The Effect of Imbalanced Data Class Distribution on Fuzzy Classifiers - Experimental Study

Sofia Visa
Department of ECECS,
University of Cincinnati,
Cincinnati, OH 45221-0030, USA
svisa@ececs.uc.edu

Anca Ralescu
Department of ECECS,
University of Cincinnati,
Cincinnati, OH 45221-0030, USA
aralescu@ececs.uc.edu

Abstract— **This study evaluates the robustness of a fuzzy classifier when class distribution of the training set varies. The analysis of the results is based on the classification accuracy and ROC curves. The experimental results reported here show that fuzzy classifiers are less variant with the class distribution and less sensitive to the imbalance factor than decision trees.**

I. Introduction

In order to evaluate correctly the performance of a given classification method on real data sets, information such as the error costs and the underlying class distribution are required [1], [2]. For learning with imbalanced class distributions - that is, for a two-class classification problem, the training data for one class (majority or negative class) greatly outnumbers the training data for the other class (minority or positive class) - such information is crucial and yet, many times not available.

Since standard methods of classification are driven by the minimization of the overall accuracy, without considering (or knowing) error costs of the two classes (minority and majority), they are not suitable for imbalanced data sets. A common practice for dealing with this problem is to rebalance classes artificially, either by up-sampling or down-sampling. As suggested in [2], up-sampling does not add information while down-sampling actually removes information. Considering this fact, the best research strategy is to concentrate on how machine learning algorithms can deal most effectively with whatever data they are given. Fuzzy classifiers, [3] and [4], derived from class frequency distributions proved effective in classifying imbalanced data sets.

II. Class Distribution in the Learning Process

In this experiment the role of class distribution in learning a fuzzy classifier from imbalanced data is investigated. A similar experiment was published in [5] using decision trees. The performance of the fuzzy classifier for multidimensional data is evaluated on five real data sets and compared with the results published in [5]. This study emerged from the fact that, *there is no guarantee that the data available for training represent (capture) the distribution of the test data*. Therefore, reduced variance of classifiers output over different training class distributions is a very important feature of a classifier.

TABLE I

Statistics about the real data sets. Second column shows the natural distribution of the data sets as the minority class percentage of the whole data set.

Name	Minority class %	# of features	Size	Train size	Test size
letter-a	3.9%	16	20000	592	5000
optDigits	9.9%	64	5620	416	1406
letter-vowel	19.4%	16	20000	2908	5001
german	30.0%	24	1000	225	250
wisconsin	35.0%	9	683	179	171

A. The Data Sets

Table I shows characteristics of the five UCI Repository domains used in this study. In the second column of the Table I are listed the natural class distributions of the data sets expressed in this paper as the minority class percentage of the whole data set.

The letter-a/letter-vowel data set was obtained from the letter data set as follows: instances of letter 'a'/of vowels represent the minority class and the remaining letters, the majority class. For the optDigits data set, the minority class is represented by the digit 0 and the remaining digits (1 - 9) represent the majority class. The wisconsin and german data sets are two-class domains: cancer versus non-cancer patients and good versus bad credit history of persons asking loans, respectively.

B. Altering the Class Distribution

To study experimentally how the class distribution affects the fuzzy classifier in learning the real domains, the distribution of the training set is varied and the classifier is evaluated, for each distribution, on the same test data (see a similar study in [5] using C4.5).

The test data set reflects the natural distribution and it is obtained by selecting randomly 25% of examples from each class (for example, for the letter-a data set, a testing set of $5,000$ points is obtained: 197 minority instances and $4,803$ majority instances). By P are denoted the remaining minority examples and by N the remaining majority examples. In order

to compare the performance of different classifiers obtained for different class distributions, the same test data is used.

The training set size($Size(TrainData)$) is equal to the $card(P)$ (number of minority examples left after forming the test data - that is 592, for the letter-a data set). The training set is altered to obtain different class distribution, as follows: for $n\%$ class distribution, $(n\% * Size(TrainData))$ random minority points are selected from P and $(Size(TrainData) - n\% * Size(TrainData))$ randomly selected majority points from N, where n is 2%, 10%, 30%, 50%, 80%, 95% and the natural distribution (listed in the second column of the Table I).

