Predicting Postgraduate Students' Performance Using Machine Learning Techniques

Maria Koutina[1], Katia Lida Kermanidis[1]

[1] Department of Informatics, Ionian University,
7 Pl. Tsirigoti, 49100 Corfu, Greece
{c09kout, kerman}@ionio.gr

Abstract. The ability to timely predict the academic performance tendency of postgraduate students is very important in MSc programs and useful for tutors. The scope of this research is to investigate which is the most efficient machine learning technique in predicting the final grade of Ionian University Informatics postgraduate students. Consequently, five academic courses are chosen, each constituting an individual dataset, and six well-known classification algorithms are experimented with. Furthermore, the datasets are enriched with demographic, in-term performance and in-class behaviour features. The small size of the datasets and the imbalance in the distribution of class values are the main research challenges of the present work. Several techniques, like resampling and feature selection, are employed to address these issues, for the first time in a performance prediction application. Naïve Bayes and 1-NN achieved the best prediction results, which are very satisfactory compared to those of similar approaches.

Keywords: Machine Learning, Student Performance Prediction, Class imbalance.

1 Introduction

The application of machine learning techniques to predicting students' performance, based on their background and their in-term performance has proved to be a helpful tool for foreseeing poor and good performances in various levels of education. Thereby tutors are enabled to timely help the weakest ones, but also, to promote the strongest. Apart from this, detecting excellent students can be very useful information for institutions and so forth for allocating scholarships. Even from the very beginning of an academic year, by using students' demographic data, the groups that might be at risk can be detected [1]. The diagnosis process of students' performance improves as new data becomes available during the academic year, such as students' achievement in written assignments and their in-class presence and participation. It has been claimed that the most accurate machine learning algorithm for predicting weak performers is the Naïve Bayes Classifier [1]. Other studies tried to detect attributes, including personality factors, intelligence and aptitude tests, academic achievement and previous college achievements, in order to make an accurate prediction about the final grade of the student. Some of the most significant factors in dropping out, as shown in these studies, are: sex, age, type of pre-university education, type of financial support, father's level of education, whether the student is a resident of the university town or not [2], [3], [4], [5]. Although these attributes may change as students move from bachelor to master studies, even more features may differ among students which come from different departments of high-level educational institutes.

This research has several contributions. First, some of the most well-known learning algorithms were applied in order to predict the performance of an MSc student in Informatics (Department of Informatics of the Ionian University in Greece) will achieve in a course taking into account not only his demographic data but also his in-term performance and in-class behaviour. Secondly, the impact of these features, and how it varies for different courses is studied with interesting findings.

Thirdly, an intriguing research aspect of the generated data is the disproportion that arises among the instances of the various class values. This problem, known as class imbalance, is often reported as on obstacle for classifiers [6]. Based on this, a classifier will almost always produce poor accuracy results on an imbalanced dataset [7], as it will be biased in favor of the overrepresented class, against the rare class. Researchers have contemplated many techniques to overcome the class imbalance problem, including resampling, new algorithms and feature selection [7], [8], [9] Resampling, feature selection and combinations of these approaches were applied to address the imbalance. To date, and to the authors' knowledge, no previous research work in marks prediction has combined feature selection with learning algorithms in order to achieve the best performance of the classifiers and to address the class imbalance problem. The results show that the use of learning algorithms combined with feature selection and resampling techniques provide us with more accurate results than simply applying some of the most well-known algorithms.

Finally, providing this information to tutors enables them to adjust their lesson depending on students' demographic and curriculum data and take timely precautions to support them.

The rest of this paper is organized as follows: Section 2 describes the data of our study. Section 3 describes the used learning algorithms and techniques of our study. Section 4 presents the experimental results. Section 5 follows with the results analysis. Finally, Section 6 ties everything together with our concluding remarks and some future research recommendations.

2 Data Description

The Department of Informatics of the Ionian University in Corfu launched a Postgraduate program in the year 2009, under the title "Postgraduate Diploma of Specialization in Informatics". The innovation of this program is that it poses no restrictions to the candidates' previous studies, and accepts graduates from any department (e.g. psychology, physics, economics/management departments e.t.c.). The main goal of this program is to educate graduate students of universities (AEI) and Technological Educational Institutes (TEI) in specialized areas of knowledge and research, in order to enable them to conduct primary scientific research and development tasks in the field of Information Technology.

