IN YOUR INTEREST
Objective Interestingness Measures for a Generative Classifier

Dominik Fisch, Edgar Kalkowski, Bernhard Sick
Computationally Intelligent Systems Lab, University of Applied Sciences Deggendorf, Deggendorf, Germany

Seppo J. Ovaska
Aalto University, School of Science and Technology, Espoo, Finland

Keywords: Classification, Data mining, Interestingness.

Abstract: In a wide-spread definition, data mining is termed to be the "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". In real applications, however, usually only the validity of data mining results is assessed numerically. An important reason is that the other properties are highly subjective, i.e., they depend on the specific knowledge and requirements of the user. In this article we define some objective interestingness measures for a specific kind of classifier, a probabilistic classifier based on a mixture model. These measures assess the informativeness, uniqueness, importance, discrimination, comprehensibility, and representativity of rules contained in this classifier to support a user in evaluating data mining results. With some simulation experiments we demonstrate how these measures can be applied.

1 INTRODUCTION

Data mining (DM)—today typically used as a synonym of *knowledge discovery in databases* (KDD)—deals with the *detection of interesting patterns* (e.g., regularities) in often huge amounts of data and the *acquisition of knowledge* (e.g., classification rules) in application fields such as marketing, fraud detection, drug design, and many more. In a well-known definition, it is termed to be the "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et al., 1996). But, how can this "interestingness" of patterns be assessed, in particular numerically? It is obvious that attributes such as novel, useful, and understandable are highly subjective as they depend on the particular needs and the previous knowledge of the data miner. Thus, usually only the validity of patterns or extracted knowledge is assessed numerically in order to get an objective validation of DM results.

In this article, we focus on some other attributes of data mining results that can be measured numerically. They are objective on the one hand and related to attributes such as novelty, usefulness, and understandability on the other. For that purpose, we use a specific classifier, a *classifier* based on probabilistic (Gaussian) *mixture models* (CMM), see also (Fisch and Sick, 2009; Bishop, 2006). CMM contain rules that have a form similar to that of fuzzy rules but they must be interpreted in a probabilistic way. A rule premise aims at modeling a data cluster in the input space of a classifier, while the conclusion assigns that cluster to a certain class. Our new interestingness measures assess the *informativeness*, *uniqueness*, *importance*, *discrimination*, *comprehensibility*, and *representativity* of rules contained in a CMM in order to support a user in evaluating DM results.

In the remainder of the article we briefly discuss some related work in Section 2. Then, we describe the classifier and introduce the various interestingness measures in Section 3. Three case studies in Section 4 show how these measures could be applied. Finally, we briefly conclude in Section 5 and also give an outlook to future work.

2 RELATED WORK

Basically, there are subjective and objective interestingness measures that are used to assess rules extracted from data in a DM process, see, e.g., (Hilderman and Hamilton, 2001; McGarry, 2005).

Objective measures are solely based on an anal-

Fisch D., Kalkowski E., Sick B. and J. Ovaska S..
IN YOUR INTEREST - Objective Interestingness Measures for a Generative Classifier.
DOI: 10.5220/0003186404140423
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 414-423
ISBN: 978-989-8425-40-9

ysis of the extracted knowledge. These interestingness measures are based, for example, on information criteria or on data-based evaluation techniques. Typical examples are Akaike's information criterion or the Bayesian information criterion on the one hand and statistical measures such as sensitivity, specificity, precision etc. computed in a cross-validation or a bootstrapping approach on training/test data on the other (cf. (Duda et al., 2001; Tan et al., 2004), for instance). Other criteria that assess the complexity of rules or rule sets are, e.g., a rule system size measure (gives the overall number of rules in the rule system), a computational complexity measure (CPU time required for the evaluation of a rule or a rule system), a rule complexity measure (number of attributes that are tied together in a rule), a mean scoring rules measure (average number of rules that have to be applied to come to a conclusion), a fuzzy quality measure (for terms such as "bad", "average", or "very good" that are associated with rules), the information gain for association rules (Atzmueller et al., 2004; Taha and Ghosh, 1997; Nauck, 2003; Hebert and Cremilleux, 2007). Also, measures are combined (e.g., in form of a weighted sum) in some cases (Atzmueller et al., 2004; Taha and Ghosh, 1997).

Subjective measures consider additional knowledge about the application field and / or information about the user of a DM system, e.g., skills and needs (Piatetsky-Shapiro and Matheus, 1994; Padmanabhan and Tuzhilin, 1999). Subjective interestingness measures mentioned in the literature are, for example, novelty (Basu et al., 2001; Fayyad et al., 1996), usefulness (Fayyad et al., 1996), understandability (Fayyad et al., 1996), actionability (Silberschatz and Tuzhilin, 1996), and unexpectedness (Padmanabhan and Tuzhilin, 1999; Silberschatz and Tuzhilin, 1996; Di Fiore, 2002; Liu et al., 2000). The existing measures use different techniques to represent information about the human domain experts and they also greatly depend on the respective kind of knowledge representation, e.g., Bayesian networks, fuzzy classifiers, or association rules.

3 METHODOLOGICAL FOUNDATIONS

In this section we will first present the generative classifier paradigm. A *generative* classifier aims at modeling the processes underlying the "generation" of the data (Bishop, 2006). We use probabilistic techniques for that purpose. Then, we will describe our new interestingness measures.

