Qualitative Aspects in Numerical Data: Structure, Interaction, and Correlation

Aleks Jakulin

Faculty of Computer and Information Science, University of Ljubljana,
Tržaška cesta 25, SI-1001 Ljubljana, Slovenia,
jakulin@acm.org

Abstract. Entropy and information are common measures of probabilistic models of data, frequently used for discrete and discretized but more rarely for continuous data. We employ finite mixture models, which handle continuous and discrete data simultaneously, to construct a probabilistic model whose entropy can be estimated. Analytic estimates of entropy are intractable for some models, so an approximate sample estimate of entropy is used. A simplified formulation of bootstrap is employed to assess the distribution of entropy, which can then be represented with confidence intervals. We show how entropy can be used as a unified approach to quantifying three fundamental qualitative aspects of probabilistic models: the correlation aspect that corresponds to the linearities in the model, the structure aspect that helps capture model's nonlinearities, and the interaction aspect underlying the entanglements of attributes resulting from structure or correlation or both.

1 Introduction

A staggering number of modelling methods have been proposed in machine learning, data mining, statistics, pattern recognition and statistical physics over the years. While the researchers keeps pushing forward, seeking new types of problems and their solutions, there is a need to organize and integrate the underlying knowledge. This paper attempts to define a set of properties that practically all models are subject to. Instead of comparing models indiscriminately through predictive performance, it is more insightful to compare and study models with respect to these aspects at a higher level.

The loss function measures how well a model matches the data. For that purpose we employed the ubiquitous Shannon entropy [1]. For entropy to be applicable, the model should be uncertain, but handling uncertainty was found to be the necessary element of any method that attempts to model real-life data, so often noisy and incomplete. Entropy is closely related with the notion of information, which might be defined as a reduction in entropy achieved by the addition of model adjustments, additional data or additional attributes. We do not treat entropy as a constant measure of a model but maintain that entropy itself should be seen as an uncertain quantity. We employ the widely used bootstrap approach [2] to show how the uncertainty of entropy may be modelled with confidence intervals. Bootstrapping information also facilitates uncertain comparisons between uncertain models.

Given a probabilistic model we may evaluate the information it provides. The sources of information are different qualitative aspects the model exploited to reduce entropy. A frequently used aspect is correlation: allowing for a linear relationship among attributes and transforming them helps provide information. Another common aspect is structure: segmenting the data into several different groups or regions. The third aspect is interaction: taking advantage of the connections, the dependencies and the associations between attributes provides insight. There are many other aspects used by models, such as symmetry, central tendency, monotonicity, so our survey is by no means complete.

2 Probabilistic Modelling

This section will discuss the basic methodology for computing entropy from even an intractable probabilistic model of the data, connecting entropy and loss. We will interpret learning as an optimization problem with the objective of minimizing entropy of a particular probability density function. We will describe how probabilistic models for both continuous and discrete data can be learned using the finite mixture modelling approach. Finally, we will briefly review how the bootstrap procedure helps estimate the confidence intervals of uncertain quantities, such as entropy.

The terminology used will be as follows. An *instance* $\boldsymbol{i}$ corresponds to an event described with a number of attribute values. An *attribute* X is a unique property of instances that has a finite or infinite *range* $\Re_X$ of mutually exclusive values. The value of attribute X for instance $\boldsymbol{i}$ is $x_{\boldsymbol{i}} \in \Re_X$. If there are several attributes, we may represent them together as an attribute vector $\boldsymbol{X} = [X_1, X_2, \ldots, X_M]$, and we refer to $\Re_{\boldsymbol{X}}$ as the attribute space. The joint probability density function (PDF) p is a model of co-appearance of individual attribute values in an instance, and can be mathematically represented as a map $p : \Re_{\boldsymbol{X}} \to \mathbb{R}$, with the positivity $\forall \boldsymbol{x} : p(\boldsymbol{x}) \geq 0$ and the normalization condition $\int_{\Re_{\boldsymbol{X}}} p(\boldsymbol{x}) d\boldsymbol{x} = 1$. From the joint PDF we can always obtain a *marginal* PDF by removing or marginalizing one or more attributes by integrating over all the combinations of values of the removed attributes. For example, marginalizing $p(\boldsymbol{a}, \boldsymbol{b})$ over the attribute $\boldsymbol{b}$ would result in $p(\boldsymbol{a}) = \int_{\Re_{\boldsymbol{B}}} p(\boldsymbol{a}, \boldsymbol{b}) d\boldsymbol{b}$. A *conditional* PDF results when some attributes are being controlled for, and the distribution of $\boldsymbol{a}$ given $\boldsymbol{b}$ can be obtained from the joint through marginalization: $p(\boldsymbol{a}|\boldsymbol{b}) = p(\boldsymbol{a}, \boldsymbol{b})/p(\boldsymbol{a})$.

