A new contextual based feature selection.

Hafida Senoussi, Brigitte Chebel-Morello

A New Contextual Based Feature Selection

H. Senoussi, B. Chebel-Morello.

Abstract—The pre processing phase is essential in Knowledge Data Discovery process. We study too particularly the data filtering in supervised context, and more precisely the feature selection. Our objective is to permit a better use of the data set. Most of filtering algorithm use myopic measures, and give bad results in the case of the features correlated part by part. Consequently in the first time, we build two new contextual criteria. In the second part we introduce those criteria in an algorithm similar to the greedy algorithm. The algorithm is tested on a set of benchmarks and the results were compared with five reference algorithms: Relief, CFS, Wrapper (C4.5), consistancySubsetEval and GainRatio. Our experiments have shown its ability to detect the semi-correlated features. We conduct extensive experiments by using our algorithm like pre processing data for decision tree, nearest neighbours and Naïve Bays classifiers.

I. Introduction

High-dimensional data often contains redundant features. Many works underlines the performances reduction in data mining algorithms when many descriptive features or examples are treated. The presence of irrelevant and/or redundant features affects the speed and accuracy of learning algorithms [6]. [1], [7], [11], [19], [9], [21], [24], [20], [26]. Langley [12] studied the behavior of C4.5 on various data test and noted that the decision tree algorithm is not affected by irrelevant features, when the relevant ones are independent. On the other hand, a problem appears when the features' combination determines the class. Indeed taken separately none of them distinguishes better the concept. The data filtering stage, useful in this context was considered by many authors [4], [5]. It aims to privilege the data quality to the quantity by selecting a subset of relevant features.

To compensate for this lack and to take into account features relations in the selection we have defined two complementary criteria, one Myopic and the other Contextual. These criteria are built from features relevance definitions, and were established in a basic tree creation algorithm [15], [16]. This tree detects correlated features on the contrary of C4.5 and recognizes completely redundant features. However, partially redundant features are not recognized. This type of algorithm did not permit to express all criteria potentialities. Thus, it is necessary to use a pre-treatment algorithm for the data to prepare features at the data mining stage. To do so we have implanted the two suggested criteria in a greedy algorithm, which will highlight the partial dependence between features. Indeed most works in statistics make the erroneous assumption in many cases of the independence between features describing the training datasets. We release this strong assumption, and we seek algorithmically with the help of two discriminating criteria, the relations between descriptive features and class. Our article will be structured as follows:

The second section will be devoted to the contextual algorithms that take into account in a finer way than the other methods the type of features. In section 3 we will present the suggested criteria, our contribution is not in the filtering algorithm, but in the combination algorithm criteria. Section 4 will relate our algorithm evaluation compared to five algorithms identified as being the best since they can treat different features type and consequently make a finer features selection of a minimal subset. In order to highlight our algorithm effectiveness to treat partially correlated data, we tested it on artificial benchmarks known for their features interactions. In section 5 we will conclude this work and give some further directions.

II. Feature Selection Techniques

A. Introduction

As described in their paper, Liu [13], [21], Blum and Langley [2] argued that most existing feature selection algorithms consist of the following components.

1) Search procedure

The search for feature subsets could start with no features with a forward search, all features with a backward search, or random subset of features with a bidirectional search thus features could be successively added or removed by a certain procedure.

2) Evaluation Criteria

Evaluation criteria is an important component of any feature selection method, it measures the goodness of a specific subset, an evaluation criteria can be categorized into two main groups based in their dependency on mining algorithm: filters and wrappers. Filters operate independently of any mining algorithm, where undesirable features are filtered out of the data before learning begins; it relies on different measure such as distance, information, dependency and consistency. Wrapper methods use the performance of the mining algorithm.

3) Stopping Criteria

A stopping criterion determines when the feature selection process should stop.

This work was supported by the Ecole supérieure de mécanique et des mécatroniques. Laboratoire d'automatique de Besançon. 25000 Besançon, France.

