UC Irvine
ICS Technical Reports

Title
A note on correlational measures

Permalink
https://escholarship.org/uc/item/50944332

Author
Schlimmer, Jeffrey C.

Publication Date
1986-05-30

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Notice: This Material
may be protected
by Copyright Law
(Title 17 U.S.C.)

ARCHIVES
Z
699
C3
no. 86-13

A note on correlational measures

Technical Report # 86–13

Jeffrey C. Schlimmer

Department of Information and Computer Science
University of California, Irvine 92717
ArpaNet:Schlimmer@ICS.UCI.EDU

May 30, 1986

Abstract

Determining the degree to which two events are interrelated is a common subtask for artificial intelligence systems, especially learning systems. This note examines four correlational measures which allow quantization of the relationships between events. Despite the fact that the measures have diverse motivations and formulations, they all indicate irrelevance precisely at the point of statistical independence.

1 Introduction

Determining the degree to which two events are interrelated is a common subtask in artificial intelligence. Moreover, it is frequently the case that events are not associated in an all-or-none manner, and thus their relationship requires quantization. One goal of learning, for example, is to specifically determine which events are associated. This paper compares four measures that have been developed to meet these needs. Following a brief description of the motivation behind each measure and its use in machine learning, a short proof is offered which demonstrates that each measure is equivalent to the notion of statistical independence.

2 Category utility

Psychological research has indicated that some categories are easier to learn and recall than others. Examples of this behavior arise in the context of hierarchically related categories such as *animal–dog–beagle*. One category, *dog*, has been shown to be easier to verify and name. This category has been termed the *basic level* (Gluck & Corter, 1985).

Gluck and Corter (1985) have formulated a correlational measure which indicates the basic level in a hierarchy. Given a hierarchical grouping of objects and their description in terms of attribute-value pairs, this measure indicates the expected utility of each category. It is defined

as[1]

$$CU(C,A) = \frac{1}{|C|}\sum_{i=1}^{|C|} p(C_i)\sum_{j=1}^{|A|}\sum_{k=1}^{|V|}\left[p(A_j = V_k|C_i)^2 - p(A_j = V_k)^2\right] \quad (1)$$

Where C is a level of categories in the hierarchy, A is the set of attribute-values over which the objects are defined, and $|V|$ is the number of values for a particular attribute. This measure is then applied to each of the levels in a hierarchy. The level with the highest expected category utility is the one the measure predicts will be the basic level.

Fisher (1986) has incorporated this measure in a conceptual clustering program called COBWEB. As each new object is processed, the algorithm examines the possibilities of including the new object in an existing category or creating a new one; the action which results in the highest category utility is performed. This heuristic is employed as the algorithm traverses down the category hierarchy to process the new instance.

The category utility measure indicates non-utility always and only when the category and attributes are statistically independent. In order to simplify the following proof, consider a restricted form of category utility for only one category and one attribute. For a single, binary valued attribute equation 1 reduces to:

$$CU(pos,A) = \frac{1}{2}p(pos)\left[p(A = V|pos)^2 + p(A \neq V|pos)^2 - p(A = V)^2 - p(A \neq V)^2\right] \quad (2)$$

The point at which there is no expected utility resulting from this categorization is when $CU(pos,A) = 0$. Setting equation 2 to zero yields:

$$\frac{1}{2}p(pos)\left[p(A = V|pos)^2 + p(A \neq V|pos)^2 - p(A = V)^2 - p(A \neq V)^2\right] = 0$$

Canceling out $\frac{1}{2}p(pos)$

$$p(A = V|pos)^2 + p(A \neq V|pos)^2 - p(A = V)^2 - p(A \neq V)^2 = 0$$

Expanding $p(A = V|pos)$ according to the definition of conditional probability

$$\frac{p(pos \wedge A = V)^2}{p(pos)^2} + \frac{p(pos \wedge A \neq V)^2}{p(pos)^2} - p(A = V)^2 - p(A \neq V)^2 = 0$$

[1] A slightly more complex form of this equation is also presented by Gluck and Corter.

