# AN HYBRID APPROACH TO FEATURE SELECTION FOR MIXED CATEGORICAL AND CONTINUOUS DATA

Gauthier Doquire and Michel Verleysen
*ICTEAM Institute, Machine Learning Group, Université catholique de Louvain*
*pl. du Levant 3, 1348 Louvain-la-Neuve, Belgium*

Keywords: Feature selection, Categorical features, Continuous features, Mutual information.

Abstract: This paper proposes an algorithm for feature selection in the case of mixed data. It consists in ranking independently the categorical and the continuous features before recombining them according to the accuracy of a classifier. The popular mutual information criterion is used in both ranking procedures. The proposed algorithm thus avoids the use of any similarity measure between samples described by continuous and categorical attributes, which can be unadapted to many real-world problems. It is able to effectively detect the most useful features of each type and its effectiveness is experimentally demonstrated on four real-world data sets.

## 1 INTRODUCTION

Feature selection is a key problem in many machine learning, pattern recognition or data mining applications. Indeed the ways to acquire and store data increase every day. A lot of features are thus typically gathered for a specific problem while many of them can be either redundant or irrelevant. These useless features often tend to decrease the performances of the learning (classification or regression) algorithms (Guyon and Elisseeff, 2003) and slower the whole learning process. Moreover, reducing the number of attributes leads to a better interpretability of the problem and of the models, which is of crucial importance in many industrial and medical applications. Feature selection thus plays a major role both from a learning and from an application point of view.

Due to the importance of the problem, many feature selection algorithms have been proposed in the past few years. However, the great majority of them are designed to work only with continuous *or* categorical features and are thus not well suited to handle data sets with *both* type of features, while mixed data are encountered in many real-world situations. To illustrate this, two examples are given. First, the results of medical surveys can include continuous attributes as the size or the blood pressure of a patient, together with categorical ones as the sex or the presence or absence of a symptom. In another field, socio-economic data can contain discrete variables about individuals such as their kind of job or the city they come from, as well as continuous ones like their income.

Algorithms dealing with continuous *and* discrete attributes are thus needed. Two obvious ways to handle problems with mixed attributes are turning the problem into a categorical or a continuous one. Unfortunately, both approaches have strong drawbacks.

The first idea would consist in coding the categorical attributes into discrete numerical values. It would then be possible to compute distances between observations as if all features were continuous. However, this approach is not likely to work well. Indeed, permuting the code for two categorical values could lead to different values of distance. To circumvent this problem, Bar-Hen and Daudin (1995) proposed to use a generalized Mahalanobis distance, while Kononenko (1994) employs the Euclidean distance for continuous features and the Hamming distance for categorical ones. The second idea is to discretize continuous features before running an algorithm designed for discrete data (Hall, 2000). Even if appealing, this approach may lead to a loss of information and makes the feature selection efficiency extremely dependant on the discretization technique.

Recently, Tang and Mao proposed a method based on the error probability (Tang and Mao, 2007) while Hu et al. reported very satisfactory results using rough set models generalized to the mixed case (Hu et al., 2008). In this last paper, the authors base their work on neighboorood relationships between mixed samples, defined in the following way. First, to be considered as neighbors, two samples must have the same

Doquire G. and Verleysen M. (2011).
AN HYBRID APPROACH TO FEATURE SELECTION FOR MIXED CATEGORICAL AND CONTINUOUS DATA.
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pages 394-401
DOI: 10.5220/0003634903940401

values for all their discrete attributes. Then, depending on the approach chosen and according to the continuous features, the Euclidean distance between the samples has to be below a fixed treshold or one of the sample has to belong to the $k$ nearest neighbors of the other. The method thus makes a strong hypothesis about the notion of proximity between samples, which can be totally inconsistent with some problems as will be illustrated later in this work.

In contrast, the approach proposed in this paper does not consider any notion of relationship between mixed samples. Instead, the objective is to correctly detect the most useful features of each kind and to combine them to optimize the performance of prediction models. More precisely, the features of each type are first ranked independently; two independent lists are produced. The lists are then combined according to the accuracy of a classifier. Mutual information (MI) based feature selection is employed for the ranking of both continuous and categorical features.

