Feature Selection with Mutual Information for Uncertain Data

Gauthier Doquire* and Michel Verleysen

Université catholique de Louvain, Machine Learning Group - ICTEAM
Place du Levant, 3, 1348 Louvain-la-Neuve, Belgium
{gauthier.doquire,michel.verleysen}@uclouvain.be
http://www.ucl.ac.be/mlg

Abstract. In many real-world situations, the data cannot be assumed to be precise. Indeed uncertain data are often encountered, due for example to the imprecision of measurement devices or to continuously moving objects for which the exact position is impossible to obtain. One way to model this uncertainty is to represent each data value as a probability distribution function; recent works show that adequately taking the uncertainty into account generally leads to improved classification performances. Working with such a representation, this paper proposes to achieve feature selection based on mutual information. Experiments on 8 UCI data sets show that the proposed approach is effective to select relevant features.

Keywords: Uncertain data, feature selection, mutual information.

1 Introduction

Nowadays, many machine learning and data mining applications have to cope with data that are inherently uncertain. This uncertainty can be caused by many different factors. As an example, measurement errors from unprecise devices or sensors with a too low resolution typically produce uncertain data. Moreover, in some applications involving continuously moving devices, the exact location of the objects is not always available or is not transmitted precisely due to privacy reasons. Eventually, data quantization or averaging from multiple measurements also lead to uncertainty.

All these reasons explain the recent interest in the development of data mining tools for uncertain data such as classification [1,2,3], clustering [4,5,6,7] or outlier detection [8] to name a few. [9] gives a nice overview on recent developments about uncertain data

A convenient way to model the uncertainty of the data is to represent any value in the data set as an uncertainty region and to define a probability density function (pdf) over it. Using this approach, [1,2] showed that adequately taking the uncertainty into account leads to better classification performances for the

* Gauthier Doquire is funded by a Belgian FRIA grant.

A. Cuzzocrea and U. Dayal (Eds.): DaWaK 2011, LNCS 6862, pp. 330–341, 2011.

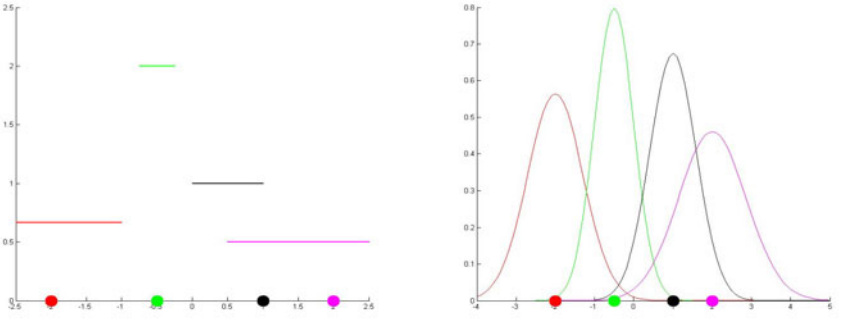


Fig. 1. Examples of modelling of the uncertainty on data with uniform (left) and Gaussian (right) pdf. The curves describe the pdf of the actual values given the observed values (shown by big dots).

decision tree and the Naive Bayes classifiers compared to the case where the values are used directly. In particular, choosing a Gaussian pdf centered in the value and with a well-chosen variance led to very satisfactory results. That is the reason why the same strategy is adopted throughout this paper. However, the proposed methodology can easily be extended to the uniform distribution or to uncertain data described by samples drawn from an underlying unknown distribution. Figure 1 illustrates the modelling of the uncertainty on data with uniform and Gaussian pdf. In this work, the problem of feature selection with uncertain data is considered; it is, to the best of our knowledge, the first time this problem is adressed. Feature selection is a very important preprocessing step for many pattern recognition problems, including classification. Its goal is to determine which (small) subset of features is the most relevant for a given task. Its benefits for classification can be numerous. First, it helps understanding the problem and interpreting the model by determining which factors really influence the output to be predicted. This is of crucial importance for many industrial and medical applications. As an example, in the context of microarray data, feature selection can help discovering a small set of genes linked to a particular disease or pathology. Secondly it generally leads to improved classification performances by removing irrelevant and/or redundant features and by preventing the classification models to suffer from the curse of dimensionality. By decreasing the number of features considered, feature selection also makes the classifiers faster. Eventually, it has also practical advantages in terms of data acquisition and warehousing. Indeed, useless features do not need to be gathered and stored anymore. See [10] for a detailed introduction on feature selection.

