# Generation of Globally Relevant Continuous Features for Classification

Sylvain Létourneau[1], Stan Matwin[2,3], and A. Fazel Famili[1]

[1] Institute for Information Technology, National Research Council Canada, Ottawa
sylvain.letourneau@nrc-cnrc.gc.ca, fazel.famili@nrc-cnrc.gc.ca
[2] School of Information Technology and Engineering, University of Ottawa, Canada
[3] Institute for Computer Science, Polish Academy of Sciences, Warsaw, Poland
stan@site.uottawa.ca

**Abstract.** All learning algorithms perform very well when provided with a small number of highly relevant features. This paper proposes a constructive induction method to automatically construct such features. The method, named GLOREF (GLObally RElevant Features), exploits low-level interactions between the attributes in order to generate globally relevant features. The usefulness of the approach is demonstrated empirically through a large scale experiment involving 13 classifiers and 24 datasets. Results demonstrate the ability of the method in generating highly informative features and a strong positive effect on the accuracy of the classifiers.

**Keywords:** Machine Learning, Attribute Interactions, Feature Extraction.

## 1 Introduction

Attribute interactions may increase the complexity of a classification task by *dispersing* the instances that belong to the same class across the attribute space. In such cases, the initial attributes, when taken individually, appear to be only remotely related to the class attribute. To uncover the predictive power of such data, the learning systems need to analyze the interacting attributes simultaneously and then build a model that takes into account the interactions observed. As explained by several researchers, this is a complex task that surpasses the ability of many existing machine learning systems.

In particular, Rendell & Seshu [12] emphasizes the fact that current machine-learning techniques rely on the assumption of simple attribute interactions which make them sub-optimal in domains with important attribute interactions. Focusing on the attribute evaluation process, Kononenko & Hong [6] and Bloedorn & Michalski [1] explain that all learning approaches that evaluate the usefulness of each attribute individually using quality measures such as the information gain, the gini-index, the distance measure, or the j-measure are likely to generate inaccurate or too complex models whenever there are important attribute interactions. There have been numerous works on trying to improve the ability of the naive-Bayes with respect to attribute dependencies (e.g., [7]). Specific issues

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 196–208, 2008.

such as the *replication* and the *fragmentation* problems with decision trees are also directly related to the lack of capacity of current techniques to deal with attribute interactions [13]. From an applied perspective, it has been argued that attribute interactions are becoming the norm in KDD applications and failing to address this problem adequately has important consequences on the performance obtained [2]. All of these observations call for novel practical techniques that can facilitate learning in domains with important attribute interactions.

This paper proposes such a technique. It is a constructive induction technique that augments the initial representation with new features which make explicit the important information hidden in the interactions among the initial attributes. The new features are self-contained globally relevant features that are suitable for learning algorithms assuming independence. As it will be shown experimentally, the new features can also increase the performance of more complex learning algorithms.

After presenting motivation and related work, the paper introduces the method to derive the new globally relevant features. Sect. 5 offers a large-scale experiment illustrating the usefulness of the approach and the last section concludes the paper.

## 2 Motivation

In this research, the concept of *relevance* designates the usefulness of a given attribute to predict the values of the class attribute. We assume that relevance is computed through a univariate measure such as the gain ratio [11]. Moreover, we use the term *globally relevant attribute* to designate an attribute that is relevant over the full training set.

To illustrate the potential effects of attribute interactions on relevance and the usefulness of globally relevant features, let us consider a simple binary classification task with three attributes $X_1, X_2$, and $X_3$ that follow a multivariate normal distribution defined by the following class-conditioned mean vectors and variance-covariance matrix (same for both class values):

$$\mathbf{u_0} = \begin{bmatrix} 70 \\ 70 \\ 40 \end{bmatrix} \quad \mathbf{u_1} = \begin{bmatrix} 70 \\ 70 \\ 55 \end{bmatrix} \quad \mathbf{\Sigma} = \begin{bmatrix} 650.0 & 0 & -160 \\ 0 & 50 & -115 \\ -160 & -115 & 350 \end{bmatrix}$$

From the mean vectors ($\mathbf{u_0}$ and $\mathbf{u_1}$), we conclude that $X_3$ is the only relevant attribute for this task while $\mathbf{\Sigma}$ indicates that $X_1$ and $X_2$ interact with $X_3$. Fig. 1 (a) shows a simple dataset generated from the above distribution. As seen from the scatter plots of $X_3$ versus $X_1$ and $X_3$ versus $X_2$, it is difficult to separate the positive from the negative instances. This difficulty is further illustrated by the class-conditional density curves for $X_3$; the great overlap between the two curves clearly indicates that any decisions based on $X_3$ will be highly error-proned. The null gain ratio and $\chi^2$ values confirm that, from a univariate global perspective, $X_3$ appears powerless in predicting the class values.

