COMPUTER DATABASE 'NEVI' ON ENDANGERMENT BY MELANOMA

ZDZISŁAW S. HIPPE

Department of Computer Chemistry, Rzeszow University of Technology, 6 Powstancow Warszawy Ave.,35–041 Rzeszow, Poland
zshippe@prz.rzeszow.pl

Abstract: The database containing statistically meaningful number of carefully verified cases of nevi pigmentosi (in four categories: benign nevus, blue nevus, suspicious nevus, and melanoma malignant) has been described. Then, various experiments of controlled data mining were performed to get an insight for the new interpretation of the TDS coefficient, broadly used in the initial classification of endangerment by melanoma tumor.

Keywords: nevi pigmentosi–, classification, virtual visualization, TDS

1. Introduction

Data about cases of skin cancer among the population of south–east part of Poland have been completed for several years by Outpatient Advice Center for Dermatology in Rzeszow [1]. These data, carefully verified (inter alia, by extended histological investigations), contained **personal informaton** (<*identity_number*>, <*serial_number*>, <*age*>, <*sex*>), **information about apparent origin of the disease** (<*nevus_origin*>, <*family_inheritance*>, <*exposure_to_sun*>), and **description of melanoid marks on the skin** (<*location*>, <*diameter*>, <*elevation*>, <*roughness*>, <*asymmetry*>, <*border*>, <*structure*>, <*color*>, and <***TDS**–index*>). Each of the registered cases was classified into one of four categories of nevi pigmentosi: benign nevus, blue nevus, suspicious nevus, or melanoma malignant.

The **TDS**–index mentioned here (**TDS** — **T**otal **D**ermatoscopic **S**core), according to German sources [2], may be conveniently used as a measure for non–invasive diagnosing of *melanocytic nevi and melanoma malignant*. The numerical value of this index is computed using **ABCD**–rules, where **A** — stands for asymmetry, **B** — for border, **C** — for color, and **D** — for diversity of the structural

changes on the skin:

$$\begin{aligned} \mathbf{TDS} = \quad & \mathbf{1.3} \cdot \mathbf{Asymmetry} + \\ & \mathbf{0.1} \cdot \mathbf{Border} + \\ & \mathbf{0.5} \cdot \mathbf{Color} + \\ & \mathbf{0.5} \cdot \mathbf{Density\ of\ structure} \end{aligned}$$

Braun–Falco [2] assigned the following values of the **TDS**–index to the previously mentioned categories of melanoid marks:

<benign_nevus>, *<blue_nevus>*	**1.0**	$\leq$	**TDS**	$<$	**4.7**
<suspicious_nevus>	**4.8**	$\leq$	**TDS**	$\leq$	**5.45**
<melanoma_malignant>			**TDS**	$>$	**5.45**

However, the detailed inspection of literature data revealed, that the designation of the **TDS**–values to a given class of melanoma nevi was executed using traditional (hence, obsolete) statistical analysis. Moreover, the selection of features (symptoms) for diagnosing of the disease was based on a limited population of cases (about 175 registered cases altogether). The entire data set (according to [2]) used for finding statistical regularities contained in fact only $175/2 \approx 86$ cases, because the basic set was split evenly into a training set (86 cases), and a prediction set (86 cases). The results of our preliminary examinations along these lines [3] forced us to start an extended research, aimed at a development of a new measure of endangerment by melanoma. Also, we would like to search for statistically strong dependencies: cause(s) $\Rightarrow$ consequence(s), allowing an early and non–invasive identification of the kind of tumor. It seemed necessary, therefore, to elucidate the following problems: (i) how the hierarchy of importance of coefficients in the **ABCD**–rules should be changed for more reliable classification of skin marks? (ii) would the assignment of other values to these coefficients improve diagnosing of melanoma endangerment? (iii) would other symptoms of melanoma permit to increase the identification certainty of the tumor?

To achieve these goals a database 'NEVI' and some derived databases were created. In general, attempts to solve (i) — (iii) were made applying selected machine learning and data mining methods, using licensed tools [4–6] and in–house developed computer programs for generating decision rules [7].

