these criteria. The current procedure when a suspicious skin lesion appears is to excise and to analyse it by means of biopsy. Commonly, the result of the biopsy allows to determine the accurate malignity of the lesion.

Dermoscopy is a non-invasive technique for a more accurate evaluation of skin lesions introduced by dermatologists two decades ago. Dermoscopy provides the opportunity to avoid the excision of benign skin lesions. However, dermatologists need to achieve a good dermoscopic classification of lesions previously to extraction [2]. Hofmann-Wellenhof et al [3] suggested a classification of benign melanocytic lesions. Recently, Argenziano et al [4] hypothesized that dermoscopic classification may be better than the classical clinico pathological classification of benign melenocytic lesions (nevi). Currently, there is no dermoscopic classification of melanoma located in trunk and extremitis. In the era of genetic profiling, molecular studies including microarrays suggest that there is more than one type of melanoma in these locations. The aim of the present work is to help dermatologists in the classification of early melanoma (*in situ melanoma*) based on dermoscopy characteristics. For this reason, dermatologists define several dermoscopic classes of in situ melanoma based on their dermoscopic features. Dermatopathologies also suggest another classification based on histological features.

The goal of this work is twofold: on one hand we want to confirm that the dermoscopic classes are well defined and, on the other hand, we want to relate these classes to the histological classes of melanomas from the histopathological analysis of biopsies. The present paper describes a method called SOMEX to help dermatologists in their research. SOMEX is a combination of two machine learning approaches: clustering and generalization. In a first step, a Self-Organizing Map [5] clusters a set of skin lesions in patterns according to their similar characteristics. In a second step, a generalization method based on the notion of anti-unification [6] is used to explain clustering results. Results should help dermatologists to discover what fails in defining classes and why lesions that they consider belong to different classes have been clustered together.

The paper is organized as follows. The next section describes the combination of clustering and generalization in SOMEX. Section 3 explains briefly the melanoma domain and it also describes some particular results achieved with SOMEX application. Section 4 describes some related work. Finally, section 5 summarizes the article with conclusions and future work.

2 SOMEX

Let us suppose the following scenario: there is a set of objects belonging to several classes and we want to test whether or not these classes are correctly defined. The first idea is to apply some clustering technique in order to achieve natural groups of similar objects. By testing these groups taking into account the classes we can determine their commonalties. This is exactly what SOMEX achieves by means of generalization of the clusters defined by the clustering technique called Self-Organizing Maps. Next sections explain in detail how SOMEX works.

2.1 Self-Organizing Maps

Self-Organizing Map (SOM) [5] is one of the major unsupervised learning paradigms in the family of artificial neural networks. It has many important properties which make it useful for clustering [7]: (1) It preserves the original topology; (2) It works well even though the original space has a high number of dimensions; (3) It incorporates the selection feature approach; (4) Although one class has few examples they are not lost; (5) It provides an easy way to show data; (6) It is organized in an autonomous way to be better adjusted to data. Moreover, SOM is a soft-computing technique that allows the management of uncertain, approximate, partial truth and complex knowledge. These capabilities are useful in order to manage real domains, which are often complex and uncertain.

Because SOM is a no supervised technique it has to discover by itself which commonalities, correlations and classes of the objects are. SOM projects the original space from an input layer of N neurons (a neuron for each feature describing the input data) to an output layer of a new space with less dimensions (a neuron for each expected cluster) with the aim of identifying groups of similar elements. Figure 1 shows a typical 2-dimensional grid of $M \times M$ neurons, where each one is represented by a *director vector* of N dimensions (v_m). A director vector can be described as the expected value for each one of the N input neuron (feature). Moreover, each input neuron is connected to all the output neurons. The definition of clusters can be summarized in the next steps:

1. Director vectors of each neuron are randomly initialized.
2. Given a new input example e, the distance between e and each director vector is computed with the aim of identifying the most suitable neuron where the e should be mapped. For example, the winner neuron is the one with the value most similar to 1 if the normalized Euclidean distance (see Eq. 1).

$$similarity(e, m) = |d(\boldsymbol{e}, \boldsymbol{v_m})| = \left| \sqrt{\frac{\sum_{n:1..N}(e(n) - v_m(n))^2}{N}} \right| \quad (1)$$

3. Directors vectors are adjusted. The director vector of the winner neuron is adjusted for improving the match with new objects similar to the current one. In contrast, the rest of directors vectors are modified to weakly represent the current example.

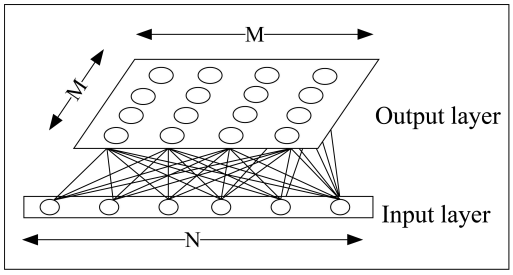

Fig. 1. SOM groups similar elements according to their data features.