A total of 117 instances have been collected. The demographic data were gathered from MSc students using questionnaires. Moreover, the in-term performance data of every student were given by the tutor of every course. The data of this study came from three courses of the first semester during the year 2009-2010, namely "Advanced Language Technology" (ALT) (11 instances), "Computer Networks" (CN) (35 instances) and "Information Systems Management" (ISM) (35 instances). Furthermore, the data were enriched with two more courses from the first semester of the year 2010-2011, namely "Advanced Language Technology" (ALT2) (8 instances) and "Computer Networks" (CN2) (28 instances),. Every course is an independent dataset, considering that in-term performance estimation differed among the courses. In some, a student had to submit at least one written midterm assignment, while, in others, midterm assignments were more than one. At this point it should be stressed that some written assignments were team-based while others were not.

The attributes (features) of the datasets are presented in Table 1 and 2 along with the values of every attribute. The demographic attributes represent information collected through the questionnaires from the MSc students themselves, concerning sex, age, marital status, number of children and occupation. Moreover, prior education in the field of Informatics, and the association between the students' current job and computer knowledge were additionally taken into consideration. For example, if a student had an ECDL (European Computer Driving License) that clarifies that (s)he is computer literate, then (s)he would qualify as a 'yes' in computer literacy. Furthermore, students who use software packages in their work (such as a word processor) and students who were graduates of Informatics departments are signified with a 'yes' in their job association with computers, whether they work part-time or full-time.

Table 1. Demographic attributes used and their values.

Demographic attributes	Value of every attribute
Sex	male, female
Age group	A) [21-25] B) [26-30] C) [31-35] D) [36- ..]
Marital Status	single, married
Number of children	none, one, two or more
Occupation	no, part-time, fulltime
Job associated with computers	no, yes
Bachelor	University, Technological Educational Institute
Another master	no, yes
Computer literacy	no, yes
Bachelor in informatics	no, yes

In addition to the above attributes, others, denoting possession of a second MSc degree, Informatics department graduates, and students who had a four- or five-year University degree or a four-year TEI degree, were also taken into account.

In-term performance attributes were collected from tutors' records concerning students' marks on written assignments and their presence in class. Finally, the results on the final examination were grouped into three categories: grades from zero to four were considered to be bad (failing grades), grades from five to seven were noted as good and grades from 8 to 10 were marked as very good.

Table 2. In-term performance attributes and their values.

In-term performance attributes	Value of every attribute
1st written assignment	0-10
2nd written assignment	0-10
3rd written assignment	0-10
Presence in class	none, mediocre, good
Final grade	bad, good, very good

3 Learning

Six classification algorithms have been used in our study, which are widely-used among the machine learning community [8]. All learners were built using WEKA (Waikato Environment for Knowledge Analysis[1]), a popular suite of machine learning software.

C4.5 decision tree learner was constructed [10] using J48 in WEKA, with the pruning parameter set both on and off. Three different k-nearest neighbors classifiers (denoted IBk in WEKA) were constructed, using k=1, 3 and 5 denoted 1NN 3NN, and 5NN. For these classifiers, all their parameters were left at default values. Experiments were also run using the Naïve-Bayes (NB) classifier, RIPPER (Repeated Incremental Pruning to Produce Error Reduction), which is a rule-based learner (JRIP in WEKA), Random Forest and, finally, a Support Vector Machines (SVMs) learner that uses the Sequential Minimal Optimization (SMO) algorithm for training and a polynomial kernel function.

3.1 The class imbalance problem

During the experimental procedure the problem of class imbalance arose, which is a challenge to machine learning and it has attracted significant research in the last 10 years [7]. As mentioned above, every course is an individual dataset, thus, the smallest dataset has eight instances whereas the biggest has thirty five. Additionally, in some courses it was noticed that final grades either between zero to four or eight to ten had less instances than grades between five to seven. Consequently, classifiers produced poor accuracy results on the minority classes.

