Learning Imbalanced and Overlapping Classes using Fuzzy Sets

Sofia Visa SVISA@ECECS.UC.EDU
Anca Ralescu ARALESCU@ECECS.UC.EDU
Machine Learning and Computational Intelligence Laboratory, Department of ECECS, University of Cincinnati, Cincinnati, OH 45221-0030, USA

Abstract

This paper describes work in progress on the problem of concept learning in the presence of overlap and imbalance in the training set. A fuzzy set representation of the concept is adopted and class discrimination is achieved using a fuzzy classifier.

1. Introduction

Current classification algorithms assume that the data needed for classifier training are balanced, that is, that for a two class problem, about the same amount of data are available for each class. However, this is not always true. Moreover, as discussed in (Japkowicz, 2000) data imbalance poses some potentially important problems for deriving at least some types of classifiers.

A typical situation that leads to imbalance relates to the functioning of a system, whose failure may cause serious consequences. An important issue in the design of such a system is that of minimizing the probability of failures. Consequently, the events of smallest probability become the most important. Yet, failures do occur and data can be collected about the conditions of the system that resulted in failure. As described in (Narazaki & Ralescu, 1994) in some cases, human operators of such systems learn to recognize conditions leading to failure and act before failure occurs and, automatic systems can be trained to recognize them too.

Traditionally, a learning algorithm seeks to optimize some criterion. The system is usually evaluated according to various criteria, including learning curve, and prediction error which is based on the percentage of errors made on test data presented to the system.

Table 1. Variability of data between and within label for variable HD.

Label	Value
S	30 - 40
	40 - 55
	28 - 32
M	40 - 65
	50 - 60
	35 - 45
B	33 - 45
	50 - 59
	48 - 63

However, in the case of imbalanced data measuring the prediction error only by the percentage of errors may not sufficient: if it were, one would not need any learning and instead assume that all data belong to the large class, ensuring then a small percentage of error, the smaller in fact, the more the data is imbalanced. Other measures for performance, such as the F-measure may be be needed.

An additional starting point of this study is the observation that often classification errors occur near class boundaries. In addition, in many real life problems classes may overlap in the sense that some data points may appear as (valid) examples in both classes. Errors in this case, classifying such a data point to the big class, may have serious consequences (Pazzani et al., 1994), (Fawcett & Provost, 1997).

The above remarks lead to the idea that prediction should depend on a penalty or cost associated with an error. In the general situation described above, an error on a data point belonging to the small class would have a larger penalty than one for the large class. The extent to which this penalty should be larger is also

to be determined from the data. However, it can be argued that the effect of assessing penalties is equivalent to changing the relative data distribution in the two classes, or, in other words, to balancing the data.

2. Problem Description

Two main approaches are currently used for learning to classify imbalanced data:

- discrimination between the classes;
- recognition (learning) of one class (ignoring the other one).

In the first approach, examples and counter examples are used to train the system to discriminate between classes; the second method works like an associative memory, learning only the instances of one class. The problem is to decide which approach of the two mentioned above is more suitable for what type of data. Recent work done on this direction, (Japkowicz, 1999), concludes that the discrimination classifier performs better for the cases when the class to be recognized requires particularly strong specialization: if the examples from the class have large variance within the class, then the information gained by using the counter examples helps to discriminate between the classes. The second method is more suitable when the class to be learned is more 'tight' (i.e., there is not high variability between the members of the same class).

To deal with the imbalance problem, two methods are mostly used to rebalance them artificially:

- *up-sampling*: resample from the smaller class until meets the same number of data as the big class;
- *down-sampling*: eliminate data from the big class until the classes are balanced.

The current work proposes a *fuzzy set approach* to the problem of learning from imbalanced data. Fuzzy sets for class representation - *as a collection of data points and their corresponding membership values* - can be used to implement an approach in which the class membership functions are derived in a way that captures the contribution of each example to the corresponding class in a way that is correctly reflected by the prediction error. In addition, the fuzzy set based approach allows (but does not require) a setting in which classes may overlap. Related previous work includes (Narazaki & Ralescu, 1994), (Inoue & Ralescu, 1999), (Visa et al., 2003).

