# A Practical Approach for Knowledge-Driven Constructive Induction

Huei Diana Lee[1], Maria Carolina Monard/ILTC[2], and José Augusto Baranauskas[2]

[1] UNIOESTE – State University of West Paraná
Department of Computer Science
P.O. Box 961, 85857-970 - Foz do Iguaçu, PR, Brazil
huei@dcc.unioeste-foz.br
[2] USP – University of São Paulo
Institute of Mathematics and Computer Sciences
Department of Computer Science and Statistics
P.O. Box 668, 13560-970 - São Carlos, SP, Brazil
{mcmonard, jaugusto}@icmc.sc.usp.br

**Abstract** Learning problems can be difficult for many reasons, one of them is inadequate representation space or description language. Features can be considered as a representational language; when this language contains more features than necessary, subset selection helps simplify the language. On the other hand, when this language is not sufficient to describe the problem, Feature Construction helps enrich the language. Feature Construction, also known as Constructive Induction, aims to discover missing information about the relationships between features and augments the space of features by inferring additional features. Thus, feature selection reduces the feature space while Feature Construction expands the feature space. In both cases, the main idea is to improve the representation space before searching for concept description, in order to improve the overall prediction accuracy of the generated concept description. This work is concerned with knowledge-driven Constructive Induction, which uses domain knowledge provided by the expert to search for a better representational space. The objective of this work is to propose an approach for practical Feature Construction when this is done with the aid of the user or the expert. We describe a series of experiments performed on four real world datasets using inducers $\mathcal{C}4.5$rules and $\mathcal{CN}2$.

**Keywords:** *Constructive Induction, Machine Learning*

## 1 Introduction

Conventional inductive-learning algorithms rely on existing, generally user provided, data to build their descriptions. Inadequate representation space or description language as well as errors in training instances can make learning problems difficult.

Features can be considered inadequate for the learning task when they are weakly or indirectly relevant, conditionally relevant or inappropriately measured (Langley, 1996; Blum and Langley, 1997). If the provided features for describing the training instances are inadequate, the learning algorithms are likely to create excessively complex and inaccurate descriptions (Bloedorn and Michalski, 1998).

However, these individually inadequate features can sometimes be combined conveniently, generating new features which can turn out to be highly representative to the description of a concept. The process of constructing new features is called Feature Construction or Constructive Induction (Michalski, 1978).

The objective of this work is to propose an approach for practical Feature Construction when this is done with the aid of the user/expert. We also describe a series of experiments performed using our approach on real world datasets using the $\mathcal{C}4.5$rules and $\mathcal{CN}2$ inducers. The reported results include, for each experiment, a description of the new features constructed, error rates, features selected by each inducer and others.

This work is organized as follows: Sect. 2 gives some background about Feature Construction. Section 3 describes the proposed approach for Constructive Induction while Sect. 4 describes the inducers and datasets used in the experiments performed using our approach. Section 5 shows the results obtained from these experiments and Sect. 6 presents some considerations about results. Finally, Section 7 gives some conclusions.

## 2 Constructive Induction

Feature Construction, also known as Constructive Induction — CI — is the process of combining primitive features (from the original dataset) producing new features possibly relevant to a concept description. In other words, CI is the application of constructive operators, *i.e.* operators used to compound features from the existing ones, resulting in the definition of one or more new features.

It is important to notice that, unlikely Feature Subset Selection where only selected features are shown to the inductive algorithm, thus decreasing feature search space (Kohavi and Sommerfield, 1995; Kohavi and John, 1997; Baranauskas and Monard, 1999; Baranauskas et al., 1999), Constructive Induction increases the feature search space.

Constructive Induction requires a decision about which constructive operators should be used as well as which primitive features should be combined using the operators.

Another important observation is that, in general, the Constructive Induction process is infeasible since the number of features which can be constructed is a combinatorial function of the number of existing features multiplied by the number of possible operators. Consequently, CI is feasible only when articulated with heuristcs that may reduce the number of possible features and the number of constructive operators which are going to be used to construct new features.

Constructive Induction methods can be grouped according to the information used to search for the best representation space as follows (Bloedorn and Michalski, 1998; Wnek and Michalski, 1994; Wnek and Michalski, 1993):

1. data-driven constructive induction, based on analysis of the training data;
2. hypothesis-driven constructive induction, based on analysis of inductive hypothesis. In this approach, useful concepts in the rules can be extracted and used to define new features;
3. knowledge-driven constructive induction, based on domain knowledge provided by an expert;
4. multi-strategy constructive induction, based on two or more of the other methods.

The Constructive Induction process can be guided and controlled by the user/expert or can be automatically conducted by the learning system. In this work, we focus on Constructive Induction guided by user/expert using thus the knowledge-driven approach.

