



Institute of Computer Science
Academy of Sciences of the Czech Republic

Trends in random forests parameters for classification of imbalanced data

Marko Robnik-Šikonja, Petr Savický

Technical report No. 1153

February 2012



Institute of Computer Science
Academy of Sciences of the Czech Republic

Trends in random forests parameters for classification of imbalanced data

Marko Robnik-Šikonja¹, Petr Savicky²

Technical report No. 1153

February 2012

Abstract:

The classification with imbalanced class proportions is a particularly difficult problem. We investigate the classification of imbalanced data sets with random forests method, which is one of the state-of-the-art classifiers. To this aim we created a semi-artificial data generation engine which for a supplied real world data set estimates its joint probability distribution with an RBF network and generates new data from this distribution at any proportion of imbalance required. This engine allowed us to systematically and consistently vary level of imbalance for approximations of several real world data sets and to study the parameters of the learning algorithms which influence the classification performance. The results show that consistently across different data sets and imbalance levels, there are notable trends in settings of the stopping criterion and the level of smoothing for an improved performance. These findings are confirmed on large UCI data sets, where imbalance can be observed naturally without modeling.

Keywords:

Machine learning, imbalanced data, classification trees, random forests

¹University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, 1001 Ljubljana, Slovenia, Marko.Robnik@fri.uni-lj.si

²Institute of Computer Science, Academy of Sciences of Czech Republic Pod Vodarenskou Vezi 2, 182 07 Praha 8, Czech Republic, savicky@cs.cas.cz

Trends in random forests parameters for classification of imbalanced data

Marko Robnik-Šikonja¹, Petr Savicky²

¹ University of Ljubljana,
Faculty of Computer and Information Science,
Tržaška 25, 1001 Ljubljana, Slovenia
Marko.Robnik@fri.uni-lj.si

² Institute of Computer Science,
Academy of Sciences of Czech Republic
Pod Vodarenskou Vezi 2, 182 07 Praha 8, Czech Republic
savicky@cs.cas.cz

Abstract

The classification with imbalanced class proportions is a particularly difficult problem. We investigate the classification of imbalanced data sets with random forests method, which is one of the state-of-the-art classifiers. To this aim we created a semi-artificial data generation engine which for a supplied real world data set estimates its joint probability distribution with an RBF network and generates new data from this distribution at any proportion of imbalance required. This engine allowed us to systematically and consistently vary level of imbalance for approximations of several real world data sets and to study the parameters of the learning algorithms which influence the classification performance. The results show that consistently across different data sets and imbalance levels, there are notable trends in settings of the stopping criterion and the level of smoothing for an improved performance. These findings are confirmed on large UCI data sets, where imbalance can be observed naturally without modeling.

1 Introduction

A two-class data set for classification is called imbalanced, if most of the cases belong to one of the classes, called the majority class, and only a small fraction of cases belongs to the other class, called the minority class. It is well-known that learning a classification tree or an ensemble of trees with an imbalanced training set is not effective. Different over and under sampling strategies are used in order to balance the training

set and thus improve the quality of the obtained model, see [2] and the references therein. A more recent study of classification trees for imbalanced data is [3].

In this paper, we investigate the setting of parameters of random forest classifier, which influence its performance on imbalanced data as well as the balancing techniques. A proper setting of these parameters can be combined with the sampling strategies to provide an efficient learning algorithm. The investigated parameters include the stopping criterion and the level of smoothing. A more detailed description of the parameters is in Section 3.

1.1 Related work

In order to assess the interaction of the investigated parameters with the imbalance of the training data, we construct classifiers for different class counts for training. The list of the pairs of class counts used is presented in Section 2.3. Using the default setting of the stopping criterion and without smoothing, the quality of the obtained classifier deteriorates with increasing imbalance. It appears, however, that the effect of large imbalance may be compensated by an appropriate setting of the stopping criterion and smoothing. These results are presented in Section 4.

2 Data sets

The main scientific instrument of empirical research in machine learning and data mining is the availability of appropriate data sets. For our study, where realistic data sets with large imbalances are rare, it is particularly important to assure the experiments will be relevant to the practitioners. To this aim we used two types of data sets. The first choice are the actual publicly available highly imbalanced data sets. Unfortunately not many of them exist, so we created a data generation engine which creates semi-artificial data with many characteristics of real world data sets but also the ability to select level of class imbalance. In this section we describe both types of data sets we used in our study.