3. The Data Set

The data set used in this study is a real data set obtained for a study on assessing the perception of lifting tasks by manual workers. The variability within a class and the overlap between classes are not artificial and cannot (should not) be dismissed as they convey both the variability within each concept to be learned, and the variability among subjects.

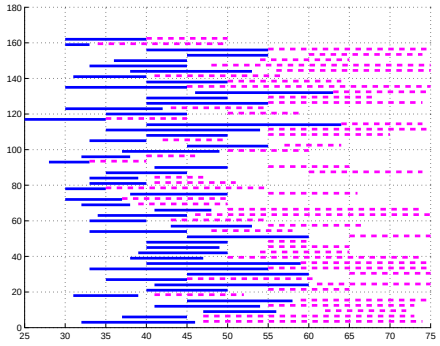

Figure 1. The high overlap between classes M and B: 54 random data from each of the classes M and B are plotted 1D.

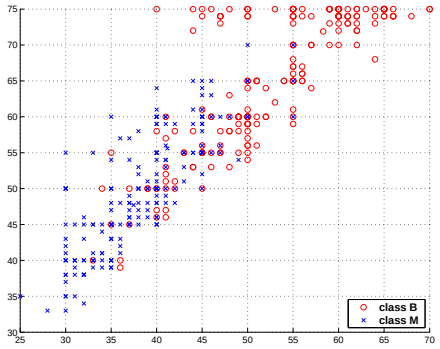

Figure 2. The high overlap between classes M and B: all data plotted 2D.

The data set comes from a survey on manual workers from Hong Kong and express the perceived difficulty level on lifting weights. A full and detailed description of the data set can be found in (Visa, 2002). Two hundred seventeen manual workers were required to imagine various lifting conditions (described verbally) on several aspects (variables) of the lifting task. Seven variables were used to describe the lifting task as follows: *Floor Weight* (FW), *Waist Weight* (W), *Horizontal Distance* (HD), *Twisting Angle* (TA), *Frequency*

(F), *Work Duration* (WD), *Vertical Distance* (VD). The values of the lifting tasks variables were assessed by each individual and labeled as *Small* (S), *Medium* (M) or *Big* (B) as illustrated in Table 1. In the current study only the data for class M and B for the variable HD were used. The total number of data used is 432: 216 from class M and 216 from class B. From a more complete study done for all the seven variable (Visa, 2002), it was observed that the data for variable HD had the highest overlap between classes and therefore the lowest accuracy in prediction (see Figure 1 for data plotted as segments and Figure 2 for data plotted as points in a two-dimensional space).

Throughout this study 108 data points from each of the two classes, labeled M and B are used for testing. **The training data are artificially imbalanced**: Class M is considered the 'Big' class and class B will be the 'Small' class. The overlap is controlled by eliminating points and resampling from the those which respect that degree of overlap which varies from 0% - 64% (64% is the real overlap between the classes over all data). For example, when the overlap is fixed to 42% only those data points for class B with first coordinate larger than 50 can be used and for a fixed value of 7% of overlap only those points from B for which the first coordinate has a value larger than 57 can be used. If there are not enough points for class B which respect the overlapping degree resampling from the ones which satisfy it is done until the desired size of the imbalanced class B is reached. For a given, fixed degree of overlap between the two classes, various degrees of imbalance are generated. The imbalance degree varies from 0% to 99% (in this case only one data point was used for class B to model its corresponding fuzzy set).

4. Current Approach

In a nutshell, the current approach consists of the following steps: given training data for each class, the membership functions of the corresponding fuzzy sets are derived under various conditions of imbalance and overlap. For each such derivation the F-measure and prediction errors of the fuzzy sets on the test data are assessed.

4.1. Deriving the Fuzzy Sets

Each class is represented as a fuzzy set on the data points used as examples of that class. The fuzzy sets are obtained from the relative frequency distributions for each class according to equation (1),

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

where $f_{(k)}$ and $\mu_{(k)}$ denote the kth largest value of the frequency distribution and membership function respectively. Equation (1) is derived as a particular case of a general procedure converting a relative frequency distribution into a fuzzy set (Ralescu, 1997), (Visa et al., 2003). Example 1 illustrates this procedure.