## 3 The Proposed Approach

We assume that the original dataset $\mathsf{O}$ is composed by $m$ features $\{X_0, X_1, \ldots, X_{m-1}\}$ and that $k$ new features $\{f_1, f_2, \ldots, f_k\}$ are constructed with the help of the user/expert based on the original $\{X_0, X_2, \ldots, X_{m-1}\}$ features. The proposed approach can be divided into the following three steps:

**First step:** Considering each one of the $k$ new features $\{f_1, f_2, \ldots, f_k\}$ the expert may suggest, the original dataset $\mathsf{O}$ is then augmented with each one of them, thus given a set of $k$ new datasets labeled as $\{\mathsf{O}+f_1, \mathsf{O}+f_2, \ldots, \mathsf{O}+f_k\}$ containing $m+1$ features each one. Therefore, each augmented dataset $\mathsf{O}+f_i$ is composed by the original features $\{X_0, X_1, \ldots, X_{m-1}\}$ and the new feature $f_i$ $(1 \leq i \leq k)$ suggested by the expert *i.e.* $\mathsf{O}+f_i = \{X_0, X_1, \ldots, X_{m-1}, f_i\}$.

**Second step:** Each new dataset from the set $\{\mathsf{O}+f_1, \mathsf{O}+f_2, \ldots, \mathsf{O}+f_k\}$ is then given to one (or more) inducer that generates a classifier. The idea is that if the new feature $f_i$ is not present in the extracted classifier, this is an indication that $f_i$ is not essential to the concept being learned by the inducer (Michalski and Kaufman, 1998). In this case, dataset $\mathsf{O}+f_i$ is not considered for the next step. On the other hand, if the feature $f_i$ appears in the classifier, then the augmented dataset $\mathsf{O}+f_i$ is used in the next step. Consequently, after this step a subset of augmented datasets that have been generated in the first step and fulfills the second condition will be considered in the third step.

**Third step:** In this step, the error estimation is performed as following described. Despite any error estimate can be used, we suggest the $K$-fold stratified cross-validation (SCV) since it is a very well known measure for error estimation in the research community and it allows assuming a normal distribution for error comparison as shown below.

First of all, a $K$-fold stratified cross-validation is performed once in the original dataset $\mathsf{O}$ to estimate its error $E_{\mathsf{O}}$. After that, a $K$-fold stratified cross-validation is performed in each augmented dataset that fulfills the second step test. Labeling the error for dataset $\mathsf{O}+f_i$ as $E_i$, then only augmented datasets that have an error difference $\mathsf{ad}(E_i - E_{\mathsf{O}}) < 0$ are selected for further investigation, since this suggests that the new feature, besides appearing in the classifier, has also increased the classifier accuracy. The difference between $E_i$ and $E_{\mathsf{O}}$ in standard deviations ($\mathsf{ad}$) is given by (1), where $\mathsf{mean}$ is the mean error of the $K$-folds stratified cross-validation and $\mathsf{sd}$ is its corresponding standard deviation.

$$\mathsf{ad}(E_i - E_{\mathsf{O}}) = \frac{\mathsf{mean}(E_i) - \mathsf{mean}(E_{\mathsf{O}})}{\sqrt{\frac{\mathsf{sd}(E_i)^2 + \mathsf{sd}(E_{\mathsf{O}})^2}{2}}} \tag{1}$$

Considering this whole process as a general methodology for applying Constructive Induction as shown in Fig. 1, the ideal situation would be when the primitive features, used to construct the new ones, are not selected during the second step by the inducer. However, this may not be the case. A possible reason for this would be that the constructed features do not capture perfectly the information embedded in each individual feature for the specific inducer or is equivalent in predictive power to (some of) the original ones. Another reason for this would be that the datasets used have already been worked out, so that the original features are, on its own, the most relevant ones.

## 4 Experimental Setup

Using the proposed approach, a series of experiments were performed, in order to evaluate the effectiveness of the new constructed features, using inducers and datasets described in the next sections (Lee and Monard, 2000). It is important to observe that the original dataset has not been preprocessed in any way, for example by removing or replacing missing values or transforming nominal to numerical features. Furthermore, each individual inducer was run with default options setting for all parameters, *i.e.* no attempt was made to tune any inducer.

### 4.1 Inducers

Two inducers, $\mathcal{C}4.5\mathsf{rules}$ and $\mathcal{CN}2$, present in the $\mathcal{MLC}$++ library (Kohavi et al., 1996), have been used in this work. These inducers are well known in the Machine Learning community and they belong to the eager learning approach where the algorithms greedily compile the training data into an intentional concept description (Aha, 1997).

$\mathcal{C}4.5\mathsf{rules}$ (Quinlan, 1993) examines the original decision tree produced by $\mathcal{C}4.5$ and derives from it a set of rules of the form $L \rightarrow R$. The left-hand side $L$ is a conjunction of feature-based tests and the right-hand side is a class. One of the classes is also designated as the default class.