III. The Fuzzy Classifier

The main problem in designing a fuzzy classifier is to construct the fuzzy sets, more precisely their membership functions. Approaches to construct fuzzy classifiers range from quite ad-hoc to more formal approaches, in which the membership function is constructed directly from data without any intervention of the designer. The current approach relies on the interpretation of a fuzzy set as a family of probability distributions and therefore, a particular membership function is the result of selecting one of the probability distributions in this family. The mechanism of deriving a fuzzy set membership function makes use of mass assignment theory(MAT) [6] and is presented shortly next (for in depth presentation, please see [7], [8] and [4]).

Given a collection of data, and the relative frequency distribution corresponding to it, $\{f_{(k)}; k = 1, \ldots, n; 1 \geq f_{(1)} \geq f_{(2)} \geq \cdots \geq f_{(n)} \geq 0, \sum_{i=1}^{n} f_{(i)} = 1\}$, the corresponding fuzzy set is obtained from the Equation 1:

$$\mu_{(k)}^{lpd} = k f_{(k)} + f_{(k+1)} + f_{(k+2)} + ... + f_{(n)} \qquad (1)$$

where $\mu_{(k)}^{lpd}$ denotes the kth largest value of the membership function corresponding to the general, *lpd(least prejudiced selection rule)* selection rule [6].

Example 1 illustrates the complete mechanism of converting a simple artificial data set into a fuzzy classifier, corresponding to the lpd selection rule [9].

Example 1: Let Maj and Min denote respectively the majority and minority classes given as:

$$Maj = \{x_1, x_1, x_1, x_1, x_2, x_2, x_2, x_2, x_2, x_2, x_2,$$

$$x_3, x_3, x_3, x_3, x_4, x_4, x_4\}$$

$$Min = \{x_4, x_4, x_5, x_5, x_5, x_6\}$$

Their relative frequency distributions (in nonincreasing order) corresponding to Maj/Min are:

$$Maj = \{(x_2, 7/18), (x_1, 4/18), (x_3, 4/18), (x_4, 3/18)\}$$

$$Min = \{(x_5, 3/6), (x_4, 2/6), (x_6, 1/6)\}$$

The membership values for each fuzzy set are computed (in decreasing order of the relative distributions) as shown in Table II. The obtained fuzzy sets (each class is mapped into a fuzzy set) are displayed in the Figure 1.

TABLE II

$FuzzySet_{lpd}$ FOR THE Maj AND Min CLASSES OF EXAMPLE 1.

	$FuzzySet_{lpd}$
Maj	$\mu_{Maj}(x_2) = 1\frac{7}{18} + \frac{4}{18} + \frac{4}{18} + \frac{3}{18} = 1$
	$\mu_{Maj}(x_1) = 2\frac{4}{18} + \frac{4}{18} + \frac{3}{18} = \frac{15}{18}$
	$\mu_{Maj}(x_3) = 3\frac{4}{18} + \frac{3}{18} = \frac{15}{18}$
	$\mu_{Maj}(x_4) = 4\frac{3}{18} = \frac{12}{18}$
Min	$\mu_{Min}(x_5) = 1\frac{3}{6} + \frac{2}{6} + \frac{1}{6} = 1$
	$\mu_{Min}(x_4) = 2\frac{2}{6} + \frac{1}{6} = \frac{5}{6}$
	$\mu_{Min}(x_6) = 3\frac{1}{6} = \frac{3}{6}$

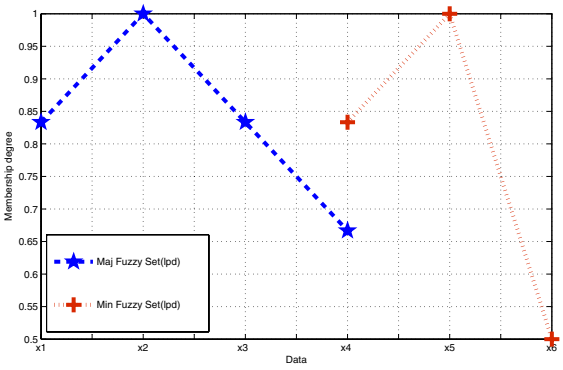

Fig. 1. The fuzzy sets obtained for the majority (left) and the minority (right) class using lpd selection rule.

