Eduardo Gade Gusmão <eggduzao@gmail.com>

Paper rejeitado

5 mensagens

Marcilio de Souto <marcilio.souto@gmail.com> 11 de novembro de 2012 15:15
Para: Eduardo Gade Gusmão <eggduzao@gmail.com>

Oi Eduardo,

O nosso paper submetido para o ACM SAC foi rejeidado --- veja comentarios abaixo. No geral, os revisores não entederam bem o q estamos fazendo, ou seja, temos q melhorar esta parte. Uma coisa q me passou completamente batido fou a mudança de titulo: não sei em que momento eu simplesmente cortei "issues on ...." do titulo. O titulo atual está completamente misleading.... O revisor3 fez um monte de criticas q diz mais respeito aos papers originais do que o nosso: ele deve ter entendido q somos nós q estamos propondo os métodos. Anyway, os comentarios "biologicos" dele são bem interessantes para analisarmos/discutirmos os nossos resultdados. Com relação ao revisor4, eu não entendi muito o q ele queria/quer.

Enfim, acho q nosso e-mail agora deve ser o iJCNN. O deadline é fevereiro. Até lá, precisamos melhorar o paper --- tem os testes estatiscos q não adcionamos e melhorar o texto.

Abraços, Marcilio

==========
SAC 2013 Reviews for Submission #1910
==========

Title: Sampling Negative Examples for Predicting Prokaryotic Promoters

Authors: Eduardo Gusmao and Marcilio de Souto
==========
REVIEWER #1
==========

Reviewer's Scores

Technical Content and Accuracy: 2
Significance of the Work: 3
Appropriate Title, Introduction, and Conclusion: 2
Overall Organization: 4
Appropriateness for SAC: 4
Style and Clarity of the Paper: 2
Originality of Content: 2
OVERALL RECOMMENDATION: 2

Comments
---

(1) The title is not appropriate: in fact, it suggests that the main topic of the paper is a method for sampling negative examples for training a classifier to predict prokaryotic promoters. As matter as fact, the main topic of the paper is an empirical study about the classification performances of different classifiers trained on different training sets (including or not negative examples).
Also the introduction is not clear.

(2) Why the paper is defined an "Extended Abstract" ?

(3) Why do the authors use ww Z curves for feature extraction ? Could they motivate their choice ? Are there other feature extractions approaches ? Please, add some references.

(4) Evaluation measures are not clear: what is the meaning of TN, TP, FN, FP, ... in table 2 ? On my opinion, the caption of the table is not adequate: correct rate, sensitivity and specificity are measures for accuracy not "statistics" ...

(5) Final conclusions are not clear: the authors claim that "using multiple datasets is essential for any study" and defining negative examples is a difficult task. At the same time, they claim that "the learners face data that do not behave so well as in wet-lab guided experimental pipeline": what does it mean ? Are the classification approaches presented in the paper unuseful for biologists ? Are the well designed in wet-lab experiments not appropriate to model a biological phenomenon ? Please, discuss better this point.

==========
REVIEWER #2
==========

Reviewer's Scores

Technical Content and Accuracy: 4
Significance of the Work: 5
Appropriate Title, Introduction, and Conclusion: 4
Overall Organization: 5
Appropriateness for SAC: 5
Style and Clarity of the Paper: 5
Originality of Content: 4
OVERALL RECOMMENDATION: 5

Comments
---

interesting work concerning the identification of the promoters in transcription process. Further investigation can be useful for a more general assessment of the problem.

==========
REVIEWER #3
==========

Reviewer's Scores

Technical Content and Accuracy: 3
Significance of the Work: 2
Appropriate Title, Introduction, and Conclusion: 4
Overall Organization: 4
Appropriateness for SAC: 4
Style and Clarity of the Paper: 5
Originality of Content: 2
OVERALL RECOMMENDATION: 2

Comments

Review of "Sampling Negative Examples for Predicting Prokaryotic Promoters"

The authors test the cross-validation accuracy of promoter prediction, using a range of different negative sets for contrasting purposes.

Although the definition of negative sets in many areas of bioinformatics is of vital importance, I found this work to be fairly elementary in design and execution, failing to take into account even some very basic ideas about nucleotide sequences and genome architecture. For example, testing of different negative sets was carried out at least as early as 1996, in GB Hutchinson, "The prediction of vertebrate promoter regions using differential hexamer frequency analysis" (CABIOS).

Another important point not considered at all in the paper is the composition of the sequences under consideration. Intergenic regions have k-mer compositions that are quite distinct from protein-coding regions, and this impacts on what the classifier actually learns during the training process. Also, the CTRL set appears to have been generated randomly, which is not appropriate as a background model for E. coli. Sequences generated in such a fashion should consider the background k-mer frequencies of the genome, rather than constructing a uniform probability distribution that does not approximate the real patterns. Since this idea underpins the widespread use of log-odds matrices in promoter detection, it should have been addressed here as well.

It is important to point out, too, that randomly selected sequences will likely be about 90% coding, since this is roughly the proportion of the E. coli genome that encodes proteins.

Also, I think trials that were dismissed as inappropriate (decision trees, COD1) should be omitted from the manuscript, or at least not addressed in detail in the results.

There is a lot of wasted space in Figures 1 and 2: starting the bars at 50% rather than 0% would make comparisons easier, as would removing the DT results.