3.1 Probabilistic Classifier CMM

3.1.1 Definition of CMM

The classifiers we are using here are probabilistic classifiers, i.e., classifiers based on mixture models (CMM). That is, for a given D-dimensional input pattern $\mathbf{x}'$ we want to compute the posterior distribution $p(c|\mathbf{x}')$, i.e., the probabilities for class membership (with classes $c \in \{1,\ldots,C\}$) given the input $\mathbf{x}'$. To minimize the risk of classification errors we then select the class with the highest posterior probability (cf. the principle of *winner-takes-all*), for instance. According to (Fisch and Sick, 2009), $p(c|\mathbf{x})$ can be decomposed as follows:

$$p(c|\mathbf{x}) = \frac{p(c)p(\mathbf{x}|c)}{p(\mathbf{x})} = \frac{p(c)\sum_{i=1}^{I_c} p(i|c)p(\mathbf{x}|c,i)}{p(\mathbf{x})} \tag{1}$$

where

$$p(\mathbf{x}) = \sum_{c'=1}^{C} p(c') \sum_{i=1}^{I_{c'}} p(i|c')p(\mathbf{x}|c',i). \tag{2}$$

This approach is based on C *mixture density models* $\sum_{i=1}^{I_c} p(\mathbf{x}|c,i)p(i|c)$, one for each class. Here, the conditional densities $p(\mathbf{x}|c,i)$ with $c \in \{1,\ldots,C\}$ and $i \in \{1,\ldots,I_c\}$ are called *components*, the $p(i|c)$ are multinomial distributions with parameters $\pi_{c,i}$ (*mixing coefficients*), and $p(c)$ is a multinomial distribution with parameters γ_c (*class priors*).

That is, we have a classifier consisting of $J = \sum_{c=1}^{C} I_c$ components, where each component is described by a distribution $p(\mathbf{x}|c,i)$. To keep the notation uncluttered, in the following a specific component is identified by a single index $j \in \{1,\ldots,J\}$ (i.e., $p(\mathbf{x}|j)$) if its class is not relevant.

Which kind of density functions can we use for the components? Basically, a D-dimensional pattern $\mathbf{x}$ may have D_{cont} continuous (i.e., real-valued) dimensions (attributes) and $D_{\text{cat}} = D - D_{\text{cont}}$ categorical ones. Without loss of generality we arrange these dimensions such that

$$\mathbf{x} = (\underbrace{x_1,\ldots,x_{D_{\text{cont}}}}_{\text{continuous}},\underbrace{\boldsymbol{x}_{D_{\text{cont}}+1},\ldots,\boldsymbol{x}_D}_{\text{categorical}}). \tag{3}$$

Note that we italicize x when we refer to single dimensions. The continuous part of this vector $\mathbf{x}^{\text{cont}} = (x_1,\ldots,x_{D_{\text{cont}}})$ with $x_d \in \mathbb{R}$ for all $d \in \{1,\ldots,D_{\text{cont}}\}$ is modeled with a multivariate *normal* (i.e., Gaussian) distribution with center $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. That is, with $det(\cdot)$ denoting the determinant of a matrix we use the model

$$\mathcal{N}(\mathbf{x}^{\text{cont}}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D_{\text{cont}}}{2}} det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-0.5\,(\Delta_{\boldsymbol{\Sigma}}(\mathbf{x}^{\text{cont}},\boldsymbol{\mu}))^2\right) \quad (4)$$

with the distance measure (matrix norm) $\Delta_{\mathbf{M}}(\mathbf{v}_1,\mathbf{v}_2)$ given by

$$\Delta_{\mathbf{M}}(\mathbf{v}_1,\mathbf{v}_2) = \sqrt{(\mathbf{v}_1-\mathbf{v}_2)^T\mathbf{M}^{-1}(\mathbf{v}_1-\mathbf{v}_2)}. \quad (5)$$

$\Delta_{\mathbf{M}}$ defines the *Mahalanobis distance* of vectors $\mathbf{v}_1,\mathbf{v}_2 \in \mathbb{R}^{D_{\text{cont}}}$ based on a $D_{\text{cont}} \times D_{\text{cont}}$ covariance matrix $\mathbf{M}$. For many practical applications, the use of Gaussian components or Gaussian mixture models can be motivated by the generalized *central limit theorem* which roughly states that the sum of independent samples from any distribution with finite mean and variance converges to a normal distribution as the sample size goes to infinity (Duda et al., 2001).

For categorical dimensions we use a 1-of-K_d coding scheme where K_d is the number of possible categories of attribute $\boldsymbol{x}_d$ ($d \in \{D_{\text{cont}}+1,\ldots,D\}$). The value of such an attribute is represented by a vector $\boldsymbol{x}_d = (x_{d_1},\ldots,x_{d_{K_d}})$ with $x_{d_k} = 1$ if $\boldsymbol{x}_d$ belongs to category k and $x_{d_k} = 0$ otherwise. The classifier models categorical dimensions by means of *multinomial* distributions. That is, for an input dimension (attribute) $\boldsymbol{x}_d \in \{\boldsymbol{x}_{D_{\text{cont}}+1},\ldots,\boldsymbol{x}_D\}$ we use

$$\mathcal{M}(\boldsymbol{x}_d|\boldsymbol{\delta}_d) = \prod_{k=1}^{K_d} \delta_k^{x_{d_k}} \quad (6)$$

with $\boldsymbol{\delta}_d = (\delta_{d_1},\ldots,\delta_{K_d})$ and the restrictions $\delta_{d_k} \geq 0$ and $\sum_{k=1}^{K_d}\delta_{d_k} = 1$.

We assume that the categorical dimensions are mutually independent and that there are no dependencies between the categorical and the continuous dimensions. Then, the component densities $p(\mathbf{x}|j)$ are defined by

$$p(\mathbf{x}|j) = \mathcal{N}(\mathbf{x}^{\text{cont}}|\mu_j,\boldsymbol{\Sigma}_j)\cdot \prod_{d=D_{\text{cont}}+1}^{D} \mathcal{M}(\mathbf{x}_d|\boldsymbol{\delta}_{j_d}). \quad (7)$$

3.1.2 Training of CMM

How can the various parameters of the classifier be determined? For a given training set $\mathbf{X}$ with N samples (patterns) $\mathbf{x}_n$ it is assumed that the $\mathbf{x}_n$ are independent and identically distributed. First, $\mathbf{X}$ is split into C subsets $\mathbf{X}_c$, each containing all samples of the corresponding class c, i.e.,

$$\mathbf{X}_c = \{\mathbf{x}_n|\mathbf{x}_n \text{ belongs to class } c\}. \quad (8)$$

Then, a mixture model is trained for each $\mathbf{X}_n$. Here, we perform the parameter estimation by means of a technique called *variational Bayesian inference* (VI) which realizes the Bayesian idea of regarding the model parameters as random variables whose distributions must be trained (Fisch and Sick, 2009). This approach has two important advantages over other methods. First, the estimation process is more robust, i.e., it avoids "collapsing" components, so-called singularities whose variance in one or more dimensions vanishes. Second, VI optimizes the number of components by its own. For a more detailed discussion on Bayesian inference, and, particularly, VI see (Bishop, 2006). More details concerning the training algorithm can be found in (Fisch and Sick, 2009).

