Instance-Based Learning Techniques of Unsupervised Feature Weighting do not perform so badly!

Héctor Núñez[1] and **Miquel Sànchez-Marrè**[1]

Abstract. The major hypothesis that we will be prove in this paper is that unsupervised learning techniques of feature weighting are not significantly worse than supervised methods, as is commonly believed in the machine learning community.
This paper tests the power of unsupervised feature weighting techniques for predictive tasks within several domains. The paper analyses several unsupervised and supervised feature weighting techniques, and proposes new unsupervised feature weighting techniques. Two unsupervised entropy-based weighting algorithms are proposed and tested against all other techniques. The techniques are evaluated in terms of predictive accuracy on unseen instances, measured by a ten-fold cross-validation process. The testing has been done using thirty-four data sets from the UCI Machine Learning Database Repository and other sources. Unsupervised weighting methods assign weights to attributes without any knowledge about class labels, so this task is considerably more difficult. It has commonly been assumed that unsupervised methods would have a substantially worse performance than supervised ones, as they do not use any domain knowledge to bias the process. The major result of the study is that unsupervised methods really are not so bad. Moreover, one of the new unsupervised learning method proposals has shown a promising behaviour when faced against domains with many irrelevant features, reaching similar performance as some of the supervised methods.

1 INTRODUCTION

A major problem in predictive tasks within instance-based algorithms is to find out which are the relevant features to be taken into account. When experts are available in a particular domain, these tasks could be easier. However, in general, when there is no expertise available, some automatic methods should be used. Many methods have been proposed and used in the literature trying to establish the relevance of attributes. One of this kind of techniques is feature weighting [1]. Feature weighting consists in the assignment of an importance degree to each one of the available features describing a domain or process. Normally weights are scaled in the range [0..1] or in an equivalent range. Thus, features with lower weights are the less important ones, while high weights mean very important features.

But there is an added problem to evaluate how well the different feature weighting techniques, and in general, the feature selection algorithms work. They must be evaluated in terms of the performance of a task. In this paper, the predictive task accuracy is used. Even making this decision, there is another key point. Which predictive method should be used? In this study, instance-based learning algorithms (IBL), and the nearest neighbour classifier will be used because they are good techniques to make predictions based on previous experience. In IBL algorithms, similarity is commonly used to decide which instance is closest to a new problem. Thus, this similarity criterion for predictive tasks will be used to evaluate the different weighting techniques in the experimental testing.

This paper aims at analysing and studying the performance of several commonly used supervised feature weighting techniques against some unsupervised feature weighting methods. Two new unsupervised weighting techniques are proposed. The techniques are evaluated in terms of predictive accuracy on unseen instances, measured by a ten-fold cross-validation process. Of course, the label class is not taken into account to find out the weights within the unsupervised methods, but it is for the predictive accuracy computation.

In recent years, many researchers are focusing on feature weighting. Feature weighting is a very important issue. It is intended to give more relevance to those features detected as important, and at the same time, it is intended to give lower importance to irrelevant features. Most general methods for feature weighting use a global scheme. It means to associate a weight to all the space of the feature. If a continuous attribute is present, a discretization pre-process is suggested to allow making a weight computation according to its interval values and its correlation with the value class. The importance of one feature will be determined by the distribution of the class values for that feature. Some research has been done such as the mutual information technique proposed in [20], the results reported by Mohri and Tanaka [15] of his QM2 method, and Creecy et al. [4] about their introduced cross-category feature importance (CCF) method, and other research in [11]. On the other hand, unsupervised weighting methods assign weights to attributes without any knowledge about class labels, so this task is considerably more difficult [6], [7]. Also, it has commonly been assumed that unsupervised methods would have a substantially worse performance than supervised ones, as they do not use any domain knowledge to bias the process [7]. To confirm this hypothesis, a performance analysis among several supervised and unsupervised methods have been done. In the unsupervised feature weighting literature, we have only found the work done by [18] on a gradient descent technique and feature selection approach [6] on unsupervised entropy-based method. The new unsupervised methods proposed derive from the last cited

[1] Knowledge Engineering & Machine Learning Group, Technical University of Catalonia, Barcelona, email: {hnunez, miquel}@lsi.upc.es

method. Some similar methods to those proposed in this paper can be found in literature, where underlying idea is to remove or add one feature at a time and use a heuristic to evaluate the new state of the order of data, as proposed Maron and Moore in [8].