==========
REVIEWER #4

==========

Reviewer's Scores

Technical Content and Accuracy: 3
Significance of the Work: 2
Appropriate Title, Introduction, and Conclusion: 3
Overall Organization: 4
Appropriateness for SAC: 6
Style and Clarity of the Paper: 5
Originality of Content: 3
OVERALL RECOMMENDATION: 3

Comments

This paper considers the problem of prediction of promoter region in the genomic sequences. The authors formulated this problem as a binary classification in machine learning (promoter and non-promoter region). Though there are many proposed methods in literature, however, one difficult issue is hard to build a strong negative example set. Several methods randomly select short genomic sequences as negative samples while others extract random sequences from coding and non-coding regions. The main aim of this paper is to perform a study of impact of diffident choices of negative examples on the classification methods. However, in my opinion, the paper has some major issues: (1) no any new constructing negative example methods are presented, as saying in Section 3, all negative example set adapted from the work of Gordon et al, 2003; (2) the idea for feature extraction from genomic sequences as well as feature selection method is obtained from the work of Song, 2012; (3) lacking the comparisons of the results with previous methods, providing only the performance of classifiers on the constructed datasets is not good enough to show the significance of the work.

Eduardo Gade Gusmão <eggduzao@gmail.com> 12 de novembro de 2012 08:45
Para: Marcilio de Souto <marcilio.souto@gmail.com>

Marcilio, perdao pela demora em responder. Eh que ainda estou sem net aqui (uso apenas no trabalho).

Sim, algumas sugestoes realmente foram interessantes (a questao de utilizar o background da E coli para criar o CTRL, por exemplo, eh crucial).
Voce quer marcar para conversar? Eu pediria para ser na semana que vem, porem como o deadline eh apertado, vou deixar para voce marcar a data que voce preferir
(menos amanha de manha, pois vou na cidade resolver algumas coisas urgentes).

Perdao tambem por nao ter te enviado a dissertacao para voce corrigir. Em parte, sei que voce esta ocupado com algumas coisas entao fiquei com
vergonha de incomodar. A outra parte foi porque ocorreram tantos eventos simultaneos e era tanta coisa pra resolver que eu acabei esquecendo
de varias coisas, inclusive de pedir a sua opiniao final sobre o texto. Depois eu te explico toda a epopeia.
Enfim, perdao por isso e estou enviando em anexo a versao final da dissertacao, como eu tinha dito.
Obrigado por tudo!

Obrigado por tudo relativo a continuacao deste estudo tambem. Creio que passei uma imagem meio displicente com a versao
que eu te mandei do ACM SAC pra voce ajeitar, mas realmente era porque era o ultimo dia para entregar a

dissertacao. Dessa
vez vou realizar os experimentos e testes estatisticos com capricho e prometo um estudo bem mais interessante.

Abracos!

Eduardo Gade Gusmão
[Texto das mensagens anteriores oculto]

Dissertacao_EduardoGadeGusmao.pdf
19602K

Marcilio de Souto <marcilio.souto@gmail.com> 12 de novembro de 2012 10:03
Para: Eduardo Gade Gusmão <eggduzao@gmail.com>

Oi Eduardo,

Sim. A gente pode marcar uma conversa. Essa semana está mio confusa pois sexta-feira começo dois novos cursos. Podemos tentar alguma dia da proxima semana. Sobre os comentarios:

(1) Intergenic regions have k-mer compositions that are quite distinct from protein-coding regions, and this impacts on what the classifier actually learns during the training process.*** Esse comentario pode ajudar na discussão dos resultados obtidos com esse tipo de exemplo negativo.

(2) Also, the CTRL set appears to have been generated randomly, which is not appropriate as a background model for E. coli. Sequences generated in such a fashion should consider the background k-mer frequencies of the genome, rather than constructing a uniform probability distribution that does not approximate the real patterns. Since this idea underpins the widespread use of log-oddsmatrices in promoter detection, it should have been addressed here as well. ***
Sobre esse comentario, se não me engano o nosso uso de CTRL (talvez tenha sido uma escolha infeliz de nome) era ver como os classificadores se comportariam qdo gerados com sequencias negativas realmente aleatórios ---- nem mesmo obdecendo a distribuição de bases (k-mer or whatever) do genoma em questão. Enfim, não sei se no nosso caso a questão levantada pelo revisor procede.

Abraços, Marcilio

Com relação aos comentarios q achei interessante,
[Texto das mensagens anteriores oculto]

Eduardo Gade Gusmão <eggduzao@gmail.com> 12 de novembro de 2012 10:41
Para: Marcilio de Souto <marcilio.souto@gmail.com>

Talvez eu consiga fazer a minha net no meu apto. nessa sexta feira. Caso eu consiga voce pode marcar qualquer dia e qualquer hora. Caso contrario, marca num dia de semana. Lembre-se que estamos no mesmo fuso horario.
abracos.

Eduardo Gade Gusmão
[Texto das mensagens anteriores oculto]

Eduardo Gade Gusmão <eggduzao@gmail.com> 24 de novembro de 2012 13:23
Para: Marcilio de Souto <marcilio.souto@gmail.com>

quer/pode conversar essa semana?

abs

Eduardo Gade Gusmão
[Texto das mensagens anteriores oculto]