Optimization of the ABCD Formula for Melanoma Diagnosis Using C4.5, a Data Mining System

Ron Andrews
Department of Electrical Engineering and Computer Science,
University of Kansas, Lawrence, KS 66045, USA

Stanislaw Bajcar
Regional Dermatology Center, 35-310 Rzeszow, Poland

Jerzy W. Grzymala-Busse
Department of Electrical Engineering and Computer Science,
University of Kansas, Lawrence, KS 66045, USA
and
Institute of Computer Science
Polish Academy of Sciences, 01-237 Warsaw, Poland

Zdzislaw S. Hippe
Department of Expert Systems and Artificial Intelligence,
University of Information Technology and Management, 35-225 Rzeszow, Poland

Chris Whiteley
Department of Electrical Engineering and Computer Science,
University of Kansas, Lawrence, KS 66045, USA

Abstract. Our main objective was to improve the diagnosis of melanoma by optimizing the ABCD formula, used by dermatologists in melanoma identification. In our previous research, an attempt to optimize the ABCD formula using the LEM2 rule induction algorithm was successful. This time we decided to replace LEM2 by C4.5, a tree generating data mining system. The final conclusion is that, most likely, for C4.5 the original ABCD formula is already optimal and no further improvement is possible.

Keywords: Rough set theory, data mining, rule induction, decision tree generation, C4.5 system, diagnosis of melanoma.

1 Introduction

The number of diagnosed cases of melanoma, one of the most dangerous skin cancers, is increasing. Thus any improvement of melanoma diagnosis is crucial to save human lives. Nowadays melanoma is routinely diagnosed with help of the so-called ABCD formula (A stands for Asymmetry, B for border, C for color, and D for diversity of structure) [2, 12]. Results of

successful optimization of the ABCD formula, using the LEM2 rule induction algorithm (Learning from Example Module, version 2), a component of the data mining system LERS (Learning from Examples using Rough Sets) [4, 5] were reported in [1, 3, 6, 7]. Rough set theory was initiated in 1982 [9, 10].

In this paper we report results on yet another attempt to optimize the ABCD formula, this time using a different, well-known data mining system C4.5 [11]. The data on melanoma, consisting of 410 cases, were collected at the Regional Dermatology Center in Rzeszow, Poland [8]. In our current research we evaluated all attributes from this data set, one attribute at a time, checking their significance for diagnosis using the number of errors determined by ten-fold cross validation and C4.5. Then we used sequences of 30 experiments of ten-fold cross validations, also using C4.5, in our attempt to look for the optimal ABCD formula. Note that in previous research [1, 3, 6, 7], using LERS, a substantial improvement in melanoma diagnosis was accomplished. However, this time our final conclusion is that the original ABCD formula, used for diagnosis with C4.5, is most likely, already optimal. Moreover, the sequence of 30 different experiments of ten-fold cross validation was not sufficient. This conclusion was reached using 300 and 3,000 experiments of ten-fold cross validation.

2 ABCD formula

In diagnosis of melanoma an important indicator is TDS (Total Dermatoscopic Score), computed on the basis of the ABCD formula, using four variables: *Asymmetry*, *Border*, *Color* and *Diversity*. The variable *Asymmetry* has three different values: *symmetric spot*, *one axial symmetry*, and *two axial symmetry*. *Border* is a numerical attribute, with values from 0 to 8. A lesion is partitioned into eight segments. The border of each segment is evaluated; the sharp border contributes 1 to *Border*, the gradual border contributes 0. *Color* has six possible values: *black*, *blue*, *dark brown*, *light brown*, *red* and *white*. Similarly, *Diversity* has five values: *pigment dots*, *pigment globules*, *pigment network*, *structureless areas* and *branched streaks*. In our data set *Color* and *Diversity*

were replaced by binary single-valued variables. The TDS is traditionally computed using the following formula (known as the ABCD formula):

$$TDS = 1.3 * Asymmetry + 0.1 * Border + 0.5 * \Sigma Colors + 0.5 * \Sigma Diversities,$$

where for *Asymmetry* the value *symmetric spot* counts as 0, *one axial symmetry* counts as 1, and *two axial symmetry* counts as 2, Σ Colors represents the sum of all values of the six color attributes and Σ Diversities represents the sum of all values of the five diversity attributes.