The rest of the paper is organized as follows. Section 2 briefly recalls basic notions about MI. The proposed methodology is described in Section 3 and experimental results are given in Section 4. Conclusions are drawn in Section 5 which also contains some future work perspectives.

# 2 MUTUAL INFORMATION

In this section, basic concepts about MI are introduced and a few words are given about its estimation.

## 2.1 Definitions

MI (Shannon, 1948) is a criterion from the information theory which has proven to be very efficient in feature selection (Battiti, 1994; Fleuret, 2004) mainly because it is able to detect non linear relationships between variables, while other popular criteria as the well-known correlation coefficient are limited to linear relationships. Moreover, MI can handle groups of vectors, i.e. multidimensional variables.

MI is intuitively a symmetric measure of the information two random variables $X$ and $Y$ carry about each other and is formally defined as follows:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{1}$$

where $H(X)$ is the entropy of $X$:

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

with $f_X$ being the probability density function (pdf) of $X$. $H(X,Y)$ is the entropy of the joint variable $(X,Y)$ defined in the same way.

The MI can be reformulated as:

$$I(X;Y) = \int\int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)}\, dx\, dy. \tag{3}$$

This last equation defines MI as the Kullback-Leibler divergence between the joint distribution $f_{X,Y}$ and the product of the distributions $f_X$ and $f_Y$, these quantities being equal for independant variables.

As in practice none of the pdf $f_X$, $f_Y$ and $f_{X,Y}$ are known, MI can not be computed analytically but has to be estimated from the data set.

## 2.2 MI Estimation

Traditional MI estimators are based on histograms or kernels (Parzen, 1962) density estimators which are used to approximate the value of the MI according for example to (1) (Kwak and Choi, 2002). Despites its popularity, this approach has the huge drawback that it is unreliable for high-dimensional data. Indeed, as the dimension of the space increases, if the number of available samples remains constant, these points will not be sufficient to sample the space with an acceptable resolution. For histograms, most of the boxes will be empty and the estimates are likely to be inaccurate. Things will not be different for kernel estimators which are essentially smoothed histograms. These problems are a direct consequence of the *curse of dimensionality* (Bellman, 1961; Verleysen, 2003), stating that the number of points needed to sample a space at a given resolution increases exponentially with the dimension of the space; if $p$ points are needed to sample a one-dimensional space at a given resolution, $p^n$ points will be needed if the dimension is $n$.

Since in this paper MI estimation is needed for multi-dimensional data points, other estimators have to be considered. To this end, a recently introduced family of estimators based on the principle of nearest neighbors are used (Kraskov et al., 2004; Gómez-Verdejo et al., 2009). These estimators have the advantage that they do not estimate the entropy directly and are thus expected to be more robust if the dimension of the space increases. They are inspired by the Kozachenko-Leonenko estimator of entropy (Kozachenko and Leonenko, 1987):

$$\hat{H}(X) = -\psi(k) + \psi(N) + \log(c_d) + \frac{d}{n}\sum_{n=1}^{N} \log(\varepsilon_X(n,k)) \tag{4}$$

where $k$ is the number of nearest neighbors considered, $N$ the number of samples of a random variable $X$, $d$ the dimensionality of these samples, $c_d$ the volume of a unitary ball of dimension $d$ and $\varepsilon_X(n,k)$ twice the distance from the $n^{th}$ observation in $X$ to its $k^{th}$ nearest neighbor; $\psi$ is the digamma function:

$$\psi(k) = \frac{\Gamma'(k)}{\Gamma(k)} = \frac{d}{dk}\ln\Gamma(k)\,,\,\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx.$$

Using (4), Kraskov et al. (Kraskov et al., 2004) derived two slightly different estimators for regression problems (i.e. for problems with a continuous output). The most widely used one is:

$$\begin{aligned}\hat{I}(X;Y) =& \psi(N) + \psi(K) - \frac{1}{k} \\ & - \frac{1}{N}\sum_{n=1}^{N}\left(\psi(\tau_x(n)) + \psi(\tau_y(n))\right)\end{aligned} \tag{5}$$

where $\tau_x(n)$ is the number of points located no further than $\varepsilon_X(n,k)$ from the $n^{th}$ observation in the X space; $\tau_y(n)$ is defined similarly in the Y space with $\varepsilon_Y(n,k)$.