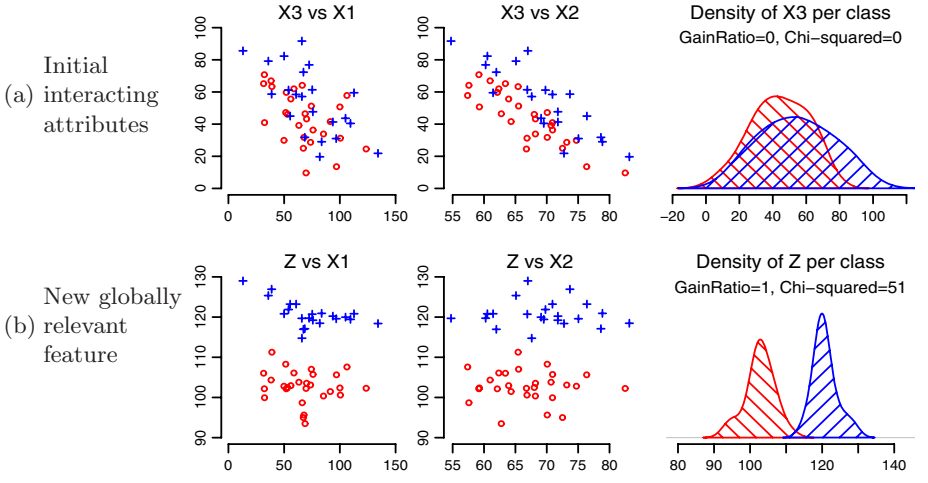


**Fig. 1.** The effects of interactions on relevance and a globally relevant feature

To uncover the power of the data, we propose a constructive induction method capable of generating a new *globally relevant feature* $Z$ that cancels the negative effects of $X_1$ and $X_2$ on $X_3$. The new feature is shown in Fig. 1 (b). We observe that the transformation removed a great proportion of the initial dispersion since the instances of the same class are now grouped together. As illustrated by the class-conditional density curves for $Z$, the new feature is highly relevant and its power is observable across the full dataset independently of the other attributes.

The data transformation approach proposed in this paper can automatically generate globally relevant features from complex interactions between any continuous attribute and an arbitrary large number of influencing attributes of possibly different types (continuous, nominal, binary). No information about the underlying distribution of the data or the nature of the interactions is required.

## 3 Related Work

Related research has been conducted in constructive induction and statistics. A large proportion of the constructive induction techniques are designed to be integrated with existing learning approaches and are not producing a new representation (e.g.:FRINGE[10], AQ17-DCI[1], and OCI[9]). With these systems, the focus is on the improvement of the accuracy of existing methods by opposition to be on the assessment and removal of the negative effects of attribute interactions. Hu [4] noticed the lack of general data pre-processing methods that are independent of specific learning algorithms. Their solution was to propose the GALA systems. These systems generate highly comprehensible features but the types of interactions that it can handle are limited to either prototypical relationships or boolean expressions. The GALA systems do not directly assess the

interactions observed in the data and do not produce a model that describes the effects of these interactions. Recent work by Jakulin & Bratko [5] introduced the notion of *interaction gain* to analyse attribute interactions along with visualization methods. They proposed an experiment showing the benefits of Cartesian product as an approach to resolve the most important interactions.

The topic of interactions has been extensively studied in statistics. PCA, ICA, and contextual normalization methods (e.g., [8]) are examples of methods that have been used in machine learning to help assess the structure of the interactions and produce new features that keep the most important information (according to some criteria). On the other hand, these methods do not rely on the class information, which limit their usefulness for classification tasks [3]. We also notice that most of them can only handle continuous attributes.

In summary, we observe a lack of paradigm-independent supervised constructive induction techniques that directly address the issues of attribute interactions while being capable of handling both continuous and discrete attributes. The GLOREF approach we propose in this paper is an attempt to fulfill this need.

## 4 The GLOREF Approach

We now describe the GLOREF (GLObally RElevant Features) approach which we propose for the construction of globally relevant features that account for the initial interactions among the attributes. GLOREF works as a pre-processor and can be used with any standard learning algorithm. The input is a training dataset which contains at least one numerical attribute. The GLOREF approach has two phases: the analysis of relevance and the generation of globally relevant features. The analysis phase computes information to characterize the interactions among the attributes along with their impact on learning. The results of this analysis are stored in data structures named *relevance matrices*. The feature generation phase uses the relevance matrices to search for transformation models. Finally, these transformation models are applied to augment the initial data representation and the learning can proceed as usual with the augmented data representation. The following subsections describe the analysis of relevance, the automatic generation of globally relevant features, and application considerations.