4. Steps 2 and 3 are repeated for all training examples until director vectors are representative enough. Usually their representativeness is determined by establishing a minimal error value computed as the global sum of distance between the set of cases of each cluster and its respectively director vector. Nevertheless, other common criteria is to establish a maximum number of algorithm iterations.
5. When the training process ends, step 2 is the procedure used to map the new input example in the most suitable clusters.

The main drawback of the method is the definition of the training parameters. First aspect is to determine the *map size*, which is related to the final number of clusters. Thus, a big size of maps will produce a high number of clusters, where each cluster will contain few objects. Conversely, small maps will produce few clusters containing a lot of objects and, consequently director vectors of clusters will be overgeneralized. A second aspect to take into account is *neighborhood factor*, which is the influence of each cluster over others. The third aspect is the learning factor, which determines the convergence of algorithm. High values of this factor could produce a random behavior of learning procedure and low values could produce slow ratio of convergence. Finally, the last aspect is the distance measure used to make comparisons.

To conclude, SOM is a smart technique to identify hidden and complex relationships between elements and also to identify the most relevant features thanks to its knowledge discovery and soft-computing capabilities. This is exactly what experts need: to discover relationships between elements to improve the precision of the classes proposed by them.

2.2 How to Explain a Cluster

Director vectors can be described as the expected values that each attribute has to satisfy to be classified as belonging to a cluster. However, from the user's point of view these tuples do not give an easy intuition of why some objects have been clustered together. Because of this in [8] we propose to build symbolic explanations of the clusters with the purpose of justifying why a set of cases have been clustered together (this is a concept similar to *characterization* used

in data mining terminology [9]). Experts found symbolic explanations more understandable than director vectors since the former are constructed using the same representation language than they used to describe the domain objects.

Thus, we propose to explain a cluster using a symbolic description that is a generalization of all objects contained in the cluster. This generalization is based on the anti-unification concept [6] although with some differences. The anti-unification (AU) of a set of objects is a description defined as their most specific generalization. The AU contains attributes shared by the set of objects and where each attribute takes as value the most specific of all the values holding in the original set. In this paper we only work with the idea of shared attributes among a set of objects.

Let M_i be a cluster and let $c_1, ..., c_n$ be the set of objects that belong to that cluster after the application of SOM to a set of objects. Each object c_j is described by a set of attributes $\mathcal{A}$. The explanation D_i of why a subset of objects have been clustered in M_i is built in the following way:

- D_i contains attributes which are common to all the objects in M_i. Attributes with *unknown* value in some object $c_j \in M_i$ are not in D_i.
- Let a_k be an attribute common to all objects in M_i such that a_k takes symbolic values on a set $\mathcal{V}_k$. The attribute a_k will not be in D_i when the union of the values that a_k takes in M_i is exactly $\mathcal{V}_k$.
- An attribute a_i takes in D_i the union of all values that a_i holds in the objects in M_i.

Let us illustrate with an example how to build explanations for a cluster. Let M_5 be the cluster formed by the three cases (see Fig. 2). Let D_5 be the explanation of why these cases are clustered (see Fig. 3). An attributes such as C_Max-Diam is not in D_5 because it is not common to all cases (i.e., C_Max-Diam is not present in obj-61). In contrast, attributes such as C_Sex, D_Pseudopigment-Network or D_Peppering are not in D_5 because they take all possible values. For instance, the feature C_Sex takes the value F in object obj-61 and the value M in both objects obj-9 and obj-61. This means that the value of this attribute is irrelevant to describe M_5.

Summarizing, explanations provide a symbolic description that contains the commonalties among all objects of a cluster. We chosen to show for each attribute the union of possible values (instead of the average or the mode as is the usual approach) because the expert finds more useful knowing all possible values. This is the explicit information that an expert extracts from SOMEX; but, our question would be if there is some implicit information from the explanations. The answer would be affirmative. Two aspects of the explanations are specially relevant from the point of view of the knowledge discovery: one is the number of attributes composing explanations and the other is the number of values that take these attributes. Both aspects give an idea of how similar the objects contained in a cluster are.

```
(Define (Object :Id Obj-68)
  (C_Sex M)
  (C_Age 69)
  (C_Max_Diam 10)
  (C_Site Back)
  (D_Pattern Multicomponent)
  (D_Pigment_Network 1)
  (D_Atypical_Pn 1)
  (D_Pseudopigment_Network 0)
  (D_Dots_And_Globules 0)
  (D_Atypical_D_And_G 0)
  (D_Streaks 1)
  (D_Irregular_Streaks 1)
  (D_Regression_Structures 1)
  (D_Peppering 1)
  (D_White_Areas 1)
  (D_Bw_Veil 0)
  (D_Bloches 0)
  (D_Irregular_Bloches 0)
  (D_Vessels 0)
  (D_Dotted_Vessels 0)
  (D_Atypical_Vessels 0)
  (D_Millia_Like_Cyst 0)
  (H_Diagnosis Mnevus))
```

```
(Define (Object :Id Obj-61)
  (C_Sex F)
  (C_Age 36)
  (C_Site Forearm)
  (D_Pattern Reticular)
  (D_Pigment_Network 1)
  (D_Atypical_Pn 1)
  (D_Pseudopigment_Network 0)
  (D_Dots_And_Globules 0)
  (D_Atypical_D_And_G 0)
  (D_Streaks 1)
  (D_Irregular_Streaks 1)
  (D-Regression_Structures 1)
  (D_Peppering 1)
  (D_White_Areas 0)
  (D_Bw_Veil 0)
  (D_Bloches 0)
  (D_Irregular_Bloches 0)
  (D_Vessels 0)
  (D_Dotted_Vessels 0)
  (D_Atypical_Vessels 0)
  (D_Millia_Like_Cyst 0)
  (H_Diagnosis P))
```

```
(Define (Object :Id Obj-9)
  (C_Sex M)
  (C_Age 64)
  (C_Max_Diam 20)
  (C_Site Upper_Extr)
  (D_Pattern Reticular)
  (D_Pigment_Network 1)
  (D_Atypical_Pn 1)
  (D_Pseudopigment_Network 0)
  (D_Dots_And_Globules 1)
  (D_Atypical_D_And_G 0)
  (D_Streaks 1)
  (D_Irregular_Streaks 1)
  (D_Regression_Structures 1)
  (D_Peppering 0)
  (D_White_Areas 1)
  (D_Bw_Veil 0)
  (D_Bloches 0)
  (D_Irregular_Bloches 0)
  (D_Vessels 0)
  (D_Dotted_Vessels 0)
  (D_Atypical_Vessels 0)
  (D_Millia_Like_Cyst 0)
  (H_Diagnosis Nonc))
```