In case of a classification problem, $Y$ is a discrete vector representing the class labels. Calling $L$ the number of classes, Gomez et al. (Gómez-Verdejo et al., 2009) took into account the fact that the probability distribution of $Y$ is estimated by $p(y = y_l) = n_l/N$, where $n_l$ is the number of points whose class value is $y_l$, and proposed to estimate the MI as:

$$\begin{aligned}\hat{I}_{cat}(X;Y) = \psi(N) - \frac{1}{N}n_l\psi(n_l) + \\ \frac{d}{N}\left[\sum_{n=1}^{N}\log(\varepsilon_X(n,K)) - \sum_{l=1}^{L}\sum_{n\in y_l}\log(\varepsilon_l(n,K))\right].\end{aligned} \tag{6}$$

In this last equation, $\varepsilon_l(n,K)$ is defined in the same way as $\varepsilon_X(n,K)$ in (4) but the neighbors are limited to the points having the class label $y_l$.

If both $X$ and $Y$ are categorical features, equations (2) and (3) become sums where the probabilities can be estimated from the samples in the learning set by simple counting and no estimator is needed. Assume $X$ (resp. $Y$) takes $s_x$ ($s_y$) different values $x_1 \dots x_{s_x}$ ($y_1 \dots y_{s_y}$), each with a probability $p_{x_i}$ ($p_{y_i}$) and denote by $p_{x_i,y_i}$ the joint probability of $x_i$ and $y_i$, then:

$$I(X;Y) = \sum_{i=1}^{s_x}\sum_{j=1}^{s_y} p_{x_i,y_j}\log\frac{p_{x_i,y_j}}{p_{x_i}p_{y_j}}. \tag{7}$$

# 3 METHODOLOGY

This section presents the proposed feature selection procedure. It ends with a few comments on the filter / wrapper dilemma.

## 3.1 Lists Ranking

As already discussed, this paper suggests avoiding the use of any similarity measure between mixed data points. To this end, the proposed procedure starts by separating the continuous and the categorical features. Both groups of features are then ranked independently, according to the following strategies.

### 3.1.1 Continuous Features

For continuous features, the multivariate MI criterion is considered, meaning that the MI is directly estimated between a set $X$ of features and the output $Y$.

The nearest neighbors based MI estimators (Kraskov et al., 2004; Gómez-Verdejo et al., 2009) described previously are particularly well suited for multivariate MI estimation. Indeed, as already explained, they do not require the estimation of multivariate probability density functions. This crucial advantage allows us to evaluate robustly the MI between groups of features with a limited number of samples. As an example the estimator described in (Kraskov et al., 2004) has been used sucessfully in feature selection for regression problems (Rossi et al., 2006).

In this paper, the multivariate MI estimator is combined with a greedy forward search procedure; at each step of the selection procedure, the feature whose addition to the set of already selected features leads to the largest multivariate MI with the output is selected. This choice is never questioned again, hence the name forward. Algorithm 1 illustrates a greedy forward search procedure for a relevance criterion $c$ to be maximized, with $R\{i\}$ being the $i^{th}$ element of $R$.

Obviously, in such a procedure, the possible redundancy between the features is implicitely taken into account since the selection of a feature carrying no more information about the output than the already selected ones will result in no increase of the MI.

### 3.1.2 Categorical Features

It is important to note that the multivariate MI estimators (Kraskov et al., 2004; Gómez-Verdejo et al., 2009) should not be considered for categorical features. Indeed, for categorical data it is likely that the distances between a sample and several others are identical, especially in the first steps of the forward selection procedure. These ex-aequos could bring confusion in the determination of the nearest neighbors and harm the MI estimation. Moreover, using directly equation (7) can be untractable in practice. As an example, if $X$ consists of 20 features, each taking 3 possible discrete values, the total number of possible values $s_x$ for points in $X$ is $3^{20} > 3 \times 10^9$.

