Testing the Significance of Attribute Interactions

Aleks Jakulin JAKULIN@ACM.ORG
Ivan Bratko IVAN.BRATKO@FRI.UNI-LJ.SI
Faculty of Computer and Information Science, Tržaška cesta 25, SI-1001 Ljubljana, Slovenia

Abstract

Attribute interactions are the irreducible dependencies between attributes. Interactions underlie feature relevance and selection, the structure of joint probability and classification models: if and only if the attributes interact, they should be connected. While the issue of 2-way interactions, especially of those between an attribute and the label, has already been addressed, we introduce an operational definition of a generalized n-way interaction by highlighting two models: the reductionistic part-to-whole approximation, where the model of the whole is reconstructed from models of the parts, and the holistic reference model, where the whole is modelled directly. An interaction is deemed significant if these two models are significantly different. In this paper, we propose the Kirkwood superposition approximation for constructing part-to-whole approximations. To model data, we do not assume a particular structure of interactions, but instead construct the model by testing for the presence of interactions. The resulting map of significant interactions is a graphical model learned from the data. We confirm that the P-values computed with the assumption of the asymptotic χ^2 distribution closely match those obtained with the bootstrap.

1. Introduction

1.1. Information Shared by Attributes

We will address the problem of how much one attribute tells about another, how much information is shared between attributes. This general problem comprises both attribute relevance and attribute interactions. Before that, we need to define a few terms. Formally, an *attribute* A will be considered to be a collection of independent, but mutually exclusive attribute values $\{a_1, a_2, a_3, \ldots, a_n\}$. We will write a as an example of the value of A. An *instance* corresponds to an event that is the conjunction of attributes' values. For example, an instance is "Playing tennis in hot weather." Such instances are described with two attributes, A with the range $\Re_A = \{\text{play}, \neg\text{play}\}$, and with the attribute B : $\Re_B = \{\text{cold}, \text{warm}, \text{hot}\}$. If our task is deciding whether to play or not to play, the attribute A has the role of the *label*.

An attribute is relevant to predicting the label if it has something in common with it. To be able to estimate this commonness, we need a general model that connects both the attribute and the label that functions with uncertain and noisy data. In general, models with uncertainty can be stated in terms of the *joint probability density functions*. A joint probability density function (joint PDF) maps each possible combination of attribute values into the probability of its occurrence. The joint PDF p for this example is a map $p : \Re_A \times \Re_B \rightarrow [0, 1]$. From the joint PDF, we can always obtain a *marginal* PDF by removing or marginalizing one or more attributes. The removal is performed by summing probabilities over all the combinations of values of the removed attributes. For example, the PDF of attribute A would hence be $p(a) = \sum_b p(a, b)$.

One way of measuring uncertainty given a joint PDF p is with Shannon's *entropy* H, defined for a joint PDF of a set of attributes $\mathcal{V}$:

$$H(\mathcal{V}) \triangleq \sum_{\vec{v} \in \Re_{\mathcal{V}}} p(\vec{v}) \log_2 p(\vec{v}) \qquad (1)$$

If $\mathcal{V} = \{A, B\}$, the $\vec{v}$ would have the range of $\Re_{\mathcal{V}} = \Re_A \times \Re_B$ – the Cartesian product of ranges of individual attributes. If the uncertainty given the joint PDF is $H(AB)$, and the uncertainties given the two marginal PDFs are $H(A)$ and $H(B)$, the shared uncertainty, the *mutual information* or information gain

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

between attributes A and B is defined as $I(A;B) = H(A) + H(B) - H(AB)$. AB can also be understood as a joint attribute, a derived attribute whose domain is the Cartesian product of the domains of A and B. Mutual information is the reduction in uncertainty achieved by looking at both attributes at the same time. The higher the mutual information, the better we can predict A from B and vice versa. If mutual information is non-zero, we say that A and B are involved in a 2-way interaction.