Fig. 2. Description of three classes included in a cluster, say M_5.

```
(C_Age (36 64 69))
(C_Site (Back Forearm Upper_Extr))
(D_Pattern (Multicomponent Reticular))
(D_Pigment_Network (1))
(D_Atypical_Pn (1))
(D_Atypical_D_And_G (0))
(D_Streaks (1))
(D_Irregular_Streaks (1))
(D_Regression_Structures (1))
(D_Bw_Veil (0))
(D_Bloches (0))
(D_Irregular_Bloches (0))
(D_Vessels (0))
(D_Dotted_Vessels (0))
(D_Atypical_Vessels (0))
(D_Millia_Like_Cyst (0))
(H_Diagnosis (Mnevus P Nonc))
```

Fig. 3. Explanation of why the objects included in cluster M_5 (Fig. 2) have been clustered together.

Concerning the number of attributes, explanations with a high number of attributes represent very similar objects whereas explanations with few attributes mean that these objects have few aspects in common. Nevertheless, the number of values holding the attributes of an explanation also plays a crucial role. Thus, the more values an attribute holds the more irrelevant this attribute is. Notice that the explanation is built using common attributes and taking as values for these attributes the union of all values hold by the objects of a cluster. Thus, a common attribute that takes several values, means that has a high variability and this attribute is probably not too relevant. Conversely, attributes holding only one value represent aspects of the objects that could be taken as candidates to characterize a cluster.

In short, clusters explained by means of descriptions composed of a high number of attributes where each attribute holds one value, can be interpreted as good clusters in the sense that all the objects included in them are very similar. On the other hand, if the object class is known two situations can happen: 1) all objects of the cluster belong to the same class, or 2) objects belong to several classes. This second situation is the interesting one from the point of view of knowledge discovery since it means that objects that, in principle, belong to different classes are highly similar. For instance, in the melanomas domain such situation means that melanomas that dermatologists classified as belonging to different clusters, the clustering process of SOMEX has put them together in the same cluster. The explanation of the corresponding cluster allows the experts to assess the actual relevance of the commonalties of these *a priori* different melanomas. This should be a starting point from the expert to reconsider the definition of classes (for instance, by merging classes to which objects belong).

Similarly, clusters explained by means of descriptions with a lot of attributes holding almost all possible values, can be interpreted as imprecise clusters in the sense that objects in the cluster have not many similarities. From the knowledge discovery point of view, this situation is interesting when all objects of such clusters belong to the same class, since it means that although they have been classified as belonging to the same class, these objects are not actually similar.

These situations will be illustrated in more detail in the next section where SOMEX is applied to support dermatologists in the definition and validation of some classes of malignant skin lesions.

3 Using SOMEX for Knowledge Discovery

Dermatologists take into account dermoscopic aspects of skin lesions with the aim of determining whether or not it will become a melanoma (malignant skin lesions) prior to lesion excision. The aim of SOMEX is to support dermatologists to obtain patterns of different kinds of melanomas *in situ*. First, SOM clusters together objects (descriptions of skin lesions) that are similar independently of the class. Then, symbolic explanations show dermatologists the common features of objects clustered together, allowing them to consider some modifications in