2.1 Entropy of Continuous Attribute Models

The entropy of a discrete attribute is an elementary exercise. It is not as clear how to compute the entropy of a continuous attribute or several of them. Although most approaches today are based on discretizing the continuous attributes, e.g. [3], Shannon [1] did not define entropy for a particular set of attributes, but for a joint *model* of the attributes, a particular joint probability density or distribution function (PDF) p. Entropy should be seen as a characteristic of a model and not of an attribute or data. For a multivariate joint PDF p modelling an attribute vector $\boldsymbol{X}$, the *differential entropy* [4] can be defined as:

$$h(\boldsymbol{X}|p(\boldsymbol{X})) \triangleq -\int_{\Re_{\boldsymbol{X}}} p(\boldsymbol{x}) \log_2 p(\boldsymbol{x}) d\boldsymbol{x} \tag{1}$$

Since an analytical derivation of differential entropy with this definition has been made only for a few distributions, the *sample entropy* (also referred to as empirical entropy) will instead be employed [5]. To estimate it, we start with a training multiset of instances $\mathcal{T} \subset \Re_{\boldsymbol{X}}$. We train a probabilistic model $p(\boldsymbol{X}|\mathcal{T})$ on this data with an arbitrary probabilistic learning algorithm (we will discuss a particular approach in Sect. 2.2). If we consider $\mathcal{T}$ as a representative sample of $\Re_{\boldsymbol{X}}$, the sample entropy corresponds to the expected negative log-likelihood in predicting a training instance with the model p: $\hat{h}(\mathcal{T}|p(\boldsymbol{X})) \triangleq E_{\boldsymbol{x}\in\mathcal{T}}\{-\log_2 p(\boldsymbol{x})\}$. This expectation is based on a uniform probability distribution over the instances, and the resulting sample entropy is the average negative log-likelihood of the model p for $\mathcal{T}$. This scheme can also be employed for computing the sample entropy of conditional and marginal entropies, the only difference being that the logarithm of a different kind of model is averaged over the training multiset. Marginalizing over $\boldsymbol{C}$, conditionalizing for $\boldsymbol{B}$, the sample entropy of $\boldsymbol{A}$ on a multiset of $|\mathcal{T}| = N$ instances is:

$$\hat{h}(\mathcal{T}|p(\boldsymbol{A}|\boldsymbol{B})) = -\frac{1}{N}\sum_{\boldsymbol{i}\in\mathcal{T}} \log_2 p(\boldsymbol{a_i}|\boldsymbol{b_i}) = \hat{h}(\mathcal{T}|p(\boldsymbol{A},\boldsymbol{B})) - \hat{h}(\mathcal{T}|p(\boldsymbol{B})) \tag{2}$$

The properties of differential entropy do not fully match those of discrete entropy. For example, the differential and sample entropies may be negative, and are sensitive to the choice of the coordinate system. Nonetheless, the magnitude and the sign of changes in sample entropy remain meaningful. Entropy should be understood generally as the loss or uncertainty of the model relative to some standard, and Shannon entropy results from the choice of a logarithmic loss function. Other loss functions may be employed and a corresponding notion of entropy thus derived [6], but its properties might not match those of Shannon entropy.

