On Biases in Estimating Multi-Valued Attributes

Igor Kononenko
University of Ljubljana, Faculty of electr eng and computei sc
Tizaika 25 SI-61001 Ljubljana
Slovenia

Abstract

We analyse the biases of eleven mtasures for estimating the quality of the multivalued attributes The values of information gain *J*-measure, gini-index and relevance tend to lin early increase with the number of values of an attribute The values of gam-ratio dis tance measure, *Relief* and the weight of evidence decrease for informative attributes and increase for irrelevant attributes The bias of the statistic tests based on the chi-square distribution is similar but these functions are not able to discriminate among The attributes of different quality We also introduce a new func tion based on the MDL principle whose value slightly decreases with the increasing number of attribute s values

1 Introduction

In top down induction of decision trees various impurity functions are used to estimate the quality of attributes in order to select the "best one to split on However various heuristics tend to overestimate the multi valued attnbules One possible approach to this problem in top down induction of decision trees is the construction of binary decision trees The other approach is to introduce a kind of normalization into the selection criterion such as gam-ratio [Quinlan, 1986] and distance measure [Mantaras, 1989]

Recently White and Liu [1994] showed that, even with normalization information based heuristics still tend to overestimate the attributes with more values Their experiments indicated that $\backslash^2$ and *G* statistics are superior estimation techniques to information gain gain ratio and distance measure They used the Monte Carlo simulation technique to generate artificial data sets with at tributes with various numbers of values However, their scenario included only random attributes with the uni form distribution over attributes' values generated independently of the class

The purpose of our investigation is to verify the conclusions of White and Liu in mort realistic situa tions where attributes art informative and/or ha\< non-uniform distribution of altnbutt s values We adopted and extended their scenario m order to verify results of methods tested b\ White and Liu and to lest also some oilier well known measures gini-index [Breiman et al 1984] *J* measure [Smyth and Goodman 1990] the weight of evidence [Miclue 1989], and relevance [Baim 1988] Besides we developed and tested also one, new selection measure based on the minimum description length (MDL) principle and a meassure derived from the algorithm RELIEF [Kira and Rendell 1992]

In the following we describe all selection measures the experimental scenario and results We analyse the (dis)advanlagfs of various- selection measures

2 Selection measures

In this section we bneflv describe all selection measures and develop *A lit* w one based on the MDL principle We \ssumt that all attributes are discrete and that the prob lem is lo select the best attribute among the attributes with various numbers of possible values All selection measures are defined in a wav that the best attribute should *maximize* the measure Let *C* *A* and 1 b» the number of classes th< number of attributes and the number of values of the given attribute, respectivelv Let n denote the number of training instances, n, the number of training instances from class C_i n_j the number of instances with the j-th value of the given attribute, and nnj the number of instances from class *C,* and with the the th value of the given attribute Let furthe her $p_{ij} = n_{ij}/n$, $p_i = n_i/n$ $p_j = n_j/n$ and $p_{i|j} = n_{ij}/n_j$ denote the approximation of the probabilities from the training set

2 1 Information based measures

Let *Hc, HA, He A,* ^a^nd *Hc/A* be the entropv of the classes of the values of the given attribute, of the joint events class - attribute value, and of the classes given the value of the attribute, respectively

$$H_C = -\sum_i p_i \log p_i \qquad H_A = -\sum_j p_j \log p_j$$

$$H_{CA} = -\sum\sum p_{ij}\log p_{ij} \qquad H_{C|A} = H_{CA} - H_A$$

where all the logarithms are of the has.e two The information gain is defined as the transmitted information by a given attribute about the object s class

$$Gain = H_C + H_A - H_{CA} = H_C - H_{C|A} \qquad (1)$$

In order to avoid the overestimalion of the multivalued attribute*? Quinlan [1986] introduced the gam-ratio

$$GainR = \frac{Gain}{H_A} \qquad (2]$$

Manldras [1989] defined a distance measure *D* tliat can be rewritten as [Whitf and Liu 1994]

$$1 - D = \frac{Gain}{H_{CA}} \qquad (3;$$

Smyth and Goodman [1990] introduced The *J* measure for estimating the information content of (the rule which is appropriate for selecting a single attribut* value for rule generation

