VERSITA
DOI:10.2478/bile-2013-0011

Biometrical Letters
Vol. 49 (2012), No. 2, 149-158

The use of information and information gain in the analysis of attribute dependencies

Krzysztof Moliński, Anita Dobek, Kamila Tomaszyk

Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Poznań, Poland, e-mail: andobek@up.poznan.pl

SUMMARY

This paper demonstrates the possible conclusions which can be drawn from an analysis of entropy and information. Because of its universality, entropy can be widely used in different subjects, especially in biomedicine. Based on simulated data the similarities and differences between the grouping of attributes and testing of their independencies are shown. It follows that a complete exploration of data sets requires both of these elements. A new concept introduced in this paper is that of normed information gain, allowing the use of any logarithm in the definition of entropy.

Key words: dendrogram, entropy, information gain.

1. Introduction

The notion of entropy has been well known for many years. Introduced in the nineteenth century by Rudolf Clausius, it was initially used only in the physical sciences. With the appearance of the fundamental paper by Shannon (1948), which concerned the amount of information in a signal and the analysis of entropy as a fundamental part of the analysis of information, entropy has found applications in the life sciences. Nowadays the analysis of entropy and information is commonly used in the technical sciences, and is also utilized in environmental sciences such as ecology, or life sciences such as medicine, genetics and others (Bezzi, 2007; Brunsell, 2010; Moniz et al., 2007; Yan et al., 2008; Jakulin et al. 2003).

In the present paper we wish to demonstrate the use of analysis of entropy in exploring categorized or ordered data constituting an attribute of an examined

unit. Statistical analysis of discrete (quality) random variables is always much more laborious than the analysis of continuous (quantity) variables, mostly because the majority of known statistical methods assume the continuity of the analyzed variables. A partial solution to this problem is the well-known, and commonly used, transformation of variables. However, this procedure only gives an approximate solution.

The analysis of entropy, representing the expected value of a random variable $-\log p(a)$ called information, is based on the event's probability and in consequence does not demand the transformation of observations. Its applicability is quite extensive and depends on the research area. In this paper we focus on the analysis of information gain (Kullback and Leibler, 1951), and the function of entropy as used in the analysis of dependencies between random variables investigated in the context of a third dependent variable (McGill, 1954; Jakulin and Bratko, 2003). In different types of experiments generally there may be a question of interactions, but in a particular case it may come down to, for example, the problem of genetic epistasis (Kang et al., 2008).

2. Material and method

The starting point for the described research was the analysis of data in which the variables represented genotypes in different loci in a group of infected individuals and in a group of controls. Therefore in the following consideration the set of possible values was reduced to three values {-1, 0, 1}.

To illustrate the method and all the possible consequences and conclusions, a simulation of data was provided. Two sets of data containing 600 and 400 records were generated. Each record contained nine elements representing the observation of attributes (variables), always from the set {-1, 0, 1}. As a starting point we took the generator of random variables from the normal distribution, and the numbers obtained were transformed to the given set of values. Because a natural association with the simulated set of three-valued data points are genotypes, as an additional condition for some attributes we took the Hardy–

Weinberg equilibrium. Moreover, when generating the data we imposed strong correlations between some variables in both sets, namely for {1,2}, {3,4}, {5,6} and {7,8} in set I and for {1, 2} in set II.

As was mentioned in the introduction, in the investigation of possible relations between variables we will use some functions of entropy and information. For a given variable A the entropy is an expression (Jakulin, 2005) $H(A) = -\sum_a p(a) \ln p(a)$, where the a are the values of the variable A and $p(a)$ are the corresponding probabilities. As is well known, entropy is a measure of uncertainty. The higher the entropy, the less credible are the predictions about A. The notion of entropy may be generalized over two and more variables, for instance (Jakulin, 2005) $H(A,B) = -\sum_a \sum_b p(a,b) \ln p(a,b)$ and also to conditional entropy, namely (Jakulin, 2005) $H(A/B) = -\sum_b p(b) \sum_a p(a/b) \ln p(a/b)$, which describes the uncertainty about A given the variable B. In what follows we will use also the notion of mutual information,

$$I(A,B) = H(A) + H(B) - H(A,B) = H(A) - H(A/B) = \\ = H(B) - H(B/A),$$

which measures the dependency or correlation between two attributes (Moore et al., 2006). If the second attribute is a label variable, then $I(A,B)$ measures the amount of information about A provided by the label variable B, and is called information gain (Jakulin and Bratko, 2004a). Again, the generalization of information gain to two variables gives interaction information (McGill, 1954)

$$IG(A,B;C) = I(A,B/C) - I(A,B)$$

where

$$I(A,B/C) = H(A,C) + H(B,C) - H(C) - H(A,B,C).$$

This measures the dependencies between two attributes in the context of the label variable.