While mutual information is limited to two attributes, Jakulin and Bratko (2004), following (McGill, 1954; Han, 1980), quantify an interaction among all the attributes in $\mathcal{V}$, $|\mathcal{V}| = k$ with k-way *interaction information*:

$$I(\mathcal{V}) \triangleq -\sum_{\mathcal{T}\subseteq\mathcal{V}} (-1)^{|\mathcal{V}|-|\mathcal{T}|} H(\mathcal{T}). \quad (2)$$

Interaction information can be seen as a generalization of mutual information. Mutual information between two attributes and 2-way interaction information between them are equal. The 3-way interaction information between attributes A, B and C will be denoted as $I(A;B;C)$. There is also a link between interaction information and conditional mutual information: $I(A;B;C) = I(A;B|C) - I(A;B)$; here, $I(A;B|C)$ stands for the conditional mutual information between A and B in the context of C. Therefore, A and B are conditionally independent given C iff $I(A;B;C) = -I(A;B)$. In this case the redundancy is wholly contained in C, and we eliminate it by controlling for C.

The 3-way interaction information is the correction term in determining the mutual information between a label C and two attributes A and B: $I(AB;C) = I(A;C) + I(B;C) + I(A;B;C)$. A positive 3-way interaction information $I(A;B;C)$ indicates a *synergy* between the attributes A and B, meaning that they yield more information together than what could be expected from the two individual interactions with the label. A negative interaction information suggests a *redundancy* between them, meaning that they both provide in part the same information about the label. Based on this quantification it is possible to analyze relationships between attributes, perform attribute clustering and guide feature selection and construction (Jakulin & Bratko, 2003).

1.2. Complexity of Joint Probability Density Functions

In problems with many attributes, the joint PDF may become sparse. The objective of learning is to construct a model of the joint PDF that will avoid this sparseness. The two basic operators for compacting it are *latent attributes* and *factorization*. We may reduce the dimensionality of the attribute space by creating a latent attribute, e.g. $\theta = f(A,B,C)$, so that $p(A,B,C) = p(\theta)$, which is useful if θ has a lower cardinality (or dimensionality) than the original attribute space. With factorization, we take advantage of independencies among attributes. For example, we can factorize $p(A,B,C,D)$ into $p(A,B)p(C,D)$, reducing the original 4-dimensional space into two independent 2-dimensional spaces. Of course, the factors themselves need not be independent: we can factorize $p(A,B,C,D)$ into $p(A,B)$ and $p(B,C,D)$, and then use one of the two conditional expressions $p(A|B)p(B,C,D)$ or $p(C,D|B)p(A,B)$.

Many popular learning algorithms are instances of the above two approaches. This is best illustrated on the example of the naïve Bayesian classifier, applied to a supervised learning problem, with A,B,C as the attributes and Y as the label, but all other attributes given. The naïve Bayesian classifier is based on factorizing $p(A,B,C,Y)$ into $p(A,Y)$, $p(B,Y)$, $p(C,Y)$. This conditional independence assumption has often been found conflicting with the data, resulting in inferior predictions of outcome probabilities (and not just the most likely outcome), which is important to applications like medical risk assessment. Kononenko (1991), Pazzani (1996) and Friedman et al. (1997) proposed approaches with less aggressive factorization, making certain dependencies between attributes a part of the model. Vilalta and Rish (2003) offered an approach based on the discovery of a latent attribute.

1.3. Contributions of the Paper

Jakulin and Bratko (2003) suggested that a considerably high or low interaction information among attributes is a heuristic indication that the attributes interact and should not be factorized. In this paper, we provide the justification for relevance of this heuristic, and replace the vague notion of 'high' and 'low' with statistical significance. An interaction can be defined teleologically as: "When a group of attributes interact, we cannot factorize their joint PDF." To decide whether factorization is warranted, we investigate the loss incurred by the best approximation that we can construct without directly observing the true joint PDF. These methods will be referred to as part-to-whole approximations. In this paper we discuss one such method, Kirkwood superposition approximation (KSA) (Kirkwood & Boggs, 1942), which we define in Sect. 2.1. It turns out that interaction information is equal to Kullback-Leibler divergence between the KSA and the joint PDF.

In the second part of the paper, we investigate the techniques for determining the significance and the importance of a particular part-to-whole approximation on the basis of the evidence provided in data. For example, we should require any complex feature, such as a 10-way interaction, to be supported by plentiful evidence, otherwise we run the risk of overfitting. To determine the significance of interactions, we note that Kullback-Leibler divergence has the χ^2 distribution asymptotically. As an alternative to χ^2, we consider the nonparametric bootstrap procedure and cross-validation. It turns out that the nonparametric bootstrap procedure yields very similar results as does χ^2, but cross-validation deviates somewhat.