Researchers have crafted many techniques in order to overcome the class imbalance problem. One simple technique is to obtain more samples from the minority class, which is not the optimal solution in real-world applications because

[1] http://www.cs.waikato.ac.nz/ml/weka/

the imbalance is a congenital part of the data [11]. Other sampling techniques include randomly undersampling the majority class [12], [13], oversampling the minority class [14], or combining over- and undersampling techniques in a systematic manner. Furthermore, a wide variety of learning methods have been created especially for this problem. These learners achieve this goal by learning using only positive data points and no other background information. One of these learners are SVMs. Last but not least, feature selection is a very promising solution for class imbalance problems; the goal of feature selection is to select a subset of *j* features that allows a classifier to reach optimal performance, where *j* is a user-specified parameter [7].

4 Experimental Setup

The training phase was divided into five consecutive steps. The first step included the demographic data along with the in-term performance data and the resulting class (bad, good, very good) for all datasets. Table 3 shows the accuracy of the predictions when training the initial data. In the second step the resample function was applied to the initial data. The resample function in WEKA oversamples the minority class and undersamples the majority class in order to create a more balanced distribution for training algorithms. Table 4 shows the accuracy of the predictions when training with re-sampled datasets.

Step three is performed with attribute evaluation, using the OneR algorithm, in order to identify which attributes have the greatest impact on the class in every dataset. OneR attribute evaluation looks at the odds of a feature occurring in the positive class normalized by the odds of the feature occurring in the negative class [6]. Studies have shown that OneR attribute evaluation proved to be helpful in improving the performance of Naïve Bayes, Nearest Neighbor and SMO [7]. During this phase it was found that attributes such as Bachelor in Informatics, Presence in class, sex and age have a great impact on the class, whereas others, like Marital Status and Another Master degree, are considered redundant.

During the fourth step, the best features (according to the results of the previous step) were selected and the learning algorithms that had the best accuracy results in step 1 were run on them. Table 5 shows the accuracy of J48 (unpruned), 1-NN, NB, Random Forest and SMO in all the datasets.

Table 3. Total accuracy (%) of the initial data.

Modules	J48 (pruned)	J48 (unpruned)	1-NN	3-NN	5-NN	NB	RF	SMO	J-Rip
ALT	54.54	54.54	72.72	63.63	54.54	72.72	72.72	72.72	54.54
CN	57.41	40.00	54.28	62.85	62.85	71.42	57.14	71.42	51.42
ISM	51.42	54.28	57.14	60.00	57.14	60.00	51.42	51.42	51.42
ALT2	25.00	25.00	25.00	37.50	25.00	37.50	37.50	25.00	25.00
CN2	57.14	60.71	60.71	50.00	53.57	60.71	57.14	57.14	67.82

In the final step the resample filter combined with 7 features that had great influence in all our datasets was applied (Table 6). The features with the highest influence in our datasets are: Bachelor in Informatics, presence in class, sex, age group, first written assignment, job association with Informatics, and number of children.

Table 4. Total accuracy (%) of re-sampled data.

Modules	J48 (pruned)	J48 (unpruned)	1-NN	3-NN	5-NN	NB	RF	SMO	J-Rip
ALT	63.63	81.81	90.90	54.54	45.45	72.72	90.90	72.72	54.54
CN	65.71	71.42	85.71	80.00	74.28	80.00	88.51	82.85	71.42
ISM	80.00	82.85	85.71	60.00	42.85	85.71	88.57	82.85	80.00
ALT2	62.50	62.50	87.50	62.50	67.50	87.50	87.50	87.50	37.50
CN2	82.14	82.14	100.00	64.28	67.85	85.71	96.42	89.28	71.42

Table 5. Total accuracy (%) of best features along with the initial data.

Modules	J48 (unpruned)	1-NN	NB	RF	SMO
ALT	54.54	72.72	81.81	63.63	72.70
CN	48.57	57.14	74.28	57.14	68.57
ISM	48.57	51.42	62.85	54.28	62.85
ALT2	25.00	12.50	50.00	12.50	25.00
CN2	42.85	53.57	57.14	42.85	67.85

Table 6. Total accuracy (%) of re-sample data and feature selection.