H. Senoussi, B. Chebel- Morello are with Laboratoire d'automatique de Besançon. 25000 Besançon, France. (e-mail: senoussih@yahoo.fr, brigitte.morello@ens2m.fr)

B. Comparison from the feature's type detected

We noted that algorithms based on myopic criteria do not detect the correlations contrary to those using semi-contextual or contextual criteria. Indeed, Relief [7], [8], **CFS** [25], [19], **mRMR** [20], **FCBF** [26] **are the most powerful algorithms from this point of view, because they consider feature-feature inter-correlation. Relief** [7] **and ReliefF** [8] **algorithms score individual features rather than scoring features subsets, those features with scores exceeding a user-specified threshold are selected for the final subset.** A **useful feature should differentiate between instances from different classes and have the same value for instances from the same class. CFS** [25], [19] **this algorithm uses a heuristic for evaluating the merit of a subset of features. This heuristic is based in the hypothesis that** *"a good feature subset is one that contains features highly correlated with the class, yet uncorrelated with each other"* **(Hall** [19] **page** 3**). The algorithm is powerful as long as the interaction between features is not too large. mRMR feature selection** [20] **minimum redundancy–maximum relevance feature selection algorithm selects features that should be both minimally redundant among themselves and maximally relevant to the target classes. The optimal subset is that which maximizes the distance between the two profits. FCBF** [26] **use a correlation measure based on the information gain to detect the redundancy between features they chose the symmetrical uncertainty. The algorithm involves two steps: (**1**) calculates the** SU **value for each feature selects and orders relevant ones according to a predefined threshold, and (**2**) selects a subset of predominant features.**

The strong points of these algorithms are their effectiveness to deal with diverse problems like, modal, continuous and noised data. The three remaining algorithms (CFS, mRMR and FCBF) calculate the pair-wise feature-feature inter-correlation; this can determine the redundancy between two features (degree of equivalence) and not the redundancy of a feature compared to a subset of features. This type of redundancy that we named partial redundancy (semi-redundancy) or correlation by part (semi-correlation) can be treated only if we compare the contribution of a feature to a subset of features. Our proposed measure makes possible to detect this type of redundancy by calculating the discriminating capacity of a subset of features with or without the feature in question. This measure also detects the relevant features compared to a subset of features.

These algorithms use respectively two measures, relevance and correlation. We will take these measures as bases to work out two new criteria which must solve the redundancy or the semi-redundancy features problem.

III. EVALUATION CRITERIA

Evaluation criteria are a crucial point in feature selection. There are two approaches in the criterion development: Traditional statistical approach, and pair-wise objects comparison approach.

In this paper we will propose a new criterion which is elaborated from a pair-wise approach that seems to us promising. The fundamental idea of pair-wise comparison is assigned to A**. Condorcet since** 178**5.**

In feature selection framework, the evaluation criteria can be categorized into two main groups based in the dependency between features:

- Myopic and semi-contextual measures (individual feature evaluation) which estimate the feature quality out of the others' context as Relief [7] **and ReliefF** [8]**.**

- Contextual measures (subset evaluation) which consider the features interactions [19], [20], [26]**.**

Most of the existing measures belong to the first category, that's why Kira and Rendel [7] **and Kononenko** [9] **underlined the induction algorithms difficulties to work with correlated data.**

A. Symbolic data notation

We will formalize our problem by using the symbolic data notation given by E.Diday [5].

Let Π **the studied population.**

Ω **the observed population composed of N objects or individuals or instances.**

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

1) Attribute value modelling

Each elements of Ω **is characterized by a set of** *r* **features or attributes:**

$$L = \{y_1, \ldots\ldots, y_k, \ldots, y_r\}$$

$$\Omega \to O_k$$
$$\omega \to y_k(\omega) = m_v^k$$

Where O_k **is a set of** m_k **modalities (values) of the feature (attribute)** y_k**.**

$O_k = \{m_1^k, \ldots m_v^k, \ldots, m_{q^k}^k\}$ **where** m_v^k **is the modalities v of the feature** y_k

$y_{class} \in \Omega_{class}$ **such as** $O_{class} = \{m_1^{class}, m_2^{class}, \ldots, m_q^{class}\}$

Let $O = O_1 \times O_2 \times \ldots \times O_k \times \ldots \times O_r$ **be the workspace**

and Y **the application:**

$$Y : \Omega \to O$$
$$\omega \to Y(\omega) = (y_1(\omega), \ldots\ldots y_r(\omega))$$

2) Functional modelling

For each feature y_k **we associate the Boolean function** φ_i^k **relative to each object.**

φ_i^k **Such as** $\Omega \to \{0,1\}$

$$\omega \to \varphi_i^k(\omega_i / m_v^k) = 1 \Leftrightarrow y_k(\omega_i) = m_v^k$$
$$= 0 \Leftrightarrow otherwise \quad (1)$$

3) Pair-wise representation

We associate to a feature y_k the function φ_{ij}^k relative to each pairs of objects.

$$\varphi_{ij}^k : \Omega \times \Omega \to \{0,1\}$$

For each pair of objects $(\omega_i, \omega_j), \quad i \neq j$

$$(\omega_i, \omega_j) \quad \varphi_{i,j}^k = \varphi^k(\omega_i, \omega_j) = \begin{cases} 1 \Leftrightarrow y_k(\omega_i) = y_k(\omega_j) \\ 0 \ otherwise \end{cases}, \quad i, j = 1, \ldots, n \quad (2)$$

