Improving performance of the eXtasy model by hierarchical sampling

Dusan Popovic[1,2], Alejandro Sifrim[1,2,3], Jesse Davis[4], Yves Moreau[1,2], and Bart De Moor[1,2]

[1] KU Leuven, Department of Electrical Engineering (ESAT),
STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics,
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
[2] iMinds Medical IT
[3] The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus,
Hinxton, Cambridge, CB10 1SA, UK
[4] Department of Computer Science, KU Leuven, Belgium

1 Background

Data from biomedical domains often have implicit hierarchical structure that is usually ignored by practitioners interested in constructing and evaluating predictive models from it. A typical example is genomic data, where a single gene can harbor many mutations while it is at the same time a part of higher-order construct (e.g., chromosome). In such a case different features can be defined over distinct levels of hierarchy. In parallel, a target variable can reflect this intrinsic property of the data, potentially resulting in a biased model.

This happens if the rows (i.e., examples) are inter-dependent, even though the data consists of a single table where each example is described as a fixed-length feature vector. The interdependencies exhibit themselves on different levels of granularity, where all inter-dependent examples have an identical value for a specific feature as well as the same value for the target variable. Thus the feature value appears correlated with the target variable whereas in reality the feature value is correlated with the hierarchical structure of the data. Failing to consider the interdependencies during learning could cause the algorithm to produce a model that simply identifies a pattern that is correlated with the hierarchical structure of the data as opposed to a pattern that is correlated with the target variable.

The described issue figures in the state-of-the-art variant prioritization algorithm called eXtasy[1]. This method incorporates predictors defined over three distinct levels of data granularity - gene level, mutation level and data record level (mutation/phenotype combination), where many data instances share the same values of higher order features (e.g. genes). Here the bias materializes as learning, to a certain degree, to recognize genes which constitute the training set, rather than extracting general characteristics of disease causing mutations. This results in degradation of the performance on the test set.

2 Methods

We propose a straightforward sampling-based solution for elimination of the described bias. It is implemented and tested within the Random forest framework [2](the eXtasy core model), but it easily generalizes to other types of ensemble learners as well. In particular, instead of extracting a bootstrap[3] from the complete training set to build a single tree on, we first stratify the training examples according to the distinct values of the feature over which the coarsest granularity level is defined. In the case of the eXtasy data, that is the gene identifier. After stratification we randomly select just one data instance from each partition to form the in-bag sample. This prevents algorithm from learning to recognize a particular value of the higher-order feature, as only one example having it will be present in the sample. The procedure differs from the stratification approach that is typically used with the Random forest, where a bootstrap replicate is extracted from each strata to assure that all of them are well-represented.

To ensure a fair comparison, we test the method on the original eXtasy benchmark data using the same evaluation scheme as in the original study. That is, we randomly divide the complete benchmark data set on the gene-level such that two-thirds of the genes belong to the training set and one-third are in the test set. Furthermore, we consider two test scenarios. In the first one we compare two sampling schemes on the unaltered test set, effectively repeating the eXtasy benchmark. In the second one we randomly undersample the positives from the test set in order to mimic the class distributions we would expect to see in the wild; where only one out of 8000 non-synonymous mutation in a genome is potentially disease-causing[4]. We repeat the aforementioned procedure 100 times to stabilize the values of the performance metrics.

3 Results and Conclusions

Under the original eXtasy benchmark scenario, Random forest trained with the hierarchical sampling achieves precision of 0.84, while classical approach results in 0.71. In the same time, the sensitivity drops from 0.86 to 0.78. However, increased Matthews correlation coefficient (from 0.75 to 0.78) indicates that the same sensitivity can be achieved with the hierarchical sampling while maintaining higher precision, by setting an appropriate decision threshold. Furthermore, the realistic class balance scenario underlines this difference even more, as precision doubles (from 0.0024 to 0.0053) with hierarchical sampling, while sensitivity drops from 0.88 to 0.81. In other words, the improved version of eXtasy classifies approximately 188 out of 8000 variants as disease causing, with the probability of capturing the real one equal to 0.81 (i.e. sensitivity). In the same time, the standard eXtasy calls 417 out of 8000 variants, with the probability of hit being 0.88.

Hence, the hierarchical sampling leads to a notable improvement in the model performance in terms of the precision, especially in the most important operating regions for this particular application (i.e. top of prioritized list, see Fig 1.). Also,

as it uses less data (per single tree) than standard Random forests, it results in a more parsimonious model. Finally, we hypothesize that the gain in performance might be even bigger for certain classes of problems. That is, if the number of distinctive values of the coarsest grain concept is much smaller than total number of data records, overfitting on these concepts is more likely to occur. Therefore, such problems could potentially greatly benefit from the proposed sampling scheme.

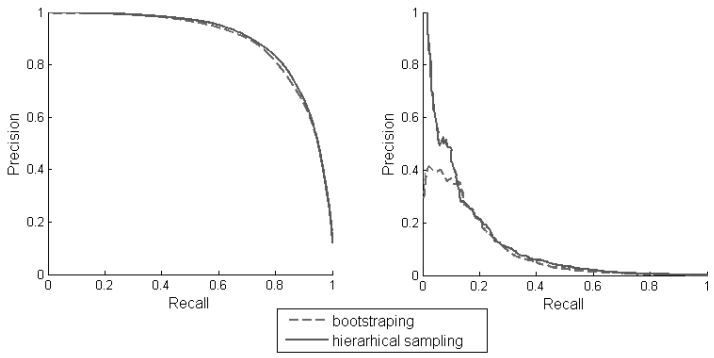

Fig. 1. Precision-recall curves obtained by the application of the eXtasy on the benchmark data (left panel) and the data with the realistic class distribution (right panel). Each panel displays two curves - the one corresponding to the standard Random forest classifier training with bootstrapping and the one corresponding to hierarchical sampling based training.

Acknowledgments. BDM and YM are full professors at the Katholieke Universiteit Leuven, Belgium. Research supported by:

- Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, CREA/11/015, OT/11/051, several PhD/postdoc & fellow grants
- Industrial Research fund (IOF): IOF/HB/13/027 Logic Insulin, IOF: HB/12/022 Endometriosis
- Flemish Government:
  - FWO: PhD/postdoc grants, projects: G.0871.12N (Neural circuits), research community MLDM
  - IWT: PhD Grants; TBM-Logic Insulin, TBM Rectal Cancer, TBM IETA, O&O ExaScience Life Pharma
  - Hercules Stichting: Hercules III PacBio RS
  - iMinds: SBO 2013; Art&D Instance; ICON b-SLIM
  - IMEC: phd grant
  - VLK Stichting E. van der Schueren: rectal cancer
  - VSC Tier 1: exome sequencing

– Federal Government: FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate)
– COST: Action BM1104: Mass Spectrometry Imaging, Action BM1006: NGS Data analysis network
– EU FP7 Marie Curie Career Integration Grant (294068)

The scientific responsibility is assumed by its authors.

References

1. Sifrim, A., Popovic, D., Tranchevent, L.C., Ardeshirdavani, A., Sakai, R., Konings, P., Vermeesch, J.R., Aerts, J., De Moor, B., Moreau, Y.: extasy: variant prioritization by genomic data fusion. Nature methods 10(11), 1083-1084 (2013)
2. Breiman, L.: Random forests. Machine learning 45(1), 5-32 (2001)
3. Efron, B.: Bootstrap methods: another look at the jackknife. The annals of Statistics pp. 1-26 (1979)
4. Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A., et al.: Whole-genome sequencing in a patient with charcot-marie-tooth neuropathy. New England Journal of Medicine 362(13), 1181-1191 (2010)