$$J_j = p_j \sum p_{i|j} \log \frac{p_{i|j}}{p_i}$$

A straightforward generalization gives the attribute selection measure

$$J = \sum_j J_j \qquad (4)$$

Another selertion measure re lated to informal ion theor\ is the *average absolute weight of enidence* [Miche 1989] It is based on *plausibility* which is an alterative to antropv from the information Theon Let *odds* $:= p/(1-p)$ The measure is defintd for only two class problems

$$WE_i = \sum_j p_j \left|\log \frac{odds_{i|j}}{odds_i}\right| \qquad i = 1\ 2$$

and it holds II E1 =r N *E* A straightforward generalization can be used for multi class problems

\\E = Ep,WE1 :s)

2 2 Gini-index and RELIEF

Breiman et al [1984] use gini-index as the (nonnegative) attribute selection measure

$$Gini = \sum_j p_j \sum_i p_{i|j}^2 - \sum_i p_i^2 \qquad (6)$$

Kira and Rendell [1992] defined the algorithm RELIEF for estimating the quality of attributes RELIEF efficiently deals with strongly dependent attributes The idea behind is the search for the nearest instances from the eame class and the nearest instances from different classes Kononenko [1994] showed that, if this "nearest" condition is omitted, the estimates of RELIEF can be viewed as the approximation of the following difference of probabilities

Relief = P(diff value of an att [different class)

—P(diff value of an at I |same class)

which can be reformulated into

$$Relief = \frac{p_j^2 \times Gini'}{p_i^2(1-p_i^2)} \qquad 7)$$

where

$$Gini' = \sum_j \left(\frac{p_j^2}{\sum_j p_j^2} \sum_i p_{i|j}^2\right) - \sum_i p_i^2 \qquad (8)$$

is highlv correlati d with the gmi-mdev The only difference to <equation(6) is tlial instead of the factor

$$\frac{p_j^2}{\sum_j p_j^2} \quad Gini \text{ uses } \quad \frac{p_j}{\sum_j p_j} = p_j$$

In our experiments besides *Gnn* and *Relief* we e\aluated *Gun* as well in order to verify whether this difference to equation (6) is significant or not

2 3 Relevance

Bairn [1988] introduced a selection measure called the *relevance* of an atlnbute Lt for a given attribute value *j* be

$$i_m(j) = arg\max_i \frac{n_{ij}}{n_i}$$

Th(relevance of the attribute is defined with

$$Relev = 1 - \frac{1}{C-1}\sum_j \sum_{i \neq i_m(j)} \frac{n_{ij}}{n_i}$$

2 4 $\backslash^2$ and *G* statistics

The measures based on the rln square distribution use the following formula

$$P(\chi_0)_D = \int_0^{\chi_0} p(x)_D dx \qquad (10)$$

where *p(z)o* ^ the chi-square distribution with *D* degrees of freedom and \o is the value of the statistic for a given attribute Press et al [1988] give the algorithms for evaluating the above formula We have two statistics that are well approximated bv the chi-square distribu tion with (\ — 1)(C — 1) degrees of freedom [White and Liu, 1994], x and G

$$\chi^2 = \sum_i\sum_j \frac{(e_{ij} - n_{ij})^2}{e_{ij}} \qquad e_{ij} = \frac{n_{\cdot j} n_{i\cdot}}{n} \qquad (11)$$

and

$$G = 2n\ Gain \log_e 2 \qquad e = 2\ 7182 \qquad (12)$$

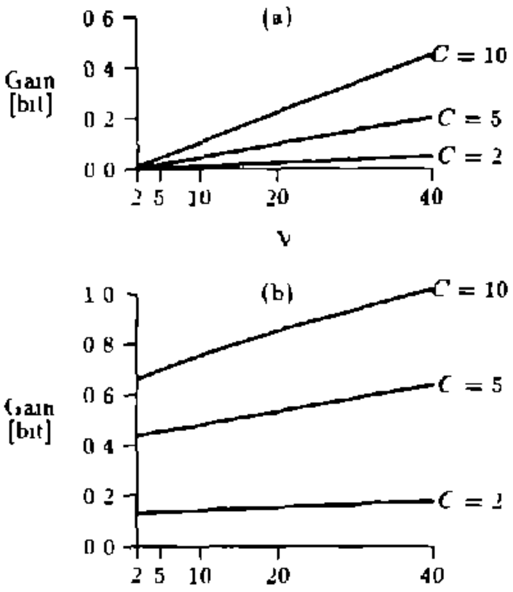


Figure 1 (a) *Gain* for uniform distribution of attribute val ue*. (b) dam for informative attributes