B. Discriminating power

The original discriminating power [18] **is the only criterion to our knowledge which is contextual but which works on pairs of concepts. We will first give some definitions resulting from E. Diday** [5] **formalism to present this criterion.**

Consider an elementary event ek = [yk = Vk] where $V_k \in O_k$. An object assertion is a conjunction of elementary events:

$$\mathbf{a} = [\, y_1 = V_1 \,] \wedge \ldots\ldots \wedge [\, y_p = V_p \,].$$

Be A **the whole assertions, and K the whole assertions couples K=AxA.**

Be the function $comp(V_{ij}, V_{jk}) = \begin{cases} 1 & si \quad V_{ij} \cap V_{jk} = \phi \\ 0 & otherwise \end{cases}$ (3)

Where V_{ik}, V_{jk} are the values taken by the feature y_k in the assertion a_i, a_j. The discriminating power of a subset L = {$y_1 \ldots, y_m$} **of features is equal to the number of assertions couples which are discriminated by at least one feature of** L.

$$\mathbf{DP(L, K)} = \sum_i \sum_j \max_{y_k \in L} \left(comp\left(V_{ik}\ V_{jk}\right) \right) \quad (4)$$

Given a subset K of objects; *the original discriminating power* **of a feature yk compared to a subset** L **of features is equal to the number of assertions couples discriminated by only yk and no other features.**

$$\mathbf{DPO(yk, L, K)} = \sum_i \sum_j \max_{(a_i, a_j) \in K} \left(comp\left(V_{ik}\ V_{jk},\right) - \max_{y_k \in L} \left(comp\left(V_{ik}\ V_{jk},\right)\ \right), 0 \right)$$

(5)

The criteria which we propose belong to those measures, and if we review the basic definitions we note that the discriminating power is a relevance criterion. Working in a supervised context, and on feature value represented objects, we develop our criteria in flow; we will name discriminating capacity, and discriminating capacity gain.

C. Suggested Criteria

Our goal is to design efficient contextual criterion from the pair-wise approach, it will be built starting from defined Boolean functions for each feature on a given pair of objects, and is aggregated on all the features in such manner to obtain a strong relevance measure and a low relevance measure.

Formalization of Relevance

Proposition 1: Relevance to the target

A feature y_k is relevant for a class y_{class} if there is a pair of objects in Ω **space such as objects differ only by the value of this feature.**

$$for\ (\omega_i, \omega_j) \in \Omega \times \Omega, \forall i, i = 1..r, \exists k / y_k(\omega_i) \neq y_k(\omega_j)$$
$$and\ y_k(\omega_i) \neq y_{class}(\omega_j)$$

The constraint of exclusiveness in a feature and only one is a very strong constraint we must release it to differentiate an irrelevant feature from a relevant one but redundant with another feature. Indeed if for example two features are identical they are not detected as relevant even if they are taken one by one. Consequently a weak relevance is considered.

Proposition n° 2: Weakly Relevant feature WR

Let consider the Ω **population and the feature y_k, the discriminating capacity of y_k on** Ω **is the number of discriminating pairs over** Ω.

$$on(\Omega \times \Omega) \qquad WR(y_k, \Omega) = \sum_{i=1}^{n} \sum_{i=1}^{n} \overline{\varphi_{i,j}^k} \bullet \overline{\varphi_{i,j}^{class}} \quad (6)$$

Proof:

A Boolean function is defined for each feature and the feature weak relevance for a pair of objects is given by WR and will be equal to 1 **when there is relevance.**

$$(\omega_i, \omega_j) \quad \varphi_{i,j}^k = \varphi_k(\omega_i, \omega_j) = 0 \Leftrightarrow y_k(\omega_i) \neq y_k(\omega_j) \Leftrightarrow \overline{\varphi_{i,j}^k} = 1$$
$$(\omega_i, \omega_j) \quad \varphi_{i,j}^{classl} = \varphi_{class}(\omega_i, \omega_j) \Leftrightarrow y_{class}(\omega_i) \neq y_{class}(\omega_j) \Leftrightarrow \overline{\varphi_{i,j}^{classl}} = 1 \quad (7)$$

$$(\omega_i, \omega_j) \qquad WR(y_k, \omega_i, \omega_j) = \overline{\varphi_{i,j}^k} \bullet \overline{\varphi_{i,j}^{class}} = 1$$

The aggregation of WR in all the pairs gives DC (discriminating capacity); DC is the number of discriminating pairs over the Ω **population.**

Proposition n°3: Weakly Relevant Subset: Discriminating Capacity DC

$$DC(L,\Omega)=\sum_{i=1}^{n}\sum_{i=1}^{n}\overline{\prod_{k=1}^{m}\varphi_{i,j}^{k}}\bullet\overline{\varphi_{i,j}^{classl}} \tag{8}$$

Giving a subset of *m* feature (L = (y_1… y_m)). Subset relevance is the number of pairs that are discriminate at least with one feature for each class.