For a test data point, the membership degree to each of these fuzzy sets are computed and compared: the point is assigned to the class to which it belongs with a higher degree. For example, the derived fuzzy classifier classifies the data as follows: $\{x_1, x_2, x_3\}$ belong to Maj class and $\{x_4, x_5, x_6\}$ belong to Min class.

Example 1 illustrates for one-dimensional data set the basic one-pass fuzzy classifier used in this study. In principle, for multidimensional data sets the approach outlined above can be applied as well. However, it should be noticed that as the dimensionality increases the data set becomes sparse, and that there may be very few data points with frequency greater than 1. Otherwise stated, this means that in order to obtain meaningful frequencies, either the data set size must increase with each new dimension, or for a given data set, preprocess it by collecting data into bins and apply the approach described to bins. The bin approach is apt to introduce errors, while increasing the data set size is not always possible (in fact,

rarely is possible).

In any case, regardless of the approach used, another problem that arises is that of interpolation for computing the membership degree to unlabeled data points. Having multidimensional fuzzy sets makes this step more complex.

The approach currently taken in this study is to derive fuzzy sets along each dimension, in effect, deriving as many classifiers as the dimension of the attribute space and to *aggregate* these classifiers in order to evaluate a data point. Several aggregation operators are proposed here but other aggregation methods such as the ones presented in [10] can be used too. The following notations are used in defining the aggregation methods (H_i, $i = 1, \ldots, 4$):
$C \in \{+, -\}$ denotes the class label of a test point x;
$C_C^i(x) = I_{\mu_C^i > \mu_{C'}^i}$ with $C, C' \in \{+, -\}$ is the indicator function;
$w_i = \frac{n_i}{\sum_{i=1}^{n} n_i}$ for $i = 1, \ldots, n$ is a set of weight characterizing the attributes (n_i is the number of correctly classified training data by the i^{th} attribute).
Then, the aggregations are defined as follows:

1) H_1: $H_1^C(x) = \sum_{i=1}^{n} C_C^i(x)$.
2) H_2: $H_2^C(x) = \sum_{i=1}^{n} \mu_C^i(x)$.
3) H_3: $H_3^C(w; x) = \sum_{i=1}^{n} w_i \cdot \mu_C^i(x)$.
4) H_4: $H_4^C(w; x) = \sum_{i=1}^{n} w_i \cdot C_C^i(x)$.

Based on the H_i, $i = 1, \ldots, 4$, the class label of x is decided by evaluating

$$D_i(x) = argmax\{H_i^C(x);\ C \in \{+, -\}\}$$

for $i = 1, \ldots, 4$.

But first, it is interesting to understand why one may expect a good performance from the fuzzy classifier applied to imbalance data. As it can be observed from Figure 1, x_4 will be assigned as belonging to the minority class since its degree to this class is 0.83 and the membership to the majority class is 0.66. Looking at the original data shows that x_4's frequency in the minority class is 2 while in the majority class it is 3. Any classifier in which x_4 is learned based on its contribution to a class relative to the whole data set, will assign x_4 to the majority class.

Classifiers such as the fuzzy classifier used in this study, which learn the classification based on the relative frequency within the class will assign x_4 to the minority class, where its relative frequency of $2/6$ is greater than its relative frequency of $2/18$ in the majority class. Otherwise stated, *within the class-size context, the point x_4 is more representative for the minority class than for the majority class. This idea is captured by the fuzzy classifier and makes it suitable for imbalanced data sets*.

IV. Performance Evaluation

When learning classes, even for balanced data sets, for which the errors coming from different classes have different costs, the overall accuracy is not a good measure of the classifier performance. Even more, when the class distribution is highly imbalanced, the accuracy is biased to favor the majority class and does not value rare cases as much as common cases. Therefore, it is more appropriate to use as performance evaluation measure the ROC (Receiving Operator Characteristic) curves. The ROC curves provide a visual representation of the trade-off between true positives (TP) and false positives (FP) as expressed in the Equations 2 and 3. The *confusion matrix* shown in Table III contains information about actual and predicted classification done by a classification system.

TABLE III
THE CONFUSION MATRIX.