Modules	J48 (unpruned)	1-NN	NB	RF	SMO
ALT	63.63	90.90	90.90	90.90	90.90
CN	65.71	71.42	82.85	74.28	60.00
ISM	68.57	80.00	80.00	80.00	74.28
ALT2	62.50	100.00	100.00	87.50	87.50
CN2	67.85	71.42	71.42	71.42	64.28

5 Results and Discussion

Unlike previous attempts related to grade prediction that focus on a single algorithm and do not perform any form of feature selection, the overall goal of this paper is to find the best combination of learning algorithms and selected features in order to achieve more accurate prediction in datasets with an imbalanced class distribution and a small number of instances.

One important conclusion that can be drawn from the results is that J48 prediction accuracy is much lower than that of NB and 1-NN, which shows that in small datasets NB and 1-NN can perform better than decision trees. Another interesting issue is that the accuracy of prediction, using either the re-sample datasets alone or feature selection combined with resampling, improves when the original data is much smaller in size (ALT, ALT2).

Comparing Table 3 with Table 4, the predictions using the re-sample dataset are significantly more accurate. Furthermore, there is a significant improvement in the accuracy of NB, 1-NN, Random Forest and SMO learning algorithms. For example, the accuracy of module ALT2 using 1-NN with the initial data is 25%, whereas, by using the re-sample technique accuracy rises up to 85.7%. The results (Table 5) show that using only the best features proves to be helpful for NB, but does not help the overall accuracy of the remaining classifiers. Additionally, comparing the results of Table 4 with Table 6, it is evident that NB is the only classifier that is being helped by feature selection. For example in module ALT2 accuracy is noticeably improved when using resampling and feature selection. The above results follow the findings of [2] on how different classifiers respond to different distributions of data.

5.1 Detailed analysis of the results

Trying to take a deeper look at the obtained results, it is presented in figure 1 to 5 in detail the f-measure for all class values in all our datasets, throughout the first, second and fifth steps, including only the learning algorithms with predominately best results.

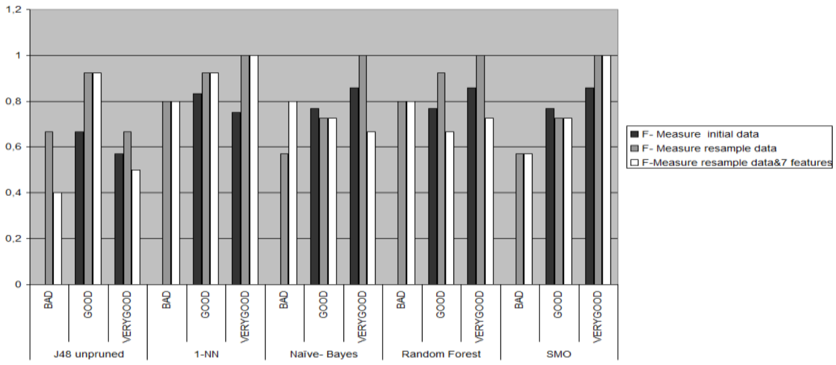

Fig. 1. F-Measure for module Advanced Language Technology (ALT)

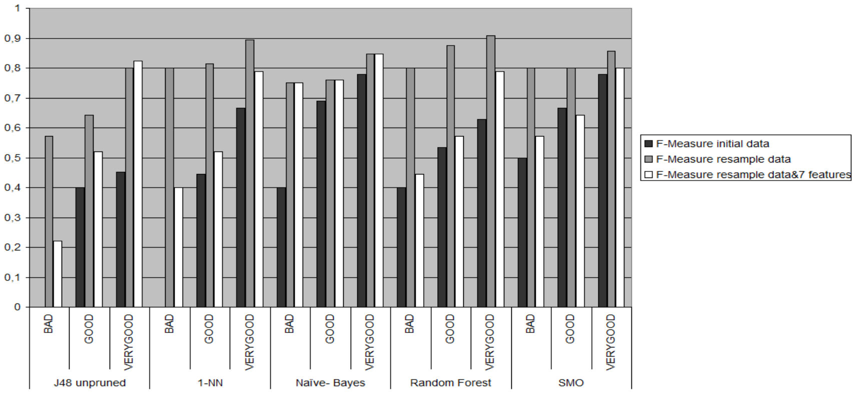

Fig. 2. F-Measure for module Computer Networks (CN)