The proposed approach is based on the well-known mutual information (MI) criterion [11], which has already been used successfully in many feature selection algorithms. A methodology to estimate MI with uncertain data is proposed and used to rank features according to their dependance to the class labels vector.

The rest of the paper is organized as follows. Section 2 recalls some concepts about MI and its estimation for classical data. Section 3 presents the proposed MI estimator for uncertain data. Section 4 is dedicated to the experimental results and Section 5 concludes the work and gives some future research perspectives.

2 Mutual Information

This section first introduces some basic notions on MI and then shows how it can be estimated since it generally cannot be computed exactly.

2.1 Basic Notions

MI, first introduced by Shannon in 1948 [11], is a quantity describing the amount of information two random variables carry about each other. It is symmetric, i.e. $I(X;Y) = I(Y;X)$ and able to detect non-linear relationships between variables. This last property has made MI a very popular criterion for feature selection [12,13,14,15] since other widely used criteria such as the correlation coefficient can only handle linear dependencies.

Formally, the MI of a pair of random variables X and Y can be defined by means of the pdf of X, Y and the joint variable (X,Y), respectively denoted as f_X, f_Y and $f_{X,Y}$:

$$I(X;Y) = \int\int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)}\, dx\, dy. \tag{1}$$

This definition can also be seen as the Kullback-Leibler divergence between the product of distributions $f_X \times f_Y$ and the joint distribution $f_{X,Y}$. If the variables are independent, then $f_{X,Y} = f_X \times f_Y$ and $I(X;Y) = 0$.

MI can also be expressed in terms of entropy, another information theoretic quantity. The entropy of a random variable is a measure of the uncertainty one has about the values taken by this variable. It is also defined in terms of pdf:

$$h(X) = -\int f_X(x) \log f_X(x) dx. \tag{2}$$

The MI is equal to:

$$I(X;Y) = h(Y) - h(Y|X) \tag{3}$$

where $h(Y|X)$ is the conditional entropy of Y given X, corresponding to the uncertainty about Y when X is known. Following (3), MI can be seen as the reduction of uncertainty about Y brought by the knowledge of X and is thus a natural criterion for feature selection assuming that Y is an output we want to predict from X, a set of possibly multivariate data points. In (3), if X and Y are independent, $h(Y|X) = h(Y)$ and again $I(X;Y) = 0$.

2.2 Estimation

As detailed previously, the MI is entirely determined by the marginal pdf f_X and f_Y and the joint pdf $f_{X,Y}$. However, in practice, these pdf are not known, meaning that the MI has to be estimated from the data set.

Traditionally, the entropy is first estimated by histograms or kernel-based estimators before the MI is computed according for instance to (1). This approach is followed in this paper, where a Parzen-window [16] density estimator is used.

Consider $x_1 \dots x_N$, N i.i.d. samples drawn from the distribution f. The estimated pdf is given by:

$$\hat{f}(x) = \frac{1}{Nb} \sum_{i=1}^{N} k\left(\frac{x - x_i}{b}\right) \tag{4}$$

where k is a kernel and b is called the bandwidth. The most popular choice for k is the Gaussian kernel with zero mean and unit variance:

$$k(x) = \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}. \tag{5}$$

The value of the bandwidth b, which acts as a smoothing parameter, is of crucial importance for the quality of the estimation. In this work, it is chosen according to the popular Silverman rule [17] for one-dimensional data points:

$$b_j = 1.06\sigma_j N \tag{6}$$

where σ_j denotes the standard deviation along the j^{th} dimension of the data set. In the next section, it will be shown how this estimator can be adapted to handle the uncertain data case.