2.2 Learning Mixture Models

It has not yet been explained learn the joint PDF p from data. In machine learning, p is rarely given a priori (except perhaps as a Bayesian prior representing background knowledge), and must be inferred from the data. Since entropy can be viewed as loss, we can phrase learning p as an optimization task, trying to minimize the entropy of p by assessing it on the training data. The objective of unsupervised minimum entropy (for Shannon entropy this is clearly equivalent to maximum likelihood) learning is to minimize the sample entropy of the model $\arg\min_p \hat{h}(\mathcal{T}|p(\boldsymbol{X}, Y))$ if $\boldsymbol{X}$ are the attributes and Y is the *label*. It is unsupervised since the label Y is an ordinary attribute and plays no distinguished role.

The optimization is intractable without specifying the structure of p. Recently, *mixture models* have received much attention. They are very general, and many machine learning models including classification and regression trees, the naïve Bayesian classifier, rules, linear discriminants, instance-based learning algorithms, and others can be represented as mixture models. Mixture models are based on a set of *components*, each component is a probability density function in the attribute space. Each component has a corresponding probability of occurrence, and a point in the attribute space may have non-zero density for several components. If the set of components is finite, the model is a finite mixture [7].

Assume an attribute X and a *latent attribute* Z having a range $\Re_Z = \{z_1, \ldots, z_K\}$. Each latent attribute value identifies a specific component. Also assume a probability density function $p(X|\phi)$, where ϕ is its parameter. The distribution of X can be described with the following finite mixture model:

$$p(X|Z) = \sum_{k=1}^{K} \pi_k p(X|\phi_k) \tag{3}$$

Since the value of Z is unknown, we infer X using a multinomial model for Z: $p(z_k) = \pi_k$, $\sum_k \pi_k = 1$.

But what if there are several attributes? The assumption of *local independence* is that the latent attribute accounts for all the dependence between the attributes $\boldsymbol{X} = [X_1, X_2, \ldots, X_M]$:

$$p(\boldsymbol{X}|Z) = \sum_{k=1}^{K} \pi_k \prod_{m=1}^{M} p(X_m|\phi_{k,m}) \tag{4}$$

The naïve Bayesian classifier (NBC) of the label Y given the attributes $\boldsymbol{X}$ is identical to the above formulation of (4), but with the non-hidden label Y playing the role of the latent attribute. An added benefit of using local independence is that for computing the marginalizations of $\boldsymbol{X}$, all that needs to be done is to compute the product for a subset of attributes.

The choice of the functions in the mixture depends on the type of the attribute. Most implementations are based on normal or Gaussian mixtures, which work for continuous attributes, e.g. [5]. Recently, multinomial mixtures for discrete or count-based attributes have been successfully utilized in information retrieval, e.g. [8]. The MULTIMIX program [9] handles both continuous and discrete attributes simultaneously with the local independence assumption, adopting the multinomial distribution for any discrete attribute X_d (5) and the normal distribution for any continuous attribute X_c (6):

$$X_d \sim \text{Multinomial}(\boldsymbol{\lambda}, 1) \qquad p(X_d = x_j|\boldsymbol{\lambda}) = \lambda_j,\ \sum_j \lambda_j = 1 \tag{5}$$

$$X_c \sim \text{Normal}(\mu, \sigma) \qquad p(X_c = x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{6}$$

The model structure itself is only a part of the problem of modelling. We employed the expectation-maximization algorithm to determine the parameters

π and ϕ in (4). The EM algorithm is an iterative procedure for improving the fit of the model by interleaving two separate optimization steps. In the *expectation* step we compute the latent attribute value probabilities for each instance of the training multiset, while keeping π and ϕ constant. In the *maximization* step, we compute the maximum likelihood (and therefore also minimum sample entropy) parameter values for each component, given the set of instances having the latent attribute value which corresponds to the component: each instance is weighted with the probability that it belongs to the component. Because the distributions we use in the mixture are simple, the maximum likelihood equations can be solved analytically [9].

Instead of the common practice of using random values as initial parameter settings, each instance was assigned crisply to one of the clusters as found by *pam*, a robust greedy medoid-based clustering algorithm [10], instead of the first E step. To prevent correlated attributes from skewing the metric, the instances were presented to *pam* projected to their eigenspace using principal component analysis (PCA). Since PCA is sensitive to the scale of attributes, each attribute was standardized to have the mean of 0 and the variance of 1 beforehand.