Example 1 *Suppose that a class C contains the following values: $C = \{x_1, x_2, x_1, x_3, x_2, x_1, x_1, x_3, x_3\}$. Written as a frequency distribution, $C = \{(x_1, 4), (x_2, 2), (x_3, 3)\}$ and again as a* relative frequency distribution *(in nonincreasing order) as $C = \{(x_1, 4/9), (x_3, 3/9), (x_2, 2/9)\}$. Then the membership values for x_i (also on nonincreasing order) are obtained as follows:*

$$\mu_{(1)} = \mu_C(x_1) = 1(4/9) + 3/9 + 2/9 = 1$$

$$\mu_{(2)} = \mu_C(x_2) = 2(3/9) + 2/9 = 8/9$$

$$\mu_{(3)} = \mu_C(x_3) = 3(2/9) = 6/9$$

4.2. Performance evaluation

In the testing phase, a data point x is classified to class $pred(x)$ given by (2).

$$pred(x) = argmax\{\mu_C(x);\ C \in \{B, M\}\} \qquad (2)$$

Two error models are used to evaluate the performance of the classifier (taking into account the overlap and imbalanced degrees).

First, for a test point x whose true class is $true(x)$ (either B or M) the simple error model of equation (3) is used to compute the prediction error:

$$error(x) = \begin{cases} 0 & \text{if } true(x) = pred(x) \\ 1 & \text{otherwise} \end{cases} \qquad (3)$$

It should be noted that, the class assignment strategy of (2) and error model of (3) are very conservative, and produce an upper bound on the prediction error. In the fuzzy set approach other error models can be devised based on the membership values $\mu_B(x)$ and $\mu_M(x)$ which are less drastic (for example, taking into account not only the ranking but the relative magnitude of these values).

A frequently used tool for assessing a classifier's performance for the small class is the confusion matrix associated with the classifier, shown here in Table 2. For a two-class classification problem (negative and

positive class) the *confusion matrix* contains information about actual and predicted classification done by a classification system. The entries of the confusion matrix have the following meanings:

- a is the number of correct negative predictions;
- b is the number of incorrect positive predictions;
- c is the number of incorrect negative predictions;
- d is the number of correct positive predictions.

Of interest here are the quantities P (*Precision*) and R (*Recall*) computed from the confusion matrix using equations (4) and (5) respectively.

$$P = \frac{d}{b+d} \tag{4}$$

$$R = \frac{d}{c+d} \tag{5}$$

Noticing that P and R cannot both increase or decrease together, they are combined through a convex combination with parameter $0 \leq \lambda \leq 1$, that is

$$\lambda \frac{d}{c+d} + (1-\lambda)\frac{d}{b+d}$$

Letting $\lambda = \frac{\alpha^2}{1+\alpha^2}$ the above can further be written as

$$\frac{\frac{\alpha^2}{1+\alpha^2}\frac{1}{P} + \frac{1}{1+\alpha^2}\frac{1}{R}}{\frac{1}{PR}}$$

which leads to equation (6).

$$F_\alpha = \frac{(\alpha^2+1)PR}{\alpha^2 P + R} \tag{6}$$

Selecting a particular value for α is problem-dependent and can be decided based on the cost of each type of errors. In this study F_1, when recall and precision are considered equally important, F_2, when recall is twice as imortant as precision, and $F_{0.5}$, when precision is twice as important as recall, are calculated.

4.3. Measuring the Overlap of Fuzzy Sets

The overlap between two (fuzzy) sets can be defined in several ways. In this study a measure of overlap known as *index of intersection* is used. In general, given two sets A and B, their index of intersection, $I(A,B)$, is defined as the size of their intersection relative to the size of their union, as described in equation (7).

Table 2. The Confusion Matrix.