### 4.1 Analysis of Relevance

The analysis of relevance takes as input the training dataset and, optionally, two lists defining the *explanatory* and the *response* attributes. If these lists are not provided, we simply generate default lists of explanatory and response attributes containing all initial attributes and all initial continuous attributes, respectively[1]. As output, the analysis of relevance returns a set of relevance

[1] The use of the terms *response* and *explanatory* attributes follows statistical nomenclature for the analysis of interactions. On the other hand, it is important to notice that the end objective of the proposed method is not to generate new features that approximate the response attributes but instead generate new features that have higher global relevance than any of the initial attributes.

matrices. These matrices provide information on the relevance of the response attributes over partitions based on the explanatory attributes.

The analysis starts by creating a partition of the training dataset $S = \{s_1, s_2, \ldots, s_N\}$ for each explanatory attribute. For example, a partition based on a nominal explanatory attribute $X$ with $m$ possible values, noted $\{X(1), X(2), \ldots, X(m)\}$[2], generates $m$ subsets $S_1, S_2, \ldots, S_m$ where each $S_i = \{s \in S \mid \text{val}_X(s) = X(i)\}$ for $i \in \{1, 2, \ldots, m\}$. If the explanatory attribute is continuous, we first discretize it and then partition based on the discretized values instead of the original ones. Since the discretized attributes produced are not going to be used for classification, there is no need to use a supervised discretization technique in this step. A simple unsupervised method such as equal-width or equal-frequency is more appropriate. By default, we use three intervalls for discretization. As shown by the experimental results in Sect. 5, this seems to be an adequate choice accross a variety of domains although it is likely that even better performance could be obtained by increasing the number of intervals.

The next step computes the relevance information. This step considers one explanatory and one response attributes at a time. To evaluate the effect of the explanatory attribute, we evaluate the relevance of the response attribute in each of the subsets ($S_i$) and in the full training dataset ($S$). Following the standard approach to characterize relevance of continuous attributes in decision tree building, we first sort the instances along the response attribute. We then define a split for each observed value of the response attribute in the given set and compute how many examples of each class would fall on each side of the split. Using these numbers, we compute the gain ratio for each possible split. Finally, we define two additional values noted $\lambda_1$ and $\lambda_2$ that identify the majority class on each side of the split. We name these two values *compatibility characteristics* since they will be used to determine whether the subsets of the partitions interact in a compatible manner or not (i.e., if they reduce the global relevance or not). All information computed during this step is stored in a set of relevance matrices noted $\text{RM}_1, \text{RM}_2, \ldots, \text{RM}_m$, and RM, where $\text{RM}_i$ contains the information computed using subset $S_i$, and RM the information from $S$.

To illustrate, let us consider the analysis of the effects of $X_2$ on the relevance of $X_3$ for the domain introduced above. First, the partitioning step needs to discretize $X_2$. Let us suppose that this discretization did lead to a new attribute $X_2\_discretized$ with 5 possible values (0, 1, 2, 3, and 4). In this case, 6 relevance matrices would be generated (one for each subset and one for the global dataset). The table on the left hand side in Fig. 2 shows part of the relevance matrix for the subset $S_1$, which includes all instances $s$ such that $\text{val}_{X_2\_discretized}(s) = 0$. There are 18 entries in this relevance matrix which corresponds to the number of distinct values observed for the response attribute $X_3$ in the given subset. For each cut point, the relevance matrix shows the threshold value, the number of instances per class in each side of the split (columns 'Cumul.' and 'Bal.'), the compatibility characteristics $\lambda_1$ and $\lambda_2$[3], and the relevance in terms of gain ratio.

[2] The missing value (indicated by '?') is considered like any other possible values.