2.3 The Bootstrap and Confidence Intervals

In Sect. 2.1 we treated sample entropy $\hat{h}$ as a fixed scalar value. As such, it might appear that a model achieving lower entropy is always better than a model that does not. Since this is unrealistic a great number of methods have been developed to detect sensitivity, complexity, instability, overfitting and other undesirable properties of models. The fundamental idea is that the model's performance on the data varies because the data is just a random sample (a further source of variation is that several models may have generated the same data, an epitome of Bayesian analysis). We will employ a scheme based on nonparametric bootstrap which may be distinguished from others by its simplicity.

Most learning algorithms make a few assumptions. One such assumption, especially important for probabilistic modelling, is that the instances were sampled independently one of another. This latter assumption is known as IID and makes the foundation for the bootstrap procedure [2]. The nonparametric bootstrap procedure is based on using sampling with replacement to construct a number of bootstrap samples or *resamples* of the original training multiset. If a particular parameter is computed on a set of resamples, this yields its bootstrap distribution.

What is of interest here is the bootstrap distribution of the sample entropy given p and the training multiset $\mathcal{T}$. If we train the PDF p on $\mathcal{T}$, we may compute the sample entropy on a set of resamples from $\mathcal{T}$ but keeping p fixed. Thereby we have a *bootstrap distribution* of bootstrap replications of entropy given $\mathcal{T}$ and p: $\Pr\{\hat{h}(\mathcal{T}^*|p)\}$, where $\mathcal{T}^*$ is a resample of $\mathcal{T}$. This formulation differs from the usual practice where p^* is relearned for each resample, but for our purposes this would be too time-consuming. The two important parameters to bootstrap are the size of each resample, which is usually kept to match the training multiset cardinality, and the number of bootstrap resamples, where we used 5000. The variance of sample entropy bootstrap distribution is connected with the variance of instance losses. The more uniformly consistent the model, the lower is the variance of the entropy distribution.

The bootstrap distribution can also be used to obtain confidence intervals of sample entropy. The 95% confidence interval by the percentile method is between the 2.5th and the 97.5th percentile of the bootstrap distribution. We may also compute the bootstrap distribution of the difference between two models p and q: $\Pr\{\hat{h}(\mathcal{T}^*|p) - \hat{h}(\mathcal{T}^*|q)\}$. If $\mathcal{T}$ was a representative sample from p, the result would correspond to the distribution of Kullback-Leibler divergence $D(p||q)$.

3 Qualitative Aspects

After we have prepared the statistical tools to perform probabilistic modelling, we can address the needs of data mining procedures: the discovery of interesting patterns in data. Human mind seems to prefer qualitative patterns more than the quantitative ones. In this section we will show that the tools of Sect. 2 can be employed to measure and test different hypotheses about the qualitative properties of a possibly black box predictive model. Although our analysis was based on finite mixture models, any other probabilistic model would be applicable instead.

An *aspect* is a particular qualitative property of a model. For example, the assumption of dependence between a group of attributes in a model can be understood as the interaction aspect. The assumption of structure in data, the need to distinguish groups, is another aspect. The assumption of correlation between two attributes is yet another aspect. These aspects all increase the complexity of the model. Other aspects have been studied, such as monotonic dependencies [11], related to the nonparametric correlation coefficient. Monotonic dependencies instead reduce the model complexity.

We quantify an aspect α in a unified way with the information gained from the data $\mathcal{T}$ in an attribute space $\Re_{\boldsymbol{X}}$ facilitated by allowing the aspect. Assume a probabilistic model $p(\boldsymbol{x}|\mathcal{T}, \alpha)$ which allows the aspect, and another similar model without the aspect $p(\boldsymbol{x}|\mathcal{T}, \neg\alpha)$. The aspect's *information gain* can be defined as the reduction in entropy it facilitates $I_\alpha \triangleq H(p|\neg\alpha) - H(p|\alpha)$. I_α should not be understood as a number but as a variable which has its own probability distribution. Wielding the IID assumption we use the bootstrap procedure of Sect. 2.3 to estimate the probability distribution of both sample entropy and aspect sample information gain: $\hat{\imath}_\alpha(\mathcal{T}|p) \triangleq \hat{h}(\mathcal{T}|p, \neg\alpha) - \hat{h}(\mathcal{T}|p, \alpha)$. Since the distribution is usually unimodal, we describe it with its mean and its 95% confidence interval.