2. General information about created databases

The main database ('NEVI'), critically reviewed in [8], was derived at the Department of Computer Chemistry (Rzeszow University of Technology) from the source data collected by Bajcar, Grzegorczyk, and others [1]. It was prepared with a format required by a program for development of decision trees [4]. Some copies of the base, with evenly distributed classes of melanoid marks were generated with Microsoft Excel format to apply specialized software for searching hidden regularities and knowledge structures (e.g. searching the most suitable — from the point of view of diagnosing certainty — combination of principal symptoms of the

disease). Additionally, the main base was transformed to ASCII–file, used then for visual inspection of available data and for disclosing the spatial distribution of the investigated cases. This part of the research was required for critical evaluation of selected symptoms, and for estimation of their hierarchy and discriminating strength. As visualization engine, **V**irtual **V**isualization **T**ools (**VVT1** for supervised learning, and **VVT2** for unsupervised learning) elaborated recently in our group [9], were applied. These research tools permit to combine case–based reasoning with application of iterative **SAHN**–procedures [10], supported recently by the capability of controlled visualization of multidimensional data ($d_{max.} = 70$) in the 3D–format. Also, an extended version of the discussed database was developed with special structures for loading of digital images of melanocytic nevi. This version of the database will be used in further steps of the research, based on the application of our own in–house developed software for processing of domain–oriented knowledge, particularly for processing of knowledge images [7].

Currently the database 'NEVI' contains a statistically significant number of cases (250), with equal distribution of all classes of melanoid marks mentioned in Section 1. All cases are described by means of 16 descriptive attributes; some of them are intentionally redundant. For example, the <***TDS–index***> conveys in fact the same information as the union of the following four attributes: <*asymmetry*>, <*border*>, <*color*>, and <*structure*>. Thus, the solution space in the research has 17–dimensions.

3. Virtual visualization of data: a problem of the TDS–index

The results of our initial experiments on searching of hidden regularities in the database 'NEVI' were obtained with the use of a programming tool [4] for the development of decision trees. These results, owing to the restricted allowance for the size of papers, are not presented here. However, it may be stated that the analysis of various decision trees obtained led to an important conclusion about location of the selected attributes in particular nodes of a tree. Namely, in almost all cases tested the attribute <*color*> was placed in the root of each tree. Moreover, estimation of the location (within the decision tree) of other attributes used for calculation of the **TDS**–index, allowed to appraise the hierarchy of importance of successive attributes (symptoms) applied for classification of melanocytic nevi.

Referring to our preliminary experiments it may be assumed that the most relevant attribute to classify melanoid marks on the skin seems to be <*color*>. Two other attributes, commonly used in calculation of the **TDS**–index, namely <*asymmetry*> and <*border*> have smaller diagnostic power. They also display similar importance in the classification of the investigated cases. On the other hand, the attribute <*structure*>, scored by Braun–Falco [2] with coefficient on the level 0.5, according to our observation has negligible influence on the classification process. These findings pointed out that coefficients currently accepted for calculation of the **TDS**–index from **ABCD**–rules have been erroneously selected. It may be expected that assignment of other numerical values carefully elucidated

(e.g. trying to preserve the discovered hierarchy of attributes) to these coefficients, would cause more effective discrimination of the investigated cases, hence, more reliable diagnosing of melanoma tumor. This conclusion was confirmed directly or indirectly in further steps of our research, especially by virtual visualization of data stored in the database. Application of this technique was in fact aimed at the inspection of available data, to disclose the spatial distribution of the investigated cases. The results obtained along these lines are presented and discussed in the further part of the text.

The main stream of the experiments with virtual visualization of data contained in the database followed very careful comparison of the classification processes, conducted partially in the form of supervised learning, and partially as unsupervised learning. The programming tools used allowed the selection of a required combination of: (i) metrics for evaluation of the similarity of cases (Euclidean, City Block, Tshebyshev's), with (ii) any of eight procedures designed for searching of clusters within data. These procedures settled the main body of the **VVT1** and **VVT2** programs, both based on modified **SAHN**–algorithms (**S**equential, **A**gglomerative, **H**ierarchical and **N**on–overlapping) [10]. In this series of experiments we went beyond the standard **TDS**–formula, and focused our attention on a *combination* of particular symptoms characteristic for the investigated disease. Complete results are shown in Table 1. Each of the five columns of the table contains values of distribution error, computed for unsupervised learning with various measures of similarity (distance) among single–class cases. First, it was confirmed that the four descriptive attributes defined in the standard **TDS**–formula are not suitable for clear and errorless disclosing of all investigated cases. Inclusion of an additional attribute (*<elevation>*) does not improve the capability to separate better the visualized clusters. Similarly, classification based on standard **ABCD**–rules and two or three additional attributes (the last column) made the spatial distribution even worse. However, the data mining process based on virtual visualization method revealed that the combination of *<**ABCD**–rules>* and *<roughness>* as the descriptive attributes, seems to be most promising.