In the analysis of dependencies between variables we will use normed mutual information, namely $I(A, B)/H(A, B)$. The higher the normed mutual

information, the stronger the correlation of attributes. As a result, the inverse of this parameter is a measure of distances of variables (Rajski, 1961) and therefore may be used for building a dendrogram illustrating the mutual connectedness of attributes. Similarly, we can interpret the absolute value of the inverse of information gain as a distance and build a dendrogram based on it. It seems to us, however, that similarly to normed mutual information, a better measure of distance is the inverse of normed interaction information, namely

$$IG_w(A,B;C) = \frac{IG(A,B;C)}{I(A,C)+I(B,C)}$$

The advantage of the normed distance is that it is not dependent on the type of logarithm used in the definition of entropy.

In connection with the estimated parameters it is possible to verify two hypotheses: one about the independence of attributes, and one about the independence of attributes in the context of the label variable. The first hypothesis has the form $H_0 : I(A,B) = 0$ (against $H_1 : I(A,B) > 0$) and we will use the χ^2 test to test its truth. As was shown by Kang et al. (2008) the statistic $2n\ I(A, B)$ has a $\chi^2_{n(A)n(B)-5}$ distribution where n is the number of observations and $n(A)$ and $n(B)$ are the numbers of possible values for the two variables tested.

To verify the hypothesis $H_0 : IG(A,B;C) = 0$ (against $H_1 : IG(A,B;C) \neq 0$) we will again use the χ^2 test. From the results of Matsuda (2000) and Jakulin and Bratko (2004b) we know that the statistic $2n\ IG(A, B;C)$ has a $\chi^2_{n(A)n(B)n(C)-1}$ distribution.

3. Results and discussion

The analysis of simulated data was performed using Microsoft Excel and some procedures from the R package. In the first step, the normed mutual information about each attribute and the label was calculated. As is shown in Table 1, the most informative attribute about the label value is variable 8, followed by variable 7. At the other end of the scale is variable 1.

Table 1. Normed mutual information about attributes and label

Attribute	1	2	3	4	5	6	7	8	9
Information	0.000	0.045	0.039	0.074	0.030	0.078	0.102	0.159	0.017

In the next step we calculated the matrices of normed mutual information for Set I, Set II, and their integrity, extended by a 10th variable, namely the label. In this way we wish to recognize the label on the basis of the remaining attributes. The dendrograms built upon these distances are represented by Figures 1–3.

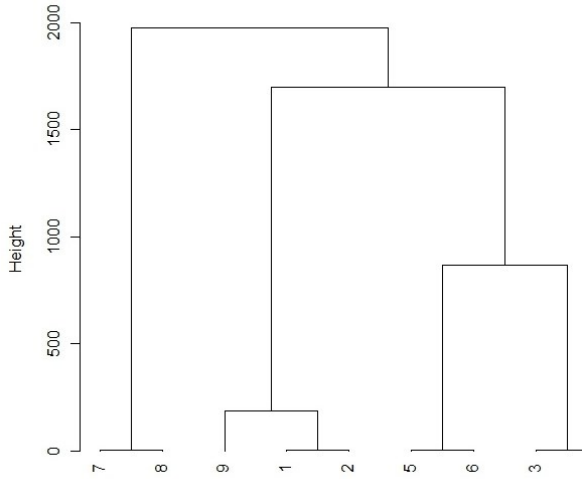

Figure 1. Grouping of attributes for data from Set I

Firstly, what can be seen in the Figures are the consequences of simulation rules and the dependencies assumed from the beginning. Strongly correlated attributes are close together on the dendrograms. In Figure 1 and Figure 3 these are attributes {1, 2}, {3, 4}, {5, 6} and {7, 8}. In Figure 2, which concerns Set II, it is only the pair {5, 6}. The dendrogram in Figure 3 also shows a strong connection between the label variable {10} and attributes 7 and 8, which carry most information about it.

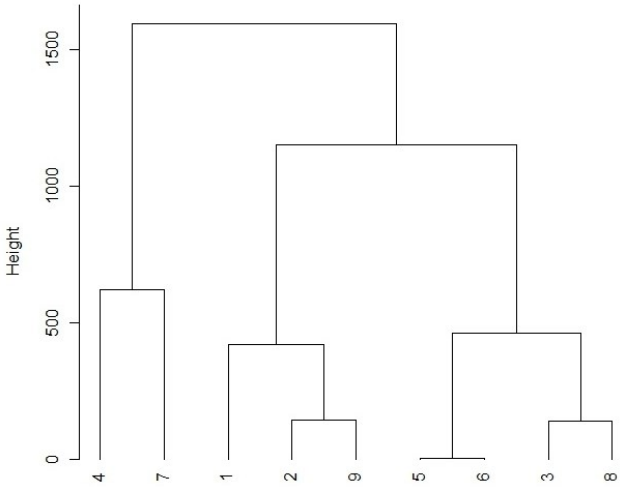

Figure 2. Grouping of attributes for data from Set II

The dendrograms for Set I and Set II group the variables into three sets, but their combination forms only two clusters. Figure 3 shows also that attributes {1, 2, 9} carry less possible information about the label.