[3] The symbol `NA` indicates that no class is in majority in the given side of the split.

| Num | Thresh. | Cumul. | Bal. | $\lambda_1$ | $\lambda_2$ | Rel. |
|---|---|---|---|---|---|---|
| 1 | 97.36 | {7,10} | {0,1} | 1 | 1 | 0 |
| 2 | 91.67 | {7,9} | {0,2} | 1 | 1 | 0 |
| 3 | 89.07 | {7,7} | {0,4} | NA | 1 | .0 |
| 4 | 84.45 | {7,6} | {0,5} | 0 | 1 | .02 |
| | | ... | | | | |
| 11 | 76.32 | {7,1} | {0,10} | 0 | 1 | .5 ◄ |
| | | ... | | | | |
| 17 | 57.83 | {2,0} | {5,11} | 0 | 1 | .0 |
| 18 | 43.43 | {1,0} | {6,11} | 0 | 1 | .0 |

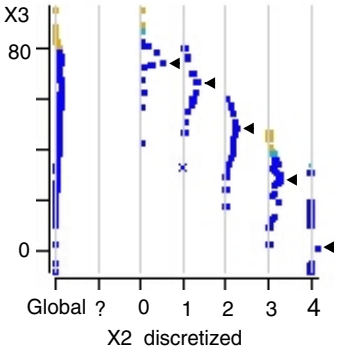


**Fig. 2.** A relevance matrix and the relevance graph to analyze the effects of $X_2$ on $X_3$

The best cut point for this subset (denoted by ◄) is at threshold 76.32 which splits the dataset into two subsets of 8 (7 from $1^{st}$ class and 1 from $2^{nd}$ class) and 10 (all from $2^{st}$ class) instances, respectively.

**Visualizing Relevance Matrices and Detecting Harmful Interactions.** The information contained in the relevance matrices for a given pair of attributes can be effeciently visualized through a *Relevance Graph*. For example, let us consider the graph in Fig. 2 which shows the effects of $X_2$ on the relevance of $X_3$ for the same example. This relevance graph is composed of 6 curves, one for each relevance matrix. The one on the left (named *global relevance curve*) accounts for the global relevance matrix (i.e., RM) while the following ones (named *local relevance curves*) are for the relevance matrices corresponding to the subsets of the partition based on $X_2$*_discretized* (i.e., $\mathrm{RM}_1, \mathrm{RM}_2, \ldots, \mathrm{RM}_5$). In particular, the first local relevance curve (labeled '0') corresponds to the relevance matrix shown on the left side. Each point on a given curve represents one entry in the corresponding relevance matrix. The threshold values for the response attribute are shown along the vertical axis. The color (or gray scale) and symbol (e.g., square, cross, plus) of each point designate the compatibility characteristics $\lambda_1$ and $\lambda_2$, respectively. There is one color (symbol) for each possible value of $\lambda_1$ ($\lambda_2$). The relevance of a given point is shown by the horizontal distance that separates it from the vertical reference line located on the left side of each curve. The larger the distance; the better is the cut point in producing pure partitions.

The effect of a given interaction on the global relevance is directly assessed by comparing the relevance of the best cut points (the ones that are the farthest away from their vertical reference line) in the local relevance curves with the relevance of the best cut point in the global relevance curve. If one or more best cut points in local curves are more relevant than the best global cut point, then the interaction has a negative effect on the global relevance of the response attribute. The relevance graph in Fig. 2 illustrates this situation since several of the most relevant cut points in the local relevance curves (indicated on the graph by ◄) are more relevant than the best global cut point.

### 4.2 Automatic Generation of Globally Relevant Features

A key idea behind the GLOREF approach comes from the observation that the global relevance of the response attribute can be modified by altering the alignement of the local relevance matrices. Such a re-alignement can be accomplished by modifying the values of the response attribute within each local relevance matrix by a value $\omega_i$ for $i = 1, \ldots, m$. The result is a new feature $Z$ defined as

$$Z = \Gamma(X, Y) = \begin{cases} Y + \omega_1 & \text{if } X = X(1) \\ Y + \omega_2 & \text{if } X = X(2) \\ \vdots & \\ Y + \omega_m & \text{if } X = X(m) \end{cases} \quad (1)$$

where $Y$ is the response attribute, $X(1), X(2), \ldots, X(m)$ are the distinct values for the explanatory attribute $X$, and $\{\omega_1, \omega_2, \ldots, \omega_m\}$ are the parameter values of the model. The objective is to set the $\omega_i$ values in a way that maximizes the global relevance of the new feature. We first present the algorithm developed to resolve this optimization problem and then introduce the approach to cope with interactions involving several explanatory attributes.

**Univariate Transformations.** A brute force solution to select the parameter values $\{\omega_1, \omega_2, \ldots, \omega_m\}$ is to evaluate all possible alignments of the local relevance curves and select the alignment with the best global relevance. Recognizing that the number of possible alignments is exponential in the number of relevance curves, this solution would not be practical in most real world applications. We therefore introduce the heuristic approach described in Fig. 3.