It is worth noting that such density estimators should only be used with low-dimensional data. Indeed, when the dimensionality increases, histograms and kernel based estimators suffer from the curse of dimensionality and from the empty space phenomenon. This phenomenon denotes the fact that the number of points needed to sample a space at a given precision grows exponentially with the dimension of the space [18]. Thus, when working in a high-dimensional space, most of the boxes of an histogram are likely to be empty and the estimated density to be innacurate. Kernel-based estimators are generally smoother but are also dramatically affected by these problems.

One possible way to alleviate the curse of dimensionality is to use nearest-neighbors based MI estimators which do not directly estimate the pdf and are thus expected to be more robust in high-dimensional spaces [19,20].

3 MI Estimation with Uncertain Data

This section shows how the MI can be estimated from uncertain data by using the previously described kernel-based density estimator.

This paper considers classification problems; Given a data set X containing N samples described by d attributes, the goal is to predict the class (a discrete value) of these samples based on previously observed input/output pairs. This means that the MI $I(X;Y)$ has to be estimated between continuous (X) and discrete (Y) random variables, the latter corresponding to the classes we want to predict.

More precisely, we are interested in evaluating $I(X_j;Y)$ for $j = 1 \dots d$, where X_j denotes the j^{th} attribute or feature of X. The pdf of this j^{th} attribute is denoted by f_{X_j}.

Assume that Y takes k different values $y_1 \dots y_k$, each y_i being represented by n_i samples ($\sum_i n_i = N$); Denote by $\hat{p}(y_i)$ the probability that $Y = y_i$, estimated by $\frac{n_i}{N}$. All that is needed to estimate the MI by (3) is:

$$\hat{h}(Y) = -\sum_{i=1}^{k} \hat{p}(y_i) \log \hat{p}(y_i) \tag{7}$$

and

$$\hat{h}(Y|X_j) = -\int_{X_j} \hat{f}_{X_j}(x) \sum_{i=1}^{k} \hat{f}_{Y_i|X_j}(y_i|x) \log \hat{f}_{Y_i|X_j}(y_i|x) dx. \tag{8}$$

Equation (7) is the discrete version of (2) and $\hat{f}_{Y_i|X_j}$ is the estimated density of the i^{th} class conditional to the j^{th} feature. As (7) will be equal for all features, it can be omitted when comparing the individual MI of the features.

According to the Bayes theorem, it is possible the rewrite $\hat{f}_{Y_i|X_j}(y_i|x)$ as:

$$\hat{f}_{Y_i|X_j}(y_i|x) = \frac{\hat{f}_{X_j|Y_i}(x|y_i)\hat{p}(y_i)}{\hat{f}_{X_j}(x)}. \tag{9}$$

We have then:

$$\hat{h}(Y|X_j) = -\int_{X_j} \hat{f}_{X_j}(x) \sum_{i=1}^{k} \frac{\hat{f}_{X_j|Y_i}(x|y_i)\hat{p}(y_i)}{\hat{f}_{X_j}(x)} \log \frac{\hat{f}_{X_j|Y_i}(x|y_i)\hat{p}(y_i)}{\hat{f}_{X_j}(x)} dx. \tag{10}$$

This last equation implies that the MI can be entirely determined by the pdf of the variable X_j, possibly limited to the points with a particular class label y_i. In the following, we show how this pdf can be estimated.

Recall that $X_j = [x_{j1} \dots x_{jN}]$ is described by Gaussian pdf to model the uncertainty in the data, i.e. $x_{j1} \sim N(\mu_{j1}, \sigma_{j1}) \dots x_{jN} \sim N(\mu_{jN}, \sigma_{jN})$. μ_{ji} is the observed value for the j^{th} dimension of the i^{th} sample and σ_{ji} is the variance that is determined by the user, following the confidence he has on the precision of the data.