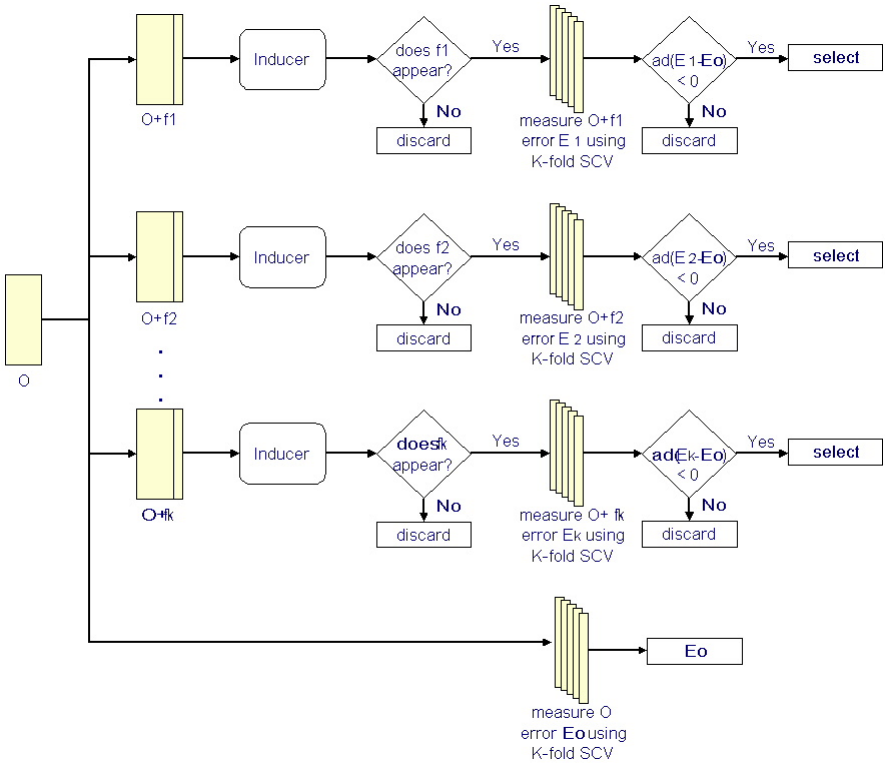


**Figure1.** The Proposed Approach

It is important to note that $\mathcal{C}$4.5rules does not simply rewrite the tree to a collection of rules. In fact, it generalizes the rules by deleting superfluous conditions — irrelevant conditions that do not affect the conclusion — without affecting its accuracy, leaving the more appealing rules.

$\mathcal{CN}2$ (Clark and Niblett, 1987; Clark and Niblett, 1989; Clark and Boswell, 1991) is a Machine Learning algorithm that directly induces '*if* <complex> *then* <class>' rules in domains where there might be noise. Each <complex> is a disjunction of conjunctions. For the experiments we have used the unordered rules $\mathcal{CN}2$ algorithm. $\mathcal{C}$4.5rules as well as $\mathcal{CN}2$ can handle missing (or unknown) values.

### 4.2 Datasets

Experiments were conducted on four real world domains. Datasets pima, cmc and hepatitis are from the UCI Irvine Repository (Blake et al., 1998). The smoke dataset was obtained from the URL http://lib.stat.cmu.edu/datasets/csb/. The criterion used to choose these four datasets is related to our user/expert domain since we are interested in his/her assistance to construct new features.

Dataset pima is a subset of a larger database maintained by the National Institute of Diabetes and Digestive and Kidney Diseases. All patients are females at least 21 years old of Pima Indian heritage living near Phoenix, Arizona, USA. The problem is to predict whether a patient would test positive for diabetes according to World Health Organization (WHO) criteria — if the 2-hour post-load plasma glucose is at least 200 mg/dl at any survey examination or if found during routine medical care — given a number of physiological measurements and medical test results.

Dataset cmc is composed by a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of the interview. The problem is to predict the current contraceptive method choice (no use, long-term methods or short-term methods) of a woman based on her demographic and socioeconomic characteristics.

Smoke is a survey dataset (Bull, 1994) concerned with the problem of predicting attitude toward restrictions on smoking in the workplace (prohibited, restricted or unrestricted) based on by-law-related, smoking-related and sociodemographic covariates.

Dataset hepatitis is for predicting life expectation of patients with hepatitis.

Table 1 summarizes the datasets used in this work. It shows, for each dataset, the number of instances (#Instances), number and percentage of duplicate (appearing more than once) or conflicting (same feature values but different class) instances, number of features (#Features) continuous and nominal, class distribution, the majority error and if the dataset have at least one missing value. Datasets are presented in ascending order of the number of features.

### 4.3 Experiments

The performed experiments follow the three basic steps of our approach. In the first step, after analyzing each dataset, the user/expert suggested two new features for datasets pima, cmc and smoke and just one new feature for dataset hepatitis as Table 2 shows.

