Improving the Interpretability of Classification Rules in Sparse Bioinformatics Datasets

James Smaldon and Alex A. Freitas
Computing Laboratory, University of Kent, Canterbury, CT2 7NF, UK
James.Smaldon@gmail.com, A.A.Freitas@kent.ac.uk

Abstract

This paper proposes a modification in rule induction algorithms aimed at improving the interpretability of the discovered rules. This modification is proposed in the context of sparse bioinformatics data sets where the presence of a feature is much less common than its absence, so that rule conditions with positive values of the feature tend to be more informative than rule conditions with negative values of that feature. The proposed modification consists of inducing only rules having positive values of the features, rather than rules using both positive and negative values of the features.

1. Introduction

The motivation for this paper came from a case study in bioinformatics reported in [6], where a biologist had difficulty in interpreting many rules discovered by a data mining algorithm. In that application the vast majority of the predictor attributes denoted whether or not a protein had a certain biological motif. For each motif (attribute), the value "present" was much less frequent in the data than the value "absent", i.e., the dataset was very sparse. Hence, a rule with conditions of the form "IF a protein has biological motif X" was easier to be interpreted by the biologist than a rule with conditions of the form "IF a protein does not have biological motif X", because the latter is much less informative.

The central idea of this paper is to modify two rule induction algorithms to discover rules having in their antecedent only conditions of the form "IF a protein has biological motif X", and not conditions of the form "IF a protein does not have biological motif X", in order to improve the interpretability of the discovered rules. Rule interpretability is often important in data mining [4], [8].

2. Rule Induction with Modified CN2 and Ant-Miner

The two rule induction algorithms modified in this work are CN2 and Ant-Miner. CN2 is a well-known rule induction algorithm [2]. Ant-Miner is based on the relatively new paradigm of ant colony optimisation [7]. Both CN2 and Ant-Miner are sequential covering algorithms, where a classification rule is discovered, examples covered by the discovered rule are removed from the training set and the process is repeated until (almost) all training examples are covered. Both algorithms construct a classification rule by adding one condition at a time to the rule, and they discover rules whose antecedent can include both conditions of the form "IF a protein *has* biological motif X" – called *present-motif* conditions – and conditions of the form "IF a protein *does not have* biological motif X" – called *absent-motif* conditions. In order to improve the rule interpretability, we modify these algorithms to discover rules having *present-motif-only* conditions.

In the original CN2 and Ant-Miner algorithms the set of candidate conditions is initialized with all conditions of the form (A_i = V_{ij}), where V_{ij} is the j-th value of the i-th attribute, $\forall$i,j. By contrast, in the proposed modification (for both algorithms) the set of candidate conditions is initialized only with *present-motif-only* conditions, i.e., conditions of the form (A_i = "present"). Once a condition is added to a rule, the system removes just that condition from the set of candidate conditions to be considered in the next iteration of the rule construction procedure. By contrast, in the original algorithms, when a condition like (A_i = "present") is added to a rule, the system has to remove both that condition and the condition (A_i = "absent") from the set of candidate rules.

3. Datasets and Experimental Setup

Experiments were done with four bioinformatics datasets involving two protein function prediction problems. The first problem consists of predicting whether or not a protein has post-synaptic activity, based on the biological motifs found in the protein primary sequence [6]. Each example (record) corresponds to a protein. Each predictor attribute corresponds to a Prosite pattern (a biological motif). An attribute can take on the value "present" or "absent", indicating whether or not the Prosite pattern occurs in a protein. The class attribute is post-synaptic activity, which can take on "yes" or "no". The second problem is the classification of G-Protein Coupled Receptors (GPCRs). In the 3 GPCR datasets used in our experiments [5], each example (record) corresponds to a protein. However, different kinds of predictor attributes (motifs) were used in the 3 datasets, viz.: Interpro entries, Prints motifs, and Prosite patterns. All these attributes are binary, indicating whether or not a protein has a motif.