The paper is organized in the following way. Some supervised weighting techniques are described in Section 2. Section 3 outlines main features about unsupervised global weighting algorithms, including the new ones proposed. Section 4 shown the experimental set-up and the results comparing the performance of all feature weighting. Finally, in Section 5 conclusions and future research directions are outlined.

2 SUPERVISED FEATURE WEIGHTING

There is some research in supervised feature weighting such as those by [4, 9, 15, 16, 20]. In this study, global and local supervised methods showing the best performance have been selected for the comparison with unsupervised ones. These methods are the Information Gain methods (IG [20], IG-DB [5]), the Feature Projection method (PRO) [2], the RELIEF-F method (RELF) [12], the global Class Distribution weighting method (CDWG) [8] the Correlation-Based methods (CD, VD, CVD) [16], and the local methods (CDWL [8], VDM [19] and EBL [16]). Due to space constraints they are nor detailed here, but further details can be found in the references.

3 UNSUPERVISED FEATURE WEIGHTING

Feature selection and feature weighting methods in supervised domains have been thoroughly discussed in the literature, (see [20] and [6]). On the other hand, very little work has been done for unsupervised domains, probably due to the assumed hypothesis that their performance would necessary be substantially worse than the supervised method performance. The two methods found in the literature, in addition to the new methods proposed are detailed in this section.

3.1 Gradient-Descent Technique (GD)

The proposal of Shiu *et al*. in [18] consists mainly on defining a feature-evaluation index *E* defined as:

$$E(w) = \frac{2\left\{\sum_{p}\sum_{q(p<q)}\left[SM_{pq}^{(w)}(1-SM_{pq}^{(1)}) + SM_{pq}^{(1)}(1-SM_{pq}^{(w)})\right]\right\}}{N(N-1)}$$

Where N is the number of instances in the data set. $SM_{pq}^{(1)}$ is the similarity between instances p and q evaluated with weight 1 for all the attributes and $SM_{pq}^{(w)}$ is the similarity evaluated with weight attributes different from 1. Noticing that the feature-evaluation $E(w)$ will gradually become zero when $SM_{pq}^{(w)} \rightarrow 0$ or 1, the main idea is to find a weight set so that the feature-evaluation function attains its minimum.

The method uses gradient-descendent technique to minimize $E(w)$. The change in w_j (i.e. Dw_j) is computed as:

$$\Delta w_f = -\eta \frac{\partial E}{\partial w_f}$$

where η is the learning rate, and for the computation of $\partial E/\partial w_f$ the following expressions are used:

$$\frac{\partial E(w)}{\partial w_f} = \frac{2\left[\sum_{p}\sum_{q(q<p)}\left(1-2SM_{pq}^{1}\right)\frac{\partial SM_{pq}^{(w)}}{\partial d_{pq}^{(w)}}\frac{\partial d_{pq}^{(w)}}{\partial w_f}\right]}{N(N-1)}$$

$$\frac{\partial SM_{pq}^{(w)}}{\partial d_{pq}^{(w)}} = \frac{-\alpha}{\left(1+\alpha d_{pq}^{(w)}\right)^2}$$

$$\frac{\partial d_{pq}^{(w)}}{\partial w_f} = \frac{w_j \gamma_j^2}{\left(\sum_{j=1}^{n} w_j^2 \gamma_j^2\right)^{1/2}}$$