The algorithm starts by handling a special case that happens when all the most relevant cut points in the various relevance matrices are compatible (equal values for both compatibility characteristics $\lambda_1$ and $\lambda_2$). In this case, the algorithm directly returns the optimal solution which aligns these most relevant cut points on an arbitrary threshold noted $T^*$[4]. The relevance graph in Fig. 2 illustrates this situation since all maximally relevant cut points are compatible (same color and same symbol). When the most relevant cut points are not all compatible, the algorithm proceeds with a gready search. This search gradually builds the complete solution by combining local solutions. It starts by finding the best alignment between the first two relevance matrices and store the result into a temporary relevance matrix noted $RM_{cum}$. In the following iteration, it combines $RM_{cum}$ with the third relevance matrix and so on until all local relevance matrices have been processed. There are three steps in each iteration of the search procedure: reduction of the two relevance matrices to be considered (`ReduceRM`), search for the best local alignment (`UnivExhaustiveSearch`), and update of the current solution (`ComputeGlobalRM`). In the first step, `ReduceRM` removes many of the cut points from the two relevance matrices considered in order to reduce

[4] By default, the algorithm sets the $\omega_i$ values such that the best global cut point will be at threshold value 0.

**Algorithm UnivGLOREF**

*Input*: The set of relevance matrices $\mathrm{RM}_1, \mathrm{RM}_2, \ldots, \mathrm{RM}_m$ for pair of attributes

*Output*: A set of values $\{\omega_1, \omega_2, \ldots, \omega_m\}$ maximizing the relevance of a new feature.

if best cut points from all subsets $S_i$ have identical $\lambda_1$ and $\lambda_2$

  $\{\mathrm{T}_1^*, \mathrm{T}_2^*, \ldots, \mathrm{T}_m^*\} \leftarrow$ best cut point thresholds in $\{\mathrm{RM}_1, \mathrm{RM}_2, \ldots, \mathrm{RM}_m\}$

  $\Omega^* \leftarrow \{-\mathrm{T}_1^*, -\mathrm{T}_2^*, \ldots, -\mathrm{T}_m^*\}$

else

  $\{\omega_1, \omega_2, \ldots, \omega_m\} \leftarrow \{0, 0, \ldots, 0\}$, $\mathrm{RM}_{cum} \leftarrow \mathrm{RM}_1$

  For $i = 2$ to $m$

    /* Simplify current relevance matrices */

    $\mathrm{RM}_{cum} \leftarrow$ `ReduceRM`$(\mathrm{RM}_{cum})$, $\mathrm{RM}_i \leftarrow$ `ReduceRM`$(\mathrm{RM}_i)$

    /* Find current best solution and update previous solution */

    $\{\omega, \omega_i\} \leftarrow$ `UnivExhaustiveSearch`$(\mathrm{RM}_{cum}, \mathrm{RM}_i)$

    For $j = 1$ to $i - 1$ $\omega_j \leftarrow \omega_j + \omega$

    /* Compute global relevance info for current partial solution */

    $\mathrm{RM}_{cum} \leftarrow$ `ComputeGlobalRM`$(\{\mathrm{RM}'_{i-1}, \mathrm{RM}'_i\}, \{\omega_{i-1}, \omega_i\})$

  $\Omega^* \leftarrow \{\omega_1, \omega_2, \ldots, \omega_m\}$

return $\Omega^*$

**Fig. 3.** Heuristic to efficiently generate univariate GLOREF features

the number of potential alignments to evaluate. Precisely, it removes all entries except the most relevant cut point for each observed combination of $\lambda_1$ and $\lambda_2$ and the two points with minimal and maximal thresholds. In the second step, `UnivExhaustiveSearch` evaluates all potential alignments of the two reduced relevance matrices and returns the two $\omega$ values that maximize the global relevance of a new feature that would be created by combining the subsets considered. Finally, `ComputeGlobalRM` updates the current solution by adding the new $\omega$ values to the previous global solution. Once all relevance matrices have been considered, the heuristic returns the set of parameter values $\{\omega_1, \omega_2, \ldots, \omega_m\}$ selected for the generation of a new globally relevant feature (Eq. 1).

**Multivariate Transformations.** The direct extension of the univariate solution to handle the multivariate case would require a multivariate partionning of the initial dataset along with the analysis of the resulting combinatorial number of subsets. Efficiency concerns and the risk of having to proceed with insufficient data in the various subsets call for an alternative method. Accordingly, we propose an inductive process where each phase has two steps: *Feature Generation* and *Feature Selection*.