2 5 MDL

According to the minimal description length principle [Rissanen 1983 Li and Vitanvi 1993] the problem of selecting the best attribute can be stated as the problem of selecting the most compressive attribute Let us have the following transmission problem Both tin. sender and the receiver have tht description of (he number of attributes 4 the number of possible values for each attnhutt V the number of possible classes C and the description of the training examples in terms of attribute-values But only the sender knows the correcl classifi cation of examples This information should be transmitted bv minimizing the length (the number of bits) of the message The sender may either explicitly code the class for each training example or may select the best attribute and encode, for each value of the selected attribute the classes of the examples having that value of the attribute Therefore either we have one coding scheme for the prior distribution of classes, or we have a separate coding scheme for each value of the attribute with the associated posterior distribution For each coding scheme a decoder has to be transmitted as well

The number of bits, that are needed to explicitly encode classes of examples for a given probability distribution, can be approximated with entropy *He* times the number of training examples n plus the number of bits needed to encode the decoder For any coding rule the sufficient information to reconstruct the decoder is the probability distribution of events, i e classes Therefore, if n is known, to reconstruct the decoder the receiver needs to reconstruct only n_i , $i = 1$ $C - 1$ (n_C can be then uniquely determined) There is n + *C* — 1 over *C* — 1 possible distributions Therefore, the approximation of the total number of bits that we need to encode the classes of *n* examples is

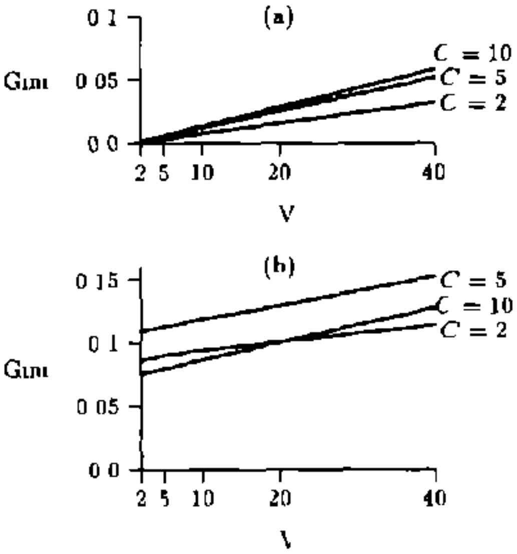


Figure 2 (a) Gini for uniform distribution of attribute val ues (b) Gini [or informative attributes

$$Prior\_MDL' = n\ H_C + \log\binom{n\ + C - 1}{C - 1}$$

and the approximation of the number of bits to encode the classes of examples in all subsets corresponding to all values of the selected attribute is

$$Post\_MDL' = \sum_j n_j H_{C|j} + \sum_j \log\binom{n_j + C - 1}{C - 1} + \log A$$

The last term (log .A) is needed to encode the selection of an attribute among 4 attributes However this term is constant for a given selection problem and can be ignored The first term is equal to *n HC\A* Therefore, the *MDL* measure that evaluates the average compression (per instance) of the message bv an attribute is defined with the following difference *Pnor.MDL* - *Post-MDL* normalized with n *MDL'* = Gain+

$$+\frac{1}{n}\left(\log\binom{n\ + C - 1}{C - 1}\right. \qquad C -$$
- *,> (" < + - - '))

(13)