4. Summary

Advanced data mining with the use of virtual visualization technique executed in the form of unsupervised learning, often brings out ideas that we would not normally arrive at. For example, examination of data clusters developed pointed out clearly that the recognition of a given class of melanoma nevus with high accuracy and reliability — contrary to statements published by Braun–Falco [2] — is extremely difficult, and may now be possible with considerable error, which equals roughly 40%. Currently, the database 'NEVI' is stepwise extended by new cases, belonging mainly to the class *<benign nevus>* or *<suspicious nevus>*. The influence of values of coefficients (used in **ABCD**–rules for the calculation of the **TDS**–index) on reliability of non–invasive diagnosing of melanoma tumor is now being extensively tested. These investigations, although still in an early stage, suggest a complete

Table 1. Results of mining various bases derived from 'NEVI' database

Metrics: Euclidean (a), City Block (b), Tshebyshev (c)	Method	Descriptive attributes in a given database														
		<*ABCD-rules*>			<*ABCD-rules*> <*elevation*>			<*ABCD-rules*> <*roughness*>			<*ABCD-rules*> <*elevation*> <*roughness*>			<*ABCD-rules*> <*elevation*> <*roughness*> <*diameter*>		
		Distribution error [%]														
		(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
	Single Average	65.0	72.5	*	88.8	77.5	*	55.0	77.5	*	63.8	97.5	*	75.0	82..5	*
	Group Average	62.5	75.0	92.5	96.3	82.5	92.5	61.3	72.5	91.3	95.0	43.8	63.8	65.0	61.3	82.5
	Weighted Average	75.0	50.0	35.0	47.5	71.3	35.0	85.0	85.0	63.8	76.3	70.0	58.8	66.3	82.5	77.5
	Complete Linkage	68.8	85.0	51.3	91.3	50.0	66.3	51.3	41.3	63.8	97.5	82.5	77.5	70.0	76.3	68.8
	Unweighted Centroid	97.5	41.3	95.0	82.5	58.8	83.8	63.8	85.0	78.8	82.5	87.5	92.5	73.8	66.3	82.5
	Weighted Centroid	97.5	47.5	*	63.8	72.5	*	63.8	85.0	*	56.3	70.0	*	70.0	86.3	*
	Minimum Variance	77.5	63.8	73.8	55.0	75.0	90.0	77.5	32.5	95.0	92.5	61.3	93.8	86.3	70.0	48.8
	FSS: $\alpha = 0.3$	78.8	67.5	*	78.8	76.3	*	57.5	85.0	60.0	73.8	75.0	80.0	70.0	65.0	*
	FSS: $\alpha = 0.6$	60.0	36.3	76.3	65.0	80.0	86.3	85.0	37.5	95.0	68.8	75.0	88.8	73.8	72.5	81.3
	FSS: $\alpha = 0.9$	70.0	69.8	77.5	67.5	52.5	91.2	55.0	97.5	38.8	57.5	46.3	96.3	36.3	57.5	52.5
Arithmetic mean (from best results):		43.8			44.2			40.9			52.9			48.8		

* — denotes invalid combination of metrics and method

change of the general concept of evaluation of color for **ABCD**–rules. Namely, it may be concluded that the global symptom *<color>* should be split into six structured attributes, representing all real colors met in melanoma nevi, and additionally having assigned different fractional coefficients.

References

[1] S. Bajcar, L. Grzegorczyk, *Endangerment by skin cancer among population of south–east part of Poland*, Hospital #1, Res. Report, Rzeszow 1997–1999

[2] O. Braun–Falco, W. Stolz, P. Bilek, T. Merkle and M. Landthaler, *Das Dermatoskop. Eine Vereinfachung der Auflichtmikroskopie von pigmentierten Hautveränderungen*, Hautartzt 1990 (40) 131–135

[3] P. Krzyś and G. Ryzner, Diploma Thesis, Department of Computer Chemistry, University of Technology, Rzeszow 1999

[4] W. Hapgood, 1stClass, Programs in Motions, Inc., Wayland (MA) 1989

[5] DataMind Professional, DataMind Corporation, San Mateo (CA) 1998

[6] NeuralWorks Predict, NeuralWare, Inc., Pittsburgh (PA) 1998

[7] Z.S. Hippe, *Design and application of an intelligent information system SCANKEE for solving selected chemical problems*. Computer Chem. 1998 (22,1) 133–140

[8] Z.S. Hippe, *Data Mining in Medical Diagnosis*, Intern. Conference on Computers in Medicine, Lodz (Poland), 23–25 September 1999

[9] M. Mazur, *Virtual Visualization Tools*. In: Research Report for the State Committee for Scientific Research (Warsaw), Grant No. 8 T11C 004 09, Rzeszow 1999

[10] Z.S. Hippe, *New Data Mining Strategy Combining SAHN Visualization and Case–Based Reasoning*, Proc. Joint Conference on Information Sciences (JCIS'98), Research Triangle Park (NC), 23–29 October 1998, Vol. II, pp. 320–322