The generation step constructs features in progressive order of complexity by combining pairs of features from the previous phase. In the first phase, it

uses the univariate transformations to create multivariate features with two explanatory attributes. For instance, if there are two univariate transformations $Z_1 = \Gamma(X_1, Y)$ and $Z_2 = \Gamma(X_2, Y)$, then the generation step in the inital phase would create a new multivariate feature $Z = \Gamma(\{X_1, X_2\}, Y)$. In the second phase, the generation step uses the selected features from the first phase to create features involving either three or four explanatory attributes, and so forth. Each multivariate feature is constructed through an iterative optimization process. Precisely, to construct a multivariate feature $Z$ involving $l$ explanatory attributes $X_1, \ldots, X_l$ and a response attribute $Y$ we repeat the following steps

1. Using the univariate procedure described above, compute for each $X_i$ a set of parameter values noted $\Omega_i^t$ that optimizes the relevance of $\Gamma(X_i, Z^{t-1})$.
2. Update the values of the new feature using

$$Z^t = \sum_{i=1}^{l} \Gamma(X_i, Z^{t-1}; \Omega_i^t) - (l-1) * Y \tag{2}$$

where $t > 1$, $Z^0 = Y$, and $\Gamma(X_i, Z^{t-1}; \Omega_i^t)$ is equivalent to (1) with the parameter values specified by the set $\Omega_i^t$. The repeated summation allows us to jointly realign the univariate relevance curves in a way that maximize the relevance of the new feature. The process stops when there is no significant improvements in the global relevance of $Z$ between two iterations or when a maximal number of iterations has been performed. In practice, only a few iterations are required to converge (between two and five in most cases). This process ensures that the number of parameters to estimate grows linearly with the number of explanatory attributes and avoids the multivariate partitioning issues mentioned above. The reuse of the efficient univariate heuristic presented above further improve the performance of the approach.

The feature selection determines which features are allowed to proceed to the next phase of the inductive process. To be selected, a new multivariate feature must have a higher global relevance than any of the attributes involved in its creation. To control the risk of overfitting, we use only 70% of the training data during the creation of the features and keep the remaining part for the feature selection step. The inductive process stops when less than two new features are selected for the following iteration. Finally, all univariate transformation models and all selected multivariate ones are applied to augment the initial representation with globally relevant features. We notice that the overall computational complexity of the approach is polynomial in the number of features provided as input to each iteration. By applying feature selection prior to each iteration, we ensure that the approach stays practical regardless of the number of initial attributes.

### 4.3 Application Issues and Smoothing of Transformations

When computing the values for the new features, two issues may arise: missing values and unseen values. Missing values might be observed for one or more of

**Table 1.** Global relevance of the best initial attribute and GLOREF feature

| Dataset | Initial GR | GLOREF Type | GR | Diff (%) | Dataset | Initial GR | GLOREF Type | GR | Diff (%) |
|---|---|---|---|---|---|---|---|---|---|
| autos | .55 | Mul | .92 | .37 (69 %) | N1F1 | .29 | Mul | .63 | .34 (117 %) |
| balance-scale | .17 | Mul | .67 | .49 (286 %) | N1MN | .15 | Mul | .29 | .14 (91 %) |
| breast-w | .55 | Mul | .86 | .31 (57 %) | N2F1 | .59 | Mul | .79 | .20 (34 %) |
| cars | .44 | Mul | .54 | .10 (23 %) | N2MN | .17 | Uni | .32 | .15 (88 %) |
| colic | .28 | Mul | .38 | .10 (36 %) | N3F1 | .56 | Mul | .90 | .34 (60 %) |
| credit-a | .42 | Mul | .47 | .05 (12 %) | N3MN | .17 | Mul | .41 | .24 (145 %) |
| diabetes | .18 | Mul | .30 | .12 (64 %) | N4F1 | .31 | Mul | .84 | .53 (169 %) |
| glass | .80 | Mul | .98 | .17 (21 %) | N4MN | .17 | Mul | .98 | .81 (488 %) |
| heart-statlog | .35 | Mul | .59 | .24 (68 %) | N5F1 | .38 | Mul | .58 | .20 (53 %) |
| hepatitis | .33 | Mul | .55 | .22 (65 %) | N5MN | .19 | Mul | 1.0 | .81 (435 %) |
| ionosphere | .50 | Mul | .73 | .22 (44 %) | N6F1 | .27 | Mul | .53 | .26 (97 %) |
| liver | .05 | Mul | .31 | .26 (482 %) | N6MN | .16 | Mul | .41 | .25 (163 %) |

the explanatory attributes or for the response attribute. The former case does not cause any problem as our implementation treats this situation explicitly by including the missing value as one of the potential values for all explanatory attributes. However, if the response attribute has a missing value then the new feature would also need to have a missing value. The problem of unseen values arises when the model tries to process an instance for which the observed explanatory attribute value has not been seen during the generation of the transformation model. Since the given value was not part of the training dataset, the models do not include an entry for this value and therefore there is no corresponding $\omega$ parameter value. In this case, the value of the new feature equals the value of the response attribute (i.e., no transformation).