The sample entropy of the data given a model represents the model's *bias*, and by introducing an aspect into the model we normally reduce the entropy and thus also the model's bias. However, introducing an aspect may also increase model's *variance*, the dispersion of the new model's entropy. We assess the variance by looking at the confidence intervals of aspect information gain and the new model's entropy, thus determining the unpredictability of the model performance, and the probability that the aspect-augmented model will perform worse than the non-augmented one.

We will now examine these three aspects through a series of experiments. The examples are vignettes describing the use of sample entropy in machine learning. They are not, however, systematic experimental evaluations of different methods. Our approach to analysis will be through *local analysis*, where we build a separate model for each subset of attributes and investigate its aspects. This should be distinguished from global analysis, where a model including all attributes is built and then analyzed by marginalization. The resulting global aspects do not provide similar quality of insight: by building a single global model, certain attributes' aspects may be neglected at the expense of properly representing others'.

3.1 Structure

Mixture modelling can be seen as an approach to constructive induction if the latent attribute Z of (3) and (4) is understood as a constructed nominal attribute, which allows capturing the intricacies of the data in the probabilistic model. In fact, it is only due to the latent attribute that any dependencies between different attributes can be represented. By increasing the number of components, we increase the complexity of the model, but decrease the sample entropy. Thereby we can define the structure aspect:

Definition 1. *The structure aspect is the reduction in entropy achieved by using K instead of K' components in a finite mixture model, $K > K'$.*

Each component can also be viewed as a separate rule, as a leaf in a classification tree, as a prototypical instance, or as a support vector. For example, the component identifying living people can be described with $temperature = 37℃ \pm 10$, while the component identifying healthy people is $temperature = 37℃ \pm 2$. By the principle of local analysis, we may investigate the structure aspect in small subsets of attributes, seeking useful patterns and trying to localize the complexity. The results of such an analysis are illustrated in Fig. 1.

Supervised, unsupervised and informative learning. The primary target in classification is predicting the label. Since the *unsupervised* learning approach of maximizing the joint likelihood is not directly concerned with this aim, a separate mixture model can be built for each label value. This is referred to as *informative* learning [12], and the objective is to minimize the entropy of the attributes given the label $\arg\min_p \hat{h}(\mathcal{T}|p(\boldsymbol{X}|Y))$. Class-conditional modelling of attributes is not, however, discriminative modelling of class boundaries, the true goal of pure supervised learning. In our entropic context we can formalize the objective of non-Bayesian supervised learning as minimization of the entropy of the predictive distribution of the label given the attributes $\arg\min_p \hat{h}(\mathcal{T}|p(Y|\boldsymbol{X}))$, fulfilling the fundamental task of minimizing the loss in predicting the label from the attributes. It is important to distinguish these three learning objectives, as they all differ one from another, in spite of the apparent similarity between supervised and informative learning. Fig. 2 illustrates that neither unsupervised