Proof:

$$(\omega_i,\omega_j)\qquad WR\ (y_k,\omega_i,\omega_j)=\overline{\varphi_{i,j}^{k}}\bullet\overline{\varphi_{i,j}^{class}} \tag{9}$$

We can aggregate this measure over the subset L

$$(\omega_i,\omega_j)\quad WR(L,\omega_i,\omega_j)=\overline{\prod_{k=1}^{m}\varphi_{i,j}^{k}}\bullet\overline{\varphi_{i,j}^{class}} \tag{10}$$

For a given pair, the above expression is equal to 1 **if there is at least one feature which discriminates among the *m* ones.**

For two different classes $\overline{\varphi_{i,j}^{class}}=1$ **, only one feature** y_l **can discriminate a data pair witch can be express as:** $\overline{\varphi_{i,j}^{l}}=1$ **and** $\varphi_{i,j}^{l}=0$ **. And the other features have the same value, what results in:** $\forall k\in[1,m]\ k\neq l\quad \varphi_{ij}^{k}=1$

$$\prod_{\substack{k=1\\k\neq l}}^{m}\varphi_{ij}^{k}=1\ \Rightarrow\ \prod_{k=1}^{m}\varphi_{ij}^{k}=0\ \Rightarrow\ \overline{\prod_{k=1}^{m}\varphi_{ij}^{k}}=1\ \Rightarrow\ \overline{\varphi_{ij}^{class}}*\overline{\prod_{k=1}^{m}\varphi_{ij}^{k}}=1$$

$$(\omega_i,\omega_j)\quad WR(L,\omega_i,\omega_j)=\overline{\prod_{k=1}^{m}\varphi_{i,j}^{k}}\bullet\overline{\varphi_{i,j}^{class}}$$

$$on(\Omega\times\Omega)\quad DP(L,\Omega)=\sum_{i=1}^{n}\sum_{i=1}^{n}WR(y_k,\omega_i,\omega_j)=\sum_{i=1}^{n}\sum_{i=1}^{n}\overline{\prod_{k=1}^{m}\varphi_{i,j}^{k}}\bullet\overline{\varphi_{i,j}^{class}} \tag{11}$$

DC (L, Ω) **measures the features group relevance and does not consider the feature exclusiveness. To take into account this exclusiveness we will define the equivalent of a "relevance gain" related to a feature compared to a feature subset, noted with WR for each pair of objects, and with DC for the aggregation on all the pairs.**

Proposition n° 4: Contextual criterion: Strong relevance DCG

The feature y_k relevance compared to a feature subset L **on a sample of data pairs $\omega_i\,\omega_j$ is given by the following equation:**

$$on(\Omega\times\Omega)\quad DCG(y_k,L,\Omega)=\sum_{i=1}^{n}\sum_{i=1}^{n}\overline{\varphi_{i,j}^{class}}\bullet\overline{\varphi_{i,j}^{k}}\bullet\prod_{l=1}^{m}\varphi_{i,j}^{l} \tag{12}$$

To take into account the y_k feature discrimination represented by $A=\overline{\varphi_{i,j}^{k}}$ **and also of the previously selected features (y_1… y_m) represented by** $B=\overline{\prod_{l=1}^{m}\varphi_{i,j}^{l}}$ **, we build a Boolean function. It will be equal to** 1 **only if there is only y_k able to discriminate a concept. For that we introduce the following product** $AB=\overline{\varphi_{i,j}^{k}}\bullet\prod_{l=1}^{m}\varphi_{i,j}^{l}$ **. Thus this product will be equal to** 1 **when only y_k is relevant compared to a set of features (y_1… y_m) for a data pair $\omega_i\,\omega_j$.**

If a feature y_l is discriminating the product becomes zero.