3 C4.5 testing of single attributes

The significance of individual attributes, or testing the importance of specific attributes as part of the ABCD formula, was conducted by changing the coefficient associated with an attribute from 0 to 2, by 0.05 increments, and keeping values of all twelve remaining coefficients equal to one.

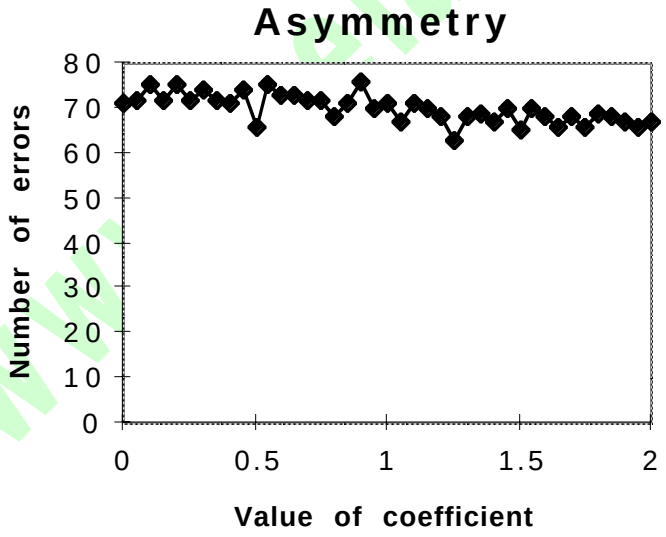

Figure 1. Number of errors for Asymmetry

Therefore, the original data set was transformed into a new data set, without TDS, and with values of all attributes, except one attribute, equal to one. For all attributes, except *Border*, the total number of errors, a result of ten-fold cross validation, was between 70 and 80. Note the total

number of errors, again determined by ten-fold cross validation for the original data set without TDS (with values of all remaining attributes unchanged), was equal to 85. A typical graphic, for *Asymmetry*, is presented in Figure 1.

For Border the number of errors was between 12 and 73 when its coefficient was between 0 and 1, and then leveled out to between 70 and 80 when its coefficient was between 1 and 2, see Figure 2. Intuitively, this test shows that when the coefficient associated with *Border* is much smaller than all other coefficients the number of errors is smaller. Obviously, creators of the ABCD formula were familiar with this fact since in the ABCD formula the coefficient for *Border* is much smaller than for other attributes.

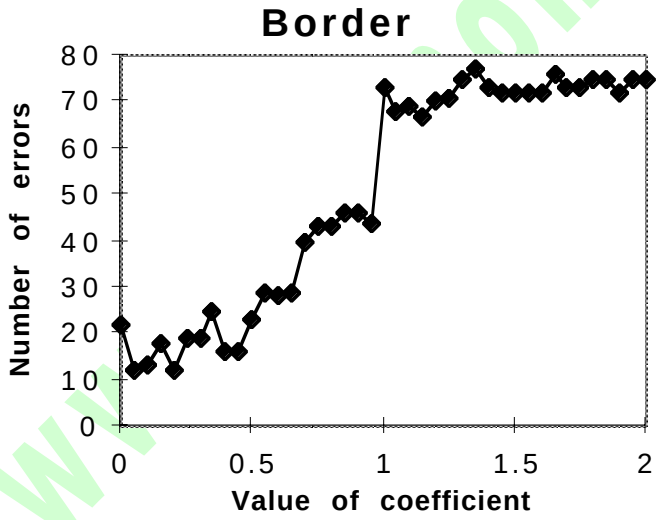

Figure 2. Number of errors for Border

4. Main Experiments

The most important performance criterion for all methods of data mining is the total number of errors. To discover the error number we used ten-fold cross validation: all cases were randomly re-ordered, and then the set of all cases was divided into ten mutually disjoint subsets of approximately equal size. For each subset, all remaining cases were used for training, i.e., for rule induction, while the subset was used for testing. Thus, each case was used nine times for training

and once for testing. Note that using different re-orderings of cases causes slightly different error numbers. The original C4.5 system is not equipped with any way to randomly re-order a data set, so we added a mechanism to accomplish this task.