The discretization of continuous explanatory attributes may introduce unnecessary discontinuities in the new features. We avoid this problem by smoothing the $\omega$ values when applying transformations that involve one or more continuous explanatory attributes. We use the inverse distance weighting smoothing method to adjust the $\omega$ values based on the observed values of the explanatory attribute(s).

## 5 Experimental Evaluation

To evaluate the feasibility of the GLOREF approach, we propose a large-scale experiment involving 24 datasets (12 artificial and 12 from the UCI repository) and 13 classifiers implemented in the WEKA package. The artificial datasets contain numerical attributes only with pre-defined univariate and simple multivariate interactions. Several of the UCI datasets contain a mix of continuous and discrete attributes. The maximal number of attributes is 35. We followed the 10-fold cross-validation methodology. In each fold, we performed the following tasks: (1) apply GLOREF on the training data to learn univariate and multivariate transformation models, (2) use these models to augment the initial representation with GLOREF features, (3) for each learning system, learn a model using only the initial attributes and another model using the augmented representation, and (4) evaluate the accuracy of the two models on test data.

| Classifier | Accuracy improvement | | | | Better | Worse |
|---|---|---|---|---|---|---|
| | Avg | Std | Min | Max | (Significant) | |
| BaggingDT | 1.10 | 3.67 | -3.8 | 11.2 | 16 (6) | 7 (0) |
| BoostingDT | 1.01 | 3.82 | -4.1 | 14.6 | 13 (5) | 10 (0) |
| DecisionStump | 9.87 | 9.81 | -.45 | 32.8 | 22 (13) | 2 (0) |
| DecisionTable | 5.33 | 7.41 | -3.5 | 21.5 | 18 (12) | 6 (0) |
| HyperPipes | 24.4 | 17.4 | -6.0 | 61.2 | 22 (22) | 2 (1) |
| IB1 | 3.01 | 3.49 | -1.9 | 10.5 | 18 (7) | 4 (0) |
| IB5 | 2.70 | 3.10 | -.30 | 8.40 | 9 (5) | 3 (0) |
| J48 | 2.87 | 4.89 | -5.6 | 12.9 | 17 (10) | 7 (0) |
| KernelDensity | 2.00 | 3.97 | -5.8 | 10.1 | 13 (8) | 10 (0) |
| NaiveBayes | 4.82 | 6.21 | -3.7 | 18.7 | 19 (11) | 5 (1) |
| OneR | 10.5 | 10.9 | -1.3 | 34.9 | 19 (16) | 5 (0) |
| PART | 3.27 | 3.73 | -4.6 | 10.0 | 20 (8) | 4 (0) |
| SMO | 1.78 | 4.09 | -3.0 | 16.3 | 12 (2) | 6 (0) |
| Artificial | 7.54 | 9.75 | -1.2 | 47.5 | 135(95) | 17(0) |
| UCI | 3.79 | 9.54 | -6.0 | 61.2 | 83(30) | 54(2) |
| All | 5.78 | 9.82 | -6.0 | 61.2 | 218(125) | 71(2) |

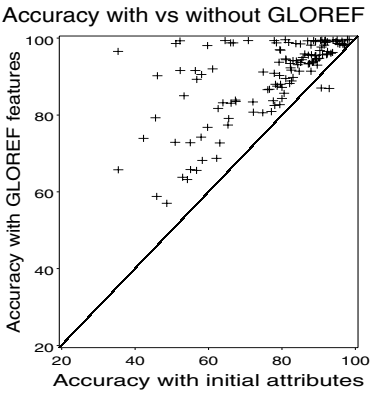


**Fig. 4.** The effects of GLOREF features on accuracies

We first consider the ability of GLOREF to produce new globally relevant features by comparing the expected gain-ratio of the best initial attribute and the best GLOREF feature. We compute the expected gain-ratio of the best initial (resp. GLOREF) attribute by averaging the gain-ratios of the best initial (resp. GLOREF) attribute based on test data from the various folds of the cross-validation procedure. Table 1 presents the results. The gain ratio for the best GLOREF feature is systematically higher that the one for the best initial attribute. The standard t-test to compare group means reveals that all increases are statistically significant at the 0.05 level. The relatively large variation in percentage of increase (from 12% to 488%) suggests that the datasets are not all equally affected by the problem of attribute interactions. We repeated the analysis using the $\chi^2$ measure and obtained consistent results. Therefore, we conclude that the GLOREF approach succeeded in producing new highly globally relevant features.