	Predicted Negative	Predicted Positive
Actual Negative	a	b
Actual Positive	c	d

$$I(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

where $|A|$ denotes the cardinality, that is, the number of elements in the set A. Obviously,

$$I(A,B) = \begin{cases} 0 & \text{if } A \cap B = \emptyset \\ 1 & \text{if } A = B \\ < 1 & \text{if } A \subset B \text{ or } B \subset A \end{cases}$$

For fuzzy sets, such as B and M, the right hand side of (7) is defined in terms of the corresponding membership functions in two steps as follows:

1. Local calculation of the index of overlap:

$$\mu_{overlap}(x) = \frac{min(\mu_M(x), \mu_B(x))}{max(\mu_M(x), \mu_B(x))} \tag{8}$$

   where min and max are the standard fuzzy sets operators for intersection and union. $\mu_{overlap}(x)$ is the degree to which the fuzzy sets B and M overlap at x. The result of computing $\mu_{overlap}$ at all the data points is a fuzzy set.

2. Calculation of the cardinality of the *overlap* fuzzy set. The cardinality of a fuzzy set with n elements, is computed according to equation (9) first derived in (Ralescu, 1986).

$$\mu_{card}(k) = min(\mu_{overlap,k}, 1 - \mu_{overlap,k+1}) \tag{9}$$

   where $\mu_{overlap,k}$ denotes the kth largest value of the membership values for the set *overlap*, and the dummy values, $\mu_{overlap,0} = 1$ and $\mu_{overlap,n+1} = 0$ are introduced. The value $\mu_{card}(k)$ represents the degree to which B and M overlap on k points.

Example 2 illustrates the computation of the fuzzy sets for two classes (where data points are actually intervals) and of the overlap between these fuzzy sets.

Example 2 *Let the two classes be* $M = \{x_1, x_2, x_3\}$ *and* $B = \{x_4, x_5, x_6\}$, *where* $x_1 = [2,7]$, $x_2 = [2,5]$,

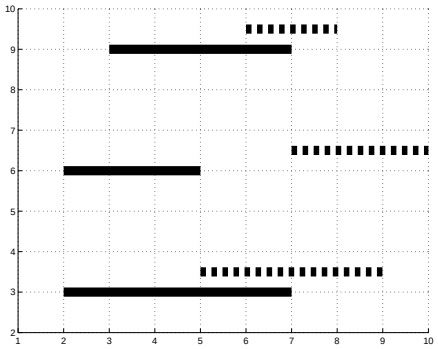

Figure 3. Data points of example 2.

Table 3. Membership function for classes M and B.

k	μ_M	μ_B
2	0.80	0
3	1	0
4	1	0
5	1	0.5
6	0.80	0.83
7	0.8	1
8	0	1
9	0	0.83
10	0	0.5

$x_3 = [3, 7]$, $x_4 = [5, 9]$, $x_5 = [7, 10]$ *and* $x_6 = [6, 8]$ *(see Figure 3). The computed frequencies for class* M *and* B *on the points* $\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ *are* $[2, 3, 3, 3, 2, 2, 0, 0, 0]$ *and* $[0, 0, 0, 1, 2, 3, 3, 2, 1]$ *respectively. The corresponding relative frequency distributions are* $[\frac{2}{15}, \frac{3}{15}, \frac{3}{15}, \frac{3}{15}, \frac{2}{15}, 0, 0, 0]$ *and* $[0, 0, 0, \frac{1}{12}, \frac{2}{12}, \frac{3}{12}, \frac{3}{12}, \frac{2}{12}, \frac{1}{12}]$.

The values of the membership functions for the fuzzy sets derived according to equation (1) and their assignments to the individual data points are shown in Table 3. The values of the membership functions for the interval data points, x_i *obtained according to equation (10) (this is a standard procedure in fuzzy sets theory to compute the degree* - possibility measure - *of a set based on the degrees of its elements) are as shown in Table 4 along with the degree of overlap.*

$$\mu(x_i) = sup_{k \in x_i} \mu(k) \tag{10}$$

For the data in example 2 the values of $\mu_{overlap}$ sorted in nonincreasing order (after $\mu_{overlap,0} = 1$, and $\mu_{overlap,7} = 0$ were added), are $[1, 1, 1, 1, 0.8, 0.8, 0.5, 0]$. Then the fuzzy cardinality of the *overlap* set is given by $[0, 0, 0, 0.2, 0.2, 0.5, 0.5]$ as shown in Table 5. According to this, the degree to