Another criterion than multivariate MI has thus to be thought of. In this paper, the minimal-Redundancy maximal-Relevance (mRmR) principle is used since it has proven to be very efficient in feature selection when combined with the MI criterion (Peng et al., 2005). The idea is to select a set of maximally informative but not redundant features.

This principle is also combined with a greedy forward search strategy: suppose a subset of features has already been selected; one searches for the unselected feature which maximises $D-R$ where $D$, the estimated relevance, is the MI between the new feature and the output. $R$ is the estimated redundancy and can be measured by the average MI between the new feature and each of the already selected features. Denote by $S$ the set of indices of already selected features; the mRmR criterion $D-R$ for feature $i$ ($i \notin S$) given an output vector $Y$ is:

$$mRmR(f_i) = I(f_i;Y) - \frac{1}{|S|}\sum_{j\in S} I(f_i;f_j). \quad (8)$$

All MI estimations or computations in the mRmR procedure are thus bivariate (i.e. involve only two variables). Of course, bivariate methods are not expected to perform as well as multivariate ones since they only consider pairwise redundancy or relevance. A simple example showing this is the well-known XOR problem. It consists of two random binary vectors $X_1$, $X_2$ and an output $Y$ whose $i^{th}$ element is 1 if the $i^{th}$ elements of $X_1$ and $X_2$ are different and 0 otherwise. Individually, both vectors carry no information about the output $Y$. However, together they entirely determine it. Thus, even if $X_1$ is selected, a mRmR procedure will not be able to determine $X_2$ as relevant while a multivariate approach will.

## 3.2 Combination of the Lists

Once established, the two lists are combined according to the accuracy (the percentage of well-classified samples) of a classification model. First, the accuracies of a model built on the first continuous or the first categorical feature are compared. The feature leading to the best result is chosen and removed from the list it belongs to. The selected feature is then combined with the best continuous or with the best categorical feature that still belong to their respective lists, i.e. that has not been selected yet; the subset for which a model performs the best is selected, and so on until all features have been selected. The whole feature selection procedure is described in Algorithm 2.

```
Input:     A set $F$ of features $i$, $i = 1 : n_f$
           A class labels vector $Y$
Output: A list $L$ of sorted indices of features.
begin
    $R \longleftarrow 1 : n_f$
    //R is the set of indices of not yet
     selected features
    $L \longleftarrow \emptyset$
    for $k = 1 : n_f$ do
        foreach $i \in R$ do
            $set \longleftarrow L \cup R\{i\}$
            $score\{i\} \longleftarrow c(set, Y)$
        end
        $winner \longleftarrow \arg\max_j score\{j\}$
        $L \longleftarrow [L; R\{winner\}]$
        $R \longleftarrow R \setminus R\{winner\}$
        clear $score$
    end
end
```

Algorithm 1: Forward search procedure to maximize a criterion $c$.

## 3.3 Filter or Wrapper Feature Selection

As can be seen from the previous developments, two different approaches to feature selection are successively used to produce a global algorithm. First, building the two lists is made by using filter methods. This means that no classification model is used and that the selection is rather based on a relevance criterion, such as MI in this paper. On the other hand, the combination of the lists does require a specific classifier and is thus a wrapper procedure.

Wrappers are generally expected to lead to better results than filters since they are designed to optimize the performances of a specific classifier. Of course, wrappers are also usually much slower than filters, precisely because of the fact they have to build a huge number of classification models with possible hyperparameters to tune.

As an example, the exhaustive wrapper approach consisting in testing all the possible feature subsets would require building $2^{n_f}$ models, $n_f$ being the number of features. If there are 20 features, about $10^6$ classifiers must be built. This method thus becomes quickly untractable as the number of features grows. An alternative is to use heuristics such as the greedy forward search presented in Algorithm 1. The number of models to build is then $\frac{n_f(n_f+1)}{2} - 1$ which is still unrealistic for complex models.