2.1 Large UCI data sets

We use the five data sets from [4], whose abbreviations and full names are as follows.

abbreviation	full name
MiniBooNE	MiniBooNE particle identification
RecordLinkage	Record Linkage Comparison Patterns
census-income	Census-Income (KDD)
census1990	US Census Data (1990)
covtype	Covertypes

The used data sets have the following numbers of attributes.

name	attributes
MiniBooNE	50
RecordLinkage	11
census-income	42
census1990	67
covtype	54

The class counts for the data sets and the way, how each class was used is specified in the next table. Some of the data sets are not imbalanced. We used stronger subsampling of one of the classes in order to make it the minority class.

data set	class	count	usage
MiniBooNE	background	93565	majority
	signal	36499	minority
RecordLinkage	FALSE	5728201	majority
	TRUE	20931	minority
census-income	50000+.	18568	minority
	- 50000.	280717	majority
census1990	0	2168997	majority
	1	236403	minority
	2	52885	minority
covtype	1	211840	minority
	2	283301	majority
	3	35754	
	4	2747	
	5	9493	
	6	17367	
	7	20510	

2.2 Semiartificial data sets

In order to obtain generators for arbitrarily large data sets, which are similar to some of the standard benchmark data sets, say D_1 , we use the following strategy. The original two-class data set D_1 is approximated using PRBF model [8, 9], which is a mixture of gaussians. The PRBF model is used to generate new data D_2 of the same size. Then, each of the data sets D_1 , D_2 , resp. is used to construct a random forest classifier. The error of the classifier trained on D_1 is measured on D_2 as a test set on each class separately. Similarly, the error of the classifier trained on D_2 is measured on D_1 . If all the four obtained errors are at most 0.15, the PRBF model is considered an adequate approximation for our experiments.

The following table summarizes the names of the standard data sets used for PRBF approximation and their origin. The data sets were obtained from [4], [1] and [5] repositories and mostly the data sets used also in [3].

name	repository
pima	UCI
page	UCI
segment	UCI
letter	UCI
pendigits	UCI
german_numeric	UCI
breast_wdbc	UCI
satimage	UCI
fourclass	LIBSVM
svmguidel	LIBSVM
splice	LIBSVM
phoneme	ELENA

For each data set, the list of all classes as they appear in the file is presented in the next table together with the choice of the minority and the majority class. This choice was obtained by considering all pairs among the first at most seven classes and the hardest combination was used. In some cases, this leads to a reversed assignment of the minority and majority class than in the real problem.

name	classes	minority	majority
pima	0, 1	0	1
page	1, 2, 3, 4, 5	5	1
segment	1, 2, 3, 4, 5, 6, 7	5	3
letter	A, B, C, D, E, F, G, ...	G	C
pendigits	0, 1, 2, 3, 4, 5, 6, ...	3	4
german_numeric	1, 2	1	2
breast_wdbc	B, M	M	B
satimage	1, 2, 3, 4, 5, 7	3	7
fourclass	-1, +1	-1	+1
svmguidel	0, 1	0	1
splice	-1, +1	+1	-1
phoneme	0, 1	0	1

2.3 Sampling strategy

The training data sets for our experiments with large UCI data were obtained by subsampling of the original data set. For the semiartificial data, we generate a new independent sample for each run. The class counts for the training sets were chosen from the following table. Note that in the abbreviations dxy , the x and y are exponents of the number of cases in minority and majority class, respectively, e.g., $d23$ stand for $10^2 = 100$ minority class instances and $10^3 = 1000$ majority class instances.

The original UCI data sets were mostly quite easy for random forest classifier, with AUC close to 1. In order to make the problems harder, we used a random subset of the attributes. The number of selected attributes was chosen empirically so that the typical AUC is between 0.7 and 0.9.

Table 1: The abbreviations for class counts used in our study. In the abbreviations dxy , the variables x and y are exponents of number of cases in minority and majority class.

abbreviation	minority	majority
d22	100	100
d23	100	1000
d24	100	10000
d25	100	100000
d33	1000	1000
d34	1000	10000
d35	1000	100000

3 Investigated parameters of random forest

For our experiments, we use CORElearn [7] extension package to R Environment for Statistical Computing [6].

We consider stopping criterion defined by limiting the number of cases in any node of each tree of the forest from below by a parameter W . This criterion is implemented by “minNodeWeight” option of the function CoreModel(). See “help(helpCore)” for more detail.

We used m-estimate smoothing where m is determined, so that $S = m \cdot p_c$, where S is a smoothing parameter and p_c is the prior probability of the least probable class, see [10]. This is implemented in the predict method for CoreModel objects (predict.CoreModel()) using option “smoothingType=4”. The level of smoothing S is then specified by the option “smoothingValue”.