Previous experiments attempted at looking for the optimal ABCD formula while using LEM2, an algorithm of LERS, were successful [1, 3, 6, 7]. Our current experiments were aimed towards the same goal: to find the optimal ABCD formula, however, this time we used the C4.5 system. Similarly as in [1, 3, 6, 7], we assumed that the optimal ABCD formula, for computing a new TDS, should be also a linear combination of 13 attributes:

$$\text{new\_TDS} = c_1 * \text{Asymmetry} + c_2 * \text{Border} + c_3 * \text{Color\_black} + c_4 * \text{Color\_blue} + c_5 * \text{Color\_dark\_brown} + c_6 * \text{Color\_light\_brown} + c_7 * \text{Color\_red} + c_8 * \text{Color\_white} + c_9 * \text{Diversity\_pigment\_dots} + c_{10} * \text{Diversity\_pigment\_globules} + c_{11} * \text{Diversity\_pigment\_network} + c_{12} * \text{Diversity\_structureless\_areas} + c_{13} * \text{Diversity\_branched\_streaks}.$$

Our objective was to find optimal values for coefficients $c_1, c_2,..., c_{13}$. The criterion of optimality was the smallest total number errors for sequences of 30 ten-fold cross validations with different re-ordering of examples in the data set.

Thus for each vector $(c_1, c_2,..., c_{13})$ the corresponding new_TDS was computed, the sequence of 30 re-orderings of the data set was performed, and then for each new data set ten-fold cross validation was used for the evaluation of the number of errors.

Since the original ABCD formula yielded relatively small number of errors, we set the base value of coefficients to the same value as in the original ABCD formula. Then we run sequences of 30 experiments of ten-fold cross validation for vectors $(c_1, c_2,..., c_{13})$ of coefficient values close to original, with increments of 0.01, running altogether over 73,000 experiments.

The smallest error obtained from such a sequence of 30 ten-fold cross validation experiments indicated the optimal choice of $(c_1, c_2,..., c_{13})$. A special script was created to compute the new_TDS given ranges for all 13 coefficients $c_1, c_2,..., c_{13}$, see Table 1. Due to

computational complexity, not all combinations of coefficients that are implied by Table 1 were tested.

Table 1. Explored coefficient ranges for thirteen attributes from the melanoma data set

Attribute	Tested Range
Asymmetry	1.28 – 1.38
Border	0.02 – 0.12
Color_black	0.41 – 0.59
Color_blue	0.42 – 0.53
Color_dark_brown	0.41 – 0.55
Color_light_brown	0.41 – 0.59
Color_red	0.51 – 0.59
Color_white	0.50 – 0.50
Diversity_pigment_dots	0.42 – 0.57
Diversity_pigment_globules	0.41 –0.59
Diversity_pigment_network	0.44 – 0.50
Diversity_structureless_areas	0.41 – 0.59
Diversity_branched_streaks	0.52 – 0.58

During testing with C4.5 using ten- fold cross-validation, we discovered that certain orderings of the data set could cause the system to core dump. This fault did not seem to have a single definitive cause, but during initial testing this issue was a cause for concern with respect to automating the system. Not wanting to spend time debugging the problem in the decision tree generation system, we opted to work around it by computing averages of successful runs of C4.5.