AB **associated with** $\overline{\varphi_{i,j}^{class}}$ **corresponds then to strong relevance SR of the feature y_k on the data pair $\omega_i\,\omega_j$.**

The aggregation of this expression on the whole pairs will define the strong relevance, noted with DCG: *discriminating capacity gain* **(see equation** 13**).**

$$on(\omega_i,\omega_j)\quad SR(y_k,L,\omega_i,\omega_j)=\overline{\varphi_{i,j}^{class}}\bullet\overline{\varphi_{i,j}^{k}}\bullet\prod_{l=1}^{m}\varphi_{i,j}^{l}$$

$$on(\Omega\times\Omega)\quad DCG(y_k,L,\Omega)=\sum_{i=1}^{n}\sum_{j=1}^{n}SR(y_k,L,\omega_i,\omega_j)$$

$$on(\Omega\times\Omega)\quad DCG(y_k,L,\Omega)=\sum_{i=1}^{n}\sum_{j=1}^{n}\overline{\varphi_{i,j}^{class}}\bullet\overline{\varphi_{i,j}^{k}}\bullet\prod_{l=1}^{m}\varphi_{i,j}^{l} \tag{13}$$

The DCG (discriminate capacity gain) is a selection criterion for strongly relevant features. It is different from zero when there is not any feature or combination of descriptive features, except for the studied feature being able to discriminate the class.

IV. FEATURE SELECTION ALGORITHM

In order to select the optimal subset we propose an algorithm related to the greedy type algorithms resulting from [1], [**4**], [18]**. So the research will be a sequential bidirectional generation, i.e. a core of features is composed from an empty set S which is built gradually until obtaining a subset having the same degree of relevance as the starting subset. The feature subset is progressively computed and revaluated at every feature addition.**

The research strategy adopted is not a complete generation but a heuristic generation. Indeed, the complete combinative

generation impose problem with huge datasets and thus with ECD process. The stopping criterion is defined when the

E = E - { discriminated pairs }

TABLE IV
SELECTED FEATURE ON SYNTHETIC DATA

sets	Relevant Features	STRASS	ReliefF	CFS Ranker (GI)	Wrapper(C4.5)	Consistancy (GA)	GainRatio
LED7	y_1 à y_5	y_1 à y_5	y_1 à y_7	y_1 à y_7	not enough instances	y_1 à y_5	y_1 à y_7
LED24	y_1 à y_5	y_1 à y_5	y_1 à y_7	y_1 à y_7	y_1,y_2,y_3,y_4,y_5,y_7, y_{16},y_{17},y_{18}	y_3, y_4, y_5, y_9, y_{10},y_{11},y_{15},y_{16} ,y_{17},y_{18},y_{19},y_{20},y_{22},y_{23}	y_1 à y_7
Parity	y_1 à y_3	y_1 à y_3	y_1 à y_3	y_{10}, y_8, y_5	y_1 à y_3	y_1 à y_4 et y_6	y_{10}, y_8, y_5
Parity 2	y_1 à y_3	y_1 à y_3	y_1 à y_3, y_{11}, y_{12}	y_{10}, y_8, y_5	y_1 à y_3 et y_6	y_1 à y_3 et y_5,	y_{10}, y_8, y_5
Corral	y_1 à y_4	y_1 à y_4	y_1 à y_4 et y_6	y_1 à y_4 et y_6	y_1 à y_4	y_1 à y_4	y_3,y_2,y_1,y_4, y_6
Bool	y_1 à y_6	y_1 à y_6	y_1 à y_6, y_7 , y_{12}	y_3 à y_6	y_1 à y_6 et y_{10}	y_1 à y_6	y_4,y_3,y_6,y_5
F1	y_3	y_3	y_3	y_3	y_3	y_3	y_3
F2	y_1 , y_3	y_1, y_3	y_1 à y_3	y_1	y_2 , y_3	y1	y_1, y_2
F3	y1, y3, y4	y1, y3, y4	all	y2, y3, y4	y1, y3, y4	y2, y3, y4	y1 à y4
F4	y1, y2,y9	y1, y5, y9	y1, y2,y9	y9	y1, y2,y9	y1, y2, y9	y1, y2,y9

selected features subset S discriminating capacity is equal to the discriminating capacity of the initial features set.