At this point, we have found parameter estimates for the $p(\mathbf{x}|c,i)$ and $p(i|c)$, cf. Eq. (1). The parameters for the class priors $p(c)$ are estimated with

$$\gamma_c = \frac{|\mathbf{X}_c|}{|\mathbf{X}|} \quad (9)$$

where $|\mathbf{S}|$ denotes the cardinality of the set $\mathbf{S}$.

3.1.3 Rule Extraction from CMM

In some applications it is desirable to extract human-readable rules from the trained classifier. This is possible with CMM if they are parametrized accordingly. For the moment we focus on a single component $p(\mathbf{x}|j)$ and omit the identifying index j. Basically, there are no restrictions necessary concerning the covariance matrix $\boldsymbol{\Sigma}$ or the number of categories K_d. However, if the covariance matrix is forced to be diagonal (i.e., assuming that there are no dependencies between continuous input dimensions), the multivariate Gaussian $\mathcal{N}(\mathbf{x}^{\text{cont}}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ can be split into a product consisting of D_{cont} univariate Gaussians ψ_d with $d \in \{1,\ldots,D_{\text{cont}}\}$. A categorical dimension $d \in \{D_{\text{cont}}+1,\ldots,D\}$ can be simplified by only considering the n_d "important" categories k_{d_i} ($i = 1,\ldots,n_d$), i.e., those with a probability δ_{d_i} above the average $1/K_d$. The probabilities of these categories are renormalized and the remaining categories are discarded. Then, a rule like the following can be extracted from a component:

if x_1 is ψ_1 and … and $x_{D_{\text{cont}}}$ is $\psi_{D_{\text{cont}}}$
  and ($x_{D_{\text{cont}}+1} = k_{(D_{\text{cont}}+1)_1}$ or …
    or $x_{D_{\text{cont}}+1} = k_{(D_{\text{cont}}+1)_{n_{D_{\text{cont}}+1}}}$)
  ⋮
  and ($x_D = k_{D_1}$ or … or $x_D = k_{D_{n_D}}$)
then c_1 is 0 and c_2 is 1 and …

The whole CMM can, thus, be transformed into a rule set whose variables are the input variables (di-

mensions of the input variable $\mathbf{x}$) and the output variable c which represents the classes. The rule premises are realized by conjunctions of the univariate Gaussians ψ_d ($d = 1,\ldots,D_{\text{cont}}$) and the simplified categorical dimensions. The latter are modeled by disjunctions of the categories. The conclusions (i.e., the class membership) are given by the class-dependent GMM to which the component $p(\mathbf{x}|j)$ belongs.

These rules enable reasoning based on uncertain observations as shown in Eq. (1). The final classification decision is obtained by superimposing the rule conclusions weighted with the degree of membership given by the rule premises, the mixing coefficients and the class priors. The extracted rules have a form which is very similar to that of fuzzy rules, but they have a very different (i.e., probabilistic) interpretation, cf. (Fisch et al., 2010).

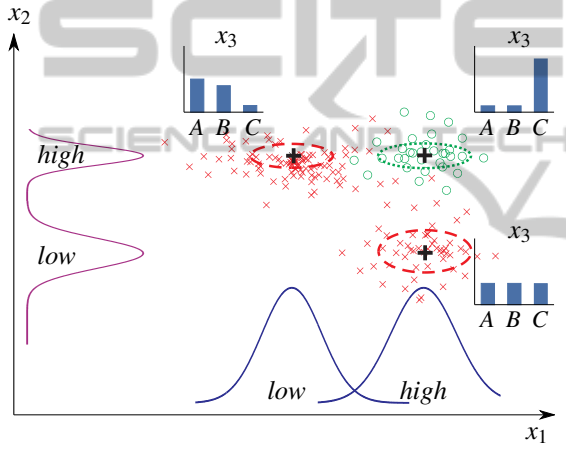


Figure 1: Example of a Classifier Consisting of Three Rules.

Fig. 1 gives an example for such a classifier consisting of three components in a three-dimensional input space. The first two dimensions x_1, x_2 are continuous and, thus, modeled by bivariate Gaussians whose centers are described by the crosses (+). The ellipses are level curves (surfaces of constant density) with shapes defined by the covariance matrices. Due to the diagonality of the matrices these ellipses are axes-oriented and the projection of the corresponding bivariate Gaussian onto the axes is also shown. The third dimension x_3 is categorical. The trained distribution of categories is illustrated by the histogram next to every component. For this CMM, the following rule set can be extracted:

if x_1 is *low* and x_2 is *high* and (x_3 is A or x_3 is B)
then $c_1 = 1$ and $c_2 = 0$
if x_1 is *high* and x_2 is *high* and (x_3 is C)
then $c_1 = 0$ and $c_2 = 1$
if x_1 is *high* and x_2 is *low*
then $c_1 = 1$ and $c_2 = 0$

Of course, this readability is accomplished at the cost of a limited modeling capability of the classifier (i.e., restricted covariance matrices and simplified categorical dimensions) and should, thus, only be used if the application demands this kind of human-readable rules.

3.2 Objective Interestingness Measures for CMM

In the following we describe some new interestingness measures that can be taken to assess a classifier based on CMM in an objective way. If the class a component belongs to is not relevant for the assessment, the component is identified by a single index $j \in \{1,\ldots,J\}$, i.e., $p(\mathbf{x}|j)$. Otherwise, it is explicitely denoted with $p(\mathbf{x}|c,i)$. If sample data are needed to evaluate a measure, we use the training data for that purpose. In addition, classical performance measures (e.g., classification error on independent test data) should be used. The knowledge we want to assess is represented by the components of which the CMM consists. We will use the term rule instead of component only if we wish to explicitely extract human-readable rules from the CMM.