Since the total number of errors for trees was larger than the total number of errors for rules, we used the latter as a guide for identification the best ABCD formula. The best results were obtained from the following formula

$$\begin{aligned} \text{new\_TDS} = {} & 1.3 * \text{Asymmetry} + 0.03 * \text{Border} + 0.5 * \Sigma\,\text{Colors} + \\ & 0.5 * \text{Diversity\_pigment\_dots} + 0.5 * \text{Diversity\_pigment\_globules} + \\ & 0.47 * \text{Diversity\_pigment\_network} + 0.5 * \text{Diversity\_structureless\_areas} + \\ & 0.5 * \text{Diversity\_branched\_streaks}. \end{aligned}$$

Results of running our experiments are presented in Tables 2-3.

Using the well-known statistical test for the difference between two averages, with the level of significance specified at 0.05, initially we concluded that new_TDS was better than the original, mostly due to small standard deviations. However, with a difference between averages being so small, we decided to run additionally 300 and then 3,000 experiments to test the same hypothesis. Surprisingly, the same test for the difference between two averages, with the same level of significance equal to 0.05, yielded quite opposite conclusions: the difference between the new_TDS and original one was not significant. Since the test with more experiments is more reliable, our final conclusion is that there is no significant difference in performance between the new_TDS and original.

Table 2. Number of errors

TDS	Rules	Unpruned trees	Pruned trees
Original TDS	9	9	9
New TDS	11	13	14
No TDS	85	89	88

Table 3. Average umber of errors

TDS	Length of a sequence of 10-fold cross validations	Number of errors	Standard deviation
Original TDS	30	6.20	0.41
New TDS	30	5.97	0.18
Original TDS	300	8.51	1.55
New TDS	300	8.45	2.00
Original TDS	3,000	8.51	9.32
New TDS	3,000	8.45	10.13

The pruned decision tree generated by C4.5 from the data with TDS computed by the original ABCD formula is presented in Figure 3. As a result of pruning of that tree by C4.5, only two attributes are used, TDS and Color_blue, see Figure 4.

```
TDS <= 4.8 :
|   C_BLUE = 1: Blue_nevus (66.0/1.0)
|   C_BLUE = 0:
|   |   D_b_STREAKS = 1: Benign_nev (71.0)
|   |   D_b_STREAKS = 0:
|   |   |   D_PIGM_DOTS = 1: Benign_nev (48.0/1.0)
|   |   |   D_PIGM_DOTS = 0:
|   |   |   |   D_PIGM_NETW = 1: Benign_nev (8.0)
|   |   |   |   D_PIGM_NETW = 0:
|   |   |   |   |   C_WHITE = 0: Blue_nevus (5.0/1.0)
|   |   |   |   |   C_WHITE = 1: Benign_nev (2.0)
TDS > 4.8 :
|   TDS <= 5.4 : Suspicious (81.0)
|   TDS > 5.4 :
|   |   C_BLUE = 1: Malignant (16.0/1.0)
|   |   C_BLUE = 0:
|   |   |   C_RED = 1: Malignant (41.0)
|   |   |   C_RED = 0:
|   |   |   |   C_WHITE = 1: Malignant (17.0)
|   |   |   |   C_WHITE = 0:
|   |   |   |   |   ASYMMETRY = 0: Malignant (0.0)
|   |   |   |   |   ASYMMETRY = 2: Malignant (5.0)
|   |   |   |   |   ASYMMETRY = 1:
|   |   |   |   |   |   TDS > 5.6 : Malignant (4.0)
|   |   |   |   |   |   TDS <= 5.6 :
|   |   |   |   |   |   |   TDS <= 5.5 : Malignant (2.0)
|   |   |   |   |   |   |   TDS > 5.5 : Suspicious (3.0/1.0)
```

Figure 3. Unpruned decision tree generated by C4.5 from the data set with TDS computed using the original ABCD formula

```
TDS <= 4.8 :
|   C_BLUE = 0: Benign_nev (134.0/7.3)
|   C_BLUE = 1: Blue_nevus (66.0/2.6)
TDS > 4.8 :
|   TDS <= 5.4 : Suspicious (81.0/1.4)
|   TDS > 5.4 : Malignant (88.0/5.0)
```