Table 4. Membership values to classes M and B and their overlap.

interval	μ_M	μ_B	$\mu_{overlap}$
x_1	1	1	1
x_2	1	0.5	0.5
x_3	1	1	1
x_4	1	1	1
x_5	0.8	1	0.8
x_6	0.8	1	0.8

Table 5. Fuzzy cardinality of the overlapped region.

no. of elements	0	1	2	3	4	5	6
μ_{card}	0	0	0	0.2	0.2	0.5	0.5

which B and M have exactly one, two or no points in common is 0; the degree to which they have exactly three or exactly four points in common is 0.2, and the degree that they have exactly five or exactly six points in common is 0.5.

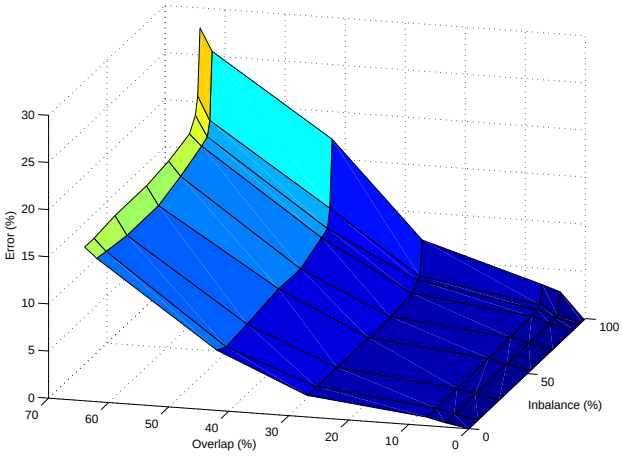

Figure 4. The error in classification as function of overlap and imbalance degree.

5. Results

The error prediction results for the fuzzy sets for the classes M and B respectively, computed according to the equation (7), under various degrees of overlap computed according to the equation (9) are summarized in Table 6. All the results are computed in average over 100 runs for each combination of the nine different levels of imbalance and seven levels of overlap.

It can be observed from the Table 6 and also from

Table 6. Error in classification over varying degrees of overlap(O) and imbalance(I).

I(%)	O (%)						
	0	4	7	27	42	62	64
0	0	.46	.9	2.3	6.4	15.2	16.3
8	0	.46	.92	2.3	5.8	15	16.3
26	0	.46	.92	2.2	6.1	14.6	16.6
53	0	.46	.95	2.2	6.7	14.6	16.6
72	0	.46	.96	2.2	6.6	15.5	17
90	0	.46	1.1	2.6	7.8	16.6	17.8
95	0	.46	1.2	3.4	8.2	17	19.2
97	0	.46	1.5	3.6	10.1	18.5	21
99	0	2.7	3.4	7.2	17.2	25.6	28

the Figure 4, that when the overlap and imbalance are high the error increased rapidly (see the region for overlap degree greater than 50% and for imbalance greater than 90%). In this region, for a fixed value of overlap the error grows fast with a small imbalance change. This suggests that, for this region, the degree of imbalance affects the classification performance more than the overlap of data. In the low region (low values of imbalance and overlap) the effect is exactly opposite: the overlap affects more than the imbalance at a lower rate.

Figure 5 shows results for the F_1-measure: it can be observed that F_1 decreases faster when imbalance degree increases than when the overlap increases. This confirms the previous observation, based on error rate, that the imbalance has a higer impact on errors than the overlap.

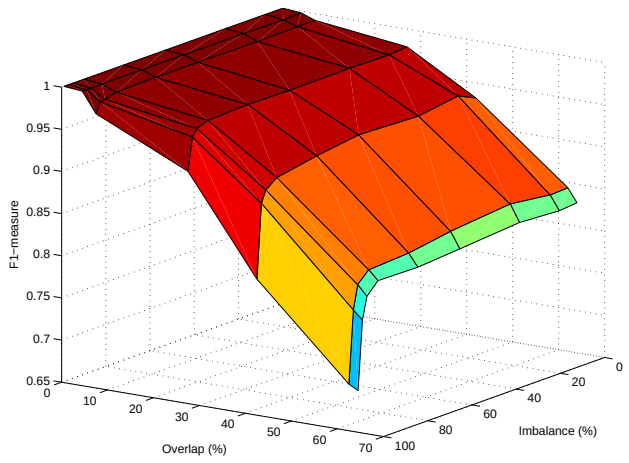

Figure 5. The error in classification as function of overlap and imbalance degree.