In the second step, $\mathcal{C}4.5\mathsf{rules}$ and $\mathcal{CN}2$ were run once using as training set all instances in each of these seven new datasets. In our experiments, the rules induced by $\mathcal{C}4.5\mathsf{rules}$ and $\mathcal{CN}2$ used features $f_1$ and $f_2$ for all datasets (for dataset hepatitis, the only one feature constructed, $f_1$, also appeared in the extracted classifier). Therefore, no augmented dataset was discarded in this step.

Next, in the third step, the two inducers $\mathcal{C}4.5\mathsf{rules}$ and $\mathcal{CN}2$ were run on each original dataset as well as on the seven new datasets and the error rate was measured using 10-fold stratified cross-validation. After this, we selected for further investigation only the datasets whose accuracies improved comparing to the original datasets accuracies.

**Table1.** Datasets Summary Descriptions

| Dataset | # Instances | Duplicate or conflicting | # Features (cont., nom.) | Class | Class % | Majority Error | Missing Values |
|---|---|---|---|---|---|---|---|
| pima | 769 | 1 (0.13%) | 8 (8, 0) | 0 | 65.02% | 34.98% | N |
| | | | | 1 | 34.98% | on value 0 | |
| cmc | 1473 | 115 (7.81%) | 9 (2, 7) | 1 | 42.70% | 57.30% | N |
| | | | | 2 | 22.61% | on value 1 | |
| | | | | 3 | 34.69% | | |
| smoke | 2855 | 29 (1.02%) | 13 (2, 11) | 0 | 5.29% | 30.47% | N |
| | | | | 1 | 25.18% | on value 2 | |
| | | | | 2 | 69.53% | | |
| hepatitis | 155 | 0 (0%) | 19 (6, 13) | die | 20.65% | 20.65% | Y |
| | | | | live | 79.35% | on value live | |

**Table2.** Original Datasets Augmented with Constructed Features

| Original Dataset | Augmented Datasets | |
|---|---|---|
| pima | pima+$f_1$ | pima+$f_2$ |
| cmc | cmc+$f_1$ | cmc+$f_2$ |
| smoke | smoke+$f_1$ | smoke+$f_2$ |
| hepatitis | hepatitis+$f_1$ | — |

## 5 Experimental Results

In this section, experimental results obtained are presented in one table for each dataset. It shows, for each original/augmented dataset and inducer, the dimensionality of the dataset given to the inducer (Total #F); the individual features used by the inducer to represent the concept[1]; the number of features selected by the classifier (#F) and the proportion of the selected features (%F). The $\mathcal{C}4.5$rules inducer is represented as $\mathcal{C}4.5$r.

Also, the table shows the error rate (Error) of each inducer (mean and standard deviation) using 10-fold stratified cross-validation, including the difference in standard deviations (ad) — computed by Equation 1 — between the classifier extracted using the derived datasets and the one extracted using original dataset. Thus, if this value is positive (negative) the classifier constructed using the original (derived) dataset outperforms the one constructed using the derived (original) dataset. However, for one classifier to outperform the other at 95% confidence level this value should be greater than 2, or less than $-2$, respectively. The results for datasets pima, cmc, smoke and hepatitis are presented in Tables 3, 4, 5 and 6, respectively.

[1] Note that features indicated with '$\circ$' refer to original dataset features while the ones indicated with '$\bullet$' refer to new constructed features.

### 5.1 Pima and Derived Datasets

Two new features were constructed for this dataset with the help of the expert:

- $f_1$: this features verifies if glucose and diastolic blood preassure are out of normal levels. It combines two primitive features: *plasma* (feature #1) and *diastolic* (feature #2).
- $f_2$: this feature verifies if glucose is out of the normal level and if 2-hour serum insulin is not present. It combines two primitive features: *plasma* (feature #1) and *two* (feature #4).

**Table3.** Pima Before and after Constructive Induction

| Feature Number | (pima, $\mathcal{C}$4.5r) | (pima, $\mathcal{CN}2$) | (pima+$f_1$, $\mathcal{C}$4.5r) | (pima+$f_1$, $\mathcal{CN}2$) | (pima+$f_2$, $\mathcal{C}$4.5r) | (pima+$f_2$, $\mathcal{CN}2$) |
|---|---|---|---|---|---|---|
| #0 | | ○ | | | | ○ |
| #1 | ○ | ○ | ○ | ○ | ○ | ○ |
| #2 | ○ | ○ | ○ | ○ | ○ | ○ |
| #3 | | ○ | | ○ | | ○ |
| #4 | | ○ | | ○ | ○ | ○ |
| #5 | ○ | ○ | ○ | ○ | ○ | ○ |
| #6 | ○ | ○ | ○ | ○ | ○ | ○ |
| #7 | ○ | ○ | ○ | ○ | ○ | ○ |
| #8 ($f_1$) | | | • | | | |
| #9 ($f_2$) | | | | | • | • |
| Total #F | 8 | 8 | 9 | 9 | 9 | 9 |
| #F | 5 | 8 | 6 | 7 | 7 | 9 |
| %F | 62.50% | 100.00% | 66.67% | 77.78% | 77.78% | 100.00% |
| Error | 26.00±1.03 | 25.38±1.38 | 25.61±1.12 | 25.90±1.15 | 26.52±1.14 | 25.77±1.33 |
| ad | | | −0.36 | 0.41 | 0.48 | 0.29 |