$SM_{pq}^{(1)}$ is the similarity between instances p and q evaluated with all weights equal to 1. $SM_{pq}^{(w)}$ is the similarity between instances p and q evaluated with the weights computed in the previous step. α and η *are* positive parameters between 0 and 1. γ^2 is the distance between values of attribute j. N is the number of instances in the data set. The stopping condition of the algorithm is that E becomes less than or equal to a given threshold or until the number of iterations exceeds a certain predefined number. The prospective result is that, on the average, the similarity values $\{SM_{pq}^{(w)}, p=1,N, q<p\}$ with trained weights are closer to 0 or to 1, than those without trained weights such as $\{SM_{pq}^{(1)}, p=1,N, q<p\}$.

3.2 Unsupervised Entropy-Based methods

In [6], the authors present a *feature selection* method for unsupervised domains based in entropy computations. Starting from this approach (UEB), we made an extension to obtain two *feature weighting* algorithms (UEB-1 and UEB-2), trying to obtain a superior performance assigning real-valued weights instead of binary-value weights. The underlying idea is that data have orderly configurations if they have distinct clusters, and have disorderly or chaotic configurations otherwise [6]. From entropy theory, it can be stated that entropy is lower for ordered configurations, and higher for disordered configurations. The feature selection method is based on the observation that removing an irrelevant feature from the feature set may not change the underlying concept of the data, but not so otherwise. Following this idea, the first step consist in compute the entropy between two instances:

$$E = -S\log_2 S - (1-S)\log_2(1-S)$$

where S is the similarity measure based on a distance concept, and assumes a very small value (close to 0.0) for very close pairs of instances, and a very large (close to 1.0) for very distant pairs. For the entire data set of N instances the entropy measure is given as:

$$E = -\sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij} * \log_2(S_{ij}) + (1-S_{ij}) * \log_2(1-S_{ij}))$$

where S_{ij} is the similarity value between the instance i and the instance j normalized to [0,1]. When all features are numeric or ordinal, the similarity of two instances is:

$S_{ij} = e^{-\alpha * D_{ij}}$, where D_{ij} is the distance between the instances i and j. The value of α is computed automatically by $\alpha = \frac{-\ln 0.5}{\bar{D}}$, where

$\bar{D}$ is the average distance among all the instances. Euclidean distance is used to compute the distance D_{ij}. If all the attributes are nominal, the similarity between two instances is: $S_{ij} = \frac{\sum_{k=1}^{M} |x_{ik} = x_{jk}|}{M}$ where $|x_{ik} = x_{jk}|$ is 1 if x_{ik} equals x_{jk} and 0 otherwise, and M is the number of features.

3.2.1 Unsupervised Entropy-Based method 1 (UEB-1)

The algorithm computes the entropy of data by removing a feature. For M features this is repeated M times. Features are ranked in descending order of relevance by finding the descending order of the entropy after removing each of the M features one at a time. Feature selection algorithms focuses on deciding if one attribute is relevant or not. On the other hand, feature weighting algorithms focus on giving a relevance measure for each attribute. In our method (UEB-1) that is the first extension of the UEB algorithm, to obtain feature weights instead of feature selection, the approach takes the entropy values computed for each one of the attributes, and applies a scaling process to assign weights. To obtain weights in [0,1] range for each attribute k, the following computation is done:

$$w_k = \frac{Entropy_k - Arg\min(Entropy)}{Arg\max(Entropy) - Arg\min(Entropy)}$$

In Figure 1, an outline of the algorithm is described. CompEnt(i) computes the entropy of the data after discarding the i^{th} feature.

```
P = Entropy values for M features
For i = 1 to M
        P_i = CompEnt(i)
Endfor
For i = 1 to M
        w_i = (P_i - Arg min(P)) / (Arg max(P) - Arg min(P))
Endfor
```

Figure 1. UEB-1 Algorithm outline.