3.2.1 Informativeness

A component of the CMM is considered as being very informative if it describes a really distinct kind of process "generating" data. To assess the *informativeness* of a component numerically we use the *Hellinger* distance $\mathrm{H}(p(\mathbf{x}),q(\mathbf{x}))$ of two probability densities $p(\mathbf{x})$ and $q(\mathbf{x})$. Compared to other statistical distance measures such as the Kullback-Leibler divergence it has the advantage of being bounded between 0 and 1. It is defined by

$$\mathrm{H}(p(\mathbf{x}),q(\mathbf{x})) = \sqrt{1-\mathrm{BC}(p(\mathbf{x}),q(\mathbf{x}))}, \qquad (10)$$

where $\mathrm{BC}(p(\mathbf{x}),q(\mathbf{x}))$ denotes the *Bhattacharyya coefficient* defined by

$$\mathrm{BC}(p(\mathbf{x}),q(\mathbf{x})) = \int \sqrt{p(\mathbf{x})q(\mathbf{x})}\,\mathrm{d}\mathbf{x}. \qquad (11)$$

$\mathrm{H}(p(\mathbf{x}),q(\mathbf{x}))$ is 0 if $p(\mathbf{x})$ and $q(\mathbf{x})$ describe the same distribution and it approaches 1 when $p(\mathbf{x})$ places most of its probability mass in regions where $q(\mathbf{x})$ assigns a probability of nearly zero and vice versa.

Using Fubini's theorem and considering the discrete nature of the multinomial distribution, the Bhattacharyya coefficient of two components $p(\mathbf{x}|j)$ and $p(\mathbf{x}|j')$, as defined in Eq. (7), can be computed by

$$\begin{aligned}&\mathrm{BC}(p(\mathbf{x}|j), p(\mathbf{x}|j')) = \\ &\int \sqrt{\mathcal{N}(\mathbf{x}^{\mathrm{cont}}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\mathcal{N}(\mathbf{x}^{\mathrm{cont}}|\boldsymbol{\mu}_{j'}, \boldsymbol{\Sigma}_{j'})}\, \mathrm{d}\mathbf{x}^{\mathrm{cont}} \\ &\cdot \prod_{d=D_{\mathrm{cont}}+1}^{D} \sum_{k=0}^{K_d} \sqrt{\mathcal{M}(\boldsymbol{e}_k|\boldsymbol{\delta}_{j_d})\mathcal{M}(\boldsymbol{e}_k|\boldsymbol{\delta}_{j'_d})}\end{aligned} \tag{12}$$

with $\boldsymbol{e}_k$ being the k-th row of a $K_d \times K_d$ identity matrix (i.e., we are iterating over all K_d possible categories of dimension d). The integral can be solved analytically by

$$\begin{aligned}&\int \sqrt{\mathcal{N}(\mathbf{x}^{\mathrm{cont}}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\mathcal{N}(\mathbf{x}^{\mathrm{cont}}|\boldsymbol{\mu}_{j'}, \boldsymbol{\Sigma}_{j'})}\, \mathrm{d}\mathbf{x}^{\mathrm{cont}} = \\ &\exp\left(-\frac{1}{8}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{j'})^{\mathrm{T}}\left(\frac{\boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_{j'}}{2}\right)^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{j'})\right) \\ &\cdot \frac{\sqrt[4]{det(\boldsymbol{\Sigma}_j)det(\boldsymbol{\Sigma}_{j'})}}{\sqrt{det\left(\frac{\boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_{j'}}{2}\right)}}.\end{aligned} \tag{13}$$

The informativeness of a component $p(\mathbf{x}|j)$ is then determined by its Hellinger distance calculated with respect to the "closest" component $p(\mathbf{x}|j')$ ($j' \neq j$) contained in the CMM:

$$\mathrm{info}(p(\mathbf{x}|j)) = \min_{j' \neq j}(\mathrm{H}(p(\mathbf{x}|j'), p(\mathbf{x}|j)). \tag{14}$$

3.2.2 Uniqueness

The knowledge modeled by the components within a CMM should be unambiguous. This is measured by the *uniqueness* of a component $p(\mathbf{x}|c,i)$ which reflects to which degree samples belonging to different classes are covered by that component. Let $\rho_{c,i}(\mathbf{x}_n)$ denote the *responsibility* of component $p(\mathbf{x}|c,i)$ for the generation of sample $\mathbf{x}_n$

$$\rho_{c,i}(\mathbf{x}_n) = \frac{p(c)p(i|c)p(\mathbf{x}_n|c,i)}{p(\mathbf{x}_n)}. \tag{15}$$

Then, we define the uniqueness of rule $p(\mathbf{x}|c,i)$ by

$$\mathrm{uniq}(p(\mathbf{x}|c,i)) = \frac{\sum_{\mathbf{x}_n \in \mathbf{X}_c} \rho_{c,i}(\mathbf{x}_n)}{\sum_{\mathbf{x}_n \in \mathbf{X}} \rho_{c,i}(\mathbf{x}_n)}. \tag{16}$$

3.2.3 Importance

The *importance* of a component measures the relative weight of a component within the classifier. A component $p(\mathbf{x}|c,i)$ is regarded as very important if its mixing coefficient weighted with the class prior $\pi_{c,i} \cdot p(c)$ is far above the average mixing coefficient $\overline{\pi} = \frac{1}{J}$. To scale the importance of a component to the interval $[0,1]$ we use a boundary function that is comprised of two linear functions. One maps all mixing coefficients that are smaller than the average to the interval $[0,0.5]$ and the other maps all mixing coefficients that are larger than the average to $[0.5,1]$. The importance of component $p(\mathbf{x}|c,i)$ is then computed by

$$\mathrm{impo}(p(\mathbf{x}|c,i)) = \begin{cases} \frac{1}{2} \cdot \frac{\pi_{c,i} \cdot p(c)}{\overline{\pi}}, & \pi \leq \overline{\pi} \\ \frac{1}{2} \cdot \left(\frac{\pi_{c,i} \cdot p(c)}{1-\overline{\pi}} - \frac{\overline{\pi}}{1-\overline{\pi}} + 1\right), & \pi > \overline{\pi} \end{cases}. \tag{17}$$