The datasets used in our experiments are somewhat modified versions of the datasets used in [6], [5], as follows. First, the post-synaptic dataset described in [6] included 2 continuous attributes (sequence length and molecular weight). In our experiments these 2 attributes were removed – only the Prosite pattern attributes were used. Second, in the GPCR datasets described in [5] the classes to be predicted are arranged in a four-level hierarchy. Our experiments involved only the prediction of classes at the first level of the hierarchy. Third, both the post-synaptic dataset [6] and the GPCR datasets [5] had a large number of attributes. In order to greatly reduce the time taken by the rule induction algorithms, we worked only with the set of the 50 best attributes for each dataset. To perform this attribute selection we used the attribute selection algorithm described in [3]. The reduced post-synaptic dataset had 2081 examples. The 3 reduced GPCR datasets (with Interpro, Prints and Prosite motifs) had 540, 323 and 177 examples, respectively.

We used the default parameters of CN2 [1], [2]. Ant-Miner was used with its default parameters [7], [9], with the exception that the parameter Max_uncovered_cases was set to 5 in the unordered rule set version. All the results reported in the paper were obtained by performing a well-known 10-fold cross-validation experiment.

4. Computational Results

The results concerning predictive accuracy are shown in Table 1. The numbers after "±" are standard deviations. Experiments were done with the ordered rule list [2] and unordered rule set [1] versions of CN2; as well as the ordered rule list [7] and unordered rule set [9] versions of Ant-Miner.

Table 1: Comparing predictive accuracy (%) using present motif only (Pres.) vs both present and absent (Pres/Abs) motifs

			Dataset			
Algor.	Unordered vs. ordered rules	Pres/Abs vs. Pres. motifs	Post-synapt.	GPCR Interp.	GPCR Prints	GPCR Prosite
CN2	Ordered	Pres/Abs	96.92 ±0.33	90.75 ±0.85	92.25 ±0.70	81.13 ±2.62
		Pres	96.88 ±0.36	90.71 ±0.90	92.56 ±0.51	80.75 ±3.30
	Unordered	Pres/Abs	96.83 ±0.37	90.20 ±0.87	93.20 ±0.59	84.48 ±2.48
		Pres	96.78 ±0.34	**85.75 ±0.70**	93.50 ±0.29	**63.20 ±1.49**
Ant-Miner	Ordered	Pres/Abs	96.73 ±0.36	87.98 ±0.52	87.60 ±1.73	66.13 ±3.03
		Pres	**88.23 ±0.17**	**78.89 ±0.45**	85.80 ±1.73	**49.52 ±2.69**
	Unordered	Pres/Abs	96.44 ±0.47	87.02 ±0.65	96.59 ±0.56	79.74 ±1.49
		Pres	96.73 ±0.31	86.30 ±0.54	**92.29 ±0.79**	**61.67 ±0.57**

Out of 16 cases (2 algorithms × 2 kinds of rule ordering × 4 datasets), there are 7 cases (in bold in Table 1) where the use of present motifs only led to a significant drop in accuracy, by comparison with the use of both present and absent motifs. A difference in two accuracy values was considered significant if the corresponding confidence intervals – taking into account the standard deviations – do not overlap. In the other 9 cases there was no significant difference between the accuracies with present motifs only and the accuracies with both present and absent motifs. Results concerning rule simplicity (measured by the total number of conditions in all rules) are shown in Table 2. In all 16 cases, the use of present motifs only led to a significant improvement in simplicity (reduction in rule set/list size).