On the contrary, the approach proposed in this paper only requires the construction of at most $n_f - 1$ classifiers. Indeed, Algorithm 2 emphasizes the fact that the use of a classification model is needed only if none of the lists are empty; in practice the number of models to build will thus often be smaller than

$n_f - 1$. In addition $\frac{n_{cont}(n_{cont}+1)}{2} + \frac{n_{cat}(n_{cat}+1)}{2} - 2$ evaluations of the MI are necessary, $n_{cat}$ and $n_{cont}$ being respectively the number of categorical and continuous features. A compromise between both approaches is thus found, which prevents us to use any similarity measure between mixed samples, while keeping the computional burden of the procedure relatively low.

```
Input: A set of categorical features $F_{cat}\{i\}$,
       $i = 1 : n_{cat}$
       A set of continuous features $F_{cont}\{i\}$,
       $i = 1 : n_{cont}$
       A class labels vector $Y$.
Output: A list $L$ of sorted indices of features.
begin
    $InCat \longleftarrow SortCat(F_{cat}, Y)$
    //Get the sorted list of indices for
      categorical features
    $InCon \longleftarrow SortCon(F_{cont}, Y)$
    //Get the sorted list of indices for
      continuous features
    $L \longleftarrow \emptyset$

    for $k = 1 : n_{cat} + n_{cont}$ do
        if $InCat \neq \emptyset$ and $InCon \neq \emptyset$ then
            $AccCat \longleftarrow Acc(L \cup InCat\{1\}, Y)$
            $AccCon \longleftarrow Acc(L \cup InCon\{1\}, Y)$
            //Function Acc(.)  gives the
              accuracy of a classifier.
            if $AccCat < AccCon$ then
                $L \longleftarrow L \cup InCon\{1\}$
                delete $InCon\{1\}$
            else
                $L \longleftarrow L \cup InCat\{1\}$
                delete $InCat\{1\}$
            end
        else
            if $InCat = \emptyset$ then
                $L \longleftarrow L \cup InCon\{1\}$
                delete $InCon\{1\}$
            else
                $L \longleftarrow L \cup InCat\{1\}$
                delete $InCat\{1\}$
            end
        end
    end
end
```

Algorithm 2: Proposed feature selection algorithm.

## 4 EXPERIMENTAL RESULTS

To assess the performance of the proposed feature selection algorithms, experiments are conducted on artificial and real-world data sets. The limitations of methods based on a given similarity measure between mixed samples are first emphasized on a very simple toy problem. Results obtained on four UCI (Asuncion and Newman, 2007) data sets then confirm the interest of the proposed approach.

Table 1: Description of the datasests used in the experiments.

| Name | samples | cont. features | cat. features | classes |
|---|---|---|---|---|
| Heart | 270 | 6 | 7 | 2 |
| Hepatitis | 80 | 6 | 13 | 2 |
| Australian Credit | 690 | 6 | 8 | 2 |
| Contraception | 1473 | 2 | 7 | 3 |

Two classification models are used in this study. The first one is a Naive Bayes classifier with probabilities for continuous attributes estimated using Parzen window density estimation (Parzen, 1962) and those for categorical attributes estimated by counting.

The second one is a 5-nearest neighbors classifier, with distances between samples computed by the Heterogeneous Euclidean-Overlap Metric (HEOM) (Wilson and Martinez, 1997) while other choices could as well have been made (see e.g. (Boriah et al., 2008)). This metric uses different distance functions for categorical and continuous attributes. It is defined for two vectors $X = [X_1 \ldots X_m]$ and $Y = [Y_1 \ldots Y_m]$ as $d_{heom}(X,Y) = \sqrt{\sum_{a=1}^{m} d_a(X_a,Y_a)^2}$ where

$$d_a(x,y) := \begin{cases} overlap(x,y) & \text{if a is categorial} \\ \frac{|x-y|}{max_a - min_a} & \text{if a is continuous} \end{cases}$$

with $max_a$ and $min_a$, respectively the maximal and minimal values observed for the $a^{th}$ feature in the training set, and $overlap(x,y) = 1 - \delta(x,y)$ ($\delta$ denoting the Kronecker delta, $\delta(x,y) = 1$ if $x = y$ and 0 otherwise). These models have mainly been chosen because they are both known to suffer dramatically from the presence of irrelevant features in comparison with, for example, decision trees.