However, entropy *He* can be used Lo derive *MDL* if the messages are of arbitrary length If the length of the message is known the more optimal coding uses the logarithm of all possible combinations of class labels for given probability distribution

$$Prior\_MDL = \log\binom{n}{n_1, \quad , n_C} + \log\binom{n\ + C - 1}{C - 1}$$

This gives better definition for *MDL*

$$MDL = \frac{1}{n}\left(\log\binom{n}{n_1, \quad , n_C} - \sum_j \log\binom{n_j}{n_{1j}, \quad , n_{Cj}} + \right.$$

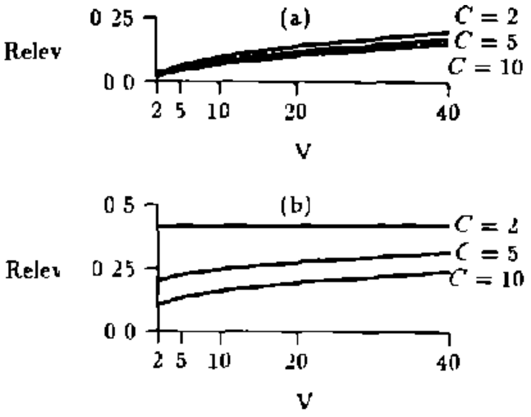


Figure 3 (a) Relevance for uniform distribution of attribute values (b) Relevance for informative attributes

$$+\log\binom{n + C - 1}{C - 1} - \sum_j \log\binom{n_j + C - 1}{C - 1}\Bigg) \qquad (14)$$

3 Experimental scenario

We adopted and extended the scenario of White and Liu [1994] Their scenario included the following variations of settings the number of classes was 2 and 5 the distribution of classes was uniform (except in out. experiment when they used a non-uniform distribution) and there were three attributes with 2 5, and 10 possible values which wtre randomly generated from the uniform distribution independently of the class They performed the Monte Carlo simulation by 1000 times randomly generating 600 training instances with the above properties The quailtv of all attributes was <eslimated using all measures described in Section 2 and the results of each measure were averaged over 1000 trials

We extended the scenario m the following directions

1 We tried the following numbers of classes 2,5 10 and the following numbers of attribute values 2, 5, 10 20, 40

2 We used also informative attributes the attributes with different number of values are made equally infor mative by joining the values of the attribute into two subsets $\{1 \ldots (V\ div\ 2)\}$ and $\{(V\ div\ 2 + 1) \ldots V\}$ corresponding to two values of the binary attribute The probability that the value is from the subset depends on the class while the selection of one particular value in side the subset is random from the uniform distribution The probability that the value is in one subset is defined with

$$P(j \in \{1, \ldots (\lfloor \tfrac{V}{2} \rfloor)\}|i) = \begin{cases} \frac{1}{i+kC} & i\ mod\ 2 = 0 \\ 1 - \frac{1}{i+kC} & i\ mod\ 2 \neq 0 \end{cases} \qquad (15)$$

We tried various values for k $(k = 0\ 1, 2)$ which determines how informative is the attribute For example for 2 possible classes with uniform distribution, the information gain of the two-valued attribute is 0 13 bits if k = 1 and 0 32 bits if $k = 2$, and for 10 possible classes

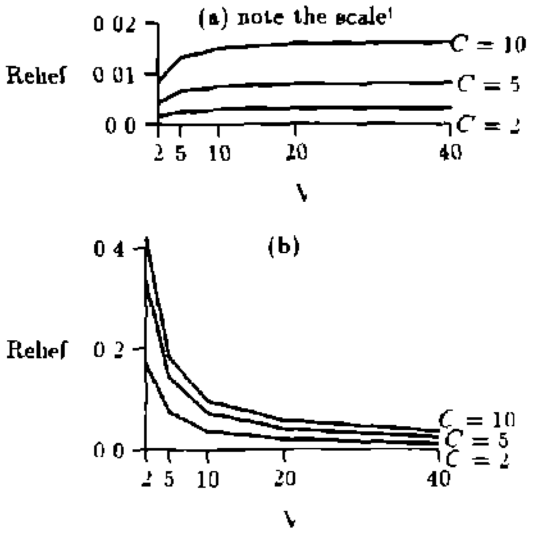


Figure 4 (a) *Relief* for uniform distribution of attribute values (b) *Relief* for informative attributes

i(is 0 60 bits if $k = 1$ and 0 77 bits if $k = 2$ However the biases of all measures were not very sensitive to the value of L In the next section we give the results for $k = 1$