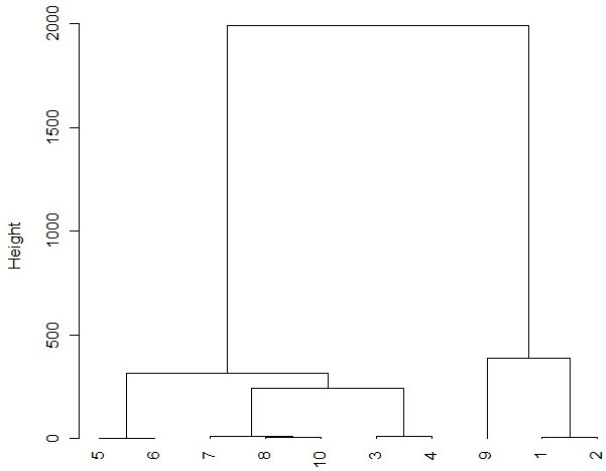

Figure 3. Grouping of attributes for data from Set I and Set II

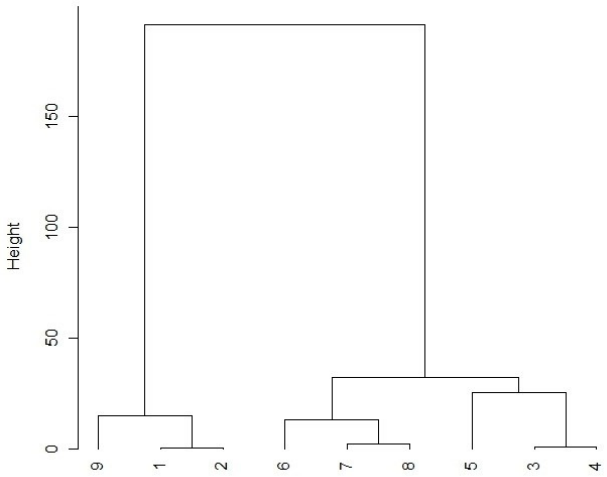

Figure 4. Grouping of attributes based on information gain

The last calculated distance matrix is the matrix of normed information gain, which was used in the construction of the last dendrogram in Figure 4. The analyzed attributes are grouped in two clusters. It should be noted that the variables {1, 2}, {3, 4} and {7, 8}, which are strongly correlated in Set I and are independent in Set II, carry a large amount of information about the label. By contrast, the pair {5, 6}, correlated in both sets, is separated on this dendrogram because this relation does not provide any information about the label attribute. Some comments are necessary in relation to the cluster {1, 2, 9}, which appears on all dendrograms. Its presence in Figure 4, concerning the information gain, may be explained by the fact that the tree attributes say the least possible about the label.

Proceeding to the verification of statistical hypotheses connected with this experiment, we began with verification of the Hardy–Weinberg equilibrium, because some attributes were generated according to this rule. It appears that in both sets, also in cases of variables assumed not to be part of this rule, the statistical test does not allow us to reject the hypothesis about the equilibrium. This proves that for large samples the χ^2 test is not reliable.

Table 2. The χ^2 values for attribute independence tests

Set I + Set II									
Attributes		2	3	4	5	6	7	8	9
	1	**1844.0**	**9.9**	**14.1**	2.6	3.2	1.2	3.7	**29.5**
	2		**11.8**	**16.6**	3.0	3.5	2.1	7.2	**28.7**
	3			**1185.1**	8.7	**12.3**	6.3	5.2	4.1
	4				5.3	**12.2**	2.0	2.4	5.5
	5					**1113.9**	8.5	**9.5**	**13.7**
	6						2.4	3.5	**10.5**
	7							**1334.4**	4.3
	8								7.0
Set I									
Attributes		2	3	4	5	6	7	8	9
	1	**1106.4**	6.0	8.4	1.6	1.9	0.7	2.2	**17.7**
	2		7.1	**10.0**	1.8	2.1	1.2	4.3	**17.2**
	3			**711.0**	5.2	7.4	3.8	3.1	2.4
	4				3.2	7.3	1.2	1.4	3.3
	5					**668.3**	5.1	5.7	8.2
	6						1.4	2.1	6.3
	7							**800.6**	2.6
	8								4.2
Set II									
Attributes		2	3	4	5	6	7	8	9
	1	9.2	4.1	1.6	4.5	4.9	2.4	2.2	3.3
	2		**9.7**	1.6	2.4	1.8	0.6	1.5	**12.0**
	3			1.4	6.6	5.3	2.9	**11.7**	3.5
	4				2.8	1.9	2.7	1.9	3.6
	5					**845.3**	1.2	6.6	5.4
	6						1.4	7.3	5.4
	7							2.7	1.8
	8								7.2