3.2.2 Unsupervised Entropy-Based method 2 (UEB-2)

The second extension (UEB-2) performs a wrapper approach in the sense that implements an update of weights in each step of the cycle taking into account the last values of computed weights. Taking this into account, UEB-1 can be seen as a filter approach. In UEB-2, an initial weight of 0.5 is assigned to each attribute and the entropy for the entire database is computed. Then, it computes the new entropy value after removing one attribute at a time. If the new entropy value after removing one attribute is less than the entropy of the entire database, then the weight of that attribute is decreased by 0.1. If the new entropy value after removing one attribute is greater than the entropy of the entire database, the weight of that attribute is increased by 0.1. This increasing/decreasing parameter was set to 0.1 after an empirical study. This cycle is performed several times allowing the weights to reach a minimum or maximum value in the [0,1] range. After an empirical evaluation, this parameter was set to 6. An outline of the algorithm is presented in Figure 2. Total_Entropy is the entropy for the entire database taking into account all the attributes. CompEnt(i) computes the entropy of the data after discarding the i^{th} feature.

```
For i = 1 to M
  w_i = 0.5
Endfor
P_T = Total_Entropy
For j = 1 to 6
    For i = 1 to M
        P_i=CompEnt(i)
        If P_i< P_T then
            W_i=W_i-0.1
        else
            W_i=W_i+0.1
        Endif
    Endfor
    P_T = Total_Entropy
Endfor
```

Figure 2. UEB-2 Algorithm outline.

Table 1. Major properties of databases considered in the experimentation

DB	SN	#I	C	OD	NOD	#C	%M
Air pollution	AP	365	5	0	0	4	0
Annealing	AN	798	6	29	3	6	64.9%
Audiology	AD	200	0	8	61	24	0
Australian	AS	690	6	4	4	2	0
Auto	AU	205	15	0	8	7	0.004
Bands	BA	512	20	0	20	2	4.87
Breast Cancer	BC	699	0	9	0	2	0
Bridges	BR	108	3	0	8	3	0.06
Cleveland	CL	303	5	2	6	2	0
Contracept MC	CM	1473	2	4	3	3	0
Credit	CR	690	6	0	9	2	0.64%
Dermatology	DE	366	1	0	33	6	0
Ecoli	EC	336	7	0	0	8	0
Flag	FL	194	3	7	18	8	0
German	GE	1000	7	0	13	2	0
Glass	GL	214	9	0	0	7	0
Hayes-Roth	HR	132	0	0	4	3	0
Hepatitis	HE	155	6	0	13	2	5.7
Horse-Colic	HC	301	7	0	16	2	30
Ionosphere	IO	351	34	0	0	2	0
Iris	IR	150	4	0	0	3	0
LED	LD	300	0	0	7	10	0
LED-17	LI	200	0	0	24	10	0
Liver Disord.	LD	345	6	0	0	2	0
Machine	MA	209	6	0	0	8	0
Pima In. Diab.	PI	768	8	0	0	2	0
Post-Operative	PO	90	1	7	0	3	0
Soybean(large)	SL	307	0	6	29	19	21.7
Vehicle	VE	946	18	0	0	4	0
Votes	VO	435	0	0	16	2	7.3
Waveform	WF	300	21	0	0	3	0
Waveform-40	WA	300	40	0	0	3	0
Wine	WI	178	13	0	0	3	0
Zoo	ZO	90	0	0	16	7	0

4 EMPIRICAL EVALUATION

To test the performance of unsupervised feature weighting methods, a nearest neighbour classifier was implemented. *L'Eixample* similarity measure [17], and CAIM discretization method [13] have been used when required in order to calculate feature weights or similarity between attribute value pairs. This selection was made on the basis of preliminary performance tests on similarity measures and discretization methods. Anyway, similar results were obtained with other discretization and

similarity approaches, which are not described here, due to space constraints.

We did not directly compare UEB-1 and UEB-2 with UEB itself due to the fact that the last one is a feature selection method and we propose feature weighting algorithms.

Table 2. Accuracy results with no weights (NW), global and local supervised weighting schemes and unsupervised weighting schemes.