Finally, we apply these instruments to identify the significant interactions in data, and illustrate them in the form of an interaction graph, revealing interesting dependencies between attributes, and describing the information about the label provided by individual attributes and the correction factors arising from their interactions with the label. These graphs can be a useful exploratory data analysis tool, but can also serve as an initial approximation in constructing predictive models.

2. Modelling and Interactions

An interaction can be understood as an irreducible whole. This is where an interaction differs from a mere dependency. A dependency may be based on several interactions, but the interaction itself is that dependency that cannot be broken down. To dispel the haze, we need a practical definition of the difference between the whole and its reduction. One view is that the whole is reducible if we can predict it *without observing all the involved variables at the same time*. We do not observe it directly if every measurement of the system is limited to a part of the system. In the language of probability, a view of a part of the system results from marginalization: the removal of one or more attributes, achieved by summing it out or integrating it out from the joint PDF.

Not to favor any attribute in particular, we will observe the system from all sides, but always with one or more attributes missing. Formally, to verify whether $P(A, B, C)$ can be factorized, we should attempt to approximate it using the set of all the attainable marginals: $\mathcal{M} = \{P(A, B), P(A, C), P(B, C), P(A), P(B), P(C)\}$, but not $P(A, B, C)$ itself. Such approximations will be referred to as *part-to-whole approximations*. If the approximation of $P(A, B, C)$ so obtained from these marginal densities fits $P(A, B, C)$ well, there is no interaction. Otherwise, we have to accept an interaction, and possibly seek latent attributes to simplify it.

2.1. Kirkwood Superposition Approximation

Kirkwood superposition approximation (Kirkwood & Boggs, 1942) uses all the available pairwise dependencies in order to construct a complete model. Matsuda (2000) phrased KSA in terms of an approximation $\hat{p}_K(A, B, C)$ to the joint probability density function $p(A, B, C)$ as follows:

$$\hat{p}_K(a, b, c) \triangleq \frac{p(a, b)p(a, c)p(b, c)}{p(a)p(b)p(c)} = p(a|b)p(b|c)p(c|a). \tag{3}$$

Kirkwood superposition approximation does not always result in a normalized PDF: $\tau = \sum_{a,b,c} \hat{p}_K(a, b, c)$ may be more or less than 1, thus violating the normalization condition. We define the normalized KSA as $\hat{p}_K(a, b, c)/\tau$.

Other approximations in closed form that do not violate the normalization condition are the models built upon the assumption of conditional independence. For three attributes, there are three such models: $P(B|A)P(C|A)P(A)$, $P(A|B)P(C|B)P(B)$, $P(A|C)P(B|C)P(C)$. For example, the first model assumes that B and C are independent in the context of A. However, we do not find these approximations to be proper part-to-whole approximations because they do not employ all the available parts; for example, the first model disregards the dependence between B and C. Moreover, the choice of the conditioning attribute is arbitrary, and we have to keep considering several models instead of a single, part-to-whole one: we do not know in advance which of them is best. However, because of the normalization the Kirkwood superposition approximation is not always superior to the above models of conditional independence, as shown experimentally in Sect. 3.1.

The interaction testing methodology can be applied to loglinear models (Agresti, 2002) as well. The loglinear part-to-whole model employs $\mathcal{M}$ as the set of constraints or association terms. The advantage of loglinear models fitted by iterative scaling is that the addition of additional consistent constraints can only improve the fit of the model. On the other hand, the Kirkwood superposition approximation is not always better than a conditional independence model, even if the latter disregards a part of the information that is available to a part-to-whole approximation method. However, the Kirkwood superposition approximation is in closed form, making it very simple and efficient for use, while fitting loglinear models of this type requires iterative methods.

2.2. Kullback-Leibler Divergence as a Statistic

Our objective now is to determine whether an interaction exists among the given attributes in the domain or whether it does not. As we described earlier, if the approximation fits the data well, there is no good evidence for an interaction. We can employ a loss function to assess the similarity between the joint PDF and its part-to-whole approximation. *Kullback-Leibler divergence* is a frequently used measure of difference between two joint probability density functions $p(\vec{v})$ and $q(\vec{v})$:

$$D(p\|q) \triangleq \sum_{\vec{v}\in\Re_{\mathcal{V}}} p(\vec{v})\log_2 \frac{p(\vec{v})}{q(\vec{v})} \tag{4}$$

The unit of measure is a bit. KL-divergence has also been referred to as the 'expected log-factor' (logarithm of a Bayes factor), expected weight of evidence in favor of p as against q given p, and the cross-entropy (Good, 1963).