Looking at Table 3, it can be observed that feature $f_1$ is nonessential for $\mathcal{CN}2$ and there is only one slight improvement using pima+$f_1$ with $\mathcal{C}$4.5rules. Also, it is possible to note that in all cases where the new feature was selected, the original ones have been selected too. Besides occuring for dataset pima, this also occurs for datasets cmc and smoke but not for hepatitis. A small degradation in performance can be seen for pima+$f_1$ with $\mathcal{CN}2$ although the new feature has been not selected.

### 5.2 Cmc and Derived Datasets

Two new features were constructed for this dataset with the help of the user:

- $f_1$: this feature shows how equal or different the educational level of wife and husband are. It combines two primitive features: *wedu* (feature #1) and *hedu* (feature #2).

– $f_2$: this feature shows if the wife has a standard of living compatible with her educational level. It combines two primitive features: *wedu* (feature #1) and *stdliv* (feature #7).

**Table4.** Cmc Before and after Constructive Induction

| Feature Number | (cmc, $\mathcal{C}4.5\mathsf{r}$) | (cmc, $\mathcal{CN}2$) | (cmc+$f_1$, $\mathcal{C}4.5\mathsf{r}$) | (cmc+$f_1$, $\mathcal{CN}2$) | (cmc+$f_2$, $\mathcal{C}4.5\mathsf{r}$) | (cmc+$f_2$, $\mathcal{CN}2$) |
|---|---|---|---|---|---|---|
| #0 | ○ | ○ | ○ | ○ | ○ | ○ |
| #1 | ○ | ○ | ○ | ○ | ○ | ○ |
| #2 | ○ | ○ | ○ | ○ | ○ | ○ |
| #3 | ○ | ○ | ○ | ○ | ○ | ○ |
| #4 | ○ | ○ | ○ | ○ | ○ | ○ |
| #5 | ○ | ○ | ○ | ○ | ○ | ○ |
| #6 | ○ | ○ | ○ | ○ | ○ | ○ |
| #7 | ○ | ○ | ○ | ○ | ○ | ○ |
| #8 | ○ | ○ | ○ | ○ | ○ | ○ |
| #9 ($f_1$) | | | | • | | |
| #10 ($f_2$) | | | | | • | • |
| Total #F | 9 | 9 | 10 | 10 | 10 | 10 |
| #F | 9 | 9 | 9 | 10 | 10 | 10 |
| %F | 100.00% | 100.00% | 90.00% | 100.00% | 100.00% | 100.00% |
| Error | 45.90±1.38 | 49.64±1.01 | 47.87±1.54 | 49.50±1.04 | 46.37±0.97 | 52.22±1.09 |
| ad | | | 1.09 | −0.68 | −0.07 | 1.85 |

As can be seen in Table 4, although feature $f_1$ has not been selected for dataset cmc+$f_1$ using $\mathcal{C}4.5\mathsf{rules}$, the performance has degraded when compared to the original dataset. Dataset cmc+$f_2$ using $\mathcal{CN}2$ has degraded considerably but not at 95% confidence level. Both cmc+$f_1$ using $\mathcal{CN}2$ and cmc+$f_2$ using $\mathcal{C}4.5\mathsf{rules}$ have increased performance slightly.

### 5.3 Smoke and Derived Datasets

Two new features were constructed for this dataset with the help of the user:

– $f_1$: this feature represents the status of the interviewed person at the time of survey. It combines four primitive features: *smoking1* (feature #5), *smoking2* (feature #6), *smoking3* (feature #7) and *smoking4* (feature #8).
– $f_2$: this feature shows a comparison between the place the interviewed person works with respect to the city of Toronto-Canada, if s/he works at home or not and if s/he lives in the city of Toronto or outside it. It combines three primitive features: *work1* (feature #2), *work2* (feature #3) and *residence* (feature #4).