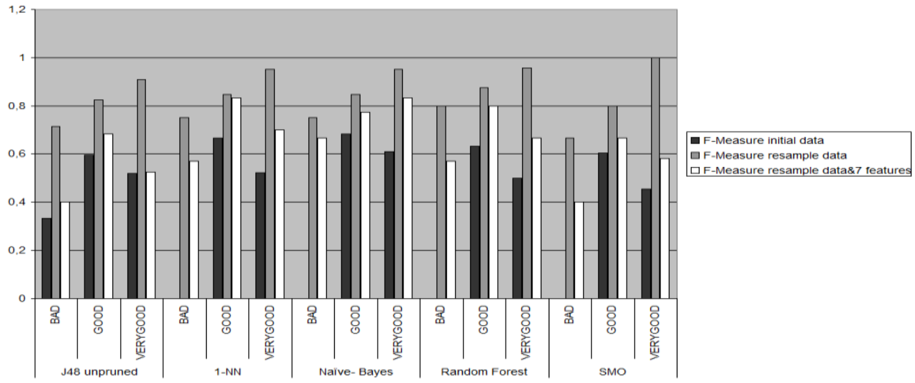

Fig. 3. F-Measure for module Information Systems Management (ISM)

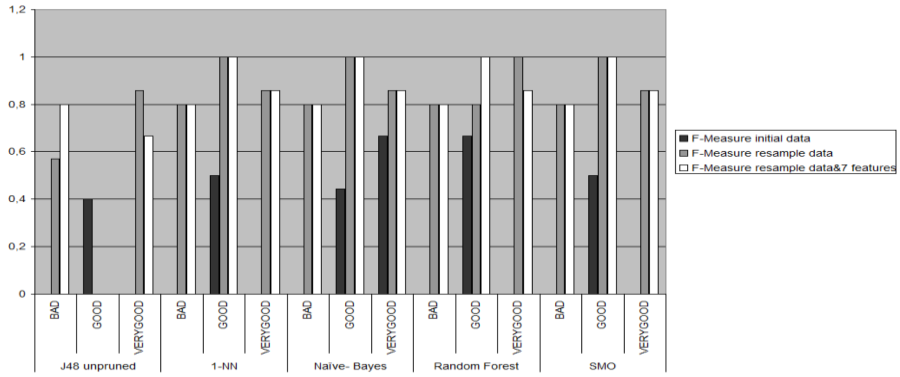

Fig. 4. F-Measure for module Advanced Language Technology 2 (ALT2)

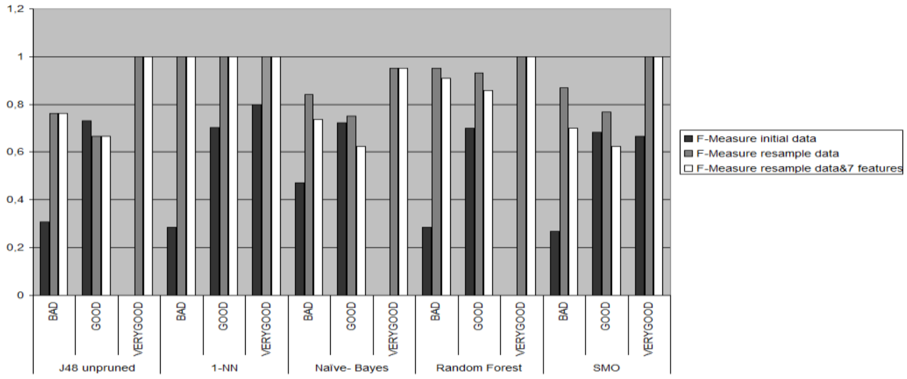

Fig. 5. F-Measure for module Computer Networks 2 (CN2)

The above figures show that in all cases both the resample technique and the combination of feature selection with resampling significantly improve the prediction of minority classes. For example, in module ALT both with NB and with 1-NN, the f-measure for class value *Bad*, using either resampling alone or in combination with feature selection, rises significantly from 0 to 0.8. Moreover, careful inspection of the f-measure in all datasets reveals that the highest scores are always achieved for the majority class label, which is *Good*.