From the equations in (Matsuda, 2000) it is clear that the 3-way interaction information, as defined in (2), is equal to Kullback-Leibler divergence (4) between the true PDF $p(A,B,C)$ and its Kirkwood superposition approximation $\hat{p}_K(A,B,C)$: $I(A;B;C) = D(p\|\hat{p}_K)$. Analogously, the divergences of the above conditional independence models from the true joint PDF are $I(B;C|A)$, $I(A;C|B)$ and $I(A;B|C)$, and these metrics have often been used for assessing whether the attributes should be assumed dependent in a model that would otherwise assume conditional independence.

The generalized Kirkwood superposition approximation for k attributes can be derived from this equality and (2). We can interpret the interaction information as the approximate weight of evidence in favor of not approximating the joint PDF with the generalized Kirkwood superposition approximation. Because the approximation is inconsistent with the normalization condition, the interaction information may be negative, and may underestimate the true loss of the approximation. Therefore, the Kirkwood superposition approximation must be normalized before computing the divergence.

If the underlying reference PDF p of categorical attributes is based on relative frequencies estimated from n instances, KL-divergence between p and an independent joint PDF $\hat{p}$ multiplied by $2n/\log_2 e$ is equal to the Wilks' likelihood ratio statistic G^2. In the context of a goodness-of-fit test for large n, G^2 has a χ^2_{df} distribution with df degrees of freedom:

$$\frac{2n}{\log_2 e} D(p\|\hat{p}) \underset{n\to\infty}{\sim} \chi^2_{|\Re_{\mathcal{V}}|-1} \tag{5}$$

Here, $df = |\Re_{\mathcal{V}}| - 1$ is based on the cardinality of the set of possible combinations of attribute values $|\Re_{\mathcal{V}}|$. $\Re_{\mathcal{V}}$ is a subset of the Cartesian product of ranges of individual attributes. Namely, certain value combinations are impossible, where the joint domain of two binary attributes A and B, where $b = \neg a$, $\mathcal{V}$ should be reduced to only two possible combinations, $\mathcal{V} = \{(a,\neg b),(\neg a,b)\}$. The impossibility of a particular value conjunction is often inferred from the zero count in the set of instances, and we followed this approach in this paper. Also, by the guideline (Agresti, 2002), the asymptotic approximation is poor when $n/df < 5$. For example, to evaluate a 3-way interaction of three 3-valued attributes, where $df = 26$, there should be least 135 instances.

The null hypothesis is that the part-to-whole approximation matches the observed data, while the alternative one is that the approximation does not fit and there is an interaction. The P-value (or the weight of evidence for accepting the null hypothesis of part-to-whole approximation) is defined to be $P\left(\chi^2_{df}(x) \geq 2nD(p\|\hat{p})/\log_2 e\right)$. The P-value can also be interpreted as the probability that the average loss incurred by p on an *independent* sample from the null Gaussian model approximating the multinomial distribution parameterized by p itself, is greater or equal to the average loss incurred by the approximation $\hat{p}$ in the original sample. In this case, the loss is measured by the KL-divergence.

We followed Pearson's approach to selecting the number of degrees of freedom, which disregards the complexity of the approximating model, assuming that the null hypothesis is p and the alternative distribution is $\hat{p}$, and that $\hat{p}$ is hypothesized and not estimated. This P-value can be interpreted as the lower bound of P-values of all the approximations. The part-to-whole approximations we discussed in the previous section instead have $df' = \prod_{X\in\vec{V}} (|\Re_X| - 1)$ residual degrees of freedom in Fisher's scheme, and we may reject them in favor of simpler approximations. Using df instead assures us that no simplification would be able to reduce the P-value, regardless of its complexity. This way, simplifying the part-to-whole approximation by means of reducing the set $\mathcal{M}$ is only performed if the P-value is low enough.

