SUBMITTED ARTICLES

SIGNIFICANCE OF CONDITION ATTRIBUTES IN CHILD WELL-BEING ANALYSIS

Ewa Adamus, Przemysław Klęsk, Joanna Kołodziejczyk, Marcin Korzeń, Andrzej Piegat, Marcin Pluciński

Faculty of Computer Science and Information Systems
West Pomeranian University of Technology
Żołnierska 49, 71-062 Szczecin, Poland
{eadamus, pklesk, jkolodziejczyk, mkorzen, apiegat, mplucinski}@wi.zut.edu.pl

Abstarct: *In this study the significance of attributes in child well-being is presented. The main goal was to find features most specific for child-well being evaluation in Poland. The dataset was obtained from a survey based on a special questionnaire. To select important attributes three filter for individual attribute rank were used χ^2, information gain and relief attribute evaluator and one filter-subset selector based on rough set theory. In the article the dataset is described in details. All the attributes are named, divided by category and for each a domain is given. Then methods of attribute selection applied in experiments are presented. Finally results on selecting attributes relevant for child well-being are discussed.*

Keywords: *Child well-being quality measurement, attribute selection.*

1. INTRODUCTION

The most common reasons for using feature selection from a dataset is preparing data for learning methods. Those methods model data dependencies and huge amount of attributes can detract model quality and accuracy. The attribute selection used before learning results in lower model dimensionality, can remove noise from data and search for correlation in attributes. Another reason is important data characteristics obtainment by ranking attributes according to their significance in predicting a decision [3,4].

Two taxonomies are used for feature selection methods. The first distinguishes techniques due to the nature of the metric used to evaluate importance of attributes and calls them “filter” and “wrapper”. Wrappers use learning algorithms to evaluate attributes. Filters are independent of the learning method and use general information from a dataset. The second taxonomy divides algorithms into those which evaluate (ranks) individual attributes and those which evaluate subset of attributes [4].

In this study the significance of attributes in child well-being is presented. The main goal was to find features most specific for child-well being evaluation in Poland. The dataset was obtained from a survey based on a special questionnaire [1,6]. A list of 81 attributes capable of influencing childhood was created and 557 records were collected. To select important attributes three filter for individual attribute rank were used χ^2, information gain and relief attribute evaluator and one filter-subset selector based on rough set theory.

In the article the dataset is described in details. All the attributes are named, divided by category and for each a domain is given. Then methods of attribute selection applied in experiments are presented. Finally results on selecting attributes relevant for child well-being are discussed.

2. DATASET ON CHILD WELL-BEING

The presented research is a part of a project on finding quality indicators of child well-being. The main idea was to gather data from which this information can be extracted by means of different data mining methods. Therefore a survey about childhood was performed. Each respondent answered 26 questions. The questionnaire was conducted mainly among students in Western Pomarania universities. The survey gathered 557 records.

There were different types of questions resulting in numerical, nominal and binary answers. As a result a dataset was obtained with 81 condition and 4 decision attributes.

Usually raw data need some preliminary processing which results in unified and better structure. Data cleaning, integration, discretisation are the most common initial process.

Mistakes or omissions during filling the questionnaire caused incomplete records. Dealing with missing values in a dataset is a part of data cleaning process. The dataset about childhood contained missing values for some attributes not exceeding 1% of the dataset size. As the number of records was not big even incomplete records were saved and missing values were replaced. Numerical attributes were substituted with the mean value and the most frequent value replaced an empty cell for nominal attributes. As preliminary experiments have shown, filling attributes in this way had practically no real influence on realised significance analysis.

All numerical attributes were discretised for two reasons: attribute standarisation and better learning algorithms performance. Discretisation was performed to ensure equal sample frequency for each nominal attribute value. The final list of 81 condition and 4 decision attributes is presented in Tablele 1.