3.2.4 Discrimination

The *discrimination* measure evaluates the influence of a component $p(\mathbf{x}|c,i)$ on the decision boundary—and, thus, on the classification performance—of the overall classifier. To calculate the discrimination of component $p(\mathbf{x}|c,i)$ we create a second CMM by removing $p(\mathbf{x}|c,i)$ from the original CMM and renormalizing the mixing coefficients of the remaining components. Then, we compare the achieved classification error on training data of the original CMM ($\mathcal{E}_{\mathrm{with}}$) to the classification error of the CMM without component $p(\mathbf{x}|c,i)$ ($\mathcal{E}_{\mathrm{without}}$) weighted with the corresponding class prior $p(c)$:

$$\mathrm{disc}(p(\mathbf{x}|c,i)) = \frac{\mathcal{E}_{\mathrm{without}} - \mathcal{E}_{\mathrm{with}}}{p(c)}. \tag{18}$$

If required by a concrete application (e.g., in some medical applications false positives are acceptable whereas false negatives could be fatal), it is also possible to use more detailed measures such as sensitivity, specificity, or precision to assess the discrimination of a component.

3.2.5 Representativity

The performance of a generative classifier highly depends on how well it "fits" the data. This kind of fitness is determined by the continuous dimensions since we explicitely assume that the data distribution can be modeled by a mixture of Gaussian distributions. For the categorical dimensions we do not assume any functional form. Therefore, the *representativity* measure only considers the continous dimensions $\mathbf{x}^{\mathrm{cont}}$. Again, we use the Hellinger distance $\mathrm{H}(p(\mathbf{x}^{\mathrm{cont}}), q(\mathbf{x}^{\mathrm{cont}}))$, cf. Eq. (10), and measure the distance between the true data distribution $q(\mathbf{x}^{\mathrm{cont}})$ and the model $p(\mathbf{x}^{\mathrm{cont}})$ (i.e., the multivariate Gaussian

part), cf. Eq. (2) and (4). As for real-world data sets the true underlying distribution $q(\mathbf{x}^{\text{cont}})$ is unknown, it is approximated with a non-parametric density estimator consisting of a *Parzen window* density estimator:

$$q(\mathbf{x}^{\text{cont}}) = \frac{1}{N} \sum_{\mathbf{x}_n \in \mathbf{X}} \frac{1}{(2\pi h^2)^{\frac{D_{\text{conf}}}{2}}} \cdot \exp\left(-\frac{1}{2} \frac{\|\mathbf{x}^{\text{cont}} - \mathbf{x}_n^{\text{cont}}\|^2}{h^2}\right) \tag{19}$$

Here, h is a user-defined parameter whose value depends on the data set $\mathbf{X}$ (Bishop, 2006). There are a number of heuristics to estimate h. For instance, in (Bishop, 1994) h is set to the average distance of the ten nearest neighbors for each sample, averaged over the whole dataset. This non-parametric approach makes no assumptions about the functional form of the underlying distribution. Therefore, we cannot use Eq. (12) to calculate the Bhattacharyya coefficient analytically. However, it can be approximated with

$$\widehat{\text{BC}}(p(\mathbf{x}), q(\mathbf{x})) \approx \frac{1}{N} \sum_{\mathbf{x}_n \in \mathbf{X}} \frac{1}{q(\mathbf{x}_n)} \sqrt{p(\mathbf{x}_n)\, q(\mathbf{x}_n)}. \tag{20}$$

Note that we sum up over samples that are distributed according to q (cf. so-called importance sampling techniques).

Representativity measures the influence of a component on the "goodness of fit" of the model with respect to the data distribution. To calculate the representativity of component $p(\mathbf{x}|j)$ we again create a second CMM without $p(\mathbf{x}|j)$ as described for the discrimination measure. Then, we compare the Hellinger distance of the CMM with ($p_{\text{with}}(\mathbf{x})$) and without ($p_{\text{without}}(\mathbf{x})$) component $p(\mathbf{x}|j)$:

$$\text{repr}(p(\mathbf{x}|j)) = \text{H}(p_{\text{without}}(\mathbf{x}), q(\mathbf{x})) - \text{H}(p_{\text{with}}(\mathbf{x}), q(\mathbf{x})). \tag{21}$$

3.2.6 Comprehensibility

Comprehensibility measures how well the components (here referred to as rules) within the classifier can be interpreted by a human domain expert.

First, we claim that in a comprehensible classifier the overall number of rules J should be be low. Therefore, we use the number of different rules as one of three measures for comprehensibility.

Second, the number of different terms τ_d for each input dimension d should be low. For a categorical dimension τ_d is given by the number of categories n_d forming the disjunctions:

$$\tau_d = \sum_{j=1}^{J} n_{d_j}. \tag{22}$$

For a continuous dimension d the number τ_d of *different* univariate Gaussians $\varphi_{d,j}$ is counted. To decide whether two Gaussians should be regarded as being different or not, we use the Hellinger distance, cf. Eq. (10), of the two Gaussians which should be clearly below 0.01, for example, to regard two Gaussians as being identical.

The overall classifier is then assessed numerically by averaging over all dimensions (both, categorical and continuous):

$$\tau = \frac{1}{D} \sum_{d=1}^{D} \tau_d. \tag{23}$$

Applying this measure to the example classifier shown in Fig. 1 gives $\tau_1 = 2, \tau_2 = 2$, and $\tau_3 = 3$ which in turn results in $\tau = 2.3$.