2.3. Obtaining P-Values by Resampling

Instead of assuming the χ^2 distribution of KL-divergence, we can simply randomly generate independent *bootstrap samples* of size n' from the original training set. Each bootstrap sample is created by randomly and independently picking instances from

the original training set with replacement. This non-parametric bootstrap corresponds to an assumption that the training instances themselves are samples from a multinomial distribution, parameterized by p. For each bootstrap sample we measure the relative frequency p', and compute the loss incurred by our prediction for the actual sample $D(p'\|p)$. We then observe where $D(p\|\hat{p})$ lies in this distribution of losses. The P-value is $P(D(p'\|p) \geq D(p\|\hat{p}))$ in the set of bootstrap estimates of p'. The bootstrap sample size n' is a nuisance parameter which affects the result: the larger value of n', the lower the deviation between p' and p. The P-value is conditional on n'. Usually, the size of the bootstrap sample is assumed to be equal to the original sample $n' = n$. The larger the n', the more likely is the rejection of an approximation with the same KL-divergence.

The most frequently used method for model selection in machine learning is cross-validation. Here, we will define a similar notion of CV-values that will be based on 2-fold cross-validation. For each replication and fold, the set of instances is partitioned into the test and training subsets. From these subsets, we estimate two joint PDFs: the training p' and the testing $\dot{p}$. On the basis of a partially observed p', we construct the part-to-whole approximation $\hat{p}'$. The CV-value is defined as $P(D(\dot{p}\|\hat{p}') \geq D(\dot{p}\|p'))$ in a large set of cross-validated estimates of $(p', \dot{p})$. As in bootstrap, the number of folds is a nuisance parameter.

2.4. Making Decisions with P-Values

On the basis of the thus obtained P-values, we can decide whether an interaction exists or not. P-value identifies the probability that the loss of $D(p\|q)$ or more is obtained by the null model predicting a sample from the null model. For example, the P-value of 0.05 means that the loss incurred by the null model will be greater or equal to the loss obtained by the approximation $\hat{p}$ on the training sample in on average 5 independent samples out of 100.

On the basis of the P-value ϕ, we may classify the situation into two types: the interaction is discovered when $\phi \leq \alpha$, and the interaction is rejected when $\phi > \alpha$. We become holistically biased towards an interaction and risk overfitting by using a high value as the threshold α, e.g., 0.95. We choose a reductionistic bias preferring a simpler, no-interaction model and risk underfitting by using a low value in α, e.g., 0.05. P-value only provides a measure of robustness of the interaction, but not its importance. We continue to employ interaction information as the measure of importance.

3. Experiments

3.1. 3-Way Attribute Interactions

We have taken 16 data sets from the UCI repository, and for each pair of attributes in each domain, we have investigated the 3-way interaction between the pair and the label. We compared the Kullback-Leibler divergence between the maximum likelihood joint probability density function $p(A, B, C)$ and its part-to-whole approximation obtained by the normalized Kirkwood superposition approximation on the basis of maximum likelihood marginals.

Kirkwood superposition approximation and conditional independence models We have compared the Kirkwood superposition approximation with best one of the three conditional independence models. It turns out that the conditional independence model was somewhat more frequently worse (1868 vs 1411), but the average error was almost 8.9 times lower than that of the Kirkwood superposition approximation (Fig. 1). This shows that an interaction that may seem significant with Kirkwood superposition approximation might not be significant if we also tried the conditionally independent approximations. On the other hand, it also shows that models that include KSA may achieve better results than those that are limited to models with conditional independence.

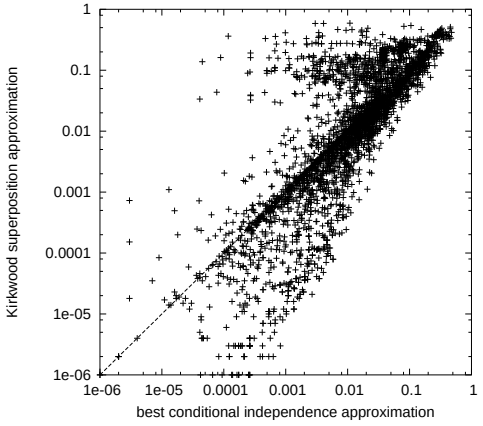

Figure 1. A comparison of the approximation divergence achieved by the Kirkwood superposition approximation and the best of the three possible conditional independence models.