The evaluation measures information gain and its uniform variant were used. In the function CoreModel(), they are set using the option selectionEstimator with values “UniformInf” or “InfGain”.

4 Results

4.1 Results for the default and the best settings

In this section, we compare the classifiers, which are obtained using different class counts and two types of the setting of the stopping criterion W (minNodeWeight) and S (smoothing parameter). By the default setting, we mean $W = 1$ and $S = 0$, which represents stopping at leaves of size 1 and no smoothing. By the best setting, we mean the setting obtained as follows. All combinations of $W = \text{minNodeWeight}$ and the smoothing parameter S from

$$W \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000\} \quad (1)$$

and

$$S \in \{0, 0.1, 0.31623, 1, 3.1623, 10, 31.623, 100\} . \quad (2)$$

are used to construct a classifier and the setting, which yields the best AUC among them is determined.

4.1.1 Large UCI data sets

In this section, we compare AUC obtained for different training counts and different settings of the parameters W and S . The next two tables present the comparisons for the default setting $W = 1$ and $S = 0$.

comparison of d22 and d25	>	=	<
RecordLinkage	10	0	0
census-income	10	0	0
census1990	6	2	2
covtype	10	0	0
total	36	2	2

comparison of d33 and d35	>	=	<
RecordLinkage	9	0	1
census-income	10	0	0
census1990	5	3	2
covtype	2	0	8
total	26	3	11

The tables demonstrate that with the default setting, training with class counts $(10^2, 10^2)$ yields typically a better AUC than class counts $(10^2, 10^5)$ despite that with the latter class counts, the training set contains more information. The large imbalance in the latter class counts does not allow to take advantage of more cases of the majority class. Similarly, the class counts $(10^3, 10^3)$ mostly yield better AUC than the class counts $(10^3, 10^5)$ except for the data set covtype.

The next two tables demonstrate that the disadvantage of large imbalance may be compensated to a large extent by an appropriate setting of the stopping criterion and smoothing. For each choice of the training class counts, the best setting of W and S among the values (1) and (2) as described at the beginning of Section 4.1 was determined and the corresponding AUC was used for comparisons.

comparison of d22 and d25	>	=	<
RecordLinkage	7	0	3
census-income	4	0	6
census1990	1	4	5
covtype	0	0	10
total	12	4	24

comparison of d33 and d35	>	=	<
RecordLinkage	3	0	7
census-income	5	0	5
census1990	2	4	4
covtype	0	0	10
total	10	4	26

For the best setting of the parameters, training class counts $(10^2, 10^5)$ provide mostly better AUC than class counts $(10^2, 10^2)$. Similarly the class counts $(10^3, 10^5)$ yield mostly better AUC than class counts $(10^3, 10^3)$.

4.1.2 Semiartificial data sets

The comparison of AUC obtained for random forests trained with d22 and d25 and with d33 and d35 demonstrates a similar dependency on the setting of the parameters W and S as the one observed for large UCI data. The results of the comparison of AUC for d22 and d25 and for d33 and d35 for semiartificial data is summarized in the following tables.

comparison of d22 and d25	>	=	<
default setting	11	0	1
best setting	0	0	12

comparison of d33 and d35	>	=	<
default setting	8	0	4
best setting	0	0	12

For the default setting, the result for balanced training is clearly better than for imbalanced training. For the best setting of the parameters, we get the opposite.

4.2 The influence of stopping criterion and smoothing

Examples of the maps of the dependence of AUC on W and S parameters.

4.3 The influence of imbalance

The plots presented in the previous section suggest that the default setting $W = 1$ and smoothing with $S = 0$ is outperformed by settings with higher W or S . In this section, we present statistics, which confirm this hypothesis. Besides of the default setting, for each training sample, a random forest was grown for each combination of the parameters $W = 1, 2, 5, 10$ and $S = 0, 0.1, 0.31623, 1$, which satisfies $W \geq 5$ or $S \geq 0.31623$. The AUC for the 12 combinations obtained in this way and AUC for the default setting are sorted in a decreasing order and the ranks are assigned to them. Since the order is decreasing, the largest AUC gets rank one. The rank of the result for the default setting is then tabulated.