In this section, we compare the proposed feature selection approach with the algorithm by Hu and al. (Hu et al., 2008). As already explained, in that paper, the authors consider two points as neighbors if their categorical attributes are equal and if one is among the $k$ nearest neighbors of the other or if the distance between them is not too large according to their numerical attributes (there are thus two versions of the algorithm). Then, they look for the features for which the largest number of points share their class label with at least a given fraction of their neighbors. The methodology is thus extremely dependent on the chosen definition of neighborhood. Even if this definition can be modified, it is not easy, given an unknown data set, to determine a priori which relation can be a good choice. Among the two versions of Hu and al.'s algorithm, only the nearest neighbors-based one will be considered in this work, since it has been shown more efficient in practice (Hu et al., 2008), which was confirmed by our experiments.

## 4.1 Toy Problem

To underline the aforementionned drawbacks, a toy problem is considered to show the limitations of methods based on dissimilarities for mixed data.

It consists in a data set $X$ containing two categorical variables $(X_1, X_2)$ taking two possible discrete values with equal probability and two continuous variables $(X_3, X_4)$ uniformly distributed over $[0;1]$. The sample size is 100. A class labels vector $Y$ is also built from $X$; the points whose value of $X_3$ is below 0.15 or above 0.85 are given the class label 1 and the other points are given the class label $-1$. The only relevant variable is thus $X_3$, which sould be selected in first place by accurate feature selection algorithms.

However, the problem with Hu et al.'s method is that many points with class label 1 do not have enough neighbors with the same label. Feature $X_3$ will not thus be detected as relevant by the algorithms. More precisely, over 50 repetitions, $X_3$ has been selected in the first place only in 28 cases. With the approach proposed in this paper, $X_3$ has been selected first in the 50 repetitions of the experiment.

Interestingly, if the problem is modified such that points for which the value of $X_3$ is between 0.4 and 0.7 have class label 1 and other points have class label $-1$, then Hu and al.'s algorithm always detect $X_3$ has the most relevant variable. Although the proportion of both classes in the two problems are the same, the second one is more compatible with the chosen definition of neighborhood, explaining the better results obtained. Of course, this definition of neighborhood could be modified to better fit the first problem but would then likely be inaccurate in other situations.

## 4.2 Real-world Data Sets

Four classification benchmark data sets from the UCI Machine Learning Repository (Asuncion and Newman, 2007) are used in the study to further illustrate the interest of the proposed approach. All contain continuous and categorical attributes and are summarized in Table 1. Two come from the medical world, one is concerned with whether an applicant should receive a credit card or not and the last one is about the choice of a contraceptive method. The data sets used in this work are not the same as those considered in (Hu et al., 2008) since many of the latest do not actually contain mixed features.

As a preprocessing, observations containing missing values are deleted. Moreover, continuous attributes are normalized by removing their mean and dividing them by their standard deviation in order to make the contribution of each attribute to the Euclidean distance equivalent in the MI estimation. As suggested in (Gómez-Verdejo et al., 2009), for continuous features, the MI is estimated over a range of different values of the parameter $k$ and then averaged to prevent strong underfitting or overfitting. In this paper, 4 to 12 neighbors are considered except for the Hepatitis data set for which 4 to 6 neighbors are taken into account. This is due to the fact that a class is represented by only a few samples in this data set.

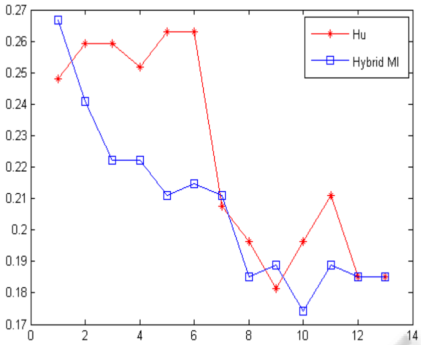

Figure 1: Error rate as a function of the number of selected features: Heart data set and naive Bayes classifier.