Constructing graphical models We employed the above interaction testing approach to construct a model of the significant 2-way and 3-way interactions for a supervised learning domain. The resulting interaction graph is a map of the dependencies between

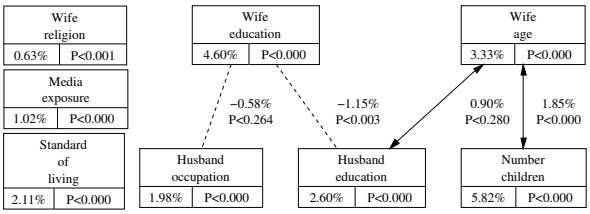

Figure 2. An interaction graph (Jakulin & Bratko, 2003) is illustrating interactions between the attributes and the label in the CMC domain. The label in this domain is the contraception method used by a couple. The chosen P-value cutoff of 0.3 also eliminated one of the attributes ('wife working'). The two dashed lines indicate redundancies, and the full arrowed lines indicate synergies. The percentages indicate the reduction in the conditional entropy of the label given the attribute or the interaction.

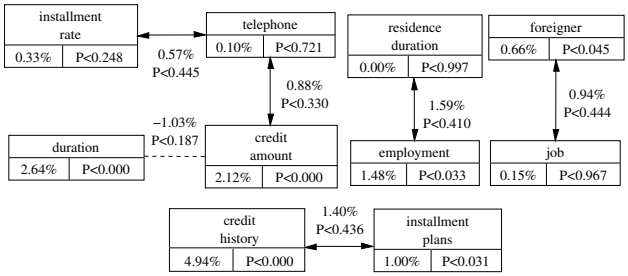

Figure 3. Only the significant interactions of the German credit domain are shown in this graph, where $\phi < 0.5$. The label in the domain is credit risk. Most notably, attributes 'telephone', 'residence duration' and 'job' are only useful as a part of a 3-way interaction, but not alone. We can consider them to be moderators.

the label and other attributes and is illustrated in Figs. 2 and 3. Kirkwood superposition approximation observed both negative and positive interactions. However, these interactions may sometimes be explained with a model assuming conditional independence: sometimes the loss of removing a negatively interacting attribute is lower than imperfectly modelling a 3-way dependence. Also, if two attributes are conditionally independent given the label, they will still appear redundant.

The interaction graph does not attempt to minimize any global fitness criterion, and should be seen as a very approximate guideline to what the model should look like. It may also turn out that some attributes may be dropped. For example, results from Fig. 1 indicate that Kirkwood superposition approximation is not uniformly better than conditional independence models. So, one of the conditional independence models for a triplet of attributes could fit the data better than Kirkwood superposition approximation, and the interaction would no longer be considered significant.

The procedure for constructing interaction graphs is not yet a complete model building procedure. P-values may be meaningful if performing a single hypothesis test, but analysis of the whole domain involves a large number of tests, and we have to account for the consequently increased risk of making an error in any of them. The best-case approach is to assume that all P-values are perfectly correlated, and we can use them without adjustment. The worst-case approach is to assume that all P-values are perfectly independent, and adjust them with Bonferroni correction. But for the proposed use of P-values in this paper, making decisions about an interaction (whether to take it into account or ignore it), it is only the ranking of P-values of the interactions that really matters.

3.2. Inferential Procedures

We compared the 2-way interactions between each attribute and the label in several standard benchmark domains with the number of instances in the order of magnitude of 100: 'soybean-small', 'lung', 'horse-colic', 'post-op', 'lymphography' and 'breast-cancer'. In domains with more instances, the 2-way interactions are practically always significant, which means that there is enough data to make it worth to model them. But on such small domains, it is sometimes better to disregard weakly relevant attributes, as they may cause overfitting.

Comparing χ^2 and bootstrap P-values We examined the similarity between the P-values obtained with the assumption of χ^2 distribution of KL-divergence, and the P-values obtained through the bootstrap procedure. The match shown in Fig. 4 is good enough to recommend using χ^2-based P-values as a reasonable heuristic which perhaps tends to slightly underestimate. The number of bootstrap samples was 10000.