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

$$TP = \frac{d}{c + d} \tag{2}$$

$$FP = \frac{b}{a + b} \tag{3}$$

However, for the purpose of comparing the results of this study with results published in [5], accuracy is also used as a measure to evaluate a classifier, in addition of the ROC curves. The fuzzy sets obtained with the procedure indicated previously in this paper are discrete fuzzy sets. However, their evaluation is required on unseen points. The standard approach to this problem is to extend the discrete fuzzy set to a continuous version by *piecewise linear interpolation*. More precisely, if x denotes a data point, and F a fuzzy set with membership μ_F, with support $S_F = \overline{\{x;\ \mu_F(x) > 0}} = \{x_i, i = 1, \ldots, n\}$, then the membership degree of x to F is given by

$$\mu_F(x) = \begin{cases} \mu_F(x_i) & x = x_i \\ \frac{\mu(x_{i+1}) - \mu(x_i)}{x_{i+1} - x_i} x + \frac{\mu(x_i) x_{i+1} - \mu_{i+1} x_i}{x_{i+1} - x_i} & x_i \le x \le x_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

V. Results and Analysis of the Study

All the results reported in this study are averaged over 30 runs and the test data reflect the natural distributions of the domains.

Figures 2 - 6 show the overall error percentage when different training class distributions are used. D_2 and D_3 outperform decision trees in four of the five domains studied here. For letter-vowel domain, D_2 and D_3 give less error only for class distributions higher than 50% (Figure 4). In Figures 7 - 11 are plotted the ROC curves of the four fuzzy classifiers, obtained for various class distributions. For all the five data sets D_2's ROC curve is dominant: it is above the other ROC curves and it is closer to the y axis.

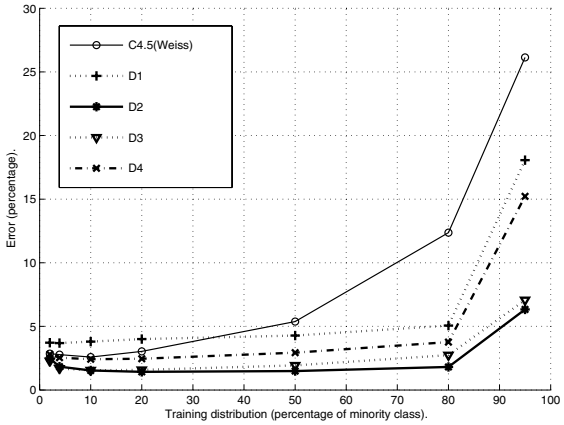

Fig. 2. Letter-a: the error in classification over various degrees of class distributions. Natural distribution is 3.9%.

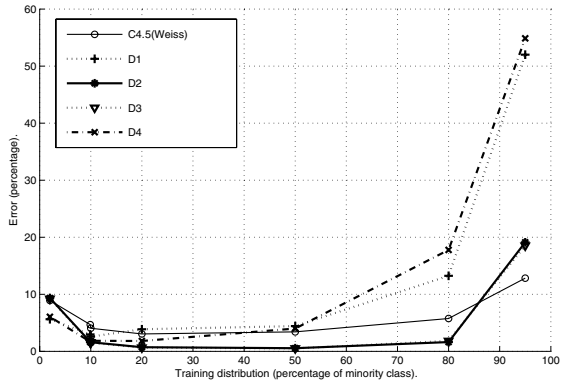

Fig. 3. OptDigits: the error in classification over various degrees of class distributions. Natural distribution is 9.9%.

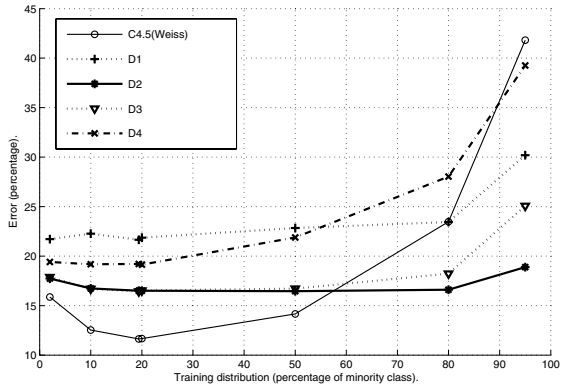

Fig. 4. Letter-vowel: the error in classification over various degrees of class distributions. Natural distribution is 19.4%.