Table 2: Comparing the total number of conditions in all discovered rules using Present motif only (Pres) vs. Present and Absent (Pres/Abs) motifs

			Dataset			
Algor.	Unordered vs. ordered rules	Pres/Abs vs. Pres. motifs	Post-synapt.	GPCR Interp.	GPCR Prints	GPCR Prosite
CN2	Ordered	Pres/Abs	50.90 ±0.40	31.80 ±0.70	39.00 ±0.60	58.90 ±0.99
		Pres	**45.80 ±0.33**	**24.80 ±0.59**	**28.10 ±0.46**	**44.20 ±0.79**
	Unordered	Pres/Abs	57.30 ±0.47	57.90 ±1.95	68.70 ±1.46	97.10 ±2.40
		Pres	**46.60 ±0.31**	**32.90 ±0.75**	**34.60 ±0.56**	**47.60 ±1.13**
Ant-Miner	Ordered	Pres/Abs	306.91 ±10.28	233.50 ±6.27	207.30 ±6.71	217.60 ±11.74
		Pres	**9.20 ±0.20**	**3.10 ±0.10**	**2.00 ±0.00**	**2.10 ±0.10**
	Unordered	Pres/Abs	355.60 ±19.20	245.80 ±1.11	188.00 ±0.00	229.70 ±5.05
		Pres	**32.00 ±0.00**	**6.0 ±0.00**	**4.00 ±0.00**	**5.00 ±0.00**

5. Conclusions

The central idea of the proposed method – aimed at improving the interpretability of discovered rules – is to modify rule induction algorithms to discover rules having in their antecedent *present-motif-only* conditions.

Concerning the simplicity of the discovered rule sets or lists, the use of *present-motif–only* conditions consistently reduced the size of the discovered rule set or list in all cases. In addition to this clear gain in *syntactical* simplicity, the use of present motifs only has the important advantage of improving the *semantic* comprehensibility of discovered rules to biologists, because in general it is easier for biologists to interpret specific conditions of the form “IF a protein has biological motif X” than to interpret much more generic conditions of the form “IF a protein does not have biological motif X”.

Concerning predictive accuracy, unfortunately the use of present-motif-only conditions led to a significant drop in accuracy in 7 out of 16 cases. On the other hand, in the majority (9 out of 16) of the cases the significant gains in syntactical and semantic simplicity were obtained without any significant drop in accuracy. This is a promising result in applications where rule interpretability is very important. However, one should be careful with the potential significant drop in predictive accuracy in applications where accuracy is very important.

Although we focused on sparse bioinformatics datasets only, the basic idea of the proposed method is also potentially useful in sparse datasets from other application domains – a possible topic for future research.

References

1. P. Clark and R. Boswell. Rule induction with CN2: some recent improvements. Proc. 5th European Working Session on Learning. 1991.
2. P. Clark and T. Niblett. The CN2 induction algorithm. Machine Learning, 3(4), 261-284. 1989.
3. E.S. Correa, A.A. Freitas and C.G. Johnson. A new discrete particle swarm algorithm applied to attribute selection in a bioinformatics data set. Proc. 2006 Genetic and Evolutionary Computation Conf. (GECCO-2006), 35-42, ACM.
4. R.J. Henery. Classification. D. Michie, D.J. Spiegelhalter, C.C. Taylor. Machine Learning, Neural and Statistical Classification, 6-16. Ellis Horwood, 1994.
5. N. Holden and A.A. Freitas. Hierarchical classification of G-protein-coupled receptors with a PSO/ACO algorithm. Proc. IEEE Swarm Intelligence Symposium (SIS-06), 77-84. IEEE, 2006.
6. G.L. Pappa, A.J. Baines and A.A. Freitas. Predicting post-synaptic activity in proteins with data mining. Bioinformatics V. 21, Supp. 2, ii19-ii25, Sep. 2005.
7. R.S. Parpinelli, H.S. Lopes and A.A. Freitas. Data Mining with an Ant Colony Optimization Algorithm. IEEE Trans. on Evolutionary Computation, 6(4), 321-332, Aug. 2002.
8. M.J. Pazzani. Knowledge discovery from data? IEEE Intellig. Systems, Mar/Apr. 2000, 10-13.
9. J. Smaldon and A.A. Freitas. A new version of the Ant-Miner algorithm discovering unordered rule sets. Proc. 2006 Genetic and Evolutionary Computation Conf. (GECCO-2006), 43-50, ACM.