A quite natural approach is to consider the expected value of the kernel k [1]. More precisely, (4) is replaced with

$$\hat{f}(x) = \frac{1}{Nb} \sum_{i=1}^{N} E\left[k\left(\frac{x - x_i}{b}\right)\right]. \tag{11}$$

The following developments then hold:

$$\begin{aligned}
\hat{f}_{X_j}(x) &= \frac{1}{Nb} \sum_{i=1}^{N} \int_{x_{ji}} k\left(\frac{x_{ji} - x}{b}\right) \frac{1}{\sqrt{2\pi}\sigma_{ji}} e^{-0.5\left(\frac{x_{ji}-\mu_{ji}}{\sigma_{ji}}\right)^2} dx_{ji} \\
&= \frac{1}{Nb} \sum_{i=1}^{N} \int_{x_{ji}} \frac{1}{\sqrt{2\pi}} e^{-0.5\left(\frac{x_{ji}-x}{b}\right)^2} \frac{1}{\sqrt{2\pi}\sigma_{ji}} e^{-0.5\left(\frac{x_{ji}-\mu_{ji}}{\sigma_{ji}}\right)^2} dx_{ji} \\
&= \frac{1}{N} \sum_{i=1}^{N} \int_{x_{ji}} \frac{1}{\sqrt{2\pi}b} e^{-0.5\left(\frac{x_{ji}-x}{b}\right)^2} \frac{1}{\sqrt{2\pi}\sigma_{ji}} e^{-0.5\left(\frac{\mu_{ji}-x_{ji}}{\sigma_{ji}}\right)^2} dx_{ji}.
\end{aligned} \tag{12}$$

Moreover, it is well-known that the convolution of two Gaussian distributions $f \sim N(\mu_f, \sigma_f)$ and $g \sim N(\mu_g, \sigma_g)$ is another Gaussian distribution $c \sim N(\mu_f + \mu_g, \sqrt{\sigma_f^2 + \sigma_g^2})$. Stated otherwise:

$$\begin{aligned}
f * g &= \int_{\tau} \frac{1}{\sqrt{2\pi}\sigma_f} e^{-0.5\left(\frac{\tau-\mu_f}{\sigma_f}\right)^2} \frac{1}{\sqrt{2\pi}\sigma_g} e^{-0.5\left(\frac{t-\tau-\mu_g}{\sigma_g}\right)^2} d\tau \\
&= \frac{1}{\sqrt{2\pi(\sigma_f^2 + \sigma_g^2)}} e^{-0.5\frac{\left(t-(\mu_f+\mu_g)\right)^2}{\sigma_f^2+\sigma_g^2}}.
\end{aligned} \tag{13}$$

By setting $\tau = x_{ji}$, $\sigma_f = b$, $\sigma_g = \sigma_{ji}$, $t = \mu_{ji}$, $\mu_f = x$ and $\mu_g = 0$, the connection between (13) and the last line of (12) is obvious. Combining these two equations, it comes:

$$\hat{f}_{X_j}(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi(\sigma_{ji}^2 + b^2)}} e^{-0.5\frac{(x-\mu_{ji})^2}{\sigma_{ji}^2+b^2}}. \tag{14}$$

With a way to estimate the pdf $\hat{f}_{X_j}(x)$, using (3), (8) and (9), it is now possible to estimate the MI between each feature X_j and the output vector Y. As already stated, evaluating the conditional pdf $f_{X_j|Y_i}$ is done exactly the same way as for f_{X_j}, except that only the samples having the output y_i are included in the computation of (14). The technical details for the numerical integration in (8) are given in the next section.