Tablele 1 List of 81 condition and 4 decision attributes

No	Attribute name	Attribute values
1.	sex	F, M
2.	age	18-20, 21-22, 23-24, 25 and over
3.	home location	village, small city, medium city, big city
4.	siblings	0, 1, 2, 3 and more
5.	family	mother and father, only mother, only father, foster family, orphanage
6.	father's age at birth	18-24, 25-27, 28-31, 32 and over
7.	mother's age at birth	18-23, 24-26, 27-30, 31 and over
8.	house	very small, small, medium, large, very large
	father's profession:	
9.	workman	yes, no
10.	office worker	yes, no
11.	freelancer	yes, no
12.	disability pensioner	yes, no
13.	pensioner	yes, no
14.	housekeeping	yes, no
15.	unemployed	yes, no
	mother's profession:	
16.	workman	yes, no
17.	office worker	yes, no
18.	freelancer	yes, no
19.	disability pensioner	yes, no
20.	pensioner	yes, no
21.	housekeeping	yes, no
22.	unemployed	yes, no
23.	job problems	very often, sometimes, rarely, never
	childcare:	
24.	kindergarten	yes, no
25.	private kindergarten	yes, no
26.	nanny	yes, no
27.	family	yes, no
28.	friends, acquaintances	yes, no
29.	none	yes, no
30.	medical care	rare, medium, often
31.	medical care quality	very weak, weak, neutral, good, very good
32.	private medical care	never, rare, often
33.	proportion of time spent with parents	only mother, more often with mother, equally, more often with father, only father
	activities with father:	
34.	reading	yes, no
35.	painting	yes, no
36.	playing	yes, no
37.	watching TV	yes, no
38.	watching films	yes, no
39.	going to cinema	yes, no
40.	watching sport games	yes, no
41.	sport activities	yes, no
42.	listening to music	yes, no
43.	walks, picnic	yes, no
44.	playing music	yes, no
45.	playing board games	yes, no
46.	playing computer games	yes, no
47.	learning	yes, no
48.	sharing a hobby	yes, no
	activities with mother:	
49.	reading	yes, no
50.	painting	yes, no
51.	playing	yes, no

52.	watching TV	yes, no
53.	watching films	yes, no
54.	going to cinema	yes, no
55.	watching sport games	yes, no
56.	sport activities	yes, no
57.	listening to music	yes, no
58.	walks, picnic	yes, no
59.	playing music	yes, no
60.	playing board games	yes, no
61.	playing computer games	yes, no
62.	learning	yes, no
63.	sharing a hobby	yes, no
64.	number of books at home	few, many, a lot
	afterschool activities:	
65.	sports	yes, no
66.	playing music	yes, no
67.	arts	yes, no
68.	dancing	yes, no
69.	studying foreign language	yes, no
70.	additional computer classes	yes, no
71.	mathematics and sciences	yes, no
72.	swimming pool	yes, no
73.	other	yes, no
74.	who chose activities?	parents, parents and me, only me
75.	pets	never, rarely, often
76.	school quality	very poor, poor, average, good, excellent
77.	school safety	very dangerous, dangerous, neutral, safe, very safe
78.	district safety	very dangerous, dangerous, neutral, safe, very safe
79.	camps	never, rarely, medium, often, very often
80.	contact with friends	never, rarely, medium, often, very often
81.	contact with family	never, rarely, medium, often, very often
82.	intensity of education	1 … 5
83.	intensity of entertainment	1 … 5
84.	health and physical fitness	1 … 5
85.	safety and living conditions	1 … 5

All 81 condition attributes are subject to attributes significance analysis. In the study, the ranking of most influential features according to each decision attribute is done separately. Using different feature selection methods various rankings can be obtained. A subset of attributes common for all rankings is the final experimental result.

3. ANALYSIS OF CONDITION ATTRIBUTES SIGNIFICANCE

Evaluation of attributes significance was performed with the application of following methods [2,3]:

- **χ^2 Attribute Evaluator (χ^2 AE):** Evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class.
- **Information Gain Attribute Evaluator (IG AE):** Evaluates the worth of an attribute by measuring the information gain with respect to the class.
- **Relief Attribute Evaluator (R AE):** Evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.

To perform the child well-being feature selection experiment, Weka (Waikato Environment for Knowledge Analysis) a powerful open-source Java-based machine learning workbench was used [2].

Moreover, attributes significance was analysed with the rough set theory application [5]. The theory defines the notion of condition attributes reducts and more exactly, in experiments the shortest reducts relative to decision attributes (i.e. such minimum subsets of condition attributes which enable a decision with the greatest quality) were searched. The quality of a relative reduct can be calculated exemplary as a value informing what part of the sample set (after reduction) includes samples that have consistent decision attributes.