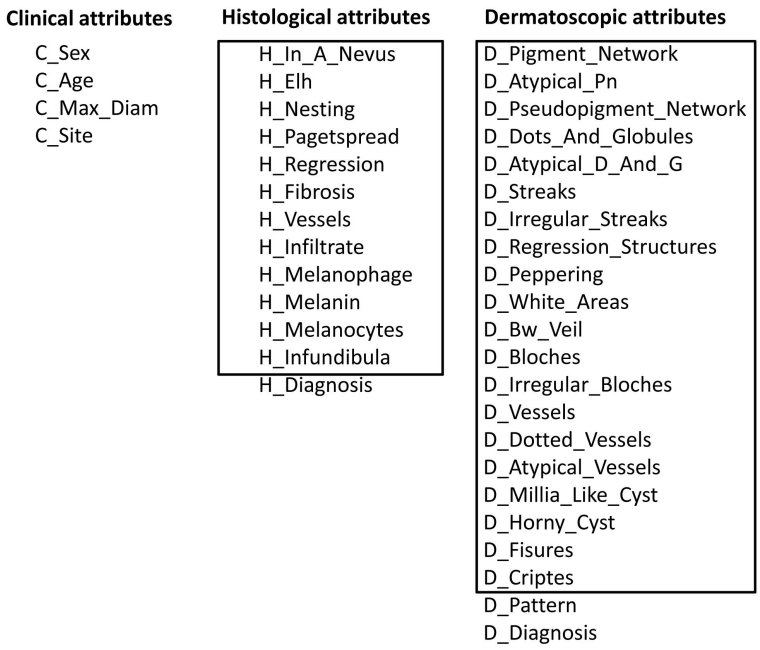

Fig. 4. Clinical, histological and dermoscopic attributes used to describe a melanoma.

the class definition. This section describes briefly the melanomas domain and results achieved by SOMEX.

3.1 Testbed: The Melanomas Domain

A skin lesion can be described from two different aspects: *dermoscopic* and *histologic*. Dermoscopic aspects are those obtained using a technique called dermoscopy. This technique combines an image magnification process (i.e. x30) with a system that decreases both the reflex ion and the refraction of the light through polarized light and polarization filters. Thus, dermoscopy allows to identify global patterns (*D_Pattern*) and local features (attributes inside the rectangle in the right part of Fig. 4), which are used for experts to suggest a hypothetical diagnosis (*D_Diagnosis*). In contrast, histological aspects (attributes inside the rectangle in the middle part of Fig. 4) are obtained from the biopsy of a excised suspicious skin lesion. Biopsy results allow experts to confirm the real diagnosis (*H_Diagnosis*). In addition to these attributes, the description of a lesion is completed with the clinical profile of the patient such as age or sex among others.

Although the experimental dataset used in this work contains only 75 melanomas, it is considered as a representative sample of the domain since it comes from a consensus among 6 experts (4 dermatologists and 2 dermatopathologists) around the description of melanomas.

With our experiments we want to support dermatologists in finding 1) how to dermoscopically describe histologic classes, and 2) to test whether or not histologic classes have been correctly defined. The next section describes the conditions under which experiments have been performed and also some interesting results obtained from SOMEX.

3.2 Experiments

Since our purpose was to support dermatologists in determining the dermoscopic features that describe the histologic classes, we only focused on the clinical and dermoscopic attributes (see Fig. 4). We also included the histological class represented by the H_Diagnosis attribute, i.e. the histologic class considered by the experts. Dermatologists defined the following histologic classes: *LTG_M, LMM, nonc, PL_M, P, P_LTG, Mnevus,* and *SKlMM.*

Bearing in mind this information, we performed several SOM configurations in order to find out interesting results. SOM was tested using several map sizes of 2-dimensions (3×3, 4×4 and 5×5 to analyze several data dispersions), two different distance measures (normalized Euclidean distance and normalized Hamming distance) and 10 random seeds (to minimize the random effects of initialization). The learning factor, the neighbor factor and maximum iterations were set respectively to values from 0.6 to 0.01, from M to 1 and 500 iterations by neuron.

3.3 Discussion of the Results

Independently of the map size and of the distance measure used to build the clusters, SOMEX results showed that the definition of histological classes should be adjusted. The reason is that most clusters include objects belonging to several histologic classes and explanations show that these objects have a lot of common aspects. This is reflected in the fact that most of explanations are very specific, i.e. they have a lot of common attributes holding a unique value. Notice that the high number of common features with only one value, the more similar the objects are. Conversely, explanations with features holding more than one value mean that, although objects are described by similar features, they have a lot of variability and they are not so similar.

The use of clustering techniques allowed a natural group of similar objects. Then by means of generalizations SOMEX explains why a subset of objects have been clustered together. Results show that some melanomas that dermatologists considered as belonging to different classes actually are not so different since they belong to the same cluster. Moreover, the explanation supports the user in discovering the common aspects and also characteristics that are different among objects of a same cluster. In fact, this provides them a clue to reconsider the definition of histological classes. Prior to SOMEX experiments, dermatologists had the hypothesis that the *pattern* (feature D_Pattern) of a skin lesion could be an important aspect to determine the classification of a lesion. As we will detail later, from SOMEX experiments we point out that the *pattern*, at least taken it isolated from other characteristics, is not enough for classification.