A. ***STRASS*** *Strong Relevant Algorithm for Subset Selection*

Initialization
E the whole set of data pairs $\Omega x\Omega$.
$L=\{y_1,y_2 \ldots y_r\}$ **a set of features to be treated**
$S = \emptyset$ **selected features**
$DCTot = DCG(S)$
$DCGmax = 0$

1. Selection of strongly relevant features "predominant"
For each feature $y_k \in L$ **do**
scan the examples space Ω
If $DCG(y_k, L\text{-}y_k) \neq 0$
Then $S=S+y_k$
$L=L-y_k$
E = E - {discriminated pairs}

2. Selection of the remaining features "weak relevant"
while $DC(S)<DCTot$ **do**
For each feature y_k L **do**
scan the examples space Ω

If $DC(y_k+S)>DCGmax$
Then $DCmax = DC(y_k+S)$
$y_{kmax}=y_k$
$S=S+y_{k\ max}$
$L=L-y_{k\ max}$

3. Suppression of the redundant features
For each feature yk $\in$ **S do**
If $DC(y_k, S\text{-}y_k)=0$
$S=S-y_k$
Return S

The complexity of the algorithm is θ**(m, c)**
m **is the number of characteristics (cardinal of** L**) and** C **is the number of pairs whose objects are not discriminated (cardinal of E).**

According to the initialization the algorithm breaks up into three stages:
Stage 1**- "Selection of strongly relevant features"**
Or predominant features impossible to avoid because they are the only ones to discriminate the classes.
Stage 2 **- " Selection of the remaining features"**
Or weakly relevant features which have the largest discriminating capacity.
Stage 3 **- " Suppression of redundant features"**
We remove the features that become redundant compared to the subset of the selected features when adding a new feature. This stage guarantees that there is no total or partial redundancy in the selected set of features.

B. Evaluation of the algorithm

1) Implementation

Our algorithm was established under MATLAB **7.5. For the filtering algorithms and classifiers we used the existing tools in WEKA machine learning platform** [23]**. The experiences were run using WEKA´s default values.**

TABLE V
C4.5 PRECISION WITH AND WITHOUT FILTERING (+ PERFORMANCE, - DEGRADATION)

datasets	C4.5	STRASS	ReliefF	CFS	Wrapper (C4.5)	Consistancy (GA)	Gain ratio
LED7	100	100	100	100	100	100	100
LED24	100	100	100	100	100	100	100
PARITY	99.4	100+	100+	49-	100+	100+	49-
PARITY2	99.4	100+	100+	49-	100+	100+	49-
CORRAL	81.25	81.25	81.25	81.25	81.25	81.25	81.25
BOOL	98.35	100+	100+	70.1-	100+	100+	70.1-
F1	100	100	100	100	100	100	100
F2	91.64	91.64	91.64	75-	78.14-	75-	75-
F3	96.42	97.02+	97.02+	97.02+	97.02+	97.02+	97.02+
F4	96.1	96.62+	93-	83-	96.62+	96.62+	96.62+

2) Direct evaluation

We evaluate STRASS on ten artificial benchmarks which we know a priori the relevant features [10]. We privileged these datasets because there features interact, and will be therefore able to test our algorithm. LED Display Domain [3], BOOL dataset [26], Parity, Parity +2, Coral and Argawal's fonctions [22]. During the experimentation, we used for each set successively 50 sets of 500 instances. We then compared our results to ReliefF, CFS, Wrapper (C4.5)[1], ConsistancySubsetEval (GA)[2] and GainRatio[3] [23] (see Table IV). However, In order to make comparable the results obtained by the five algorithms, we turn ReliefF on the whole objects and not on a certain number of instances randomly selected. Table IV shows feature selected by each algorithm. The best results are shadowed.

We can see that for LED Display Domain (Led 7, Led 24) all the algorithms fail to detect the relevance and the sufficiency of the five segments except STRASS.

For the Parity dataset, STRASS, Wrapper(C4.5) and ReliefF establish that the features y_1, y_2, y_3 are relevant whereas the others are useless.

For Parity2 dataset all the algorithms fail to remove the redundant feature y_6 and the doubled features y_{11} and y_{12}, our algorithm is the only one to have removed them.

STRASS, Wrapper(C4.5) and ConsistancySubsetEval have detected the redundancy of the feature y_6 in Corral dataset. This feature being correlated to 75% with the feature class is considered to be relevant by CFS and ReliefF because the two algorithms can not evaluate the relevance of a feature compared to the combination of the other features. In the case of Argawal's functions Wrapper(C4.5) and STRASS give the relevant features for three functions. Thus make our algorithm the most powerful compare to the whole datasets. In fact STRASS detects the relevant features as well the redundant ones.