Eliminating all occurrences of $p(A \neq V)$ by noting that

1. $$\begin{aligned} p(pos \wedge A \neq V)^2 &= [p(pos) - p(pos \wedge A = V)]^2 \\ &= p(pos \wedge A = V)^2 - 2p(pos)p(pos \wedge A = V) + p(pos)^2 \end{aligned}$$

2. $$\begin{aligned} p(A \neq V)^2 &= [1 - p(A = V)]^2 \\ &= p(A = V)^2 - 2p(A = V) + 1 \end{aligned}$$

Results in

$$\frac{2p(pos \wedge A = V)^2 - 2p(pos)p(pos \wedge A = V) + p(pos)^2}{p(pos)^2} - 2p(A = V)^2 + 2p(A = V) - 1 = 0$$

Multiplying both sides by $p(pos)^2$

$$2p(pos \wedge A = V)^2 - 2p(pos)p(pos \wedge A = V) + p(pos)^2 - 2p(pos)^2 p(A = V)^2 + 2p(pos)^2 p(A = V) - p(pos)^2 = 0$$

Canceling out $+p(pos)^2$ and $-p(pos)^2$ and then dividing both sides by 2

$$p(pos \wedge A = V)^2 - p(pos)p(pos \wedge A = V) - p(pos)^2 p(A = V)^2 + p(pos)^2 p(A = V) = 0$$

Solving for $p(pos \wedge A = V)$ as a general quadratic equation with a solution of the form

$$x = -\frac{b \pm \sqrt{b^2 - 4ac}}{2a}$$

With the following substitutions

$$\begin{aligned} x &= p(pos \wedge A = V) \\ a &= 1 \\ b &= -p(pos) \\ c &= p(pos)^2 p(A = V) - p(pos)^2 p(A = V)^2 \end{aligned}$$

We then have

$$p(pos \wedge A = V) = -\frac{-p(pos) \pm \sqrt{p(pos)^2 - 4[p(pos)^2 p(A = V) - p(pos)^2 p(A = V)^2]}}{2}$$

Multiplying both sides by 2 and expanding

$$2p(pos \wedge A = V) = p(pos) \pm \sqrt{p(pos)^2 - 4p(pos)^2 p(A = V) + 4p(pos)^2 p(A = V)^2}$$

Factoring out $p(pos)^2$

$$2p(pos \wedge A = V) = p(pos) \pm \sqrt{p(pos)^2[1 - 4p(A = V) + 4p(A = V)^2]}$$

Factoring $1 - 4p(A = V) + 4p(A = V)^2$ yields

$$\begin{aligned}
2p(pos \wedge A = V) &= p(pos) \pm \sqrt{p(pos)^2[1 - 2p(A = V)]^2} \\
&= p(pos) \pm p(pos)[1 - 2p(A = V)] \\
&= p(pos) + p(pos)[1 - 2p(A = V)] \text{ or } p(pos) - p(pos)[1 - 2p(A = V)] \\
&= 2p(pos) - 2p(pos)p(A = V) \text{ or } 2p(pos)p(A = V) \\
p(pos \wedge A = V) &= p(pos)[1 - p(A = V)] \text{ or } p(pos)p(A = V) \\
&= p(pos)p(A \neq V) \text{ or } p(pos)p(A = V)
\end{aligned}$$

Since the definition of statistical independence is $p(A \wedge B) = p(A)p(B)$ (Fine, 1973), the category and attribute are statistically independent when the category utility measure is zero. This proof is only a demonstration for a simplified form of the category utility measure. A proof for the more general form of category utility could follow the same format.

3 Logical sufficiency/necessity

The Prospector mineral exploration system (Duda, Gaschnig, & Hart, 1979) utilizes a pair of correlational measures to encode the contribution of a number pieces of evidence toward belief in a hypothesis. In mineral exploration, the presence of a particular geological formation (evidence) may indicate that the area is likely to contain a rich ore deposit (hypothesis). It may also be the case that the absence of the formation indicates that the ore is unlikely. The first measure used is termed *logical sufficiency* (LS), and it measures the degree to which the presence of evidence (E) increases belief in a hypothesis (H). The second is called *logical necessity* (LN) and measures the degree to which the absence of evidence decreases belief in a hypothesis. They are defined

as[2]

$$LS = \frac{p(E|H)}{p(E|\neg H)} \qquad LN = \frac{p(\neg E|\neg H)}{p(\neg E|H)} \tag{3}$$

Both of these measures have similar interpretations. A value of unity indicates that the evidence is irrelevant to the hypothesis. Greater than unity indicates that the evidence confirms the hypothesis, while less than unity corresponds to evidence that infirms the hypothesis.