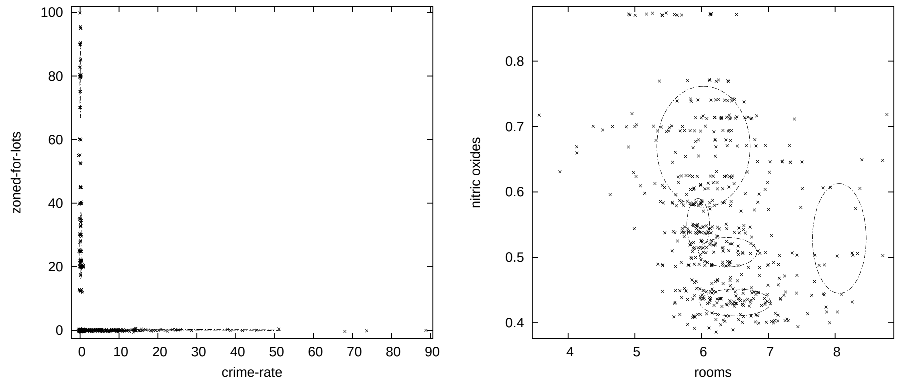

Fig. 1. On the 'housing' dataset we compared two models for each pair of attributes, one with the structure, based on five components, and an alternative with a single component but allowing correlation. Based on the structure information gain, the pairs of attributes were ranked. An axis-aligned ellipse depicts the circumference at one standard deviation for each component in the locally independent mixture. The pair with the highest structure information gain was the pair $(crime\ rate, zoned\ for\ lots)$, where only one of the two attributes may have a non-zero value for an instance (left). It is easy to see that at least two components are needed to describe this mutually-exclusive relationship, one vertical and one horizontal. On the other hand, allowing for structure yielded little benefit for the pair $(nitric\ oxides, rooms)$ (right).

nor informative learning match the supervised learning objectives, and that informative learning is not necessarily better than unsupervised learning.

3.2 Correlation

Correlation is an indication of a linear dependence among attributes. If there is a correlation, it is the linearly transformed attributes that are conditionally independent given the cluster in a locally independent mixture model.

Definition 2. *The correlation aspect is the reduction in entropy achievable by allowing a linear transformation of the attribute space within each component in the mixture model in the attribute space. The correlation aspect is defined for a particular number of components K.*

We now need a model that is able to allow for the correlation aspect. Instead of using a separate univariate normal distribution for each attribute within each component, we can use a single multivariate normal distribution, noting that this is no longer consistent with the local independence model (4): the vector of attributes $\boldsymbol{X}$ is now treated as a single multi-dimensional attribute. If a d-dimensional attribute $\boldsymbol{X} \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$p(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{7}$$

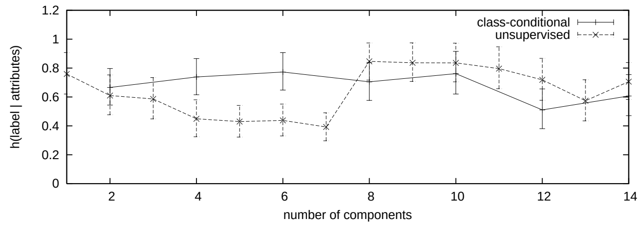

Fig. 2. This analysis on the 'voting' dataset demonstrates that increasing the number of components does not always result in better classification performance on the training set, regardless of whether an unsupervised or a class-conditional scheme is used. Furthermore, an unsupervised scheme may well yield better results than a class-conditional one, with smaller confidence intervals. The reason for this is that the maximum likelihood EM algorithm is not seeking to maximize the conditional likelihood of the label given the attributes, the goal of pure supervised learning.

The operational quantification of the correlation aspect is the information gained by using a multivariate normal distribution of attributes $\boldsymbol{X} = [X_1, X_2, \ldots, X_M]$ instead of a univariate normal distribution for each attribute X_i in a mixture model.

This scheme is not limited to two dimensions, and correlations between an arbitrary number of attributes can be investigated. Furthermore, it is possible to use the covariance matrix $\boldsymbol{\Sigma}$ to identify the either principal or independent components. From the viewpoint of constructive induction, these principal or independent components can be understood as the latent variables that model the continuous relationship between the attributes within each component. From Fig. 3 we can see that the correlation aspect information gain parallels the correlation coefficient, and how it is possible to combine both correlation and structure aspects. It must be noted that the more general structure aspect may capture the information that would otherwise be captured by correlation; therefore, preference should be given to correlation.

3.3 Interactions

An interaction is an aspect that indicates that a certain amount of entropy cannot be eliminated without seeing all the attributes at once. Although the past work in this area, e.g. [13], has been general in the sense that it was based on probability distributions, it has not been tested on continuous attributes without discretization. The general definition of a k-way interaction aspect can be phrased as:

Definition 3. *The aspect of a k-way interaction among k groups of attributes $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k\}$ is the reduction in entropy achievable by using the joint PDF of k attributes $p(\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k)$ rather than its an part-to-whole approximation reconstructed solely from the complete set of p's marginals $\mathcal{M} = \{p(\mathcal{A}_i^-);\ \mathcal{A}_i^- = \mathcal{A} \setminus \mathcal{A}_i,\ 1 \leq i \leq k\}$.*