	NW	IG	IG-DB	PROJ	RELF	CDWG	CD	VD	CVD	CDWL	VDM	EBL	GD	UEB-1	UEB-2
AP	91.17	99.47	99.47	99.75	87.61	95.58	98.67	98.39	**100.00**	95.31	97.25	99.19	91.70	64.43	77.10
AN	91.48	92.47	90.60	92.85	92.09	91.48	94.00	91.35	92.97	90.37	93.24	**94.25**	91.97	83.75	90.97
AD	77.50	**80.00**	68.50	78.00	72.50	78.00	78.00	77.50	78.00	77.00	77.50	76.00	71.00	76.50	72.50
AS	82.03	82.90	82.03	**83.19**	82.90	82.90	80.58	78.41	79.86	82.90	81.59	82.75	82.61	75.80	81.16
AU	75.33	80.79	77.40	**84.72**	75.29	75.33	73.86	77.77	81.24	75.83	80.83	77.46	76.79	35.01	50.21
BA	76.30	78.52	71.30	78.33	77.22	76.48	70.93	73.89	78.70	76.48	**81.48**	80.37	71.67	73.70	69.44
BC	95.90	95.04	95.92	96.10	95.49	96.12	94.40	94.44	94.61	96.12	96.32	**96.76**	94.40	94.19	92.69
BR	86.03	**94.79**	**94.79**	91.46	87.69	86.03	90.69	89.59	94.03	84.69	89.79	92.56	86.69	90.13	88.55
CL	75.86	77.52	**78.51**	75.88	75.51	75.84	74.19	76.86	74.89	75.84	77.81	77.18	78.46	71.54	75.86
CM	45.01	45.28	45.89	45.14	44.74	45.21	45.35	45.89	45.35	44.33	**46.71**	45.89	44.60	44.81	44.40
CR	81.16	**82.75**	81.45	82.32	82.61	82.46	80.14	79.71	80.00	82.46	81.30	82.17	80.14	78.55	81.45
DE	94.23	95.28	83.18	95.13	95.83	94.23	**97.05**	96.22	95.85	93.68	92.61	93.16	92.05	57.65	91.82
EC	80.49	80.84	80.54	**83.26**	80.49	80.49	80.79	78.41	79.67	79.58	81.75	80.58	81.66	59.74	80.49
FL	49.21	60.66	**65.47**	60.60	58.04	51.32	48.80	55.96	61.11	49.24	64.83	64.30	53.44	36.54	45.84
GE	70.50	69.90	66.70	71.10	70.50	70.70	66.40	70.40	69.20	70.70	**72.50**	70.80	66.60	63.20	69.80
GL	72.97	**80.96**	75.44	74.77	73.44	73.38	71.42	73.44	77.15	76.72	77.27	73.52	73.52	66.44	52.57
HR	70.62	70.62	70.62	73.49	72.72	70.62	70.62	70.62	70.62	78.21	85.03	**85.03**	72.26	76.56	70.62
HE	79.60	75.57	78.24	78.86	80.82	78.93	80.05	78.86	76.42	78.93	81.35	**84.86**	75.09	82.20	79.60
HC	74.22	79.36	74.77	76.86	78.03	75.19	78.74	**80.63**	76.94	75.19	78.46	79.90	71.38	73.60	72.19
IO	90.87	91.72	90.57	91.14	90.87	91.14	91.71	91.44	92.00	91.14	93.44	**94.87**	90.86	77.48	87.73
IR	94.00	94.67	95.33	94.67	94.67	95.33	95.33	95.33	94.67	94.67	**96.00**	**96.00**	93.33	92.67	94.00
led	66.67	**68.00**	67.33	66.00	**68.00**	66.67	65.67	65.00	**68.00**	66.67	66.00	64.67	67.67	66.00	66.67
LI	38.00	62.50	47.00	54.50	**64.00**	38.00	60.50	57.00	57.00	38.00	53.50	55.50	36.00	11.50	38.00
LD	64.85	64.57	63.18	63.41	64.83	65.19	66.42	64.04	65.56	65.19	69.24	**69.85**	58.76	61.79	63.54
MA	71.55	69.53	72.05	67.33	71.15	71.05	67.53	68.13	68.53	71.65	72.58	**72.68**	68.83	63.87	69.67
PI	71.35	69.74	70.02	71.19	71.21	70.29	66.14	70.14	69.01	70.29	**73.17**	72.67	70.47	62.01	69.32
PO	53.33	56.67	57.78	56.67	56.67	53.33	51.11	**66.67**	52.22	56.67	62.22	63.33	55.56	56.67	53.33
SL	91.07	91.50	71.73	**93.25**	92.68	91.07	92.84	91.21	91.37	90.47	92.97	92.97	90.04	88.71	89.74
VE	68.41	68.19	**70.21**	68.91	68.41	68.06	58.37	67.83	67.35	68.29	69.60	69.50	66.74	46.22	57.91
VO	93.58	95.16	95.05	**97.04**	96.10	94.52	96.10	95.16	94.63	94.52	96.74	95.88	94.63	88.52	92.64
WF	71.95	**76.68**	75.77	76.17	73.71	74.42	69.12	69.54	72.92	74.34	73.47	73.19	75.82	43.13	61.80
WA	69.67	81.33	72.67	79.67	71.00	71.00	59.67	72.00	79.00	71.33	81.33	**81.67**	73.00	32.67	72.33
WI	96.22	97.28	97.98	97.87	96.92	95.15	97.06	96.69	96.11	96.11	**99.52**	97.87	95.74	74.17	89.50
ZO	96.09	96.09	97.09	**98.09**	97.09	96.09	97.00	98.00	96.09	97.00	95.09	96.09	96.09	86.09	96.09
Av.	76.68	79.60	77.19	79.34	78.26	77.11	76.74	78.13	78.56	77.35	80.37	**80.40**	76.16	66.35	73.22
SD	14.92	13.37	13.74	14.01	13.03	14.99	15.38	13.50	13.84	14.88	12.93	13.26	14.97	19.41	15.90