A conclusion from the experiments is that criteria used by dermatologists when defining histologic classes (*H_Diagnosis*) do not take into account all aspects describing a melanoma. In fact, clusters almost always contain melanomas of several histological classes. thus, from the predictivity point of view, clusters are not appropriate. However, when experts analyze the explanations of clusters they find them interesting despite their entropy. Experts noted that attributes shared by melanomas into a cluster usually are those considered as important for experts (for instance, D_dots_and_globules or D_Pigment_network). For this reason we prefer to show the analysis that experts performed of the SOMEX explanations instead of giving predictivity measures.

Experiments produced three types of clusters: 1) clusters with a reasonable number of objects belonging to different classes, 2) clusters with few objects belonging all of them to the same class, and 3) clusters with few objects of several classes. SOMEX results show that there are not clusters with a high number of objects belonging all of them to the same class nor clusters with few objects with a general explanation. Let us to analyze in more detail the SOMEX results.

Example 1 Let us suppose the cluster M_{15} containing 10 objects. The explanation of this cluster can be seen in Fig. 5. Concerning the number of attributes of the explanation, we see that there is a subset of 15 attributes (from the 28 composing a complete description of an object) shared by all the objects of the cluster. Focusing on values of these common attributes, we seen that most of them have an unique value, meaning that the explanation is specific enough.

In the explanation of the cluster M_{15} there are five attributes with more than one value: C_Age, C_Max-Diam, C_Site, D_Pattern and H_Diagnosis. Two of these attributes, C_age and C_Max-Diam are numerical and currently we cannot extract any conclusion from them. This is because explanations are not able to handle with continuous attributes. Dermatologists plan to establish some kind of discretization to establish ranges of equivalent values for these attributes. Concerning the values of attributes C_Site and D_Pattern, SOMEX shown that they hold a lot of values (almost all the possible values in the case of D_Pattern). In particular, the role of a lesion *pattern* as potential relevant aspect of a melanoma seems to be compromised according to the explanation of this cluster.

Finally, an interesting analysis can be carried out from values of H_Diagnosis. This is, in fact, the classification proposed by dermatopathologists; therefore, according to their criterion objects of M_{15} belong to five different classes (*LTG_M, nonc, PL_M, P_LTG* and *Mnevus*). However SOMEX show that these objects have a high similarity and the explanation suggests to dermatologists a possible analysis of the relevance of object commonalties so as that they should reconsider the criteria used to classify objects in different histological classes. An analysis of the differences among the objects in M_{15} could also clarify the class definition.

Cluster 15

The cluster is composed of the objects :
(<Obj-69> <Obj-70> <Obj-28> <Obj-31> <Obj-56> <Obj-15> <Obj-10> <Obj-11> <Obj-37> <Obj-44>)

The explanation is the following
((C_Site (Leg Arm Lower_Extr Upper_Extr Trunk Back))
(C_Max_Diam (5 6 7 8 9 18))
(C_Age (28 43 49 50 54 65 66 68)))
(D_Millia_Like_Cyst (0))
(D_Atypical_Vessels (0))
(D_Irregular_Bloches (0))
(D_Bloches (0))
(D_Peppering (0))
(D_Regression_Structures (0))
(D_Irregular_Streaks (0))
(D_Streaks (0))
(D_Atypical_D_And_G (1))
(D_Dots_And_Globules (1))
(D_Pattern (Globular Reticular Unspecific Multicomponent))
(H_Diagnosis (Mnevus P_Ltg Pl_M Nonc Ltg_M))

Cluster 24

The cluster is composed of the objects :
(<Obj-12> <Obj-13> <Obj-16>)

The explanation is the following
((C_Sex (F))
(C_Age (38 40 62))
(C_Max_Diam (4 6 28))
(C_Site (Trunk Lower_Extr))
(D_Pattern (Unspecific Reticular))
(D_Atypical_Pn (0))
(D_Dots_And_Globules (0))
(D_Atypical_D_And_G (0))
(D_Streaks (0))
(D_Irregular_Streaks (0))
(D_Bw_Veil (0))
(D_Bloches (0))
(D_Irregular_Bloches (0))
(D_Vessels (1))
(D_Atypical_Vessels (1))
(D_Millia_Like_Cyst (0)))
(H_Diagnosis (Pl_M))

Fig. 5. Explanations justifying the clusters M_{15} and M_{24}.

Example 2 The explanation of cluster M_{24} is composed of three objects with the same histological class. There are 4 multi-valued attributes: C_Age, C_Max-Diam (both numerical), C_Site, and D_Pattern. Globally this explanation seems a good partial characterization for the class PL_M since a further analysis of the numerical values could produce a more specific explanation. An important aspect to take into account is that the attribute D_Pattern has two possible values, *unspecific* and *reticular*. The importance of this fact is that according to SOMEX results, dermatologists should consider the possibility to reject D_Pattern as relevant for classifying a melanoma, since this feature holds different values in objects of the same histological class. In our current experiments we do not consider neither the relationship among attributes nor the weight of some attributes in order to bias the clustering. A possibility is that the pattern of a melanoma could be relevant in relation to the value of any other attribute.