5.2 Feature selection results

Apart from the most suitable method in order to predict postgraduate students' performance, it was also attempted to find which attributes influence the most the selected classifiers. The most important attributes in all the datasets were *Presence in class* and *Bachelor in Informatics,* which shows, first, that in-term performance of a student highly affects his final grade, and, secondly, that students who don't have a degree in Informatics are at risk. Another important issue is that in modules, ALT-ALT2 and ISM there is no significant influence from features like Job associated with computers and Computer literacy whereas those attributes were important in modules CN and CN2. Hence, it is believed that if actual predictions are required, it would be better splitting datasets into technical and non-technical lessons and apply on them the same algorithms but with difference selected features.

6 Conclusion

Machine learning techniques can be very useful in the field of grade prediction, considering that they enable tutors to identify from the beginning of the academic year the risk groups in their classes. Hence, this will help them adjust their lesson in order to help the weakest but also to improve the performance of the stronger ones.

An interesting finding from this research work is that NB and 1-NN, combined with resampling alone, or in combination with feature selection, accurately predict the students' final performance, given our datasets, especially when these include a small number of instances.

The overall prediction accuracy in our analysis varies from 85.71% to 100%, learning a discrete class that takes three values. Results are more than promising and enable the future implementation of a student performance prediction tool for the support of the tutors in the Informatics Department of the Ionian University. Furthermore, extending the above tool using regression methods which will predict the exact grade of the student may provide even more fine-grained support to tutors.

Another interesting issue in our study is that the average accuracy of the learning algorithms can be improved by feature selection. It was found that students' occupation, type of bachelor degree (AEI or TEI), and their possession of another master degree do not improve accuracy. Thus, this information is not necessary.

However, students' presence in class and their possession of a Bachelor degree in Informatics proved to be very important for the classifiers.

Acknowledgments

The authors would like to thank the tutors and students who participated in this study for their significant contribution.

References

1. S. Kotsiantis, C. Pierrakeas ,P. Pintelas, "Predicting Students Performance in Distance Learning Using Machine Learning Techniques", Applied Artificial Intelligence (AAI), Volume 18, Number 5/ May- June 2004, pp. 411-426 (2004)
2. Parmentier, P. La reussite des etudes universitaires: facteurs structurels et processuels de la performance academique en premiere annee en medecine. PhD thesis, Catholic University of Louvain (1994)
3. Touron, J. The determination of factors related to academic achievement in the university: implications for the selection and counseling of students, Higher Education 12, p. 399-410 (1983)
4. Lassibille, G., Gomez, L. N. Why do higher education students drop out? Evidence from Spain, Education Economics 16(1), p. 89-105 (2007)
5. Herzog, S. Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. In Proc. of 44th Annual Forum of the Association for Institutional Research (AIR), (2004)
6. G. Forman and I. Cohen, "Learning from Little: Comparison of Classifiers Given Little Training," Proc. Eighth European Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 161-172 (2004)
7. M. Wasikowski, X. Chen, "Combining the Small Sample Class Imbalance Problem Using Feature Selection, IEE Computer Society, Volume 22, pp 1388-1400 (2010)
8. J.V. Hulse, T.M. Khoshgoftaar, A. Napolitano, "Experimental Perspectives on Learning from Imbalanced Data", Appearing in Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, (2007)
9. R. Barandela, R. M. Valdovinos, J. S. Sanchez, & F. J. Ferri, "The imbalanced training sample problem: Under or over sampling?" In Joint IAPR In- ternational Workshops on Structural, Syntactic, and Statistical Pattern Recognition (SSPR/SPR'04), Lec-ture Notes in Computer Science 3138, 806–814 (2004)
10. J. R. Quinlan, "C4.5: Programs for machine learning." San Mateo, California: Morgan Kaufmann (1993)
11. N. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 1-6 (2004)
12. M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Data Sets: One Sided Sampling," Proc. 14th Int'l Conf. Machine Learning, pp. 179-186 (1997)
13. X. Chen, B. Gerlach, and D. Casasent, "Pruning Support Vectors for Imbalanced Data Classification," Proc. Int'l Joint Conf. Neural Networks, pp. 1883-1888 (2005)
14. M. Kubat and S. Matwin, "Learning When Negative Examples Abound," Proc. Ninth European Conf. Machine Learning (ECML '97), pp. 146-153 (1997)