**Table5.** Smoke Before and after Constructive Induction

| Feature Number | (smoke, $\mathcal{C}4.5\mathsf{r}$) | (smoke, $\mathcal{CN}2$) | (smoke+$f_1$, $\mathcal{C}4.5\mathsf{r}$) | (smoke+$f_1$, $\mathcal{CN}2$) | (smoke+$f_2$, $\mathcal{C}4.5\mathsf{r}$) | (smoke+$f_2$, $\mathcal{CN}2$) |
|---|---|---|---|---|---|---|
| #0 | ○ | ○ | ○ | ○ | ○ | ○ |
| #1 | ○ | ○ | ○ | ○ | ○ | ○ |
| #2 | ○ | ○ | ○ | ○ | ○ | ○ |
| #3 | ○ | ○ | ○ | ○ | ○ | ○ |
| #4 | ○ | ○ | ○ | ○ | ○ | ○ |
| #5 | ○ | ○ | ○ | ○ | ○ | ○ |
| #6 | ○ | ○ | ○ | ○ | ○ | ○ |
| #7 | | ○ | ○ | ○ | ○ | ○ |
| #8 | ○ | ○ | ○ | ○ | ○ | ○ |
| #9 | ○ | ○ | ○ | ○ | ○ | ○ |
| #10 | ○ | ○ | ○ | ○ | ○ | ○ |
| #11 | ○ | ○ | ○ | ○ | ○ | ○ |
| #12 | ○ | ○ | ○ | ○ | ○ | ○ |
| #13 ($f_1$) | | | | • | | |
| #14 ($f_2$) | | | | | • | • |
| Total #F | 13 | 13 | 14 | 14 | 14 | 14 |
| #F | 12 | 13 | 13 | 14 | 14 | 14 |
| %F | 92.31% | 57.40% | 92.86% | 100.00% | 100.00% | 100.00% |
| Error | 32.71±0.65 | 31.87±0.35 | 33.28±0.80 | 31.56±0.45 | 32.93±0.49 | 31.49±0.45 |
| ad | | | 0.78 | −0.77 | 0.38 | −0.94 |

Table 5 shows that the insertion of the two new features individually has increased the accuracy for the $\mathcal{CN}2$ inducer but not for $\mathcal{C}4.5\mathsf{rules}$. The new feature $f_1$ has degraded performance for $\mathcal{C}4.5\mathsf{rules}$ even $f_1$ was not selected.

### 5.4 Hepatitis and Derived Datasets

One new feature was constructed for this dataset with the help of the expert:

- $f_1$: indicates if the pacient probably will live or die. It combines three primitive features: *liver-firm* (feature #8), *ascites* (feature #11) and *varices* (feature #12).

In Table 6 we can observe that the new feature $f_1$ was responsible for one (feature #12) original feature not appearing in the classifier for dataset hepatitis+$f_1$ with $\mathcal{C}4.5\mathsf{rules}$, even causing an increase in performance. For $\mathcal{CN}2$, the new feature has dramatically changed the subset of features used in the extracted classifier when compared with the one extracted from the original dataset.

## 6 Discussion

Table 7 presents a summary of the results obtained through the three steps performed in the experiments reported in this work. This table shows, for each one of the augmented datasets, the following information:

**Table6.** Hepatitis Before and after Constructive Induction

| Feature Number | (hepatitis, $\mathcal{C}4.5\mathsf{r}$) | (hepatitis, $\mathcal{CN}2$) | (hepatitis+$f_1$, $\mathcal{C}4.5\mathsf{r}$) | (hepatitis+$f_1$, $\mathcal{CN}2$) |
|---|---|---|---|---|
| #0 | ○ | ○ | ○ | ○ |
| #1 | ○ | | ○ | ○ |
| #2 | | | | |
| #3 | ○ | ○ | | |
| #4 | ○ | ○ | ○ | |
| #5 | ○ | ○ | ○ | |
| #6 | | | | |
| #7 | ○ | | ○ | ○ |
| #8 | ○ | ○ | ○ | |
| #9 | | | | |
| #10 | ○ | ○ | ○ | ○ |
| #11 | ○ | | ○ | ○ |
| #12 | | | | ○ |
| #13 | | | | ○ |
| #14 | | | | ○ |
| #15 | ○ | ○ | ○ | ○ |
| #16 | ○ | | ○ | ○ |
| #17 | ○ | ○ | | ○ |
| #18 | | | | ○ |
| #19 ($f_1$) | | | | • |
| Total #F | 19 | 19 | 20 | 20 |
| #F | 12 | 8 | 10 | 13 |
| %F | 63.16% | 42.11% | 50.00% | 65.00% |
| Error | 21.29±2.99 | 18.25±3.83 | 18.00±3.74 | 17.50±2.04 |
| $\mathsf{ad}$ | | | −0.97 | −0.24 |

- A - the names of the datasets;
- B - total number of features in the dataset;
- C - the first number, in brackets, indicates the number of primitive features used to construct the new one followed by the number which identifies the new constructed feature (NewF) as well as the primitive features used for that task (PrimF);
- D - features used by $\mathcal{C}4.5\mathsf{rules}$;
- E - features used by $\mathcal{CN}2$;
- F - if accuracies measured by 10-fold stratified cross-validation improved, using $\mathcal{C}4.5\mathsf{rules}$ and/or $\mathcal{CN}2$ on the augmented datasets. If so, this is indicated by the inducer that had the accuracies improved, *i.e.* where $\mathsf{ad}(E_i - E_{\mathsf{O}}) < 0$ computed by Equation 1.