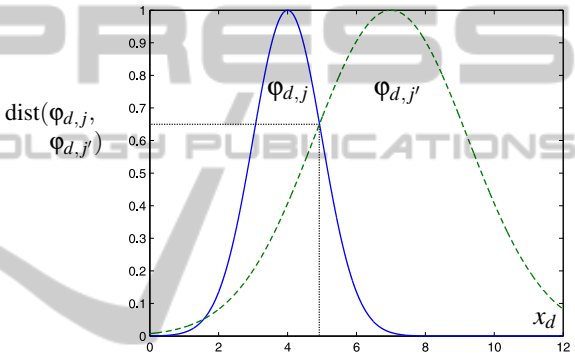


Figure 2: Example of an Assessment of the Distinguishability of two Gaussians.

Third, to simplify the understanding of a rule set, two different rules should be easy to distinguish. This *distinguishability* is only determined for the continuous dimensions $d \in \{1, \ldots, D_{\text{cont}}\}$. It is measured by the ordinate value of the intersection point of two univariate Gaussians $\varphi_{d,j}$ and $\varphi_{d,j'}$ (the y-coordinate of that intersection point which has an x-coordinate between the two means, to be precise), cf. Fig. 2. This measure is restricted to the unit interval by omitting the normalization coefficients of the Gaussians in the calculation which results in:

$$\text{dist}(\varphi_{d,j}, \varphi_{d,,j'}) = 1 - \exp\left(-\frac{(\mu_{d,j} - \mu_{d,j'})^2}{2 \cdot (\sigma_{d,j} + \sigma_{d,j'})^2}\right) \tag{24}$$

with $\text{dist}(\varphi_{d,j}, \varphi_{d,j'}) \in (0, 1]$. Values higher than 0.3, for example, could be regarded as desirable.

The distinguishability of the whole rule set is given by the pair of rules which is most difficult to distinguish, i.e., the pair with the highest value of $\text{dist}(p(\mathbf{x}|j), p(\mathbf{x}|j'))$.

4 CASE STUDIES

In this section we demonstrate the application of our proposed interestingness measures by means of three publicly available data sets. The first case study serves as an illustrative example of the general usage and characteristics of the measures. Then, we investigates how restricting the classifier in order to produce human-comprehensible rules influences the classification performance. Finally, we show how some of the interestingness measures can be used to automatically prune classifiers.

4.1 "Clouds" Data Set

The first case study uses the "clouds" data set from the UCL/MLG Elena database (UCL, 2007). This two-dimensional (both attributes are continuous) data set contains 5 000 samples belonging to two classes. Fig. 3 shows a part of the clouds samples together with a classifier trained on that data. There, the crosses (+) describe the centers of the Gaussian components of the trained GMM, the ellipses are corresponding level curves (surfaces of constant density with a Mahalanobis distance of one to the corresponding center) with shapes defined by their respective covariance matrices. The solid blue line illustrates the decision boundary of the classifier. The VI algorithm was initialized with 15 components and it pruned the final GMM down to four. The model was trained with 75% of the data set (i.e., 3 750 samples with a final training error of 10.6%) and tested on the remaining 1 250 samples resulting in a classification error of 9.9%.

Now, we assess this classifier by means of our proposed interestingness measures. Table 1 shows the evaluation of the four components (i.e., rules) with regard to uniqueness, informativeness, importance, discrimination, and representativity. First, it can be seen that the components 1 and 3 are distant to the remaining two components and, thus, their informativeness values are quite high. Additionally, they are only slightly covered with samples of a different class (i.e., the red boxes) which leads to high uniqueness values. Components 2 and 4, in contrast, belong to different classes and overlap. Thus, their uniqueness and informativeness values are lower. Note that they exhibit identical informativeness because they are mutually closest to each other and the informativeness measure is symmetric. As the class of the red boxes is only modeled by component 4, its influence on the decision boundary (i.e., its discrimination) is high. The class of the green circles is modeled by the remaining three components which results in lower discrimination values that scale with their importance (i.e., the fraction of samples they cover). Representativity is also highly correlated with importance as the more samples are covered by a component the higher its influence on the "goodness-of-fit".

Table 1: Evaluation of the Component-Based Measures for the "Clouds" Data Set.

j	uniq(j)	info(j)	imp(j)	disc(j)	repr(j)
1	0.903	0.889	0.498	0.460	0.141
2	0.657	0.771	0.246	0.133	0.048
3	0.908	0.922	0.256	0.210	0.085
4	0.853	0.771	0.667	0.803	0.234

Regarding the comprehensibility of the trained classifier it can be stated that the number of four components is certainly very low which is a good basis for a comprehensible classifier. Counting the number of different univariate Gaussians in every dimension requires to determine the projections of the components onto the axes corresponding to the different input dimensions. The unnormalized projections of the four components on the x- and y-axes are shown in Fig. 4. The projections on the x-axis, Fig. 4(a), show two nearly identical univariate Gaussians centered at -0.5 whose Hellinger distance is below 0.01. Thus, they are regarded as being identical from the viewpoint of comprehensibility which results in 3 univariate Gaussians in the x-dimension and 4 univariate Gaussians in the y-dimension. An average of 3.5 rules per dimension is a very good value for a comprehensible classifier. The minimum distinguishability of 0.0, however, deteriorates the comprehensibility. An example for two components with this low distinguishability can be seen for the y-axis projection, cf. Fig. 4(b), at the intersection point (-0.5, 1).

4.2 "Iris" Data Set

The aim of the second case study is to investigate the impact on classification performance when the classifier is restricted to generate human-comprehensible rules. For that purpose, we use the well-known "iris" data set from the UCI machine learning repository (Frank and Asuncion, 2010). This data set contains three classes, each with 50 four-dimensional (all continuous) samples.