Table 3. Accuracy results over artificial databases HR15 and IR15

	NW	IG	IGDB	PROJ	RELF	CDWG	CD	VD	CVD	CDWL	VDM	EBL	GD	EUB11	EUB2
HR	70.62	70.62	70.62	73.49	72.72	70.62	70.62	70.62	78.21	70.62	**85.03**	**85.03**	72.26	76.56	70.62
HR15B	45.33	68.24	64.45	47.80	52.20	46.15	69.01	68.19	63.68	44.62	68.35	**76.65**	38.52	26.48	45.33
IR	94	94.67	95.33	94.67	94.67	95.33	95.33	94.67	94.67	95.33	**96.00**	**96.00**	93.33	92.67	94.00
IR15C	78.83	95.33	93.33	94.67	92.04	84.13	96.00	**96.67**	95.33	85.46	94.67	95.33	75.50	27.79	95.33

Although for benchmark data used, the class information is know, the class label was hidden in the unsupervised algorithms in order to create an "artificial" unsupervised domain but keeping available the label to evaluate the methods' performance afterwards. Next, some tests were carried out to evaluate the generalisation accuracy using both supervised global and unsupervised global weighting techniques, trying to show in an empirical way that the generalisation accuracy can be maintained when using unsupervised methods. Thus, the goal was to reject the commonly agreed hypothesis that the unsupervised weighting method performance is substantially worse than supervised methods' performance. All tests were performed with thirty-four databases selected from the UCI database repository [3] and other sources. Detailed description of the databases is shown in Table 1, where short name (SN), number of instances in each database (#I), number of continuous attributes (C), ordered discrete attributes (OD), not ordered discrete attributes (NOD), number of classes (#C) and missing values percentage (%M.) are depicted. To verify the accuracy of the nearest neighbour classifier, a test by means of a 10-fold cross-validation process was implemented. The average accuracy over all 10 trials is reported for each data test and for each weighting scheme. The highest accuracy achieved in each data set is shown in boldface in Table 2. Av. is the average databases and SD is the standard deviation across all databases.