3 We trn d also various non uniform distributions of attribute values for uninformalive attributes and also various non-uniform distributions of classes for each possible distribution of attribute values (uniform, non-uniform, informative) The biases of all measures were independent of the distribution of classes and for unmfonnative attributes also independent of the distribution of attribute values The graphs m the next section are for uniform distribution of classes

4 Results

We will show here two different results for irrelevant (unmformative) and informative attributes We will present results jointly for the measures with the similar behavior

4 1 Linear bias in favor of multivalued attributes

The values of measures increase linearly with the number of values of attributes, in all different scenarios and for all different numbers of classes for the following measures Gam (1) *J* (4) *Gtm* (6), and Gini' (8) On Figures 1 (a) and (b) the values of *Gam* are depicted ./-measure has similar graphs

Note that the scale of the graph for *Gun* is not comparable to the scale for *Gam* *Gum* and Gum have prac tically identical graphs The difference to Gam is that Gini Lends to decrease with the increasing number of classes which seems to be undesired feature for selection measures This is shown on Figure 2 (b) where the values of *Gini* for higher number of classes, where attributes are even more informative (see eq (15)), are lower than the

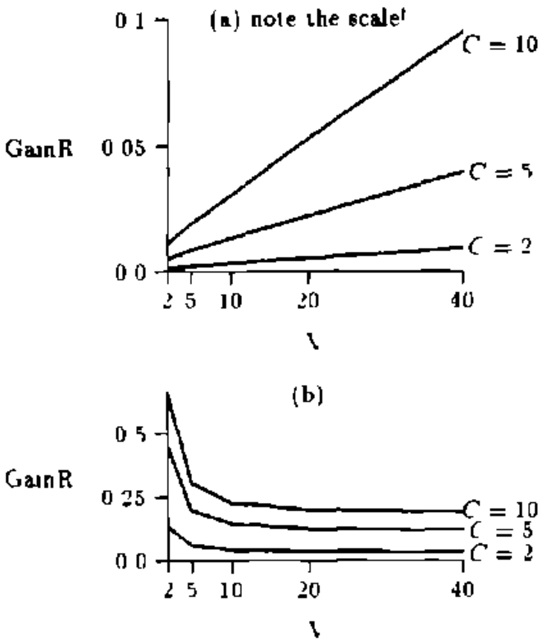


Figure 5 (a) GamR for uniform distribution of attribute values (b) *GauinR* for informative attributes

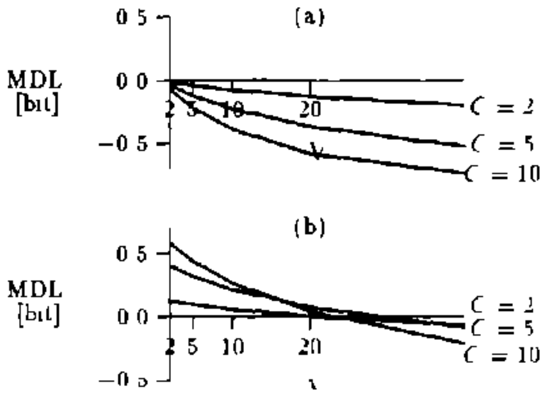


Figure 6 (a) *MDL* for uniform distribution of attribute val ues (b) *MDL* for informative attributes

values, for lower number of classes This becomes even more obvious for more informative attributes (e g for *L — 2* in eq (15)), where the graph becomes similar to that on Figure 3 (b) except that for *Gim* all curves are straight lines with higher slope

Relrv (9) has similar behavior like gmi index, except that the estimates uicrea.se less, than linearly with the number of values of attributes Figure 3 shows this Note that relevance tends to decrease with the increasing number of classes even though the attributes Tor problems with more classes are more informative