Note that features in underlined bold style correspond to the new constructed features. Although there were improvements in accuracy, the results would only fit into the perfect situation if during the second step these primitive features, used to construct the new ones, were not selected by the inducers.

**Table7.** Results Summary

| Step 1 | | Step 2 | | | Step 3 |
|---|---|---|---|---|---|
| A | B | C | D | E | F |
| Dataset | | NewF → PrimF | $\mathcal{C}$4.5rules | $\mathcal{CN}2$ | $\mathsf{ad}(E_i - E_{\mathsf{O}}) < 0$ |
| pima+$f_1$ | 9 | (2) | 1 2 5 6 | 1 2 3 4 | $\mathcal{C}$4.5rules |
| | | 8 → 1 2 | 7 **8** | 5 6 7 | |
| pima+$f_2$ | 9 | (2) | 1 2 4 5 | 0 1 2 3 | |
| | | 9 → 1 4 | 6 7 **9** | 4 5 6 7 **9** | |
| cmc+$f_1$ | 10 | (2) | 0 1 2 3 | 0 1 2 3 | $\mathcal{CN}2$ |
| | | 9 → 1 2 | 4 5 6 7 | 4 5 6 7 | |
| | | | 8 | 8 **9** | |
| cmc+$f_2$ | 10 | (2) | 0 1 2 3 | 0 1 2 3 | $\mathcal{C}$4.5rules |
| | | 10 →1 7 | 4 5 6 7 | 4 5 6 7 | |
| | | | 8 **10** | 8 **10** | |
| smoke+$f_1$ | 14 | (4) | 0 1 2 3 | 0 1 2 3 | $\mathcal{CN}2$ |
| | | 13 → 5 6 7 8 | 4 5 6 7 | 4 5 6 7 | |
| | | | 8 9 10 | 8 9 10 | |
| | | | 11 12 | 11 12 **13** | |
| smoke+$f_2$ | 14 | (3) | 0 1 2 3 | 0 1 2 3 | $\mathcal{CN}2$ |
| | | 14 → 2 3 4 | 4 5 6 7 | 4 5 6 7 | |
| | | | 8 9 10 11 | 8 9 10 11 | |
| | | | 12 **14** | 12 **14** | |
| hepatitis+$f_1$ | 20 | (3) | 0 1 4 5 | 0 1 7 10 | $\mathcal{C}$4.5rules |
| | | 19 → 8 11 12 | 7 8 10 | 11 12 13 | $\mathcal{CN}2$ |
| | | | 11 15 16 | 14 15 16 | |
| | | | | 17 18 **19** | |

Table 7 shows that from the 14 cases analyzed (7 datasets and 2 inducers), there were 7 cases (50%) where some improvement in accuracy was obtained although not at the 95% confidence level. Datasets smoke+$f_1$ and smoke+$f_2$ using $\mathcal{CN}2$ have presented the larger increase in accuracy, followed by dataset hepatitis+$f_1$ using $\mathcal{C}4.5$rules. On the other hand, datasets cmc+$f_1$ using $\mathcal{C}4.5$rules and cmc+$f_2$ using $\mathcal{CN}2$ have presented a considerable degradation in performance. Note that in the case dataset cmc+$f_1$ using $\mathcal{C}4.5$rules, the new feature was not selected but it did decrease performance. Theoretical analysis and expertimental studies indicate that many algorithms scale poorly to domains with large number of features that are irrelevant, redundant or both (Langley, 1996). This may suggest that Feature Subset Selection (Kohavi and Sommerfield, 1995; Kohavi and John, 1997) and Constructive Induction can be used together.

Also, none of the new constructed features shows a high strength on its own. If a constructed feature appears in the concept learned it can be observed that at least one of the primitive features is also present.

As the number of new features is small for all datasets, we decide to proceed the analysis, removing the original features used to compose the new feature from each augmented dataset for further investigation (results not shown). For this situation, only two datasets presented an improvement in accuracy without the primitive feature: smoke+$f_2$ using $\mathcal{C}4.5$rules ($\mathsf{ad} = -0.60$) and $\mathcal{CN}2$ ($\mathsf{ad} = -0.76$) as well as hepatitis+$f_1$ using $\mathcal{CN}2$ ($\mathsf{ad} = -0.25$). In these cases, the classifiers not only had a better performance using the new constructed feature but also when primitive features (those used to create the new feature) were removed, the accuracy still remained better than the one obtained using just the original set of features.

## 7 Conclusions

The Constructive Induction approach is based on domain knowledge provided by an user/expert: given the primitive features of the original datasets, the user/expert suggested freely the construction of some new features.