The task of reduct finding is NP-hard, so finding shortest reducts in the space of 81 condition attributes was computationally too complex. A simplified (heuristic) method was applied to find suboptimal reducts known as a quick reduct algorithm [7,8].

4. RESULTS AND DISCUSSION

First each decision attribute i.e. ’intensity of education’, ’intensity of entertainment’, ’health and physical fitness’ and ’safety and living conditions’ were analysed separately. Applied filters resulted in different significant attribute lists. It issued from the fact that each method evaluated significance from another point of view. However some results were quite similar and this allowed for more general conclusions.

First ten attributes and best-found reducts are presented in each experiment. The 'attrib.' columns from Tableles in the next sections contain the attribute number given in Tablele 1.

4.1. Attribute ranking for 'intensity of education'

χ^2 AE and IG AE placed attribute (76) i.e. 'school quality' as the first and most important feature connected with the decision attribute 'intensity of education' (Tablele 2). R AE put (76) at the second position, but the difference between the first and second attribute is less than the displayed precision. Second the most important attribute from χ^2 AE and IG AE was 'medical care quality' (31) however it was not indicated by R AE. Attributes (76) and (31) were also present in all found reducts. Another attribute that was common and ranked as the most important by R AE is 'number of books at home' (64). There are three other common attributes in all three rankings: 'after school activity connected with studying mathematics and science' (71), 'contacts with friends' (80) and 'sports activities with father' (41).

It should be noticed that differences in consecutive significance measures in all rankings are very small.

Tablele 2 Ranking of first ten most significant features that influence 'intensity of education'

No	attrib.	χ^2 AE	attrib.	IG AE	attrib.	R AE
1	76	66.8888	76	0.0641	64	0.0418
2	31	46.8861	31	0.0599	76	0.0418
3	33	44.4812	79	0.0451	71	0.0347
4	77	43.5276	77	0.0451	43	0.0312
5	80	38.3898	80	0.0431	2	0.0302
6	79	37.9393	64	0.0398	4	0.0265
7	64	30.8015	71	0.0348	39	0.0257
8	71	28.5901	41	0.0334	41	0.0253
9	78	26.7202	33	0.0330	47	0.0244
10	41	25.3914	78	0.0310	80	0.0237

reduct length	attributes	reduct quality
4	2, 31, 76, 78,	0.3806
5	2, 7, 31, 76, 78	0.6750
6	2, 7, 31, 76, 77, 81	0.8689
7	2, 3, 6, 31, 76, 77, 81	0.9587
8	2, 3, 6, 31, 64, 76, 77, 81	0.9928

4.2. Attribute ranking for 'intensity of entertainment'

Results are much more consistent (Tablele 3) for the 'intensity of entertainment' attribute. All methods evaluated that the most important attribute was: (80) 'contacts with friends'. Values of significance for the (80) attribute were several times greater than others and the attribute was present in all found reducts. That means that relationship with contemporaries is the basis in childhood entertainment evaluation.

Tablele 3 Ranking of first ten most significant features that influence 'intensity of entertainment

No	attrib.	χ^2 AE	attrib.	IG AE	attrib.	R AE
1	80	312.434	80	0.3572	80	0.2143
2	76	104.452	81	0.0701	31	0.0302
3	81	61.7215	76	0.0677	77	0.0299
4	33	53.5382	77	0.0563	54	0.0247
5	77	52.3713	79	0.0419	74	0.0243
6	79	36.0482	31	0.0402	76	0.0233
7	31	33.8612	8	0.0310	39	0.0229
8	8	24.6697	33	0.0286	17	0.0226
9	78	20.0989	78	0.0251	6	0.0211
10	65	17.8526	65	0.0228	81	0.0205

reduct length	attributes	reduct quality
4	3, 77, 79, 80	0.3824
5	3, 6, 77, 79, 80	0.7217
6	2, 3, 6, 77, 79, 80	0.9084
7	2, 7, 31, 47, 78, 79, 80	0.9641
8	2, 7, 27, 31, 47, 78, 79, 80	1

Some other attributes are common for all ranking lists. While 'contacts with family' (81) are quite intuitive other are not directly related with the studied decision attribute. For example (76) 'school quality' and (77) 'school safety' according to common sense are associated with (80) because child's friends come mostly from school. Quite unusual is the connection between (31) 'medical care quality' and 'intensity of entertainment'.