Afterwards, to analyse the performance of the unsupervised and supervised weighting methods, when faced against databases with many irrelevant features, the following experiment was done. Two new databases were artificially created. Iris database, which has all four continuous attributes, was selected and was expanded with fifteen irrelevant continuous attributes. Values for these attributes were randomly generated in the range [0,10].

A similar approach was implemented to the Hayes-Roth database, which has only discrete attributes. Fifteen binary irrelevant attributes were added, and these values were randomly selected between 0 and 1. Results are detailed in the Table 3.

5 CONCLUSIONS AND FUTURE WORK

Main conclusions after the analysis of the performance among all weighting schemes are that, in general, unsupervised weighing algorithms maintain a very close relation with supervised weighting schemes. Moreover, accuracy of unsupervised methods over all databases is very similar to accuracy obtained with supervised approaches. After a two-tailed paired *t* test, even though we cannot conclude that both supervised and unsupervised methods are indistinguishable, unsupervised method accuracy is only between a 5% and 9% lower than supervised methods. On the contrary, a commonly greater decrease was thought to be true. In spite of the initial belief or hypothesis that very low accuracy could be obtained when you do not know the class labels of instances, the accuracy achieved by unsupervised weighting methods is high enough to consider these methods as very promising tools in classification and retrieval tasks, specially, in front of new unknown and unsupervised databases. These results show in an empirical way that you can use unsupervised weighting algorithms to determine the feature relevance in unsupervised databases, as a first approach.

We think that this is due to the fact that similarity computations between instances in unsupervised methods capture the intrinsic distribution of different instances (different "classes") in a similar way than supervised methods do.

The UEB-1 method seems to a have a higher variability than the other two unsupervised methods. GD method seems to be as good as the supervised methods when facing relevant features. On the other hand, UEB-2 seems to be a good method needed to be tuned for a better performance. It has shown a very promising behaviour when faced against domains with many irrelevant features (see Table 3). Additionally, UEB-2 reaches similar or better performance than some of the supervised methods, and definitively better than using no weights at all. Thus, the UEB-2 seems to be a very promising method that should be studied further.

After analysing Table 2 and Table 3, we have observed that our unsupervised methods UEB-1 and UEB-2 perform well when facing domains where total number of discrete attributes are greater than total number of continuous attributes. Specially, in AP, AU, GL, LI, VE and WF databases, the number of ordered discrete attributes was zero, and their performance was poor. We think this situation implies that entropy variation of similarity values computed by our methods will be lower, and hence, the variation of feature weights will be lower too, leading its behaviour to a more regular feature weight distribution with worse performance.

A first step has been done in the design of suitable unsupervised feature weighting techniques, to be used in predictive tasks in instance-based algorithms. Also, the proposed unsupervised entropy-based weighting approach (UEB-2) seems to be promising, especially to detect irrelevant features. More experiments are being carried out to support this statement.

Future work will be focused both on a further tuning (0.1 increment/decrement parameter, and number of cycles) and sensitivity analysis of these methods, and on the design, study and analysis of other unsupervised weighting algorithms.

REFERENCES

[1] D. Aha. *Feature weighting for lazy learning algorithms*, In H Liu and H. Motoda (Eds.) Feature Extraction, Construction and Selection: A Data Mining Perspective. Norwell MA: Kluwer, 1998.

[2] H. A. Güvenir and A. Akkus, *Weighted K Nearest Neighbour Classification on Feature Projections*, in Proceedings of the Twelfth International Symposium on Computer and Information Sciences (XII ISCIS) S. Kuru, M.U. Caglayan and H.L. Akin (Eds.), Antalya, Turkey, (Oct. 27-29, 1997), 44-51.