Since $p(\neg E|H) = 1 - p(E|H)$ and $p(\neg E|\neg H) = 1 - p(E|\neg H)$ it is the case that the LN measure (for example) could be rewritten as:

$$LN = \frac{1 - p(E|\neg H)}{1 - p(E|H)}$$

Using this identity, it is easy to show that when $LS = 1$ (and therefore $p(E|H) = p(E|\neg H)$) then $LN = 1$. Similarly, when either LS or LN are greater than unity, the other is also; when one is less than unity so is the other. However, it is not true in general that $LS = LN$. For example, if $p(E|H) = 0.3$ and $p(E|\neg H) = 0.1$ then $LS = 3$ and $LN = \frac{7}{9}$. Maintaining a pair of measures allows differentiating between the positive and negated associations.

For the purposes of analysis, consider the conditions under which the LS measure fails to indicate relevance. As stated previously, $LS = 1$ indicates that the evidence has no relevance to the hypothesis.

$$\frac{p(E|H)}{p(E|\neg H)} = 1$$

Multiplying both sides by the denominator gives:

$$p(E|H) = p(E|\neg H)$$

By the definition of conditional probabilities we have:

$$\frac{p(E \wedge H)}{p(H)} = \frac{p(E \wedge \neg H)}{p(\neg H)}$$

Cross multiplying yields:

$$p(\neg H)p(E \wedge H) = p(H)p(E \wedge \neg H)$$

[2]In the original definition, LN has the inverse definition.

Substituting $1 - p(H)$ for $p(\neg H)$ and multiplying on the left hand side:

$$[1 - p(H)]p(E \wedge H) = p(H)p(E \wedge \neg H)$$

$$p(E \wedge H) - p(H)p(E \wedge H) =$$

Adding $p(H)p(E \wedge H)$ to both sides and factoring out $p(H)$

$$p(E \wedge H) = p(H)p(E \wedge \neg H) + p(H)p(E \wedge H)$$

$$= p(H)[p(E \wedge \neg H) + p(E \wedge H)]$$

Reducing $p(E \wedge \neg H) + p(E \wedge H)$ to $p(E)$

$$p(E \wedge H) = p(E)p(H)$$

Which is precisely the definition of statistical independence. The derivations for $LS > 1$ and $LS < 1$ are similar. When the LS and LN measures indicate relevance, the evidence and hypothesis are statistically dependent; when the Prospector measures indicate irrelevance, the evidence and hypothesis are statistically independent.

4 Contingency

In the late 1960's animal learning researchers formulated a law of learning which characterized a class of situations in which animals failed to learn that two events were associated. Specifically, in a classical conditioning experiment a subject is given repeated presentations of a novel cue (NC) and an unpleasant stimulus (US). Researchers found that animals will learn that the novel cue leads to the unpleasant stimulus only if $p(US|NC) > p(US|\neg NC)$ (Rescorla, 1968). If the probability of the unpleasant stimulus is the same with or without the novel cue, subjects fail to learn an association between them.

Recently, Granger and Schlimmer (1985) have formulated a learning model which uses the ratio of the two conditional probabilities as a correlational measure. A second measure is for-

mulated from the ratio of $p(\neg US|\neg NC)$ and $p(\neg US|NC)$.

$$C1 = \frac{p(US|NC)}{p(US|\neg NC)} \qquad C2 = \frac{p(\neg US|\neg NC)}{p(-US|NC)} \tag{4}$$

These two measures are estimated by a simple process of counting event types. Additionally, they are used in a manner similar to Prospector's LS and LN to adjust expectation of the unpleasant stimulus. If the novel cue is present, C1 is used to modify expectation; if it is absent, C2 is used. These two measures also guide formation of Boolean functions representing combinations of novel cues. C1 indicates good conjunctions while C2 indicates good disjunctions (Granger & Schlimmer, 1985).