It is obvious that the developments presented in this section assuming a Gaussian pdf can be adapted to handle other models of uncertainty. For instance, a uniform pdf could be considered instead. This would mean that we believe every observed value has been drawn from a domain of possible values, all having the same probability. In contrast, the Gaussian pdf implies that the observed

value is actually the most probable even if some imprecisions are possible. If one wishes to model the uncertainty on x_{ji} by a uniform pdf on the domain $[a; b]$ ($a < b$), then in (12), $f_{x_{ji}}(x_{ji})$ is equal to the constant $\frac{1}{b-a}$. The estimation of the density then resumes to the integration of a Gaussian function evaluated between a and b. See again Figure 1 for an illustration of the differences between both approaches.

Another way of specifiying the uncertainty is to represent each point by numerous samples drawn from its distribution. The expectation in (11) then becomes a sum where each sample contributes to the estimation with an importance weighted by its probability. However, in [1], this approach is shown to be much more time-consuming than the Gaussian pdf based approach, without leading to better classification performances. It is thus not investigated in the present work even if it can be helpful when ones wants to consider a distribution for which (12) has no closed-form solution.

4 Methodology and Experiments

To assess the effectiveness of the proposed feature selection procedure, experiments are carried out on eight data sets from the UCI machine learning repository [21]. They consist of values obtained through measurements, and have been shown to benefit well from taking their uncertainty into account [1,2].

The first part of this section describes exactly how the uncertainty is handled in this paper; technical details about the integration in (8) are also given. Experimental results obtained on the data sets are then presented and commented.

4.1 Methodology

The MI is first evaluated between each feature of the training set and the output vector: the features are then ranked according to this score. The number of selected features should either be set a priori or should be determined by cross-validation procedures on an independent validation set.

The uncertainty on the data set is modelled by a Gaussian pdf with the mean equal to the observed value and the standard deviation defined following [1,2]. If min_j and max_j are respectively the minimum and maximum values taken by the feature X_j, then the standard deviation is $\sigma_j = 0.25\,(max_j - min_j)\,w\,\%$. It is the same for all x_{ji} and w is a parameter representing the level of uncertainty we have about the values of X_j. The rationale behind this choice is thus that the uncertainty about a variable is proportional to the size of the range of values taken by this variable. In other words, the more the observed values for a given variable are close, the more the uncertainty about these values will be considered as small. The values of w chosen in this paper are those already adopted in [1] and/or [2] (except for the Parkinson data set which has not been used in these references). Indeed, even if the two classifiers introduced in these works are very different, both achieve the best performances on the same data sets with very similar values of w.

Table 1. Description of the datasests used in the experiments

Name	Samples	Features	Classes	w
glass	214	9	4	3
iris	150	4	3	20
wine	178	13	3	1
segment	2310	18	7	4
waveform	5000	21	3	3
satellite	6435	36	2	6
pageBlock	5473	10	5	1
parkinson	195	22	2	5

Moreover, the accuracy of these classifiers generally reaches a peak at the optimal value of w, meaning that if this value is slightly increased or decreased, the performances of the classifiers degrade [1,2]. Those observations conjecture the fact that the considered data sets do contain errors and have an intrinsic optimal value of w (at least for the Gaussian pdf). The data sets are described in Table 1 which also gives the corresponding value of w.

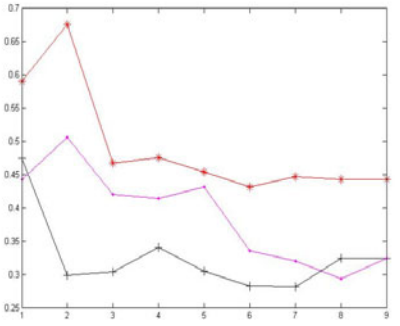
The integral in (8) is evaluated numerically using the simple trapeze rule. It consists in interpolating the function piecewise linearly by using its values in a certain number of points. To this end, 1000 equally spaced points are sampled between a_{min} and a_{max}. $a_{min} < min_j$ is the value for which a Gaussian pdf with unit variance and mean min_j equals 10^{-3}. $a_{max} > max_j$ is the value for which a Gaussian pdf with unit variance and mean max_j equals 10^{-3}. It is worth noting that the bandwidth in (14) has to be adapted to each individual feature and to the fact that the density can be conditioned to a class label.