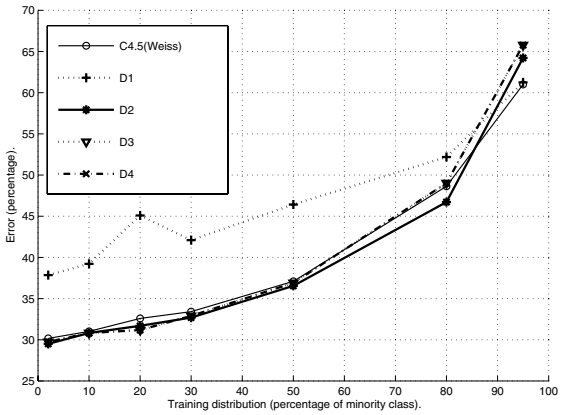

Fig. 5. German: the error in classification over various degrees of class distributions. Natural distribution is 30%.

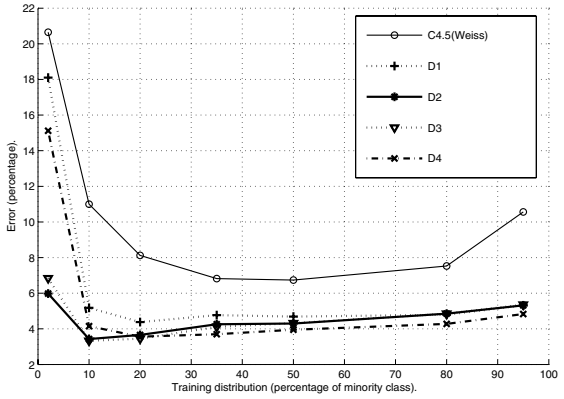

Fig. 6. Wisconsin: the error in classification over various degrees of class distributions. Natural distribution is 35%.

For the german data set, the trade-off between FP and TP is obvious (Figure 10): training with more Min examples introduces more false positives. The combination of two factors contributes to this behavior:

1) 21 attributes (out of 24) have exactly the same range of values for the Min and Maj classes (complete overlap) and the remaining three attributes overlap partially;
2) the natural class distribution (present in the test data) is 30%.

Therefore, when the classifier is trained with many Min examples, the recognition of the Min class (which makes 30% of the test set) improves, but at the cost of misclassifying much more Maj points, since the Maj class is present in testing with 70% of data. The analysis of Figure 5 (where the plain error is reported) leads to the same conclusion.

The letter-a domain presents naturally more imbalance (3.9%) than the letter-vowel domain (19.4%), though surprisingly, letter-a is better recognized (see Figures 2 and 4). This is mainly due to the fact that, the Min class for letter-a (instances of letter a) is better defined, as a concept, than the Min class for

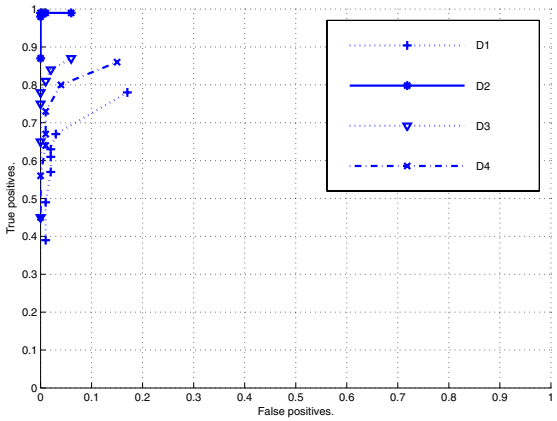

Fig. 7. Letter-a: the ROC curves obtained for the various class distributions. Natural distribution is 3.9%.

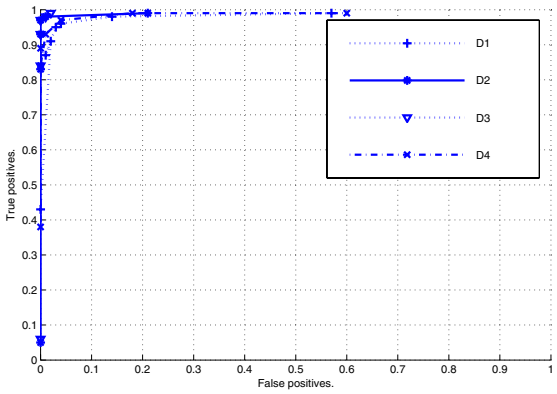

Fig. 8. OptDigits: the error in classification over various degrees of class distributions. Natural distribution is 9.9%.

letter-vowel (instances of a, e, i, o, u). In the same idea, there is more overlap between the classes in letter-vowel set than in the letter-a data set: letter-vowel domain has two attributes completely overlapped and in other 13 attributes (out of 16) has more overlap than the letter-a data set. The ROC curves are also consistent with the previous observation: they show indeed, a better (tighter) clustering of the letter-a Min class (Figure 7) than the letter-vowel (Figure 9).