4 2 Exponential bias against the informative multivalued attributes

The estimates of informative and highly informative attributes decrease exponentiallv with l he number of values of attributes for *GainR* (2) 1 - *D* (3) and *Relief (1)* However for irrelevant attributes all three measures exhibit (he bias in favor of the multivalued attributes Th]e estimates of *Relief* increase logarithm] callv with the number of values while for 1 — *D* and *GainR the* estimates increase hnearlv Figure 4 shows both biases for *Relef* and Figure 5 shows both biases for GamR Tht performance of 1 — *D* is ver\ similar to that of *GaiuR* Note the different scales for irrelevant and informative attributes This shows that tht bias in favor of the irrelevant multivalued attributes is not very strong and is the lowest for *Relief*

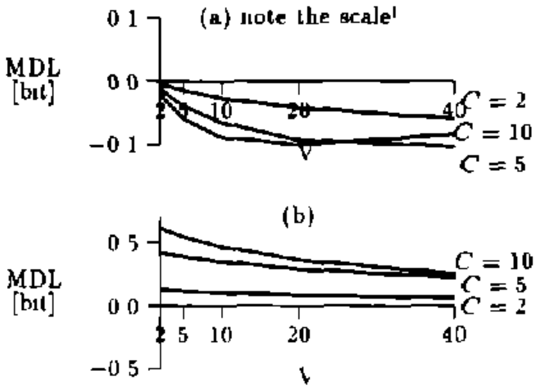


Figure 7 (a) *MDL* for uniform distribution of attribute val ues (b) *MDL* for informative attributes

4 3 Slight bias against the multivalued attributes

MDL (1 \) exhibits the bias against the multivalued at tributes *MDL* almost hnearlv decreases with the num her of values of attributes in all scenarios as is shown on Figures b (a) and (b) As expected from the definition in eq (13) the. bias (the slope of the curve) is higher for the problems witli the higher number of classes Namelv the number of classes influences the number of bits needed to encode the decoders

MDL is always negative for irrelevant attributes and therefore the irrelevant attributes are alwavs considered as non-compressive For informative attributes the compression decreases with the number of values of attributes

MDL (]4) has similar behaviour as *MDL* for irrele vant attributes, however the slope of the curves is lower and all informative attributes are considered compressive This is shown on Figure 7

The behavior of \\ *E* (5) is not stable The reason may be in the use of the non-differentiable function (absolute \alues) The behavior seems to be somewhere between *Relev* (for irrelevant attributes) and *MDL* (for in forma live attributes) This is shown on Figure 8 Like *MDL WE* decreases faster for the problems with more classes

4 4 Almost unbiased but also non-discriminating measures

Figure 9 (a) shows that P(\$^2$) defined with eq (10) and (11) is unbiased le its valuer do not show an) tendenc\

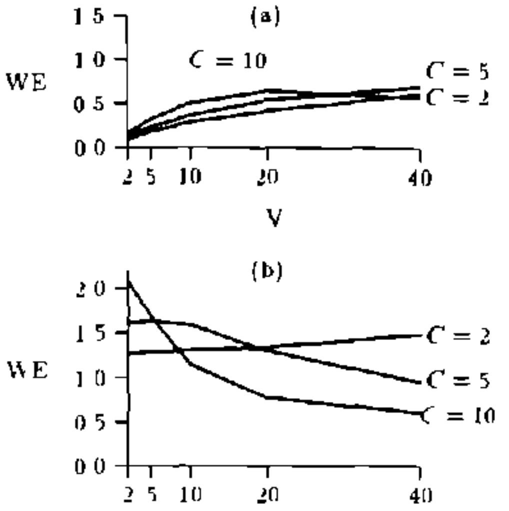
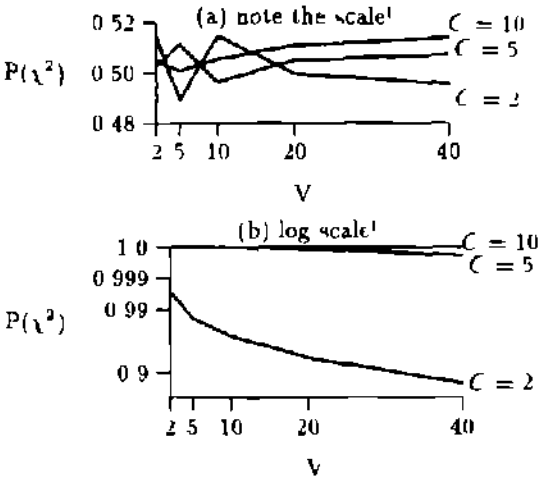