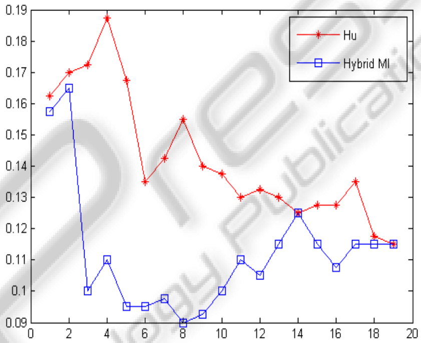

Figure 2: Error rate as a function of the number of selected features: Hepatitis data set and naive Bayes classifier.

The criterion of comparison is the classification error rate. It is estimated through a 5-fold cross validation procedure repeated 5 times with different random shufflings of the data set. The presented error rates are obtained on the test set, independant of the training set used to train the classifier. The feature to choose at each wrapper step is determined by a 5-fold cross validation on the training set.

Figures 1 to 7 show the average classification error rate achieved by the proposed method (referred to as Hybrid MI) and by Hu and al.'s one with respect to the number of features selected. As can be seen, the approach is very competitive with both classifiers and leads to the global smallest misclassification rate in 7

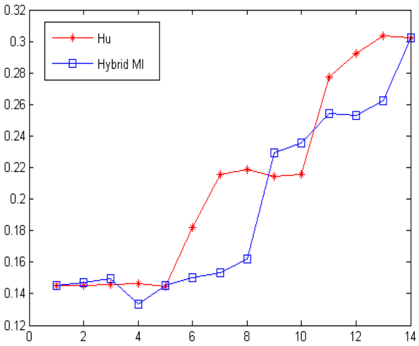

Figure 3: Error rate as a function of the number of selected features: Australian credit data set and naive Bayes classifier.

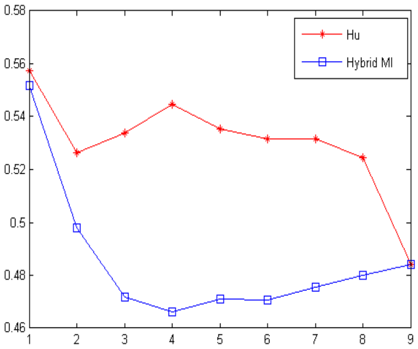

Figure 4: Error rate as a function of the number of selected features: Contraception data set and naive Bayes classifier.

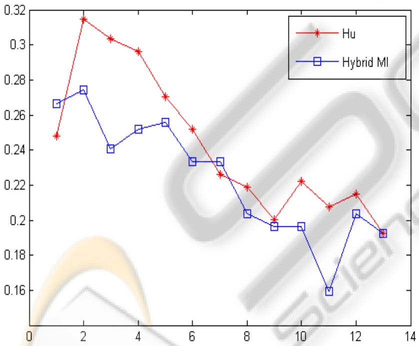

Figure 5: Error rate as a function of the number of selected features: Heart data set and k-nn classifier.

of the 8 experiments. Moreover, for the Hepatitis and the Contraception data sets, it is the only approach selecting a subset of features leading to better performances than the set of all features. The improvement in classification accuracy is thus obvious.

## 5 CONCLUSIONS

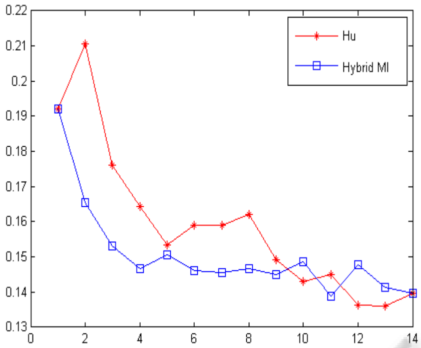

Figure 6: Error rate as a function of the number of selected features: Australian credit data set and k-nn classifier.