F_1, F_2, $F_{0.5}$-measures are computed for all combinations of the seven overlap and nine imbalance levels. Only four sets of results (others beeing similar) are shown in Figures 6-9 with imbalance in $[0\%, 99\%]$ and overlap in $[4\%, 64\%]$.

Comparing the three F-measures in figures 6 and 7, it can be concluded that the classifier performs better when using F_2 and $F_{0.5}$. The performance drops fast for overlap higer then 42% in the case of 99% imbalance (Figure 7) than when there is no imbalance (Figure 6).

For small degrees of overlap (Figure 8) all F-measures are high and vary little with the imbalance. The same small variation over imbalance degrees can be observed for highest degree of overlap (Figure 9), but with much smaller accuracy on F-measures. The largest drop occurs only at highest level of imbalance (99%) and highest overlap degree (64%).

In general, $F_{0.5}$-measure is more accurate then the other two F-measures when overlap is high. This can be observed from all plots (Figures 6 - 9).

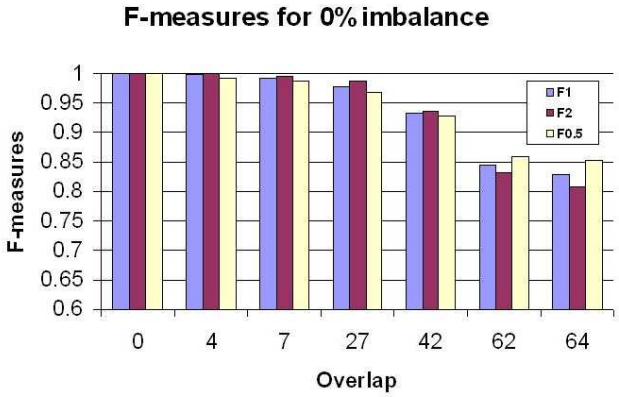

Figure 6. F-measures for 0% imbalance over the seven levels of overlap degree.

6. Conclusion

Initial results of a study on the *effect of imbalance and overlap* between classes for a fuzzy classifier were presented. These results, using prediction error and F-measures, support the idea that a fuzzy classifier can capture these features of the data set in a meaningful way. The fuzzy set approach can be viewed as a middle ground of the two main approaches mentioned in Section 2, to some extent classes are discriminated, to some extent they are each individually learned. Another feature of the approach presented is that no alteration of the data sets is required. Although, to begin with, the scenario of imbalanced data emerges

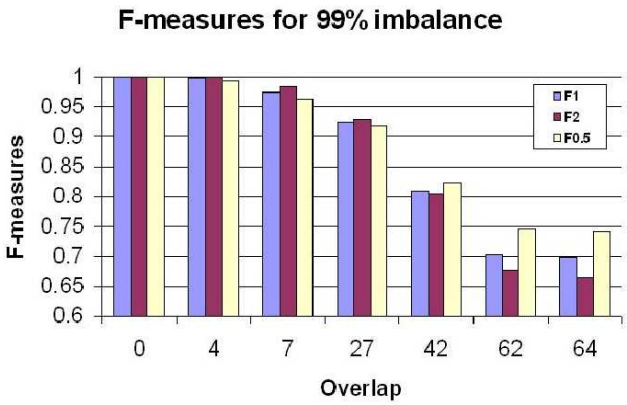

Figure 7. F-measures for 99% imbalance over the seven levels of overlap degree.