[3] C.L. Blake, and C.J. Merz. *UCI Repository of machine learning databases* [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science. 1998.

[4] R. H. Creecy, B. M. Masand, S. J. Smith and D. L. Waltz, Trading MIPS and memory for knowledge engineering. *Communications of the ACM* 35:48-64, 1992.

[5] W. Daelemans, A. Van Den Bosch. *Generalization performance of backpropagation learning on to syllabification task*. In Proceedings of TWLT3: Connectionism Natural and Language Processing, pp. 27-37. Enschede, The Netherlands. 1992.

[6] M. Dash and H. Liu. *Handling Large Unsupervised Dates via Dimensionality Reduction*. In Proc. of SIGMOD Data Mining and Knowledge Discovery Workshop (DMKD), Philadelphia, USA, May 1999.

[7] N. Howe, C. Cardie. *Feature Subset Selection and Order Identification for Unsupervised Learning*. Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA. 2000. pp. 247-254.

[8] N. Howe, C. Cardie. *Examining locally varying weights for nearest neighbour algorithms*. Proceedings of the Second International Conference on Case-Based Reasoning. 1997. pp. 455-466.

[9] J. Jarmulak, S. Craw and R. Rowe. *Self-Optimising CBR Retrieval*. Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence. pp. 376-383. 2000.

[10] K. Kira and L. Rendell. *A Practical Approach to Feature Selection*. In Proceedings of the 9th International Conference on Machine Learning, Morgan Kaufmann, pp. 249-256, 1992.

[11] R. Kohavi, P. Langley, and Y. Yun. *The utility of feature weighting in nearest neighbour algorithms*. In Proceedings of the European Conference on Machine Learning (ECML97) , 1997.

[12] I. Kononenko, *Estimating Attributes: Analysis and extensions of RELIEF*, In Proc. of European Conf. on Machine Learning, Springer, pp. 171-182, 1994.

[13] L. Kurgan and K. J. Cios. *Discretisation Algorithm that Uses Class-Attribute Interdependence Maximisation*, Proc. of the 2001 Int. Conf. on Artificial Intelligence (IC-AI 2001), pp.980-987, Las Vegas, Nevada. 2001.

[14] O. Maron and A Moore. The Racing Algorithm: Model selection for Lazy Learners. *Artificial Intelligence Review, , Special Issue on lazy learning Algorithms* 11(1-5), pp.193-225, 1997.

[15] T. Mohri and H. Tanaka. *An Optimal Weighting Criterion of Case Indexing for Both Numeric and Symbolic Attributes*, Aha, D. W., editor, Case-Based Reasoning papers from the 1994 workshop, AAAI Press, Menlo Park, CA.

[16] H. Núñez, M. Sànchez-Marrè and U. Cortés. *Improving Similarity Assessment with Entropy-Based Local Weighting*. Procc of 5th International Conference on Case-Based Reasoning (ICCBR'2003), pp.377-391, Trondheim, Norway, 2003.

[17] M. Sànchez-Marrè, U. Cortés, I. R-Roda, and M. Poch. *L'Eixample distance: a new similarity measure for case retrieval. Procc. of 1st Catalan Conference on Artificial Intelligence (CCIA'98), ACIA bulletin* 14-15 pp. 246-253.Tarragona, Catalonia, EU.

[18] S.C.K. Shiu, D.S. Yeung, C.H. Sun, X.Z. Wang. *Transferring case knowledge to adaptation knowledge: an approach for case-base maintenance. Computational Intelligence* 17(2), pp. 295-314, 2001.

[19] C. Stanfill, D. Waltz. Toward Memory-Based Reasoning, *Communications of the ACM* 29(12), pp 1213-1228, 1986.

[20] D. Wettschereck, D. W. Aha, and T. Mohri. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review, Special Issue on lazy learning Algorithms*, 11(1-5) 1997. pp 273-314