4.3. Attribute ranking for 'health and physical fitness'

Based on results from all filters (Tablele 4) the most important attributes were: 'after school activities – sports' (65) and 'contacts with friends' (80) for decision attribute 'health and physical fitness'. (65) and (80) condition attributes take the two highest places. Third attributes in all lists were scored with significant values less than 50% of the second attribute value of importance. Therefore the rest of the list did not indicate links with the decision attribute so clearly.

Other important attribute, which was in the top four in all rankings, was 'camps' (79). 'Medical care' (30) is the attribute that is present in all lists, however it is not always scored as very significant.

In all presented reducts the attributes ‘medical care quality’ (31) and ‘camps’ (79) are present. It means that they can also have significant influence on this decision attribute.

Tablele 4 Ranking of first ten most significant features that influence ’health and physical fitness’

No	attrib.	χ^2 AE	attrib.	IG AE	attrib.	R AE
1	65	141.499	65	0.1976	65	0.1687
2	80	114.865	80	0.1355	80	0.0812
3	8	55.9423	79	0.0628	30	0.0291
4	79	52.6986	8	0.0550	79	0.0273
5	76	50.1914	81	0.0508	75	0.0264
6	81	42.0446	76	0.0495	40	0.0238
7	77	36.2083	77	0.0442	41	0.0212
8	31	34.0130	31	0.0429	63	0.0207
9	30	29.5975	30	0.0354	52	0.0188
10	78	26.3601	75	0.0313	62	0.0188

reduct length	attributes	reduct quality
4	2, 31, 65, 79	0.3303
5	2, 7, 31, 65, 79	0.6589
6	2, 7, 31, 65, 79, 80	0.8671
7	2, 6, 31, 65, 79, 80, 81	0.9659
8	2, 3, 6, 31, 65, 79, 80, 81	0.9928

4.4. Attribute ranking for ’safety and living conditions’

Most various results were obtained from analysing ’safety and living conditions’ decision attribute. Definitely the most important attribute was ’district safety’ (78) which was the second most important attribute in all lists and which was present in all found reducts (Tablele 5). Other features that can be significant are ’medical care quality’ (31), ’school safety’ (77), ’the size of house/residence’ (8) and ’parents job problems’ (23). All these attributes occurred in all lists and get high scores.

The first in the list and the most significant condition attribute indicated by χ^2 AE and IG AE is ’school quality’ (76). However in the top ten attributes given by R AE this attribute is absent.

Unintuitive are connections between ’contact with friends’ (80) which is strongly supported by R AE and ’safety and living conditions’ decision attribute.

Tablele 5. Ranking of first ten most significant features that influence ’safety and living conditions’

No	attrib.	χ^2 AE	attrib.	IG AE	attrib.	R AE
1	76	140.162	76	0.1041	80	0.0479
2	78	93.8293	78	0.0889	78	0.0395
3	31	91.5940	31	0.0688	8	0.0377
4	77	74.5048	8	0.0675	69	0.0366
5	33	64.2957	77	0.0648	75	0.0321
6	8	55.4508	80	0.0553	31	0.0307
7	80	54.7900	79	0.0520	23	0.0265
8	23	46.9824	33	0.0518	17	0.0256
9	79	39.3507	23	0.0483	77	0.0251
10	64	34.8605	64	0.0455	30	0.0223

reduct length	attributes	reduct quality
4	2, 31, 78, 79	0.3788
5	2, 6, 31, 78, 79	0.6984
6	2, 3, 6, 31, 78, 79	0.8833
7	2, 3, 6, 31, 75, 78, 79	0.9641
8	2, 3, 6, 31, 52, 75, 78, 79	1

4.5. Child well-being general evaluation

All presented results analyzed four different aspects of life separately. However the question about general childhood quality evaluation is a natural consequence of the presented research. To find the answer and set of most significant attributes for child well-being integrally, four decision attributes were joint into one resultant attribute.