Example 3 The cluster M_5 shown in Fig. 2 is an example of a small one with elements of several histological classes (in fact, each object belongs to a different class). The explanation of this cluster is shown in Fig. 3. This explanation is specific since it is composed by 17 common attributes and all of them except 4 hold an unique value. Notice that as in previous examples, attributes with multiple values are C_Age, C_Site, D_Pattern and H_Diagnosis. Once again the

conclusion should be to reconsider the definition of histological classes and to analyse the relevance of attributes that dermatopathologists used to define them.

An important point from the application of SOMEX is that symbolic explanations obtained from clusters give to dermatopathologists descriptions of groups of melanomas that they commonly recognize as different. For instance, the explanation for cluster M_{15} (see Fig. 6) describes lesions that under dermoscopy presents both dots and globules and typical pigment network (notice that all other features have as value 0, meaning *absence*). This description is clearly recognized from the dermatological point of view since they provided us the picture shown in Fig. 5 left, that corresponds to a lesion belonging to cluster M_{15} (in particular, it is the object *<Obj-11>*. Similarly, the explanation of cluster M_{24} describes lesions, completely different that those of cluster M_{15}. In particular, lesions in cluster M_{24} have as unique feature the presence of typical vessels, dermatologists recognized lesions such as the shown in Fig. 5 right (corresponds to *<Obj-13>*) of cluster M_{24}). Summarizing, SOMEX provides a natural clustering of objects and the use of symbolic explanations supports dermatologists in analysing the correctness of the clusters and also in redefining some of the histological classes they propose.

4 Related Work

Clustering techniques are a smart way to extract relationships from huge data sets. Consequently, this useful property has been widely used in medical domains such as the one in which this work addresses. The focus of works found in the literature mainly depend on the data topology and the usage of extracted relations from analysis. There are melanoma studies focused in the identification of relationships between malignant melanoma and familiar or hereditary tumors (i.e. breast cancer, ovarian cancer, colon cancer, pancreatic cancer) such as in [10]. Other works analyze thousands of genes with the aim of extracting the 'guilty' genes [11, 12] related to the cancer. Anyway, both approaches help experts to be aware and detect melanoma formation in early stages. The main difference between our work and others is that we use SOMEX to help experts to improve their melanoma definition and classification. This improvement will has as a consequence an increment of the precision in melanoma diagnosis.

The idea of using symbolic descriptions for characterizing clusters can be interpreted as a memory organization. In this sense, our approach is similar to Perner's [13] and Abidi's [14] works. Perner proposes to organize the cases following a hierarchy similar to a decision tree where each node c_i is described by a symbolic description (prototype). Each symbolic description subsumes descriptions of all nodes included in the subtree rooted by c_i until reach the leaves that contain the individual cases. Somehow, nodes of that hierarchy could be interpreted as explanations, i.e. why a subset of domain objects (cases) have been grouped under a node. This work relies on the context of case-based reasoning where the main aim is to classify a new problem, therefore prototypes are used