4.3.1 Large UCI data sets

For large UCI data, the comparison was performed separately for each data set, for each of the 10 random subsets of attributes and for each

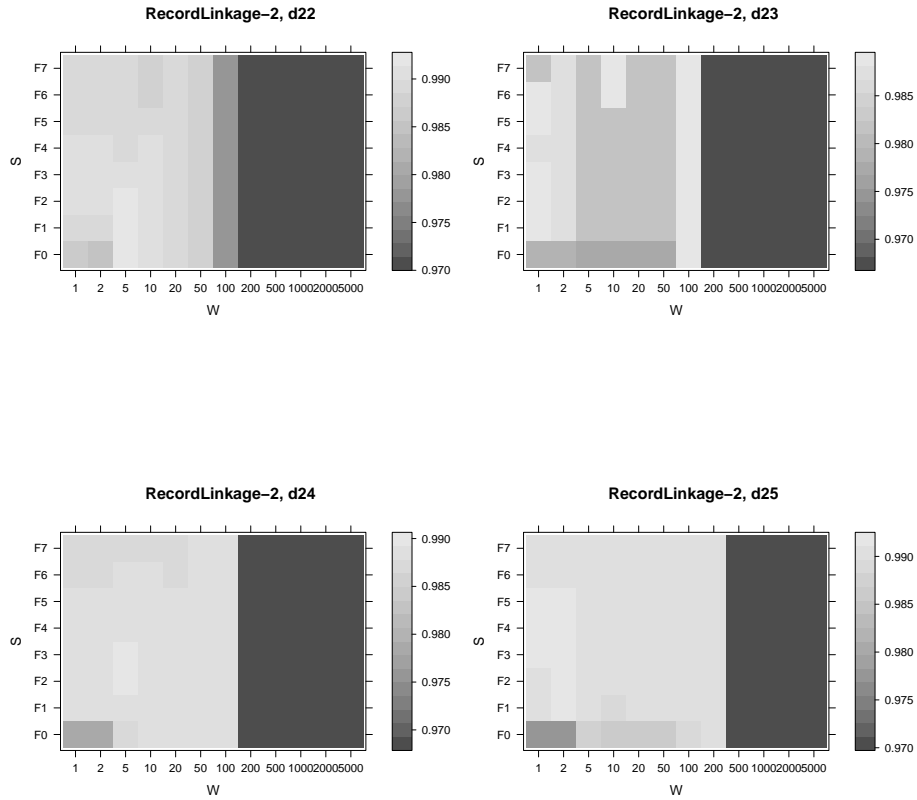


Figure 1: Levelplots of AUC for 100 training cases of the minority class.

combination of training counts. The 10 ranks of the default setting computed for different column subsets of a given data set and given training counts are averaged. The rows of the tables correspond to data sets and columns to the training counts.

Average ranks of AUC obtained for the default setting among the results obtained for other tested settings of the parameters for different data sets and training counts.

	d22	d23	d24	d25
MiniBooNE	9.20	12.3	13.0	–
RecordLinkage	10.30	12.2	12.8	12.75
census-income	11.80	13.0	13.0	13.00
census1990	8.75	10.4	10.6	10.60
covtype	9.00	11.5	13.0	13.00