An irrelevant C1 or C2 measure indicates that the two stimuli are statistically independent. The proof closely parallels the proof for the Prospector measures. By visually substituting evidence (E) for the unpleasant stimulus (US) and the hypothesis (H) for the novel cue (NC), the two sets of measures become syntactically equivalent.

As an aside, it may be interesting to note that measures seemingly similar to C1 or C2 may not indicate statistical dependence. For instance, $p(US|NC)$ fails to make the same correlational distinction, for it can range from nearly zero to unity while the two events are either statistically dependent or not. If $p(US|NC) = 0$ then all that can be said is that $p(US \wedge NC) = 0$. Conversely, if $p(US|NC) = 1$ then $p(US \wedge NC) = p(NC)$. In either case, the probability of a unpleasant stimulus may be either zero or unity, and thus this measure does not assess correlation in a manner similar to statistical independence.

5 Expected information

Another formulation for describing the correlation between two events is based on the information conveyed by one event about the other. If one event always occurs, then the other event's occurrence does not convey any information. Similarly, if the two events are completely uncorrelated, then the occurrence of one event offers little information about the second. However,

if the two events always occur together, the occurrence of one provides complete information about the occurrence of the other.

Quinlan (1985) has formulated an information theoretic measure for the purpose of assessing the correlation between attributes of concept instances and their identification as positive or negative examples of a concept to be learned. His concept attainment program, ID3, uses this measure to select the most highly correlated attribute as a root for a decision tree. Through an iterative process, ID3 constructs a complete decision tree in order to distinguish examples from nonexamples. An arbitrary process could be followed to select test attributes, but using the information theoretic measure yields smaller decision trees which capture more of the regularity inherent in the training concept.

The information theoretic measure Quinlan uses is based on the difference between the amount of information that a complete decision tree provides and the amount of information provided by a particular attribute. Given that there are p positive instances of the concept and n negative instances, the amount of information provided by a complete decision tree is

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \tag{5}$$

Given the attribute A_i with V values is known to be the root of a decision tree, the information conveyed about the instance class is the weighted average of the information conveyed by each of the resulting subtrees:

$$E(A_i) = \sum_{j=1}^{V}\frac{p_{i.j}+n_{i.j}}{p+n}I(p_{i.j},n_{i.j}) \tag{6}$$

$p_{i.j}$ is the number of known positive instances for which attribute A_i has value V_j; $n_{i.j}$, negative instances with $A_i = V_j$. The expected information gain by choosing attribute A_i as the root, therefore, is the difference between the information conveyed by the complete tree and the expected information given root attribute A_i:

$$Gain(A_i) = I(p,n) - E(A_i)$$

ID3 chooses a root attribute which maximizes this gain. As Quinlan notes, since $I(p,n)$ remains constant for each subtree it is sufficient to minimize $E(A_i)$.

If we expand equation 6 by substituting in the definition of $I(p,n)$ we have:

$$E(A_i) = \sum_{j=1}^{V} \frac{p_{i.j} + n_{i.j}}{p+n} \left[-\frac{p_{i.j}}{p_{i.j}+n_{i.j}} \log_2 \frac{p_{i.j}}{p_{i.j}+n_{i.j}} - \frac{n_{i.j}}{p_{i.j}+n_{i.j}} \log_2 \frac{n_{i.j}}{p_{i.j}+n_{i.j}} \right] \quad (7)$$

Since $p_{i.j} + n_{i.j}$ is the number of instances where attribute A_i has value V_j, and $p+n$ represents the total number of known instances, the fraction $\frac{p_{i.j}+n_{i.j}}{p+n}$ is equal to the probability that attribute A_i has value V_j, or $p(A_i = V_j)$. Likewise, the fraction $\frac{p_{i.j}}{p_{i.j}+n_{i.j}}$ may also be interpreted as a probability if we multiply by $\frac{p+n}{p+n} = 1$.