In the survey respondents were asked to specify weights (in %) for all decision attributes in the general evaluation of child well-being quality. Each decision attribute can be scored by a numeric value from 1 to 5. Therefore the general evaluation can be calculated as:

$$E_{general} = \text{round}(w_{edu}\cdot E_{edu} + w_{ent}\cdot E_{ent} + w_{health}\cdot E_{health} + w_{safety}\cdot E_{safety})\ , \qquad (1)$$

where: E – decision attribute, w – weight in %.

The general evaluation proposed in (1) takes values from a set {1,2,3,4,5}. This additional attribute is also nominal and can also be examined with feature selection methods. Results will show the ranking of the most important attributes for child well-being quality evaluation.

Tablele 6. Ranking of first ten most significant features that influence child well-being

No	attrib.	χ^2 AE	attrib.	IG AE	attrib.	R AE
1	80	155.673	80	0.1642	80	0.1105
2	76	154.880	76	0.1011	76	0.0631
3	33	76.1039	77	0.0678	31	0.0475
4	77	72.5151	31	0.061	73	0.0457
5	31	69.9018	81	0.0577	43	0.0455
6	81	44.3828	79	0.0483	7	0.0449
7	78	40.3767	78	0.0480	10	0.0447
8	79	40.1450	33	0.0412	33	0.0437
9	23	34.0800	65	0.0405	65	0.0436
10	65	29.5536	8	0.0318	29	0.0434

reduct length	attributes	reduct quality
4	2, 76, 78, 79	0.4764
5	2, 76, 78, 79, 80	0.7713
6	2, 31, 76, 78, 79, 80	0.9168
7	2, 31, 47, 76, 78, 79, 80	0.9773
8	2, 27, 31, 32, 76, 78, 79, 80	1

All filters indicated the same attributes: 'contacts with friends' (80) and 'school quality' (76) as the two most significant features for child well-being (Tablele 6). Both attributes were highly scored. They were present in almost all top ten attribute lists described previously.

Common for three lists are also attributes: 'medical care quality' (31), 'sports as after school activity' (65) and 'proportion of time spent with parents' (33). All attributes significant for general childhood evaluation were indicated previously as important for each decision attribute independently. The only exception is the attribute (33).

5. SUMMARY

Presented research was aimed at indicating the set of significant attributes for child well-being estimation. The data-mining feature selection methods were used to obtain rankings and sets of the most influential aspect of persons childhood.

Research effects occurred conformable to earlier expectations, but some lower ranked attributes presented in Tableles 2-6 can be treated as a surprise.

The most interesting result concerns qualification which condition attributes has the real and greatest influence on all aspects of child well-being. Such qualification was possible thanks to the weighted aggregation of all known decision attributes. The most important occurred 'contact with friends' and 'school quality' and the significance of these attributes was distinctly greater than others.

On the base of found significant attributes such models as decision trees or rule models can be created and thanks to the attribute reduction they can be simpler and usually have greater quality and real accuracy.

References

1. Adamus E., Klęsk P., Kołodziejczyk J., Korzeń M., Piegat A. and Pluciński M.: *Rules induction on child well being*. Metody Informatyki Stosowanej, pp. 5-10, no. 4, 2010.
2. Bouckaert R.R., Frank E., Hall M., Kirkby R., Reutemann P., Seewald A. and Scuse D.: *WEKA Manual for version 3.6.0*. University of Waikato, Hamilton, New Zealand, 2008.
3. Cichosz P.: *Learning systems*. Wydawnictwa Naukowo-Techniczne, Warszawa, 2000 [in Polish].
4. Hall M.A., Holmes G.: *Benchmarking Attribute Selection Techniques for Discrete Class Data Mining*. IEEE Trans. on Knowl. and Data Eng., pp. 1437-1447, vol. 15, no. 6, 2003.
5. Pawlak Z.: *Rough Sets*. International Journal of Computer and Information Sciences, pp. 341-356, vol. 11, no. 5, 1982.
6. Sen A.K.: *Capability and Well-Being*, pp. 30-54. The Quality of Life. Clarendon Press, Oxford, 1993.
7. Shen Q., Jensen R.: *Rough Sets, their Extensions and Applications*. International Journal of Automation and Computing, pp. 100-106, 04(1), 2007.
8. Wang X., Yang J., Teng X., Xia W., Jensen R.: *Feature selection based on rough sets and particle swarm optimization*. Pattern Recognition Letters, pp. 459-471, no. 28, 2007.