Figure 4 (a) Tlu weight of evidence for uniform distribution of at tribute values (b) The weight of evidence for informative attributes

Willi inereasing nurnber of values of attributes However, this measure is not able to distinguish between more and less informative attributes All informative attributes (for $k = 1$ and / m eq (15)) have *P{\$^{2}$)* - 1 00000 , regardless of the number of values In faat when using floating point in (the cast when $(= 2$ and $\chi > 5$ we dittcted that $P(\chi ')$ differs from 1 (I m sometime*, ex otic decimal places (like J7lh decimal plact) Of course (Ins is the problem of computer precision and numerical evaluation of eq (10) But. the fact is that on most com pulers without special algorithms this measure will not be ahle to distinguish between attributes with different quahty which makes the measure impractical Someone may argue that in the cases when *P(\$^{2}$)* = 1 0 the value of χ^2 could be used This can be done onlv when comparing (lit attributes with t\actly the same number of values (the same digre of freedom) which is not verv useful

Figure 10 (a) shows that *P((*)* defiiud with equations (10) and (12) overestimates the multivalued attributes which is not m agreement with the conclusions by White and Liu [1994] Their conclusions seem valid if the figure is limited to C = 2 and 5 and $\chi = 2, 5$ and 10 which was their original scenario Besides, *P(G)* has the same problems as *P{\$^{2}$)* with informative attributes Its values for informative attributes are all equal to 1 0

We verified this for *P{\$^{2}$)* and *P(G)* by varying the parameter in eq 25 0 50 As soon as $k > 0$ all attributes are too informative and both statistics get the value 1 0 The results in Figures 9 (b) and 10 (b) for *k* = 0 show the biases for $P(\chi^2)$ and *P{G)* against the multivalued attributes for slight!} informa ti\e attributes



Fıgurı 9 (a) $P(\chi^2)$ for unıform dıstrıbutıon of attrıbutı val ues (b) $P(\chi^2)$ for slightlv ınformatıve attrıbutes $k = 0$ ın eq (15)

5 Conclusions

While our results with the original scenario by W'hile and Liu are the same, an increase of the numher of attribute values shows a slightly different picture and further variations of the scenario reveal that their conclusions should be considered with caution $P(G)$ shows clear bias in favor of irrelevant multivalued attributes $P(\chi^2)$ and $P(G)$ measures seem to be biasc d against the informative mullivaluid attributes However the prob lem of evaluating the correct value with the given computer precision makes this functions impractical as they are not able to discriminate between the attributes of different quality

From our results we can conclude that the worst mea sures are those whose values in all different scenarios tend to increase with the number of values of an attribute information gam /-measure gini-mdex and relevance Some of the measures (gnu index and rel evanct) exhibit an undesired behavior their values decreet with tin increasing number of classes even though tht attributes for problems with more classes are more informative However the weight of evidence and *MDL* (13) show similar tendency but only with the increasing number of attributes values For *MDL* this behavior can be explained in terms of the number of bits required lo encode the decoders For uninformative attributes, the bias of *WE* is similar to the bias of relevance, however its bias is not stable

The performance of gain-ratio, distance measure and fie he/, which all use a kind of normalization, is similar The values exponentjallv decrease for informative attributes and increase for irrelevant attributes For ir relevant attributes the performance of *Relief* seems to be better, as the bias in favor of multivalued attributes is not linear but rather logarithmic However the exponen-