The graph on the right side of Fig. 4 offers a quick view of the usefulness of the new features for learning. There is one point for each combination of learning system and dataset for which the use of GLOREF features significantly changed the accuracy. All points located above the diagonal line indicate positive results and inversely for the points located below. The table on the left side details the results by classifier. The first four columns provide the statistics on increase in accuracy due to the GLOREF features while the last two columns count the number of better and worse results with statistically significantly results in parentheses (the number of datasets for which the addition of the GLOREF features did not change the results equals the difference between 24 and the sum of the 'Better' and 'Worse' columns). Out of the 312 experiments (13 classifiers * 24 datasets), 127 lead to a significant difference in accuracy and only 2 of these are on the negative side. As expected, learning systems which are powerless with respect to attribute interactions such as HyperPipes, OneR, and DecisionStump profited the most from the GLOREF features with average increase in accuracy of 24%, 10.5% and 9.8%, respectively. Focusing on statistically significant results,

we notice that all classifiers have been positively affected by the GLOREF features, with the number of statistically significant wins varying from 2 to 22 over 24. Moreover, the column 'Max' clearly shows that complex approaches such as bagging, boosting and support vector machine (SMO) can also greatly benefit from highly globally relevant features. The relatively important standard deviations tend to confirm the heterogeneousness of the selected datasets. Finally, by analyzing the results by datasets, we observe that the levels of increase in accuracy tend to match the increase of global relevance between the best GLOREF and best initial feature. In other words, large improvements in global relevance generally result in high increases in accuracy and inversely.

## 6 Conclusion

This paper links the problem of attribute interactions to the concept of attribute relevance. After discussing the potential effects of interactions on relevance, we introduce the GLOREF method to model interactions and construct new globally relevant features. The autonomous solution is evaluated through a large-scale experimentation involving 24 datasets and 13 learning systems. The analysis of the relevance of the new features shows that the GLOREF system generates highly globally relevant features for all datasets, with some increases in gain ratio that are close to 500%. Adding the GLOREF features to the initial representation significantly improved the accuracy in more than 40% of the experiments, while reducing it in only less than 1%. Although these results are strongly positive, it is possible that the heuristics proposed are not optimal. Future work will investigate alternative heuristics to further improve performance.

## References

[1] Bloedorn, E., Michalski, R.S.: Data-driven constructive induction. IEEE Intelligent Systems and their Applications 13(2), 30–37 (1998)
[2] Freitas, A.A.: Understanding the crucial role of attribute interaction in data mining. Artificial Intelligence Review 16, 177–199 (2001)
[3] Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, NY (1990)
[4] Hu, Y.-J.: Representational Transformation Through Constructive Induction. PhD thesis, University of California, Irvine (1999)
[5] Jakulin, A., Bratko, I.: Analyzing attribute dependencies. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 229–240. Springer, Heidelberg (2003)
[6] Kononenko, I., Hong, S.J.: Attribute selection for modelling. Future Generation Computer Systems 13, 181–195 (1997)
[7] Langley, P.: Induction of recursive Bayesian classifier. In: Brazdil, P.B. (ed.) ECML 1993. LNCS, vol. 667, pp. 152–164. Springer, Heidelberg (1993)

[8] Létourneau, S., Famili, A.F., Matwin, S.: A normalization method for contextual data: Experience from a large-scale application. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 49–54. Springer, Heidelberg (1998)
[9] Murthy, S.K., Kasif, S., Salzberg, S.: A system for induction of oblique decision trees. Journal of Artificial Intelligence Research 2, 1–33 (1994)
[10] Pagallo, G., Haussler, D.: Boolean feature discovery in empirical learning. Machine Learning 5(1), 71–99 (1990)
[11] Quinlan, J.R.: Induction of decision trees. Machine Learning 1, 81–106 (1986)
[12] Rendell, L.A., Seshu, R.: Learning hard concepts through constructive induction: Framework and rationale. Computational Intelligence 6(4), 247–270 (1990)
[13] Vilata, R., Blix, G., Rendell, L.A.: Global data analysis and the fragmentation problem in decision tree induction, pp. 312–326 (1997)