$$\frac{p_{i.j}}{p_{i.j}+n_{i.j}} = \frac{p_{i.j}/(p+n)}{(p_{i.j}+n_{i.j})/(p+n)}$$

$\frac{p_{i.j}}{p+n}$ is the number of positive instances with $A_i = V_j$ divided by the total number of instances and is thus equal to $p(pos \wedge A_i = V_j)$. Using these substitutions, we may rewrite equation 7 as:

$$E(A_i) = \sum_{j=1}^{V} p(A_i = V_j) \left[\begin{array}{l} -\frac{p(pos \wedge A_i = V_j)}{p(A_i = V_j)} \log_2 \frac{p(pos \wedge A_i = V_j)}{p(A_i = V_j)} \\ -\frac{p(neg \wedge A_i = V_j)}{p(A_i = V_j)} \log_2 \frac{p(neg \wedge A_i = V_j)}{p(A_i = V_j)} \end{array} \right]$$

Expressions of the form $\frac{p(A \wedge B)}{p(B)}$ may be expressed as $p(A|B)$ according to Bayes theorem. This yields:

$$E(A_i) = \sum_{j=1}^{V} p(A_i = V_j) \left[\begin{array}{l} -p(pos|A_i = V_j) \log_2 p(pos|A_i = V_j) \\ -p(neg|A_i = V_j) \log_2 p(neg|A_i = V_j) \end{array} \right] \quad (8)$$

As can be seen by the above algebraic manipulations, minimizing equation 8 is equivalent to maximizing expected gain.

By examining the maximal value possible for $E(A_i)$ we may identify the conditions under which this measure identifies an attribute as irrelevant. We may simplify the inner term of the summation by noticing that $p(neg|A_i = V_j) = 1 - p(pos|A_i = V_j)$. Thus we may consider when an equation of the form $-x \log x - (1-x)\log(1-x)$ for $x \in (0,1]$ is maximal. This function is plotted in figure 1; its maximum is at $x = 0.5$.

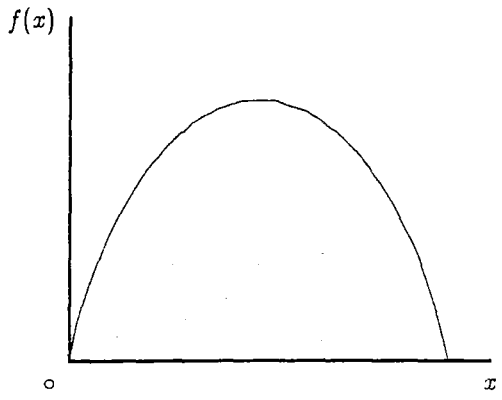

Figure 1: $f(x) = -x \log x - (1-x)\log(1-x)$.

$E(A_i)$ will therefore be maximum when each inner term is maximal or when $p(pos|A_i = V_j) = p(neg|A_i = V_j) = \frac{1}{2}$ for all $j \in [1..V]$.

Now, since the conditional probabilities for each $A_i = V_j$ do not encompass the probability of a positive instance when $A_i \neq V_j$, a proof demonstrating equivalence to statistical independence requires examining the value of $p(pos|A_i \neq V_j)$. By the definition of conditional probabilities we have

$$p(pos|A_i \neq V_j) = \frac{p(pos \wedge A_i \neq V_j)}{p(A_i \neq V_j)}$$

Knowing that all of the values for A_i are mutually exclusive we may rewrite the above as:

$$p(pos|A_i \neq V_j) = \frac{\sum_{k \neq j} p(pos \wedge A_i = V_k)}{\sum_{l \neq j} p(A_i = V_l)} \tag{9}$$

Consider each term in the numerator of the right side of equation 9.