Figure 3 shows that fuzzy classifier D_2 performs well in recognizing both the Min and Maj class for the optDigit domain. This domain has 64 attributes (of which, 24 attributes totally overlap) and a natural imbalance of 9.9%. A higher error when the training class distribution is 2%, is due to the fact that the Min class is not learned well and mainly Min class contributes to the error (for D_2, a ROC point on the y axis at 0.83). The increase in error for the class distribution of 95% is due to the fact that Maj class is under-represented in training and this time Maj class has a higher error rate. Though, the number of false positives does not grow much (Figure 8: for D_2, the ROC point is $(0.21.0.99)$).

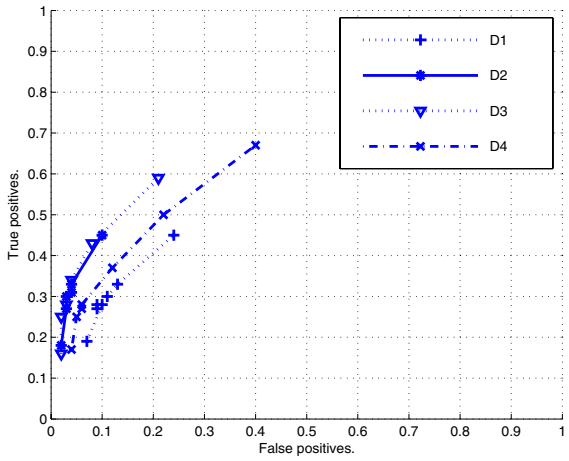

Fig. 9. Letter-vowel: the ROC curves obtained for the various class distributions. Natural distribution is 19.4%.

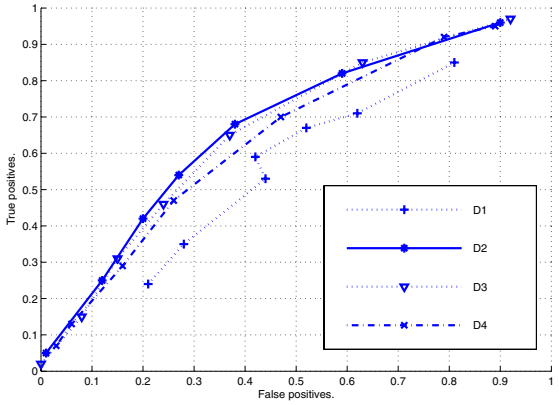

Fig. 10. German: the ROC curves obtained for the various class distributions. Natural distribution is 30%.

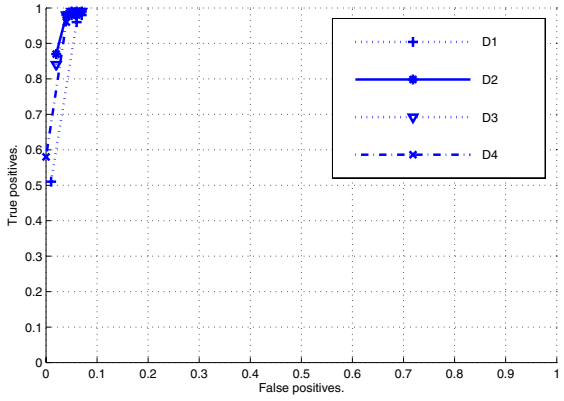

Fig. 11. Wisconsin: the ROC curves obtained for the various class distributions. Natural distribution is 35%.

TABLE IV

THE BEST CLASS DISTRIBUTIONS (AMONG THE ONES STUDIED HERE)FOR LEARNING TASK. DUE TO THE LACK OF SPACE, RESULTS FOR C4.5 AND D2 ONLY WERE REPORTED HERE.