On the difference between P-values and cross-validated CV-values We compared the P-values obtained with bootstrap with similarly obtained CV-values, using 500 replications of 2-fold cross-validation. The result is illustrated in Fig. 5 and shows that the two estimates of significance are correlated, but behave somewhat differently. P-values are more conservative, while very low and very high P-values do not guarantee an improvement or deterioration in CV performance. Although CV-values might seem intuitively more appealing (even if the number of folds is another nuisance parameter), we are not aware of suitable asymptotic approximations that would allow quick estimation.

P-values and cross-validated performance We employed cross-validation to verify whether a classifier benefits from using an attribute, as compared to a classifier based just on the prior label probability distribution. In other words, we are comparing classifiers $p(Y)$ and $p(Y|X = x) = p(Y, X = x)/p(X = x)$, where Y is the label and X is an attribute. The loss function was the expected change in negative log-likelihood of the label value of a test set instance when given the instance's attribute value. This way, we use no probabilistic model of the testing set $\dot{p}$, but instead merely consider instances as samples from it. The probabilities were estimated with the Laplacean prior to avoid zero probabilities and infinitely negative log-likelihoods. We employed 2-fold cross-validation with 500 replications. The final loss was the average loss per instance across all the instances, folds and replications.

The results in Fig. 6 show that the P-value was a very good predictor of the increase in loss. The useful attributes appear on the left hand side of the graph. If we pick the first 100 of the 173 total attributes with $\phi < 0.3$, there will not be a single one of them that would increase the loss. On the other hand, if we picked the first 100 attributes on the basis of mutual information or information gain, we would end up with a deterioration in 7 cases, which is still a two-fold improvement upon the base rate, where 14.4% of all the attributes yield a deterioration in this experiment.

On the other hand, it must be noted that 60% of the 30 most insignificant attributes with $\phi > 0.9$ also result in a decrease of prediction loss! The cut-off used for detecting overfitting through an increase in loss by cross-validation is obviously somewhat ad hoc, especially as both CV-values and P-values turned to be largely equivalent in this experiment. For that reason we should sometimes be skeptical of the performance-based results of cross-validation. Significance can be seen as a necessary condition for a model, carrying the aversion to chance and complexity, but not a sufficient one, neglecting the expected performance difference.

4. Discussion

We have shown how interaction information can be interpreted as Kullback-Leibler divergence between the 'true' joint PDF and its generalized Kirkwood superposition approximation. If the approximation is normalized, we can employ the methods of statistical hypothesis testing to the question of whether a group of attributes interact. It has been shown that KL-divergence has a χ^2 distribution asymptotically, but we have also validated this distribution with bootstrap sampling, observing a very good match, but noting that the computations given χ^2 are orders of magnitude cheaper. We also introduce the notion of a part-to-whole approximation which captures the intuition associated with irreducibility of an interaction, and Kirkwood superposition approximation is an example of such approximation.

P-values penalize attributes with many values, making an interaction often insignificant in total, even if for some subset of values, the interaction would have been significant. Perhaps latent attributes should be involved in modelling both the joint PDFs and the marginals used for constructing the part-to-whole approximation. Alternatively, we could employ the techniques of subgroup discovery (in themselves a kind of a latent attribute) and identify the subset of situations where the interaction does apply.

From experiments we made, there seems to be a difference between the bootstrap formulation and the cross-validated formulation of hypothesis testing, but the two are not considerably different when it comes to judging the risk of average deterioration. This conclusion has been disputed, but a tenable explanation for our results could be that all our evaluations were based on the Kullback-Leibler divergence, while earlier experiments tried to employ statistical testing based on probabilistic statistics for improving classification performance assessed with a conceptually different notions of classification accuracy (error rate) or instance ranking (area under the ROC).

Pearson's goodness of fit testing which we primarily used in this paper is just one of possible testing protocols, and there has been much debate on this topic, e.g., (Berger, 2003). For example, using P-values alone, we would accept a model with rare but grave errors, but reject a model with frequent but negligible ones. Similarly, we would accept a model with very frequent but negligible yields, but reject a model with rare but large benefits. Without significance testing, the average performance is insufficient to account for complexity and risk. While Pearson's approach is very close to Fisher's significance testing, differing just in the choice of the degrees of freedom, the CV-values resemble the Neyman-Pearson hypothesis testing, because both interaction and non-interaction models and their dispersion are taken into consideration. The expected loss approach is closest to Jeffreys' approach, because both models are included and because the difference in loss is actually the logarithm of the associated Bayes factor for the two models. Hence, we offer CV-values and expected decrease in loss as viable alternatives to our choice of P-values for interaction sig-

nificance testing that was influenced by the simplicity and efficiency of closed-form computations given the χ^2 distribution.