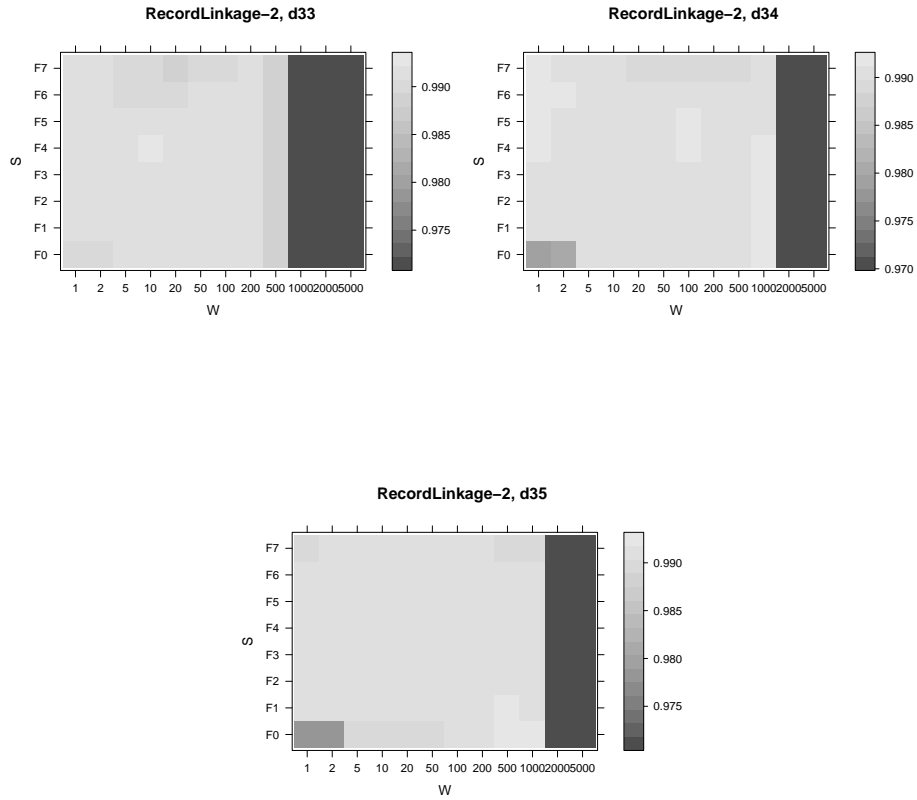


Figure 2: Levelplots of AUC for 1000 training cases of the minority class.

	d33	d34	d35
MiniBooNE	12.0	13.0	–
RecordLinkage	11.9	13.0	11.85
census-income	13.0	13.0	13.00
census1990	10.4	10.6	10.60
covtype	5.7	6.3	11.10

The rank of AUC obtained for the default settings is typically a large one and often even the largest, which is 13 and corresponds to the worst result.

4.3.2 Semiartificial data sets

The following table presents the average ranks of the AUC obtained with the default setting among the other tested setting for the samples of semiartificial data and different training counts. Larger rank means

a worse result.

d22	d23	d24	d25
5.75	7.25	12.25	13.0
<hr/>			
d33	d34	d35	
4.08	6.92	12.58	

Similarly as for large UCI data, the rank of the result for the default setting increases with increasing imbalance.

4.3.3 Summary of the influence of imbalance

It may be seen that for most of the large tested UCI data sets and also for the tested samples of semiartificial data, the average rank of AUC for the best setting is increasing with increasing imbalance in the sequences of the counts (d22, d23, d24, d25) and (d33, d34, d35). Since larger rank means a worse result, this is a further evidence that with increasing imbalance, it becomes more important to set either the stopping criterion W or smoothing parameter S or both to a higher value than the default one.

5 Conclusion

We analysed the quality of the prediction for random forest trained on imbalanced data and the influence of the setting of the parameters for random forest. The stopping criterion and the level of smoothing proved to be important for minimizing the generalization error. These parameters may be used together with subsampling of the majority class to achieve a good classifier.

Acknowledgements

The authors were supported through Slovene-Czech bilateral cooperation. Marko Robnik-Šikonja under the grant BI-CZ/10-11-008 and research program P2-0209 by Slovenian Research Agency. Petr Savický acknowledges the support by the Grant Agency and MEYS and Technology Agency of the Czech Republic under the grant numbers P202/10/1333 and MEB091008 and by Institutional Research Plan AV0Z10300504.

References

- [1] C. Chang and C. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.
- [2] N. Chawla. Data mining for imbalanced datasets: An overview. In O. Maimon and L. Rokach, editors, *Data mining and knowledge discovery handbook*, pages 853–867. Springer, 2005.

- [3] D. A. Cieslak and N. V. Chawla. Learning decision trees for unbalanced data. In W. Daelemans, B. Goethals, and K. Morik, editors, *ECML2008: Machine Learning and Knowledge Discovery in Databases*, volume 5211 of *Lecture Notes in Computer Science*, pages 241–256, Berlin / Heidelberg, 2008. Springer.
- [4] A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL <http://archive.ics.uci.edu/ml>.
- [5] C. Jutten et al. ELENA: Enhanced Learning for Evolutive Neural Architectures. Technical report, Basic Research ESPRIT project Number 6891, 2000. URL <http://www.dice.ucl.ac.be/neural-nets/Research/Projects/ELENA/elena.htm>.
- [6] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL <http://www.R-project.org>. ISBN 3-900051-07-0.
- [7] M. Robnik-Sikonja and P. Savicky. *CORElearn: CORElearn - classification, regression, feature evaluation and ordinal evaluation*, 2011. URL <http://lkm.fri.uni-lj.si/rmarko/software/>. R package version 0.9.36.
- [8] M. K. Titsias and A. Likas. Shared kernel models for class conditional density estimation. *IEEE Trans. Neural Networks*, 12(5): 987–997, 2001.
- [9] M. K. Titsias and A. Likas. Class conditional density estimation using mixtures with constrained component sharing. *IEEE Trans. Pattern Anal. and Machine Intell.*, 25(7):924–928, 2003.
- [10] B. Zadrozny and C. Elkan. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 204–213. ACM, 2001.