The selected features for the ten datasets were tested using three classifiers: decision tree (C4.5), nearest neighbours (IBk) and Naïve Bays (NB). The error rate was estimated by cross validation. The classifiers' precision are given in the table V, VI,VII. We can note that the classifier performances are improved on the whole data bases after filtering data by STRASS, ConsistancySubsetEval, Wrapper(C4.5) and ReliefF. Whereas filtering by CFS and GainRatio decreases the C4.5 precision. This is explained by the presence of strongly correlated features. Indeed

TABLE VI
IBK PRECISION WITH AND WITHOUT FILTERING (+ PERFORMANCE, - DEGRADATION)

datasets	IBK	STRASS	ReliefF	CFS	Wrapper (C4.5)	Consistancy (GA)	Gain ratio
LED7	100	100	100	100	100	100	100
LED24	82	100+	100+	100+	97.37+	51.49-	100+
PARITY	87.8	100+	100+	50-	100+	100+	50-
PARITY2	99.4	100+	100+	50-	100+	100+	50-
CORRAL	65.6	84.3+	71.8+	71.8+	84.3+	84.3+	71.8+
BOOL	92.2	100+	98.8+	69.18-	98.7+	100+	69.18-
F1	85.8	100+	100+	100+	100+	100+	100+
F2	83.7	97.7+	96.5+	73.5-	78.12-	73.5-	72.7-
F3	96.8	96.8	96.8	96.5	96.8	96.5	96.7
F4	89.4	94.5+	97.8+	79.6-	97.8+	97.8+	97.8+

GainRatio treats each feature independently of the others for discriminating objects and CFS detects the pair-wise feature-feature correlation consequently the algorithm can not

[1] Wrapper Method uses C4.5 and Genetic search
[2] Consistancy based Algorithm with Genetic search
[3] GainRatio evaluation criteria algorithm with ranker searsh

TABLE VII
NB PRECISION WITH AND WITHOUT FILTERING (+ PERFORMANCE, - DEGRADATION)

datasets	IBK	STRASS	ReliefF	CFS	Wrapper (C4.5)	Consistancy (GA)	Gain ratio
LED7	100	100	100	100	100	100	100
LED24	82	100+	100+	100+	97.37+	51.49-	100+
PARITY	87.8	100+	100+	50-	100+	100+	50-
PARITY2	99.4	100+	100+	50-	100+	100+	50-
CORRAL	65.6	84.3+	71.8+	71.8+	84.3+	84.3+	71.8+
BOOL	92.2	100+	98.8+	69.18-	98.7+	100+	69.18-
F1	85.8	100+	100+	100+	100+	100+	100+
F2	83.7	97.7+	96.5+	73.5-	78.12-	73.5-	72.7-
F3	96.8	96.8	96.8	96.5	96.8	96.5	96.7
F4	89.4	94.5+	97.8+	79.6-	97.8+	97.8+	97.8+

TABLE VIII
SUMMARY OF PERFORMANCE AND DEGRADATION

		STRASS	ReliefF	CFS	Wrapper (C4.5)	Consistancy (GA)	Gain ratio
C4.5	Perf	5	4	1	5	5	2
	Deg	0	1	5	1	1	4
IBk	Perf	8	8	3	7	6	4
	Degr	0	0	5	1	2	4
NB	Perf	2	2	4	3	2	4
	Degr	3	2	2	3	4	1
Average		5+\1-	4.6+\1-	2.6+\4-	5+\1.66-	4.3+\2.33-	3.33+\3-

identify the great interactions between features (low relevance). In fact Hall [66] specifies that CFS gives good results when there is a moderated interaction between the features.

Table VIII resumes the performances and degradations in the tree induction algorithms after filtering the data, performances and degradation superior to and less than 0.5% are considered, the best results are shadowed. STRASS on average enhances or maintains predictive accuracy better than the other algorithms.

V. CONCLUSION

We present a theoretical analysis of the *discriminating capacity gain*. The suggested contextual criterion allows detecting as well the strongly relevant features as the weakly relevant; the contextual criterion allows us too detecting the relevance or redundancy of a feature compared to a subset of features.

These criteria established in a greedy type algorithm give satisfactory performances for the selection of a minimal set of relevant features. STRASS on average selects the smallest number of features and performs better then the other filtering algorithms. However, in high-dimensional data which often contains a large portion of irrelevant and/or redundant features, our pair-wise criteria pose problem. Thus why this study constitutes only a beginning, and requires some adaptations for huge dataset. In order to not fall into a combinative problem from all the existing pairs of objects in a problem, we have simplify the criteria and express them in a contingency form, thanks to Marchotorchino [14] for his passage formulas. In some future works we will give our STRASS contingency study in real and huge datasets.

REFERENCES

[1] H.ALMUALLIM, T. G. DIETTERICH, "Learning with Many Irrelevant Features", *Proc. of the Ninth National Conference on Artificial Intelligence*, 1991, pp. 547-552.

[2] A. L. BLUM and P. LANGLEY, "Selection of relevant features and examples in machine learning", *Artificial Intelligence* 97(1-2), 1997. pp. 245-271.