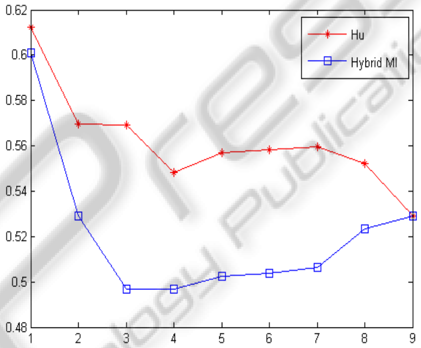

Figure 7: Error rate as a function of the number of selected features: Contraception data set and k-nn classifier.

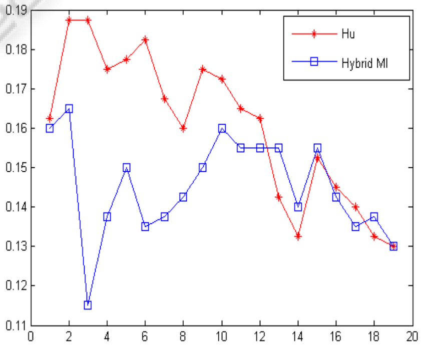

Figure 8: Error rate as a function of the number of selected features: Hepatitis data set and k-nn classifier.

This paper introduces a new feature selection methodology for mixed data, i.e. for data with both categorical and continuous attributes. The idea of the method is to independently rank both types of features before recombining them guided by the accuracy of a classifier. The proposed algorithm is thus a combination of a filter and a wrapper approach to feature selection. The well-known MI criterion is used to produce both ranked lists. For continuous features, multidimensional MI estimation is used while a mRmR approach is considered for categorical features.

One of the most problematic issues for feature selection algorithms dealing with mixed data is to chose an appropriate relationship between categorical and continuous features leading to a sound similarity measure or neighborood definition between observations. This new method simply alleviates this problem by ignoring this unknown relationship. The hope is to compensate the loss of information induced by this hypothesis by a more accurate ranking of features of each type and by the use of a classification model.

Even if the approach requires the explicit building of prediction models, the number of models to build is small compared to a pure wrapper approach. Moreover, experimental results on four data sets show the interest in terms of the accuracy of two classifiers.

All the developments presented in this paper could be applied to regression problems (problems with a continuous output); the only modifications needed would be to use the MI estimator (5) instead of (6) for the continuous features and the mRmR approach for the categorical ones. It would thus be interesting to test the proposed approach on such problems.

## REFERENCES

Asuncion, A. and Newman, D. (2007). UCI machine learning repository. *University of California, Irvine, School of Information and Computer Sciences, available at http://www.ics.uci.edu/∼mlearn/MLRepository.html*.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5:537–550.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Boriah, S., Chandola, V., and Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In *SDM'08*, pages 243–254.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.*, 5:1531–1555.

Gómez-Verdejo, V., Verleysen, M., and Fleury, J. (2009). Information-theoretic feature selection for functional data classification. *Neurocomputing*, 72:3580–3589.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of ICML 2000*, pages 359–366. Morgan Kaufmann Publishers Inc.

Hu, Q., Liu, J., and Yu, D. (2008). Mixed feature selection based on granulation and approximation. *Know.-Based Syst.*, 21:294–304.

Kozachenko, L. F. and Leonenko, N. (1987). Sample estimate of the entropy of a random vector. *Problems Inform. Transmission*, 23:95–101.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69(6 Pt 2).

Kwak, N. and Choi, C.-H. (2002). Input feature selection by mutual information based on parzen window. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:1667–1671.

Parzen, E. (1962). On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:1226–1238.

Rossi, F., Lendasse, A., François, D., Wertz, V., and Verleysen, M. (2006). Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and Intelligent Laboratory Systems*, 80(2):215–226.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27:379–423.

Tang, W. and Mao, K. Z. (2007). Feature selection algorithm for mixed data with both nominal and continuous features. *Pattern Recogn. Lett.*, 28:563–571.

Verleysen, M. (2003). Learning high-dimensional data. *Limitations and Future Trends in Neural Computation*, 186:141–162.

Wilson, D. R. and Martinez, T. R. (1997). Improved heterogeneous distance functions. *J. Artif. Int. Res.*, 6:1–34.