First, we train a classifier without any restrictions, i.e., with full covariance matrices. We run the VI algorithm on the data set in a 4-fold cross-validation (stratified data). With this parametrization, the resulting models consist of seven to ten components, depending on the random initialization. The mean classification error on test data is 1.9% (std. dev. 1.1%). Now, we evaluate the model of the last fold consisting

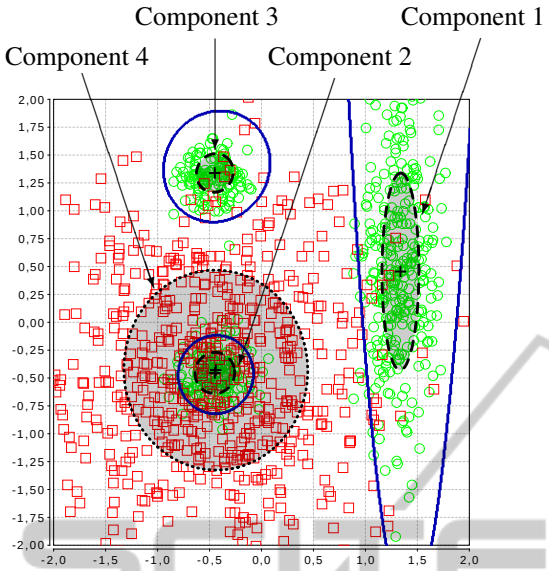


Figure 3: CMM for the "Clouds" Data Set.

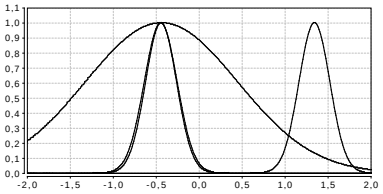
of seven components with our interestingness measures, cf. Tab. 2. The uniqueness and informativeness measures show that all components are very "tight" around their underlying samples and all components are very well localized (i.e., almost no overlap). The importance measure clearly shows four components that cover a large portion of samples. Interestingly, one of them has almost no impact on the decision boundary, as shown by the discrimination measure. As we have already stated in the first case study, representativity is highly correlated with importance.

Table 2: Evaluation of the Component-Based Measures for the Classifier with Seven Components and Full Covariance Matrices.

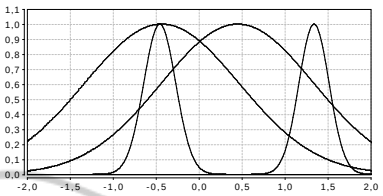
j	uniq(j)	info(j)	imp(j)	disc(j)	repr(j)
0	1.000	1.000	0.611	1.000	0.180
1	0.944	0.913	0.544	0.359	0.052
2	1.000	0.936	0.404	0.026	0.027
3	1.000	0.994	0.062	0.000	0.005
4	0.931	0.913	0.537	0.282	0.027
5	1.000	0.981	0.254	0.000	0.017
6	1.000	0.937	0.126	0.026	0.007

Regarding the classification performance it can be seen that the VI algorithm is able to generate very good classifiers for this data set. However, from the viewpoint of a data miner who wants to extract interesting rules, our evaluation reveals that the classifier is too detailed and even models regions in the input space that contain very little information.

Thus, we parametrize the VI algorithm to seek a solution with a lower number of components which



(a) x-Axis



(b) y-Axis

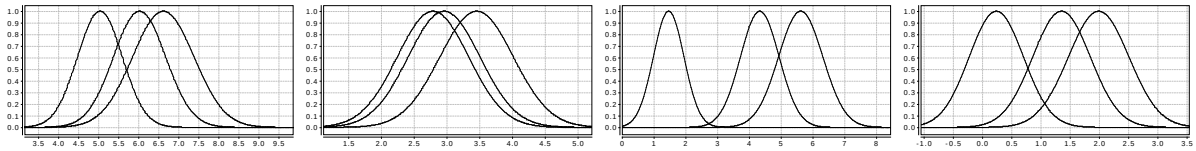
Figure 4: Projection of the Gaussians onto the x- and y-Axes for the "Clouds" Data Set.

usually results in a less detailed model. Again, we perform a 4-fold cross-valiation which results in a mean test error of 3.8% (std. dev. 3.8%). This time, all four generated classifiers consist of three components (one for each class). The results of our interestingness measures for the classifier of the last fold are listed in Tab. 3. Uniqueness and informativenes show that one class is well separated from the remaining two classes. The last two classes (and the respective components) overlap to a certain degree. As every class in this coarse-grained model is modeled with a single component, importance shows identical values for all components. The overlap of the classes is also reflected by discrimination and representativity since the well-separated component has a larger value than the remaining two components.

Table 3: Evaluation of the Component-Based Measures for the Classifier with Three Components and Full Covariance Matrices.

j	uniq(j)	info(j)	imp(j)	disc(j)	repr(j)
0	1.000	0.993	0.500	1.000	0.027
1	0.829	0.677	0.500	0.795	0.014
2	0.856	0.677	0.500	0.795	0.007

This result is a very good starting point to find a comprehensible classifier. Now, we restrict the VI algorithm to diagonal covariance matrices to enable the extraction of human-comprehensible classification rules. In a 4-fold cross-validation all models consist of three components again and yield a mean test error of 4.5% (Std. dev. 3.3%). Tab. 4 shows the assessment of the classifier of the last fold. Compared to the classifier with full covariance matrices uniqueness and informativeness show similar values. Obvi-



(a) 1st Dimension (b) 2nd Dimension (c) 3rd Dimension (d) 4th Dimension

Figure 5: Projections of the Classifier with Three Components and Diagonal Covariance Matrices onto the Four Axes.

ously, the limited modeling capability of the classifier due to the restricted covariance matrices has no severe impact for this data set.

Table 4: Evaluation of the Component-Based Measures for the Classifier with Three Components and Diagonal Covariance Matrices.

j	uniq(j)	info(j)	imp(j)	disc(j)	repr(j)
0	1.000	0.995	0.500	1.000	0.028
1	0.815	0.739	0.500	0.718	0.014
2	0.857	0.739	0.500	0.718	0.005

Fig. 5 illustrates the projections of the three components onto the four axes of the input space. From the viewpoint of comprehensibility we can state that the number of three components (i.e., three rules) is certainly very good. The average number of terms per dimension is also three since there are no two Gaussians that should be regarded as being identical. However, the distinguishability is very low (i.e., 0.01193). This is due to the projections of the second dimension, cf. Fig. 5(b), where two of the three univariate Gaussians are very close to each other and, thus, difficult to distinguish.