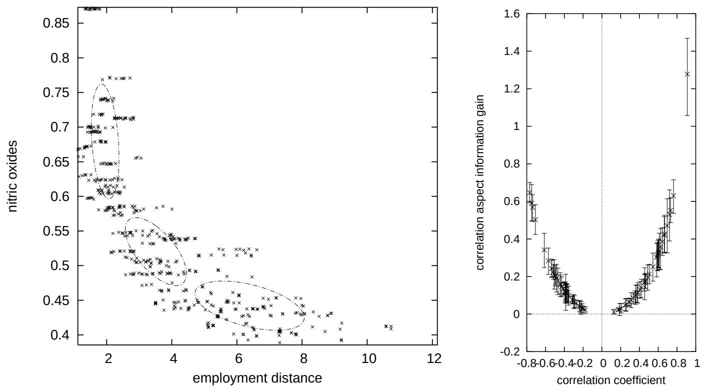

Fig. 3. Allowing for structure and correlation aspects at the same time by modelling with a multivariate normal mixture may unravel nonlinearities in the data, with each component (drawn as an ellipse) capturing localized linearity (left). Instead of using the correlation coefficient, we may express the correlation with correlation aspect information gain (right). While retaining monotonicity, the scale of information gain is more logical than that of the correlation coefficient, as correlation coefficients lower than 0.3 are known to be uninteresting. The large confidence interval on the extreme right should raise suspicion: that particular pair of attributes' high correlation is merely due to outliers.

To obtain an operational definition, without getting into the intricacies of part-to-whole approximations, all we need to define is how to compute the sample entropy for the reconstruction. The basic definition of interaction information for a set of attributes $\mathcal{A}$ can be used to that aim [13]:

$$\hat{\imath}_{\mathcal{A}}(\mathcal{T}|p) \triangleq -\sum_{\mathcal{X}\subseteq\mathcal{A}} (-1)^{|\mathcal{A}|-|\mathcal{X}|}\hat{h}(\mathcal{T}|p(\mathcal{X})) \tag{8}$$

It is possible to obtain $p(\mathcal{X})$ easily by marginalizing $p(\mathcal{A})$. Using this definition of k-way interaction aspect sample information gain for $k = 1, 2, 3$ it is possible to create interaction graphs and other visualizations [14]. The interaction graph of a regression dataset with a continuous label and a mix of nominal and continuous attributes is illustrated in Fig. 4. Only the interactions involving the label were investigated.

It is important to note the dependence of the interaction aspect on correlation and structure aspects. It is easy to see that interaction aspects only appear along with correlation and structure aspects, but not in their absence. Therefore, the joint PDF p used for interaction analysis should always include correlation and/or structure aspects.

4 Discussion

Most problems in estimating entropy of data are tied with finding a good probabilistic model of the data. Apart from having appealing properties (some of which