$$\frac{p(pos \wedge A_i = V_{k \neq j})}{\sum_{l \neq j} p(A_i = V_l)}$$

We may multiply by $p(A_i = V_{k \neq j})/p(A_i = V_{k \neq j}) = 1$

$$\frac{p(pos \wedge A_i = V_{k \neq j})p(A_i = V_{k \neq j})}{\sum_{l \neq j} p(A_i = V_l)p(A_i = V_{k \neq j})}$$

which simplifies to

$$p(pos|A_i = V_{k\neq j})\frac{p(A_i = V_{k\neq j})}{\sum_{l\neq j} p(A_i = V_l)}$$

Since we know that $p(pos|A_i = V_k) = \frac{1}{2}$ for all k we can factor each $p(pos|A_i = V_{k\neq j})$ term in the numerator of the right side of equation 9 out of the summation leaving

$$p(pos|A_i \neq V_j) = \frac{1}{2}\frac{\sum_{k\neq j} p(A_i = V_k)}{\sum_{l\neq j} p(A_i = V_l)}$$

The ratio of sums is trivially reducible to unity and therefore $p(pos|A_i \neq V_j) = \frac{1}{2}$. The remainder of the proof follows section 3 since $p(pos|A_i = V_j) = p(pos|A_i \neq V_j)$.

Thus the point at which the information measure used by Quinlan's ID3 reaches its maximum, indicating the greatest irrelevance, is precisely the point at which the attribute becomes statistically independent of the class of the instance.

6 Conclusion

This paper has briefly examined four correlational measures. Category utility was formulated to predict the basic level in a category hierarchy and is used in a conceptual clustering program. Logical sufficiency and logical necessity were formulated to represent the contribution of different types of evidence toward the belief in a hypothesis. A related pair of measures were motivated by a law formulated from animal learning research. These measures are utilized in a concept attainment program to guide prediction and the formation of descriptive Boolean functions. An information theoretic measure was developed for use in another concept attainment program which builds discrimination trees in order to characterize concepts. Use of this measure results in small trees which capture concept regularity. Though these correlational measures have diverse backgrounds and are formulated in different languages, they all indicate irrelevance precisely when the events are statistically independent.

Though this paper has shown a common feature among some correlational measures, there are other classes of correlational measures. A brief example of a conditional probability measure

was given, but there are other measures which are used by artificial intelligence systems to assess the relationships between events that do not adhere to the notion of statistical dependence. It is the underlying assumption of this paper that measures which perform some *ad hoc* measurement of correlation are inferior to those that reflect statistical independence.

What these derivations have not shown, however, is that the orderings between differentially correlated situations is preserved across the four measures. For example, if the Prospector measures indicate that one feature is more relevant than another, will the information theoretic measure also? Early empirical studies in progress indicate that this may be the case, though conclusive results will have to await further mathematical analysis.

Acknowledgements

This research was supported in part by the Office of Naval Research under grants N00014-84-K-0391 and N00014-85-K-0854, the National Science Foundation under grants IST-81-20685 and IST-85-12419, the Army Research Institute under grant MDA903-85-C-0324, and by the Naval Ocean Systems Center under contract N66001-83-C-0255. I would like to thank the entire machine learning group at Irvine for their vigorous discussions and consistent encouragement.

References

Duda, R., Gaschnig, J., & Hart, P. (1979). Model design in the Prospector consultant system for mineral exploration. In D. Michie (Ed.), *Expert systems in the micro electronic age.* Edinburgh: Edinburgh University Press.

Fine, Terrence L. (1973). *Theories of Probability.* New York: Academic Press.

Fisher, D. (1986). *An incremental algorithm for the acquisition of basic-level concepts* (Unpublished manuscript). Irvine, California: The University of California, Department of

DEC 1 6 1986

Library Use Only

Information and Computer Science.

Gluck, M. A., & Corter, J. E. (1985). Information, uncertainty, and the utility of categories. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 283–287). Irvine, California: Lawrence Erlbaum Associates.

Granger, R. H., Jr., & Schlimmer, J. C. (1985). Learning salience among features through contingency in the CEL framework. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 65–79). Irvine, California: Lawrence Erlbaum Associates.

Quinlan, J. R. (1985). *Induction of decision trees* (Technical report 85.6). New South Wales, Australia: The New South Wales Institute of Technology, School of Computing Sciences.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66*, 1–5.