This case study showed that human-comprehensible rules can be generated from a classifier if the VI algorithm is parametrized accordingly. This comprehensibility, however, comes at the cost of a reduced classification performance as the modeling capability of the classifier is restricted. A compromise between understandability and classification performance is to only reduce the number of components while still allowing full covariance matrices. Then, our proposed interestingness measures enable a higher-level analysis of the structure of the data.

4.3 "Heart" Data Set

Some of the proposed interestingness measures can be used to automatically prune components from a trained classifier. We demonstrate this with the "heart" data set from the UCI machine learning repository (Frank and Asuncion, 2010). This 12-dimensional data set (six continuous and six categorical dimensions) consists of 270 samples which are partitioned into two classes. A 4-fold cross-validation of the VI algorithm results in a mean test error of 29.13%.

We select the model from the last fold which consists of 20 components and yields a test error of 25.0%. Then, we reduce the size of this classifier by pruning all components whose discrimination measure is below 0.1. The mixture coefficients of the remaining components are renormalized. The resulting classifier consists of only two components (which corresponds to a size reduction of 90%) and still achieves the same test error of 25.0% (i.e., all pruned components had a discrimination value of 0.0).

While the reduction of the original model is certainly optimal from the viewpoint of classification performance, it does not model the structure of the data anymore and, thus, is not suitable for data analysis. Therefore, we used again the classifier from the last fold with 20 components as a starting point and pruned all components with a discrimination below 0.1 *and* an importance below 0.1 (i.e., components that only cover a few data points are deleted). The resulting model has seven components and still yields a test error of 25.0%. This is optimal regarding the classification performance and the interesting regions in the input space are modeled.

Tab. 5 summarizes the results of this case study. Certainly, it is possible to use even more interestingness measures for a more sophisticated classifier pruning. Depending on the kind of rules the data miner is interested in, informativeness (rules that are distant to the remaining rules, cf. exception mining) or uniqueness (rules representing unambiguous knowledge) can be used additionally.

Table 5: Pruning results.

Pruning	Size	Test error
none (original model)	20	25.0%
disc ≤ 0.1	2	25.0%
disc ≤ 0.1 and impo ≤ 0.1	7	25.0%

5 CONCLUSIONS AND OUTLOOK

In this article, we first presented a probabilistic classifier based on mixture models (CMM) that can be used in the field of data mining to extract classification rules from labeled sample data. Then, we defined some objective interestingness measures that are tailored to measure various aspects of the rules of which this classifier consists. These measures are also based on probabilistic methods. A data miner may use these measures to investigate the knowledge extracted from sample data in more detail. In three case studies using well-known data sets we demonstrated the application of our approach.

In our future work we will investigate the possibility to apply our objective measures to each of the C class-specific mixture models to obtain an even more detailed class-specific assessment of the components. In this work we used the measures as a post-processing step to prune a trained model. However, it is also possible to use them as side conditions in the objective functions that are used for the training of CMM in order to support certain properties of a classifier already during training. Additionally, we will investigate how the measures can be combined to perform a ranking of rules based on their interestingness. There is a close relation of CMM to certain kinds of fuzzy classifiers concerning the functional form as outlined in (Fisch et al., 2010). Thus, it would also be interesting to transfer the proposed measures to that kind of classifiers and compare them to other measures.

REFERENCES

Atzmueller, M., Baumeister, J., and Puppe, F. (2004). Rough-fuzzy MLP: modular evolution, rule generation, and evaluation. In *15th International Conference of Declarative Programming and Knowledge Management (INAP-2004)*, pages 203–213, Potsdam, Germany.

Basu, S., Mooney, R. J., Pasupuleti, K. V., and Ghosh, J. (2001). Evaluating the novelty of text-mined rules using lexical knowledge. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, pages 233–238, San Francisco, CA.

Bishop, C. (1994). Novelty detection and neural network validation. *IEE Proceedings – Vision, Image, and Signal Processing*, 141(4):217–222.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Di Fiore, F. (2002). Visualizing interestingness. In Zanasi, A., Brebbia, C., Ebecken, N., and Melli, P., editors, *Data Mining III*. WIT Press, Southampton, U.K.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons, Chichester, New York, NY.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 1996)*, pages 82–88, Portland, OR.

Fisch, D., Kühbeck, B., Ovaska, S. J., and Sick, B. (2010). So near and yet so far: New insight into properties of some well-known classifier paradigms. *Information Sciences*, 180:3381–3401.

Fisch, D. and Sick, B. (2009). Training of radial basis function classifiers with resilient propagation and variational Bayesian inference. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN '09)*, pages 838–847, Atlanta, GA.

Frank, A. and Asuncion, A. (2010). UCI machine learning repository.

Hebert, C. and Cremilleux, B. (2007). A unified view of objective interestingness measures. In Perner, P., editor, *Machine Learning and Data Mining in Pattern Recognition*, number 4571 in LNAI, pages 533–547. Springer, Berlin, Heidelberg, Germany.

Hilderman, R. J. and Hamilton, H. J. (2001). *Knowledge Discovery and Measures of Interest*. Kluwer Academic Publishers, Norwell, MA.

Liu, B., Hsu, W., Chen, S., and Ma, Y. (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5):47–55.

McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*, 20(1):39–61.

Nauck, D. D. (2003). Measuring interpretability in rule-based classification systems. In *Proceedings of the 12th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'03)*, volume 1, pages 196–201, St. Louis, MO.

Padmanabhan, B. and Tuzhilin, A. (1999). Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems*, 27(3):303–318.

Piatetsky-Shapiro, G. and Matheus, C. (1994). The interestingness of deviations. In *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases (KDD 1994)*, pages 25–36, Seattle, WA.

Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge And Data Engineering*, 8:970–974.

Taha, I. and Ghosh, J. (1997). Evaluation and ordering of rules extracted from feedforward networks. In *International Conference on Neural Networks*, volume 1, pages 408–413, Houston, TX.

Tan, P.-N., Kumar, V., and Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313.

UCL (2007). Elena Database. http://www.ucl.ac.be/mlg/index.php?page=Elena.