This work proposes a systematic approach for knowledge-driven Constructive Induction based on three steps: (1) creating and adding new features suggested by an user/expert; (2) applying an inducer to the datasets augmented with new features individually and evaluating if the new feature appears on the extracted classifier and finally (3) evaluating increase in performance in the extracted classifier due to the introduction of the new feature. Augmented dataset, and consequently associated new features, that fulfills all steps are selected for further investigation.

The ideal situation would be when the primitive features are not selected during the second step. However, this may not be the case since the new feature may not capture the information embedded in each of the original features for the specific inducer or it is equivalent in predictive power to (some of) the original ones or even the dataset may have already been worked out, so that the original features are, on its own, the most relevant ones.

This work also shows some empirical results of knowledge-driven Constructive Induction. Accuracy and the set of features selected were evaluated when given different sets of features to $\mathcal{C}4.5$rules and $\mathcal{CN}2$. A feature is considered relevant for the learning task if it is used by one of these algorithms to induce the rules.

Results show that, in spite of having an user/expert help, it is difficult to construct new features that are really relevant to learn the concept embedded in these datasets. For future work we suggest that it is necessary to go beyond the available dataset repositories and work on real world datasets that have been not pre-processed for knowledge discovery applications.

**Acknowledgments:** We are grateful to Wu Feng Chung, MD. for the advice in the construction of new features for datasets pima, hepatitis as well as valious suggestions for dataset cmc. We also wish to thank Jaqueline Brigladori Pugliesi for helpful comments on the draft of this paper. This research is partially supported by National Research Councils Finep and CAPES as well as FMRP-USP and FAEPA-HCFMRP-USP.

# References


Aha, D. W. (1997). Lazy learning. *Artificial Intelligence Review*, 11:7–10.

Baranauskas, J. A. and Monard, M. C. (1999). The $\mathcal{MLC}$++ wrapper for feature subset selection using decision tree, production rule, instance based and statistical inducers: Some experimental results. Technical Report 87, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_87.ps.zip.

Baranauskas, J. A., Monard, M. C., and Horst, P. S. (1999). Evaluation of $\mathcal{CN}2$ induced rules using feature selection. *Argentine Symposium on Artificial Intelligence (ASAI/JAIIO/SADIO)*, pages 141–154. http://www.fmrp.usp.br/∼augusto/ps/ASAI99.web.ps.zip.

Blake, C., Keogh, E., and Merz, C. (1998). Uci irvine repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Bloedorn, E. and Michalski, R. S. (1998). Data-Driven Construtive Induction. *IEEE Intelligent Systems*, 13(2):30–37. March/April 1998.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, pages 245–271.

Bull, S. (1994). Analysis of attitudes toward workplace smoking restrictions. *Case Studies in Biometry*, pages 249–271.

Clark, P. and Boswell, R. (1991). Rule induction with $\mathcal{CN}2$: Some recent improvements. In Kodratoff, Y., editor, *Proceedings of the 5th European Conference EWSL 91*, pages 151–163. Springer-Verlag.

Clark, P. and Niblett, T. (1987). Induction in noise domains. In Bratko, I. and Lavrač, N., editors, *Proceedings of the 2nd European Working Session on Learning*, pages 11–30, Wilmslow, UK. Sigma.

Clark, P. and Niblett, T. (1989). The $\mathcal{CN}2$ induction algorithm. *Machine Learning*, 3(4):261–283.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, pages 273–324.

Kohavi, R. and Sommerfield, D. (1995). Feature subset selection using the wrapper model: Overfitting and dynamic search space topology. pages 192–197.

Kohavi, R., Sommerfield, D., and Dougherty, J. (1996). Data mining using $\mathcal{MLC}$++: A amchine learning library in $\mathcal{C}$++. *Tools with IA*, pages 234–245.

Langley, P. (1996). *Elements of Machine Learning*. Morgan Kaufmann Publishers, Inc, San Francisco, CA.

Lee, H. D. and Monard, M. C. (2000). Applying knowledge-driven constructive induction: Some experimental results. Technical Report 101, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_101.ps.zip.

Michalski, R. (1978). Pattern recognition as knowledge-guided computer induction. Technical Report 927, Department of Computer Science – University of Illinois, Urbana-Champaign, Ill.

Michalski, R. S. and Kaufman, K. A. (1998). *Data Mining and Knowledge Discovery: A Review of Issues and a Multistrategy Approach*, pages 71–112.

Quinlan, J. R. (1993). *C*4.5*: Programs for Machine Learning*. Morgan Kaufmann. San Francisco, CA.

Wnek, E. B. J. and Michalski, R. S. (1993). Multistrategy constructive induction. In Kaufmann, M., editor, *Proceedings of the Second International Workshop Machine Learning – ML93*, pages 188–203, San Francisco.

Wnek, J. and Michalski, R. S. (1994). Hypothesis-Driven Constructive Induction in AQ17-HCI: A Method and Experiments. *Machine Learning*, 14(2):139–168.