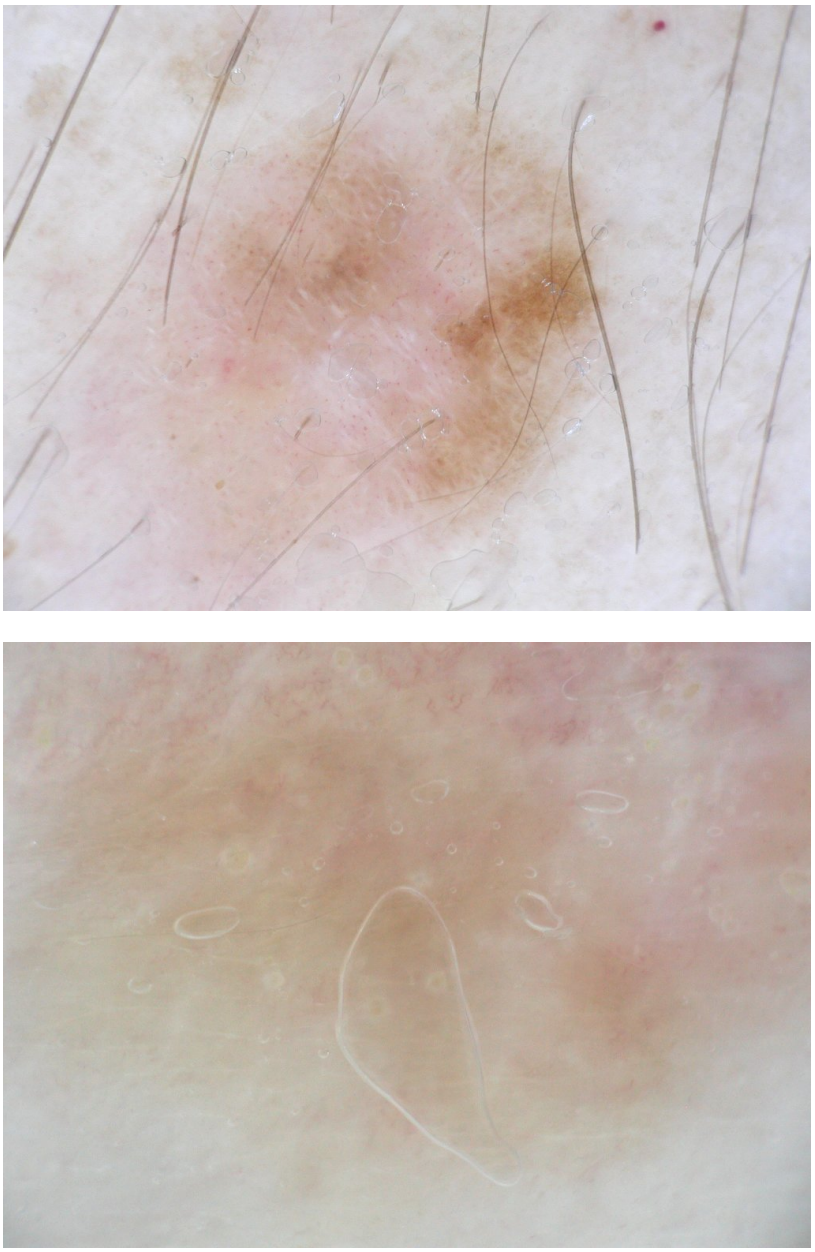

Fig. 6. Melanoma image for <*Obj-11*> and <*Obj-13*> from cluster M_{15} and M_{24} respectively. <*Obj-11*> presents under dermoscopy dots and globules (that are atypical) characteristic of this cluster of lesions, in the absence of all negative features (values 0 in the definition) and also the presence of typical pigment network and typical vessels (may be present in this cluster). <*Obj-13*> presents under dermoscopy only vessels that are atypical, all other criteria are negative.

to select a subset of cases to solve a new problem. In previous works such as [15], we also proposed the use of explanations during the retrieval phase of the case-based reasoning, nevertheless in SOMEX the use of explanations is different. SOMEX does not take into account the class class of cases (i.e. the H_Diagnosis attribute in the melanoma domain), since we assume that these classes could no be accurately defined.

The procedure proposed by Abidi et al [14] is similar to SOMEX because they produce rules that describe objects included in a cluster without using the class information. Firstly, domain objects are clustered according to their similarity, secondly continuous values are discretized, and finally they use rough sets to generate symbolic rules for each cluster. In fact, the explanation generated by SOMEX could also be interpreted as a domain rule (as we suggested in [16]).

The basic difference among SOMEX and the works above is the use of explanations. Perner uses symbolic descriptions to organize the memory of cases with the purpose of achieving a more efficient retrieval. Abidi et al. propose a procedure to obtain symbolic rules from clusters. In SOMEX, explanations are used as a basis for knowledge discovery. By analyzing the explanation of clusters, experts can compare the classification they proposed with the classification obtained by the clustering method. Because of the clustering method do no take into account the class label, differences among both expert and explanations classification give some clues for redefining the classes.

5 Conclusions and Further Work

This paper describes SOMEX and how it can be used for knowledge discovery. SOMEX is a combination of clustering and generalizations, to support dermatologists in discovering knowledge about *in situ* melanomas. The purpose of dermatologists was to define several classes of melanomas and finding dermoscopic features characterizing these classes. SOMEX supported dermatologists in focusing on groups of similar objects and commonalties among them. In particular, dermatologists can analyze the entropy of clusters, i.e. why melanomas that they consider as belonging to different histological classes are actually so similar. Dermatologists can also analyze the relevance of attributes for classification. A particular example is the melanoma *pattern*, considered as a relevant aspect prior to SOMEX application and that results proved that taken in isolation is not a good classifier due to its variability on the objects of a cluster.

As future work we plan to modify some parameters of the clustering in two ways. Firstly we want to confirm the relevance of *pattern* and we plan to weight some of these features in order to highlight the relationship of this feature with others. A second kind of experiments could be focused on enforcing the number of clusters and experimentally determining the best group of melanomas in order to empirically define their histological classes. Finally, from the point of view of the explanations, we could analyze relations between them.

Acknowledgments

We would like to thank the Spanish Government for the support in MID-CBR project under grant TIN2006-15140-C03 and the *Generalitat de Catalunya* for the support under grants 2005SGR-302 and 2007FIC-0976. We would like to thank *Enginyeria i Arquitectura La Salle* of Ramon Llull University for the support to our research group as well.

On the other hand, we also would like to thank the clinicians involved in the confection of the dataset: Dr Paolo Carli (Dermatologist), Dr Vinzenzo di Giorgi (dermatologist), Dr Daniela Massi (dermatopathologist) from the University of Firenze; Dr Josep Malvehy (dermatologist and co-author), Dr Susana Puig (dermatologist and co-author) and Dr Josep Palou (dermatopathologist) from Melanoma Unit in *Hospital Clinic i Provincial de Barcelona*. Part of the work performed by S. Puig and J. Malvehy is partially supported by: *Fondo de Investigaciones Sanitarias* (FIS), grant 0019/03 and 06/0265; Network of Excellence, 018702 GenoMel from the CE.