Figure 4. Pruned decision tree generated by C4.5 from the data set with TDS computed using the original ABCD formula

5 Conclusions

This paper presents an attempt to find the optimal ABCD formula that is widely used by physicians to diagnose melanoma. Our assumption was that the diagnosis will be supported by C4.5, a data mining system. Therefore, all experiments aimed at optimizing the ABCD formula were conducted

using C4.5. First, all thirteen attributes from our data set describing melanoma were tested for significance, with a total number of errors determined by ten-fold cross validation using C4.5. The only conclusion was that the coefficient, associated with the attribute *Border* should be small. Our main experiments were designed to look for an optimal ABCD formula while preserving the original form of linear combination of attributes, characteristic for the ABCD formula. This optimization was conducted by applying many thousands of vectors of values of the thirteen coefficients and processing each such vector by a sequence of 30 experiments of ten-fold cross validation, each with a different re-ordering of the data sets. As a result, the optimal ABCD formula was found. Nevertheless, after additional experiments, running ten-fold cross validation for the data containing TDS computed by the original ABCD formula and data containing TDS computed by the optimal ABCD formula 300 times and 3,000 times, each ten fold cross-validation with different re-ordering of both data sets, we observed that —statistically—there is no significant difference in the total number of errors for both formulas, with the level of significance = 0.05. Thus, our final conclusion is that, most likely, for C4.5 the original ABCD formula is already optimal.

References

[1] Alvarez, A., Brown, F. M., Grzymala-Busse, J. W., and Hippe, Z. S.: Optimization of the ABCD formula used for melanoma diagnosis. Proc. of the IIPWM'2003, Int. Conf. On Intelligent Information Processing and WEB Mining Systems, Zakopane, Poland, June 2–5, 2003, 233–240.

[2] Friedman, R. J., Rigel, D. S., and Kopf, A. W.: Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. *CA Cancer J. Clin*. 35, 1985, 130–151.

[3] Grzymala-Busse, J. P., Grzymala-Busse, J. W., and Hippe Z. S.: Melanoma prediction using data mining system LERS. Proceeding of the 25th Anniversary Annual International

Computer Software and Applications Conference COMPSAC 2001, October 8–12, 2001, Chicago, IL, 615–620.

[4] Grzymala-Busse, J. W.: LERS—A system for learning from examples based on rough sets. In *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory*. Slowinski, R. (ed.), Kluwer Academic Publishers, Dordrecht, Boston, London, 1992, 3–18.

[5] Grzymala-Busse J. W.: A new version of the rule induction system LERS. *Fundamenta Informaticae* 31 (1997), 27–39.

[6] Grzymala-Busse J. W. and Hippe Z. S.: Postprocessing of rule sets induced from a melanoma data set. Proc. of the COMPSAC 2002, 26th Annual International Conference on Computer Software and Applications, Oxford, England, August 26–29, 2002, 1146–1151.

[7] Grzymala-Busse J. W. and Hippe Z. S.: A search for the best data mining method to predict melanoma. Proceedings of the RSCTC 2002, Third International Conference on Rough Sets and Current Trends in Computing, Malvern, PA, October 14–16, 2002, Springer-Verlag, 538–545.

[8] Hippe, Z. S.: Computer database NEVI on endargement by melanoma. *Task Quarterly* 4, 1999, 483–488.

[9] Pawlak, Z.: Rough Sets. *International Journal of Computer and Information Sciences*, 11, 1982, 341–356.

[10] Pawlak, Z.: Rough Sets. *Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht, Boston, London, 1991.

[11] Quinlan, J. R.: *C*4.5*: Programs for Machine Learning, Morgan Kaufmann Publishers*, San Mateo, CA, 1988.

[12] Stolz, W., Braun-Falco, O., Bilek, P., Landthaler, A. B., Cogneta, A. B.: *Color Atlas of Dermatology*, Blackwell Science Inc., Cambridge, MA, 1993.