References

Agresti, A. (2002). *Categorical data analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons. 2nd edition.

Berger, J. (2003). Could Fisher, Jeffreys and Neyman have agreed upon testing? *Statistical Science*, *18*, 1–32.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*, 131–163.

Good, I. J. (1963). Maximum entropy for hypothesis formulation. *The Annals of Mathematical Statistics*, *34*, 911–934.

Han, T. S. (1980). Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, *46*, 26–45.

Jakulin, A., & Bratko, I. (2003). Analyzing attribute dependencies. *PKDD 2003* (pp. 229–240). Springer-Verlag.

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

Kirkwood, J. G., & Boggs, E. M. (1942). The radial distribution function in liquids. *Journal of Chemical Physics*, *10*, 394–402.

Kononenko, I. (1991). Semi-naive Bayesian classifier. *EWSL 1991*. Springer Verlag.

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

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, *19*, 97–116.

Monti, S., & Cooper, G. F. (1999). A Bayesian network classifier that combines a finite mixture model and a naive-Bayes model. *UAI 1999* (pp. 447–456).

Pazzani, M. J. (1996). Searching for dependencies in Bayesian classifiers. In *Learning from data: AI and statistics V*. Springer-Verlag.

Vilalta, R., & Rish, I. (2003). A decomposition of classes via clustering to explain and improve naive Bayes. *ECML 2003* (pp. 444–455). Springer-Verlag.

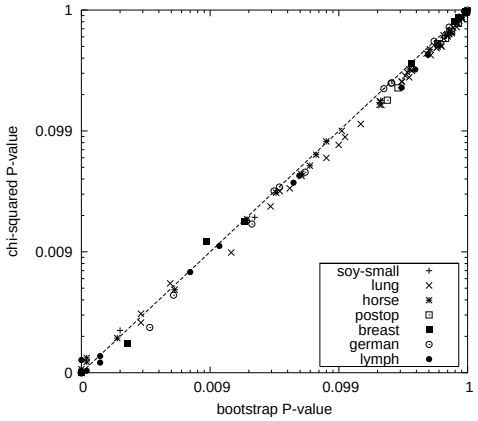

Figure 4. A comparison of P-values estimated by using the bootstrap and by assuming the χ^2 distribution.

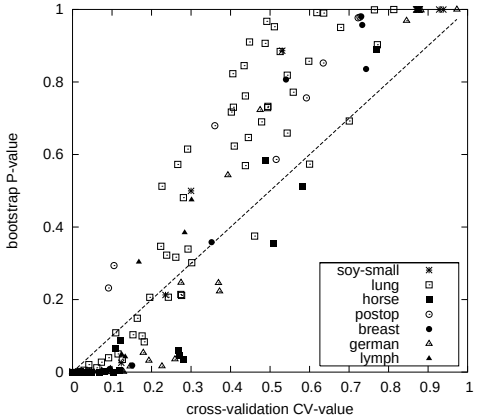

Figure 5. A comparison of P-values estimated with the bootstrap with the probability that the test set loss of the interaction-assuming model was not lower than that of the independence-assuming one in 2-fold cross-validation (CV-value).

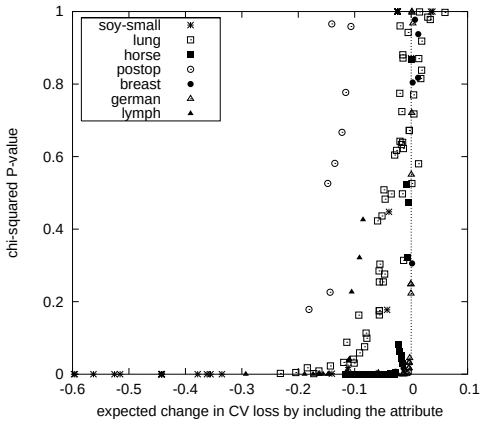

Figure 6. A comparison of P-values assuming χ^2 distribution with the average change in log-likelihood of the data given the information about the attribute value.