References

1. Friedman, R.J., Rigel, D.S., Kopf, A.W.: Early detection of malignant melanoma: The role of physician examination and self-examination of the skin. Ca-A Cancer J Clinicians **35** (1985) 130–151
2. Puig, S., Argenziano, G., Zalaudek, I., Ferrara, G., Palou, J., Massi, D., Hofmann-Wellenhof, R., Soyer, H., Malvehy, J.: Melanomas that failed dermoscopic detection: a combined clinicodermoscopic approach for not missing melanoma. Dermatol Surg **33**(10) (2007) 1262–1273
3. Hofmann-Wellenhof, R., Blum, A., Wolf, I., Zalaudek, I., Piccolo, D., Kerl, H., Garbe, C., Soyer, H.: Dermoscopic classification of clark's nevi (atypical melanocytic nevi). Clin Dermatol **20**(3) (2002) 255–258
4. Argenziano, G., Zalaudek, I., Ferrara, G., Hofmann-Wellenhof, R., Soyer, H.: Proposal of a new classification system for melanocytic naevi. Br J Dermatol **157**(2) (2007) 217–227
5. Kohonen, T.: Self-Organizing Maps. 3rd edn. Springer (2000)
6. Armengol, E., Plaza, E.: Bottom-up induction of feature terms. Machine Learning **41**(1) (2000) 259–294
7. Haykin, S.: Neural Networks: A Comprehensive Foundation. 2nd edn. Prentice Hall (1999)
8. Corral, G., Armengol, E., Fornells, A., Golobardes, E.: Data security analysis using unsupervised learning and explanations. In Corchado, E., Corchado, J., Abraham, A., eds.: Innovations in Hybrid Intelligent Systems. Volume 44., Springer-Verlag (2007) 112–119
9. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Second edn. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann (2005)
10. Stefano, P., Fabbrocini, G., Scalvenzi, M., Emanuela, B., Pensabene, M.: Malignant melanoma clustering with some familiar/hereditary tumors. Annals of Oncology **3** (2002) 100

11. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. National Academy Scientific **25**(95) (1998) 1486314868
12. Dang, H., T. Le, T.S., Levy, J.: Integrating database information in microarray expression analyses: Application to melanoma cell lines profiled in the nci60 data set. Journal of Biomolecular Techniques **13** (2002) 199–204
13. Perner, P.: Case-base maintenance by conceptual clustering graphs. Engineering Applications of Artificial Intelligence **9** (2006) 381 – 393
14. Abidi, S.S.R., Hoe, K.M., Goh, A.: Analyzing data clusters: A rough sets approach to extract cluster-defining symbolic rules. In: IDA '01: Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis, London, UK, Springer-Verlag (2001) 248–257
15. Fornells, A., Armengol, E., Golobardes, E.: Explanation of a clustered case memory organization. In: Artificial Intelligence Research and Development. Volume 160., IOS Press (2007) 153–160
16. Armengol, E.: Usages of generalization in cbr. In Weber, R., Richter, M.M., eds.: Case-based Reasoning and Development. Number 4626 in Lecture Notes in Artificial Intelligence, Springer-Verlag (2007) 31–45

Vitae

Dr. Albert Fornells received his Ph.D. degree in Computer Science by Universitat Ramon Llull, Spain, in 2007. He is a member of the Research Group in Intelligent Systems from the university since 2000, and an associate professor since 2003. His research interests include data mining and knowledge discovery, self-organizing maps, case-based reasoning, soft computing, soft-case based reasoning and Artificial Intelligence applied in Health Care (Medicine).

Dr. Eva Armengol received her Ph.D. in Computer Science by the Universitat Politecnica de Catalunya in 1997. She is now a permanent scientist of the Artificial Intelligence Institute (IIIA). Her research has been mainly focused on knowledge representation, machine learning and case-based reasoning. Currently she focuses her research on how to explain the results of learning methods.

Dr. Elisabet Golobardes received her Ph.D. in Computer Science by Universitat Ramon Llull in 1998. She is a member of the Research Group in Intelligent Systems (GRSI) of Enginyeria i Arquitectura La Salle - Universitat Ramon Llull since 1994. Her research interests are mainly focused on Case-Based Reasoning, Soft-Computing, Clustering, Machine Learning, Computer Aided Systems, Artificial Intelligence applied in Health Care (Medicine) and applied in Network Security.

Dr. S. Puig graduated in 1988 at the Faculty of Medicine of Barcelona (Spain) and obtained her specialization diploma in Dermatology and Venereology in 1992 and the doctoral degree in 2000. She is a dermatologist, the director of the research program of the melanoma Unit at the University Hospital Clinic of Barcelona. She is co-director of the Oncology and Dermoscopy groups at the

CILAD. Special research areas are dermoscopy, digital follow-up of melanocytic tumors and genetics of melanoma. She is member of the board of the International Dermoscopy Society, Genomel and EORTIC.

Dr. Josep Malvehy graduated in 1992 at the Faculty of Medicine of Barcelona in Spain and obtained his specialization diploma in Dermatology and Venereology in 1996 and the doctoral degree in 2006. He is the director of the melanoma Unit at the University Hospital Clinic of Barcelona (Spain). His main research field is the melanocytic tumors and melanoma and particularly the study of new diagnostic tools for the evaluation in vivo of skin tumors based in dermoscopy, confocal microscopy and other techniques.