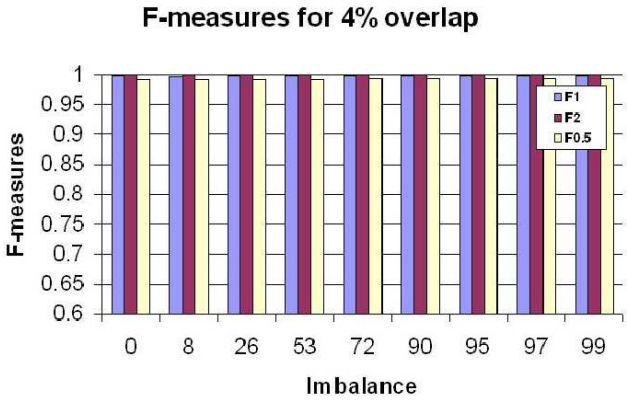

Figure 8. F-measures for 4% overlap over the nine levels of imbalance degree.

from situations in which this imbalance is inherent to the nature of the data, a possible new scenario is in connection with the concept of *web intelligence*. More precisely, in this scenario, learning is online and data may be *only temporarily imbalanced*. Additionally, over time, the imbalance may also change. This means that approaches that change the data sets may not necessarily be suitable. Instead, algorithms which can adapt the class representation (here the fuzzy set) are needed. Such adapting algorithms are part of the general approach of converting frequency (probability) distributions into fuzzy sets and updating the latter when more data is obtained are described in (Ralescu, 1997) and will be further explored in this context.

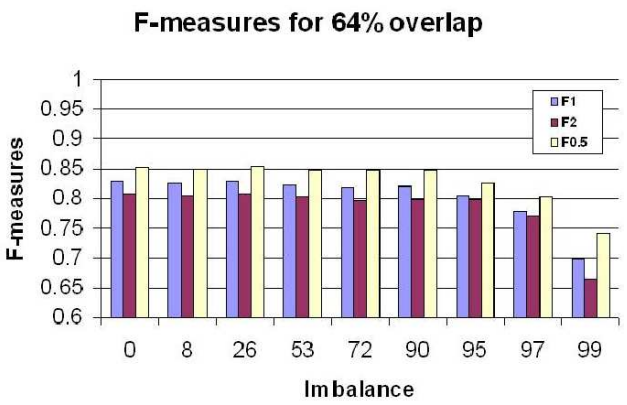

Figure 9. F-measures for 64% overlap over the nine levels of imbalance degree.

Acknowledgments

The authors wish to acknowledge the useful comments made by the three anonymous reviewers. Sofia Visa's work was partially supported by a Graduate Fellowship from Ohio Board of Regents. Anca Ralescu's work was partially supported by a JSPS Seniro Research Fellowship during May-June 2003, and by the Grant ONR N00014-03-1-0706.

References

Fawcett, T., & Provost, F. J. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery, 1*, 291–316.

Inoue, A., & Ralescu, A. (1999). Generation of mass assignment with nested focal elements. *Proceedings of 18th International Conference of the North American Fuzzy Information Processing Society (NAFIPS-99)* (pp. 208–212).

Japkowicz, N. (1999). Are we better off without counter examples? *Proceedings of the First International ICSC Congress on Computational Intelligence Methods and Applications (CIMA-99Learning from Imbalanced Data* (pp. 242–248).

Japkowicz, N. (2000). Learning from imbalanced data sets: A comparison of various strategies. *Proceedings of Learning from Imbalanced Data* (pp. 10–15).

Narazaki, H., & Ralescu, A. (1994). Iterative induction of a category membership function. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2*, 91–100.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). Reducing misclassification costs. *Proceedings of International Conference on Machine Learning (ICML-94)* (pp. 217–225).

Ralescu, A. (1986). A note on rule representation in expert systems. *Information Science*, *38.2*, 193–200.

Ralescu, A. L. (1997). Qualitative summarization of numerical data using mass assignment theory.*unpublished manuscript*.

Visa, S. (2002). Comparative study of methods for linguistic modeling of numerical data.

Visa, S., Ralescu, A., Yeung, S., & Genaidy, A. (2003). Linguistic modeling of physical task characteristics. In B. Bouchon-Meunier, L.Foulloy and R. Yager (Eds.), *Intelligent systems for information processing: from representation to applications*. Annecy: Elsevier.