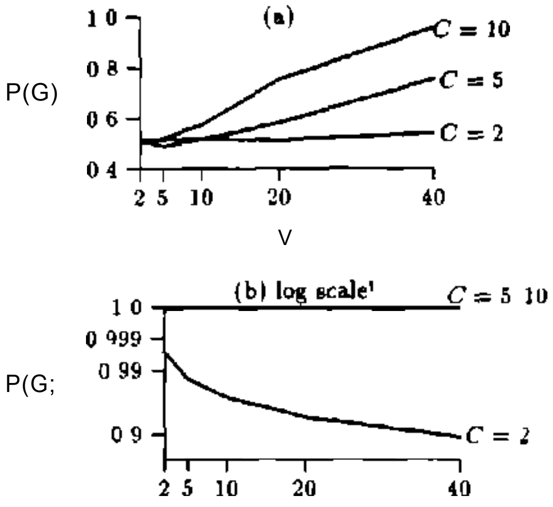


Figure *10* (a) *P{G)* for uniform distribution of attribute values (b) *P(G)* for slightly informative attributes *k = 0* in eq (15)

tial bias against the multivalued attributes can hardl) be justified and more conservative bias mav be more acceptable

The purpose of this investigation was to analvse the performance of various measures on multivalued at tributes independent to other attributes We used the name *Relief* for the function (7) wihch is derived from the original function of RELIEF bv omiting the search for the nearest instances Among all the measures only RELIEF (together with the search for the nearest instances) is non-myopic in the sense that it is able to appropriately deal with strongly dependent attributes Besides RELIEF can also efficiently, estimate continuous attributes [Kira and Rendell, 1992] The extensions introduced in the algorithm RELIEFF [kononenko, 1994] enable it to efficiently deal with noisy data missing values, and multi-class problems All these important features together with the relatively acceptable bias de scribed in this paper, make RELIEFF a promising measure

The values of two new selection measures based on the MDL principle slightly decrease with the increasing number of attribute's values This bias is stable and seems to be better than the bias of other selection measures The selection measures have natural interpretation and also show when the attribute is not useful at all if it is not compressive i e when the value of *MDL* (13) or *MDL* (14) is less than or equal to zero *MDL* seems more appropriate than *MDL'* it uses optimal coding and its graphs have lower (negative) slopes which indicates lower bias against the multivalued attributes

Acknowledgments

1 thank Matevz Kovacic, Uros Pompe, and Marko Robnik for discussions on the MDL principle and for comments on earlier drafts of the paper

References

[Bairn 1988] PW Bairn A method for attribute selection in inductive learning systems *IEEE Trans on PAMI* 10 888-896, 1988

[Breiman *tt* a/ 1984] L Breiman J H Friedman R A Olshen and C J Stone *Classification and Regression Trees* Wadsworth International Group 1984

[Kira and Rendell, 1992] K Kira and L Rendell A practiral approach to feature selection *Proc Interm Conf on Machine Learning* (Aberdeen, Jul} 1992) D Sleeman and P Edwards (eds), Morgan Kaufmann pp 249-256

[Kononenko 1994] I Kononenko Estimating attributes Analysis and extensions of RELIEF *Proc European Conf on Machine Learning* (C atania April 1994) L Dr Raedt and F Bergadano (eds) Springer Verlag pp 171-182

[Li and Yilanyi 1993] M Li and P Vilanyi *An introduction to holmogoroi Complexity and its applications* Springer Verlag 1993

[Mantaras 1989] RL Mantaras ID3 Revisited A distance based criterion for attribute selection *Proc Int Symp Methodologies for Intelligent Systems,* Char lotte, North Carolina U S A , Oct 1989

[Michie 1989] D Michie Personal Models of Rationalit} *J of Statistical Planning and Infertnct,* Special Issue on Foundations and Philosophy ol Probability and Statistics 21 1989

[Press cf *al* , 1988] W H Press, S A Teukolsky V\ T \ettenug B P Flannerv *Numerical recipes m C The art of scientific computing,* Cambridge University Press 1988

[Quinlan 196b] R Quinlan Induction of decision trees *Machine Learning,* 1 81-106, 1986

[Rissanen, 1983] J R Rissanen J A universal prior for integers and estimation by minimum description length *Annals of Statistics,* 11 416-431, 1983

[Smyth and Goodman, 1990] P Smyth and R M Goodman Rule induction using information theory In G Piatetsky Shapiro and W Frawley (eds) *Imowledgt Discoiery* in *Databases,* MIT Press, 1990

[White and Liu, 1994] A P White and W 2 Liu B las in information-based measures in decision tree induction *Machine Learning* 15 121-329, 1994