4.2 Experimental Results

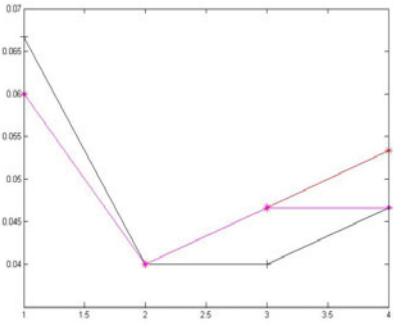
Figure 2 shows the classification error rate (the percentage of misclassified samples) as a function of the number of selected features for the Naive Bayes classifier adapted to uncertain data [1] and the first six data sets. For comparison, the error rate is also shown when no uncertainty is taken into account (neither in the feature selection process nor in the classification). Eventually, to show the interest of considering the uncertainty for feature selection, the error rate when uncertainty is only considered in the classification step is also presented.

The reported results are obtained through a 10-fold cross validation procedure. This means that the dataset is first randomly divided into ten disjoint equally sized sets of samples. Then each set is successively used to test the performances of a classifier built on the nine other sets. The ten error rates obtained this way are eventually averaged. In this paper, no additional validation set is needed since there is no parameter to tune.

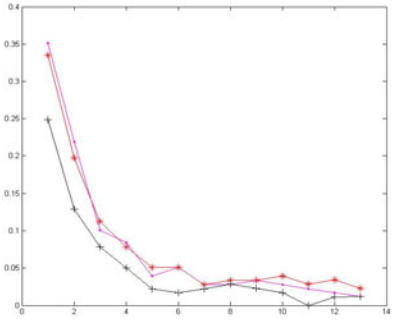
The interest of the proposed feature selection method is obvious for the considered data sets. Indeed, the first observation is that in each case, it is possible to reduce the classification error by considering only a subset of the original features. In particular for the glass, iris and segment data sets, at least half the