Name	Natural class distribution	$C4.5$ (Weiss)	Fuzzy Classifier (D2)
letter-a	3.9%	10%	20%
optDigits	9.9%	20%	50%
letter-vowel	19.4%	19.4%	50%
german	30.0%	2%	2%
wisconsin	35.0%	50%	10%

The Maj and Min classes for the wisconsin data set overlap completely on three (out of nine) attributes. For this domain, it is interesting to investigate why the largest error is obtained when training class distribution is 2% (Figure 6). For this analysis, the ROC curves from Figure 11 are useful: the Maj class is well defined, as a concept (a high ROC point for class distribution 2%). By increasing the positive training examples, the false positives do not increase much (so Maj class is still recognized well) but a better recognition of the Min class is achieved.

In Table IV are presented for each of the five data sets, the training distributions which achieved the best accuracy in testing among the distributions studied here. It is obvious that not always the natural or the 50% (that means no imbalance) distribution gives the best generalization power: the fuzzy classifier D_2 achieved best generalization for a 50% distribution for only two domains (optDigits and letter-vowel) and C4.5 achieved the best distribution of 50% for wisconsin data and the best distribution of 19.4% for letter-vowel domain. Of course, the above "best distribution" discussion addresses only the distributions investigated here: we do not know the best learning distribution among all possible ones. The Table IV raises a natural question: why distributions greater than 50% do not give good results? The ranking of the distributions is a result of evaluating accuracies. Since the testing data respect the natural distributions of the data sets (which are naturally imbalanced, please see Table I), Maj class contributes more to the accuracy than the Min class. From this experiment we can also say that the issue of "best distribution for learning the data" is both, domain and classifier dependent.

VI. CONCLUSIONS AND FUTURE WORK

This study investigates the sensitivity of the fuzzy classifier to the learning distribution. This is possible by evaluating the classifier performance on the same test data (which respects the natural distribution), for various training distributions. The results show that the fuzzy classifier D_2 is less error prone than C4.5 and outperform C4.5 for the majority of class distributions used in training (Figures 2 - 6). The fuzzy classifier is less sensitive to the class imbalance: its output varies less than C4.5 over various training class distributions. Therefore, when the class distribution of the data set (or testing set) is not known, the fuzzy classifier is more robust as learner of imbalanced data than C4.5.

Since different classifiers learn in different ways, it will be interesting to investigate the performance of classifiers such as neural network, decision trees, minimum distance classifier, support vector machines and the fuzzy classifier presented here, for various distributions. Such an experiment will show generalization ability and limitations of each classifier. In a different direction, the aggregation rules for the fuzzy classifier require further investigations.

ACKNOWLEDGMENTS

This work was partially supported by the Ohio Board of Regents, Grant N000140310706 from the Department of the Navy, and the Rindsberg Graduate Fellowship of the College of Engineering, University of Cincinnati.

REFERENCES

[1] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 445–453.

[2] F. Provost, "Machine learning from imbalanced data sets 101 (extended abstract)," 2001. [Online]. Available: citeseer.ist.psu.edu/387449.html

[3] S. Visa and A. Ralescu, "Learning imbalanced and overlapping classes using fuzzy sets," in *Proceedings of the ICML-2003 Workshop: Learning with Imbalanced Data Sets II, Washington*, 2003, pp. 97–104.

[4] ——, "Fuzzy classifiers for imbalanced, complex classes of varying size," in *Proceedings of the IPMU Conference, Perugia*, 2004, pp. 393–400.

[5] G. M. Weiss, "The effect of samall disjuncts and class distribution on decision tree learning," PhD Thesis, Rutgers University, May 2003.

[6] J. Baldwin, T. Martin, and B. Pilsworth, "Fril-fuzzy and evidential reasoning in artificial intelligence," pp. 47–95, 1995.

[7] S. Visa, "Comparative study of methods for linguistic modeling of numerical data," MS Thesis, University of Cincinnati, December 2002.

[8] S. Visa and A. Ralescu, "Linguistic modeling of physical task characteristics," *Intelligent Systems for Information Processing: From Representation to Applications*, pp. 431–442, 2003.

[9] A. Inoue and A. Ralescu, "Generation of mass assignment with nested focal elements," in *Proceedings of 18th International Conference of the North American Fuzzy Information Processing Society*, 1999, pp. 208–212.

[10] A. Ralescu and D. Ralescu, "Extensions of fuzzy aggregation," *Fuzzy Sets and Systems*, vol. 86, no. 3, pp. 321–330, 1997.