Bold denotes significance at a level of at least 0.05

The significance of normed mutual information from Table 1, tested by the χ^2 test, shows a strong dependency between the distribution of attributes and label. An exception is the first attribute, for which the distribution is not dependent on the label.

Table 3. The χ^2 values for attribute independence in the label context tests

Attributes	2	3	4	5	6	7	8	9
1	**486.1**	2.0	6.9	2.8	2.3	2.0	3.1	11.9
2		10.2	1.8	1.4	5.8	18.3	**28.3**	6.8
3			**312.4**	11.7	20.9	9.6	**41.3**	1.7
4				7.8	24.9	**28.0**	**32.2**	1.2
5					8.4	8.5	22.1	3.3
6						29.9	**48.3**	5.0
7							**242.9**	7.0
8								11.6

Bold denotes significance at a level of at least 0.05

The results from testing the independence of attributes presented in Table 2 and Table 3 confirm the statistical dependencies connected with the way of generating data, as well as some dependencies that may be observed on the dendrograms. This is particularly true for pairs of variables that are also close on the dendrograms. The dependence of variables {1, 2, 9} was also statistically proven. By forming conclusions based on dendrograms and the χ^2 test we should be conscious of the different role of these two statistical instruments. While the dendrogram shows the sets of attributes, the statistical test χ^2 is answers questions about dependencies between variables. As we may observe, the conclusions obtained by these two instruments are identical at some points, but at others they differ significantly. A complete analysis of a set of data requires both of these elements.

Finally it should be pointed out that the data analysis presented in this paper may be used in the case of variables with an unrestricted number of values. The

attribute playing the role of a level variable may not be bivariate, but may also have more values.

REFERENCES

Bezzi M. (2007): Quantifying the information transmitted in a single stimulus. Biosystems, 89: 4-9.

Brunsell N.A. (2010): A multiscale information theory approach to assess spatial–temporal variability of daily precipitation. Journal of Hydrology 385: 165-172.

Jakulin A. (2005). Machine learning based on attribute informations. PhD Dissertation. University of Ljubljana.

Jakulin A., Bratko I., Smrke D., Demsar J., Zupan B. (2003): Attribute interactions in medical data analysis. In: 9th Conference on Artificial Intelligence in Medicine in Europe (AIME 2003), October 18-22, (2003), Protaras, Cyprus.

Jakulin A., Bratko I. (2003): Analyzing attribute dependencies. In: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003), September 22-26, Cavtat, Croatia.

Jakulin A., Bratko I. (2004a): Quantifying and visualizing attribute interactions: an approach based on entropy. http://arxiv.org/abs/cs.AI/0308002 v3.

Jakulin A., Bratko I. (2004b): Testing the significance of attribute interaction. Proc. 21st International Conference on Machine Learning. Banff, Canada.

Kang G., Yue W., Zhang J., Cui Y., Zuo Y., Zhang D. (2008): An entropy-based approach for testing genetic epistasis underlying complex diseases. Journal of Theoretical Biology 250: 362-374.

Kullback S., Leibler R.A. (1951): On information and sufficiency. Annals of Mathematical Statistics 22(1): 79-86.

Matsuda H. (2000): Physical nature of higher-order mutual information. Intrinsic correlation and frustration. Physical Review E, 62: 3096-3102.

McGill W.J. (1954): Multivariate information transmission. Psychometrika 19(2): 97-116.

Moniz L.J., Cooch E.G., Ellner S.P., Nichols J.D., Nichols J.M. (2007): Application of information theory methods to food web reconstruction. Ecological Modeling 208: 145-158.

Moore J.H., Gilbert J.C., Tsai C.-T., Chiang F.-T., Holden T., Barney N., White B.C. (2006): A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. Journal of Theoretical Biology 241: 252-261.

Rajski C. (1961): A metric space of discrete probability distributions. Information and Control 4: 373-377.

Shannon C. (1948): A mathematical theory of communication. Bell System Technical Journal 27: 379-423, 623-656.

Yan Z. Wang Z., Xie H. (2008): The application of mutual information-based feature selection and fuzzy LS-SVM-based classifier in motion classification. Computer Methods and Programs in Biomedicine 90: 275-284.