[3] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., and STONE, C. J. "Classification and Regression Trees". *Belmont, CA: Wadsworth*, 1984.

[4] V. CHVATAL, "A greedy heuristic for the set-covering problem", *Mathematics of operations research*, Vol. 4, no. 3, 1979, pp 233 - 235

[5] A. Dhagat and L. Hellerstein, "PAC learning with Irrelevant Attributes" *Proceedings of the IEEE Symposium on Foundations of Computer Science* , 1994, pp. 64-74.

[6] G. H.JOHN, R.KOHAVI, and K. PFLEGER, "Irrelevant Features and the Subset Selection Problem" *Proceedings of the Eleventh International Conference on Machine Learning* . New Brunswick, NJ: Morgan Kaufmann, 1994. pp. 121-129.

[7] K. KIRA, L. A. RENDELL, "A Practical Approach to Feature Selection", *in Proc. of the Ninth International Workshop*, ML, 1992, pp. 249-255.

[8] I. KONONENKO, "Estimating Attributes: Analysis and Extensions of Relief", *in Proc. of the Seventh European Conference on Machine Learning, Italie*, 1994, pp. 171-182.

[9] I. KONONENKO, S.E. HONG, "Attribute selection for modelling ", *Future Generation Computer Systems*, 1997, 13, pp 181 – 195.

[10] P. LANGLEY, S. SAGE, "Induction of selective bayesian classifiers", *Proc of the Tenth Conference on Uncertainty in Artificial Intelligence, Seattle,* 1994 pp 399 – 406.

[11] P. LANGLEY, "Selection of relevant features in machine learning", *Proc of the AAAI, Fall Symposium on relevance*, New Orleans 1994 pp 399 – 406 .

[12] P. LANGLEY and S. SAGE, "Scaling to Domains with Irrelevant features, Computational learning theory and natural learning systems", Vol. 4, MA: MIT Press, Cambridge, 1997, pp 17-29.

[13] H. LIU and H. MOTODA, "*Feature Selection for Knowledge Discovery and Data Mining*" Kluwer Academic Publishers, 1998.

[14] F. MARCOTORCHINO, " Utilisation des comparaisons par paires en statistiques des contingences " *IBM Scientific Center French* ,1° part Etude n° F-069 and 2 part n° F-071 1984

[15] D. MICHAUT, "Filtering and Variable Selection in Learning Processes", *PHD,* University of Franche Comté, December, 1999.

[16] D. MICHAUT, P. BAPTISTE, " Selection of a Relevant Feature Subset for Induction Tasks", proc of the 11th of ISMIS 1999, Warshaw, Springer Verlag Poland, 1999.

[17] A. N. MUCCIARDI, GOSE E.E., " A Comparison of Seven Techniques for Choosing Subsets of Pattern Recognition Properties", *IEEE Transactions on Computers*, Vol. C-20, 1971, pp. 1023-1031.

[18] R.VIGNES, J. LEBBE, "Sélection d'un sous ensemble de descripteurs maximalement discriminant dans une base de connaissances", in : *3èmes journées "Symbolique-Numérique",* IPMU-92, Paris, (pp. 219-232), 1992.

[19] M. Hall, "Correlation-based feature selection of discrete and numeric class machine learning". *In Proceedings of the International Conference on Machine Learning*, pages 359-366, San Francisco, CA. Morgan Kaufmann Publishers 2000.

[20] H. Peng, F. Long, Chris Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp 1226-1238, Aug., 2005.

[21] H. Liu et L. Yu. "Toward Integrating Feature Selection Algorithms for Classification and Clustering", *IEEE Trans on Knowledge and Data Engineering*, VOL. 17, NO. 4, APRIL 2005.

[22] R. Agrawal, S. Ghosh T. Imielinski B. Iyer and A. Swami, "An interval classifier for database mining applications". *In VLDB Conference*, Aug 1992.

[23] I. H. Witten and E. Frank, "Data Mining—Practical Machine Learning Tools and Techniques with JAVA Implementations", Morgan Kaufmann, San Francisco, CA, 2000.

[24] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", *The Journal of Machine Learning Research*, 3, 3/1/2003.

[25] M. Hall, "Correlation-based feature selection for machine learning". *PhD thesis*, Department of Computer Science, University of Waikato, Hamilton, New Zealand. 1998.

[26] L.Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy". *Journal of Machine Learning Research 5* 1205–1224. 2004.

[27] P. Smyth, R.M. Goodman and C. Higgins, "A hybrid rule-based Bayesian classifier" *in Proc, ECAI*, Stockholm, pp 610-615, 1990.