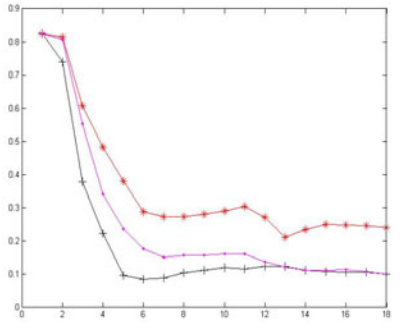
glass



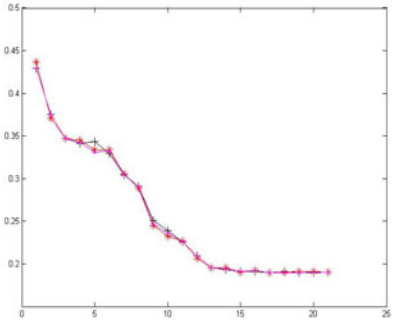
iris



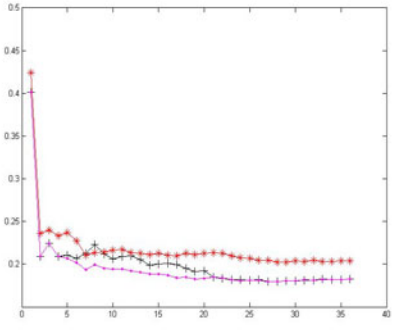
wine



segment

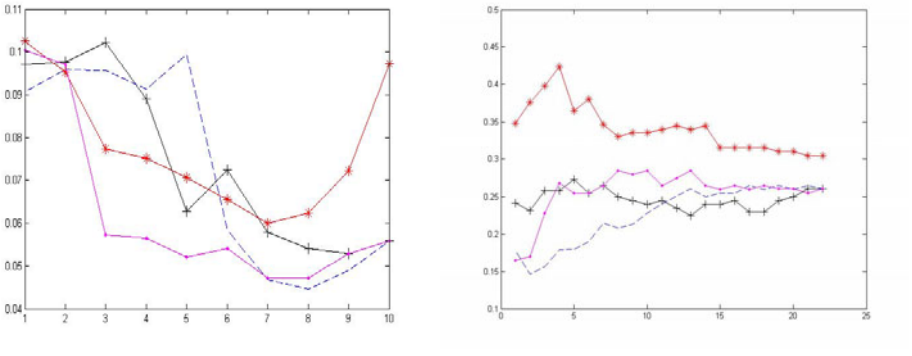


waveform



satellite

Fig. 2. Classification error rate of a Naive Bayes classifier as a function of the number of selected features for six data sets. (+) Uncertainty in the feature selection and the classification; (.) Uncertainty only in the classification; (∗) No uncertainty.



pageBlock parkinson

Fig. 3. Classification error rate of a Naive Bayes classifier as a function of the number of selected features for two data sets. (+) Uncertainty in the feature selection and the classification; (.) Uncertainty only in the classification; (∗) No uncertainty; (−) Same as (+) with normalized data.

features can be discarded without decreasing the original accuracy. In the satellite data set, more than a third of the features can be removed without harming the classifier performances.

Then, it clearly appears that considering the uncertainty allows us to increase the performances of both the feature selection and the classification. For the 6 data sets, lower error rates are obtained with the proposed approach than when no uncertainty at all is considered (except for the satellite data set where equal error rates are achieved). Moreover, taking the uncertainty into account only in the classification step (and not for feature selection) nevers allows to reach better performances than with the suggested methodology.

When applying the same methodology to the parkinson and pageBlock data sets, the results obtained at the first were not so encouraging. The reason is that some features in those data sets have very different ranges of values. The entropy (8) of features with a larger range of values is likely to be higher than for features with a smaller range of values which could bias the MI estimation procedure. As an example, the entropy of a Gaussian variable with variance σ^2 is given by $0.5 \log(2\pi e \sigma^2)$ and thus increases with the dispersion of the data. To circumvent this issue, each feature X_j was normalized by removing its mean and dividing it by its standard deviation σ_{X_j} before the feature selection step. To account for this normalization the parameter w controling the uncertainty for each feature was adapted to $w \times \sigma_{X_j}$.

Figure 3 confirms that the suggested normalization helps improving the feature selection with uncertainty for the two data sets. It is also the case for feature selection without taking the uncertainty into account but the results are not displayed for clarity reasons, since they are inferior to those obtained considering the uncertainty. With the normalization, the best results are again achieved by the proposed methodology. In particular, for the Parkinson data set, the error

rate is reduced by more than 10% with only the first two features. When applied to the six first data sets, the normalization leads to very similar results than those presented in Figure 2.

5 Conclusions

This paper is concerned with the important problem of feature selection for classification problems, in the specific context of uncertain data. To this end, it is proposed to rank the features according to their MI with the class labels vector, a widely used criterion for feature selection. The current work is motivated by recent papers showing that properly taking the uncertainty of the data into account generally increases the precision of classifiers.

Following these works, the uncertainty on the data is handled by representing the values in the data set as an uncertainty region and to define a pdf over theses regions. In this work, Gaussian pdf are considered while it is shown how the developments could be easily extended to the uniform distribution or to an arbitrary distribution defined by a collection of samples drawn from it.

A method to evaluate the MI between each uncertain feature and the output is then introduced. It is based on the traditionnal kernel density estimation which is adapted to handle points described as pdf. More precisely, the expected value of the kernel estimator is determined by exploiting the fact that a convolution between Gaussian pdf is still a Gaussian pdf with known mean and variance. A convenient way to numerically evaluate the entropy and thus the corresponding MI is also proposed.

Experimental results on eight UCI databases containing uncertainty show that the proposed approach is effectively able to select relevant features. Indeed, for all data sets, the classification performances can be improved by removing irrelevant features. Moreover, the advantage of considering the inhenrent uncertainty of the data for both feature selection and classification is clearly established. It is also shown how the normalization can help improving the feature selection when some features have large differences in their range value, harming the estimation of the entropy and consequently the estimation of MI.