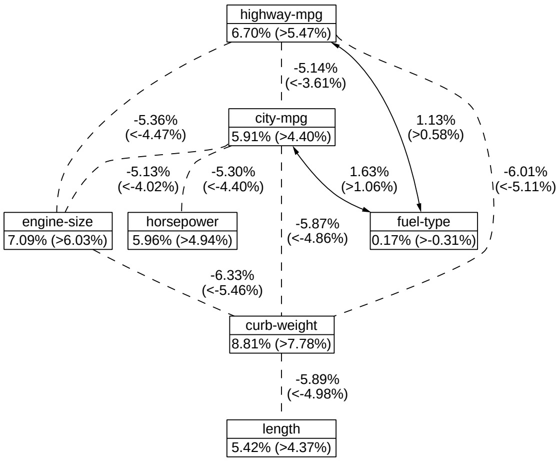

Fig. 4. This interaction graph identifies the strongest 2-way and 3-way interactions in the 'imports-85' dataset with the price of a car as the continuous label. For each potential interaction, a five-component joint mixture model was built, and the aspect sample information gain estimated with a 95% confidence interval. The sample information gain was expressed as a proportion of the label sample entropy. The numbers below each attribute indicate the proportion of label entropy the attribute eliminates, with a bottom bound. For example, *highway mpg* alone eliminates 6.7% of uncertainty about the price on average, but in 97.5% of cases more than 5.5%. *fuel type* is apparently a useless attribute on its own, eliminating only 0.2% of entropy, but there is a positive interaction or a synergy between fuel type and the fuel consumption on the highway, eliminating an additional 1.13% of label entropy. Dashed edges indicate negative interactions or redundancies, where two attributes provide partly the same information about the label. For example, should we consider the fuel consumption both on highways and in the city, the total amount of label entropy eliminated would be $6.7 + 5.9 - 5.1$ percent, accounting for their overlap. Due to the imprecision of sample entropy and the unsupervised modelling criteria, apparent illogicalities may appear: the length of the automobile is hurting the predictions of the car's price in combination the car's weight.

are not retained by sample entropy), entropy can be seen as a prototypical loss function which measures the quality of a particular model, and, as information, the worth of changes to the model. For a practical prediction task, however, entropy should be replaced with a more realistic cost function if a utility function or a cost matrix are given.

Our bootstrap scheme is fast, simple, and largely corresponds to the confidence intervals as used in statistics. In fact, because we do not retrain the model for each resample and because sampling of per-instance loss is so simple, an analytic procedure could easily replace random sampling. The disadvantage is that it does not verify the models' ability to generalize upon unseen instances. For that purpose, other means of data perturbation, such as multiply replicated cross-validation may instead be employed to construct the confidence intervals.

Alternatively, either Bayesian modelling or the usual formulation of bootstrap will match overfitting with high model variance.

There are other important issues we did not cover for the lack of space. We have not systematically evaluated mixture models in comparison with other machine learning methods, but we refer an interested reader to [15, 16], where using a a separate latent attribute Z_y for each value y of the label, and therefore using multiple components per each label value was shown to improve the classification accuracy. We have not explained how to determine the optimal number of components, but there is already a considerable bibliography on this problem.

References

1. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27** (1948) 379–423, 623–656
2. Efron, B., Gong, G.: A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician **37** (1983) 36–48
3. Paninski, L.: Estimation of entropy and mutual information. Neural Computation **15** (2003) 1191–1253
4. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley Series in Telecommunications. Wiley, New York (1991)
5. Roberts, S.J., Everson, R., Rezek, I.: Maximum certainty data partitioning. Pattern Recognition **33** (1999) 833–839
6. Grünwald, P.D., Dawid, A.P.: Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. Annals of Statistics **32** (2004)
7. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
8. Buntine, W.: Variational extensions to EM and multinomial PCA. In Elomaa, T., Mannila, H., Toivonen, H., eds.: ECML 2002. Volume 2430 of LNCS., Springer-Verlag (2002)
9. Hunt, L., Jorgensen, M.: Mixture model clustering using the MULTIMIX program. Australian & New Zealand Journal of Statistics **41** (1999)
10. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
11. Šuc, D., Bratko, I.: Induction of qualitative trees. In De Raedt, L., Flach, P., eds.: ECML 2001. Volume 2167 of LNAI., Springer-Verlag (2001) 442–453
12. Rubinstein, Y.D., Hastie, T.: Discriminative vs informative learning. In: SIGKDD 1997, AAAI Press (1997) 49–53
13. Jakulin, A., Bratko, I.: Quantifying and visualizing attribute interactions: An approach based on entropy. http://arxiv.org/abs/cs.AI/0308002 v3 (2004)
14. Jakulin, A., Bratko, I.: Analyzing attribute dependencies. In Lavrač, N., Gamberger, D., Blockeel, H., Todorovski, L., eds.: PKDD 2003. Volume 2838 of LNAI., Springer-Verlag (2003) 229–240
15. Monti, S., Cooper, G.F.: A Bayesian network classifier that combines a finite mixture model and a naive-Bayes model. In: UAI 1999. (1999) 447–456
16. Vilalta, R., Rish, I.: A decomposition of classes via clustering to explain and improve naive Bayes. In Lavrač, N., Gamberger, D., Blockeel, H., Todorovski, L., eds.: ECML 2003. Volume 2837 of LNAI., Springer-Verlag (2003) 444–455