Future work could be focused on the development of MI estimation algorithms for two uncertain *continuous* vectors. This would be helpful for feature selection in regression problems (problems with a continuous output to predict). It would also allow one to consider the redundancy between features. Indeed, only the relevance (measured by the MI) is considered as a criterion for feature selection in the present work. Taking the redundancy into account could lead to improved performances, especially if one works with highly redundant data such as in near infra-red spectra analysis [14].

References

1. Ren, M., Lee, S.D., Chen, X., Kao, B., Cheng, R., Cheung, D.: Naive Bayes Classification of Uncertain Data. In: 9th IEEE International Conference on Data Mining, ICDM 2009, pp. 944–949 (2009)

2. Tsang, S., Kao, B., Yip, K.Y., Ho, W.-S., Ho, W.-S., Lee, S.: Decision Trees for Uncertain Data. IEEE T. Knwol. Dat. En. 23, 64–78 (2011)
3. Bi, J., Zhang, T.: Support Vector Classification with Input Data Uncertainty. In: Advances in Neural Information Processing Systems, NIPS (2004)
4. Ngai, W.K., Kao, B., Chui, C.K., Cheng, R., Chau, M., Yip, K.Y.: Efficient Clustering of Uncertain Data. In: 6th IEEE International Conference on Data Mining, ICDM 2006, pp. 436–445 (2006)
5. Kriegel, H.-P., Pfeifle, M.: Hierarchical Density-Based Clustering of Uncertain Data. In: 5th IEEE International Conference on Data Mining (ICDM 2005), pp. 689–692 (2005)
6. Kao, B., Lee, D., Cheung, D.W., Ho, W.-S., Chan, K.F.: Clustering Uncertain Data using Voronoi Diagrams. In: 8th IEEE International Conference on Data Mining, ICDM 2008, pp. 333–342 (2008)
7. Cormode, G., McGregor, A.: Approximation Algorithms for Clustering Uncertain Data. In: 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2008), pp. 191–200 (2008)
8. Aggarwal, C.C., Yu, P.S.: Outlier Detection with Uncertain Data. In: SIAM International Conference on Data Mining (SDM), pp. 483–493 (2008)
9. Aggarwal, C.C., Yu, P.S.: A survey of Uncertain Data Algorithms and Applications. IEEE T. Knwol. Dat. En. 21, 609–623 (2009)
10. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. J. Mach. Lear. Res. 3, 1157–1182 (2003)
11. Shannon, C.E.: A mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423, 623–656 (1948)
12. Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE T. Neural. Networ. 5, 537–550 (1994)
13. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE T. Pattern. Anal. 27 (2005)
14. Rossi, F., Lendasse, A., François, D., Wertz, V., Verleysen, M.: Mutual Information for the Selection of Relevant Variables in Spectrometric Nonlinear Modelling. Chemometr. Intell. Lab. 80, 215–226 (2006)
15. François, D., Rossi, F., Wertz, V., Verleysen, M.: Resampling Methods for Parameter-free and Robust Feature Selection with Mutual Information. Neurocomputing 70, 1276–1288 (2007)
16. Parzen, E.: On Estimation of a Probability Density Function and Mode. Ann. Math. Statist. 33, 1065–1076 (1962)
17. Silverman, B.W.: Density Estimation. Chapman & Hall, London (1986)
18. Verleysen, M.: Learning High-Dimensional Data. In: Limitations and Future Trends in Neural Computation, pp. 141–162 (2003)
19. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating Mutual Information. Phys. Rev. E 69, 66138 (2004)
20. Gomez-Verdejo, V., Verleysen, M., Fleury, J.: Information-Theoretic Feature Selection for Functional Data Classification. Neurocomputing 72, 3580–3589 (2009)
21. Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2010), `http://archive.ics.uci.edu/ml`