Rohit Bhattacharya
03/11/2015

Seminar Paper Critical Review

Paper Citation:

Hua S, Sun Z: **Support vector machine approach for protein subcellular localization prediction.** *Bioinformatics.* 2001; **17**(8): 721-728
See more at:
http://f1000research.com/articles/10.12688/f1000research.3-17.v1/doi#sthash.ZS4uh7wL.dpuf

Introduction:

This paper explores the use of Support Vector Machines (SVMs) in the prediction of subcellular localization of proteins. It was written around the turn of the new millennium, when breakthroughs in genome sequencing was producing large amounts of raw sequence data and SVMs had a become a topic of great interest following a publication in 1995 by Vladimir Vapnik. The subcellular location of a protein is thought to be a key functional characteristic that might help identify the function of potential genes in the genome sequence. The authors hope that by the use of SVMs, they will be able to automate the process of predicting these locations and thus, allow for the conversion of large amounts of genome sequence data into biological knowledge. In terms of relevance, this paper could not have been written at a better time.

The paper cites previous attempts at solving the problem as falling into two main categories - one that makes use of N-terminal sorting signals and the other that uses amino acid composition as prediction features. von Heijne et all employed the first method to build a neural network that would predict protein subcellular localizations (PSLs). Reinhardt and Hubbard (1998) built supervised neural networks that took advantage of the second method and were able to achieve 81% accuracy for three subcellular locations in prokaryotes and 66% for four in eukaryotes. Using the same dataset as Reinhardt and Hubbard, Choud and Elrod (1999) used a covariant discrimination algorithm and Yuan(1999) used Markov chain models to perform predictions. Yuan's model was able to achieve 89% accuracy for prokaryotic sequences and 73% for eukaryotic sequences. The authors of this paper, see these techniques as being inadequate in their predictive power and not robust against flaws in the genome sequence. In light of this, they propose SVMs as a more robust and theoretically sound alternative.

Body:

The authors begin by presenting a superficial explanation of the classification technique used by SVMs. They point to the original Vapnik paper for further technical details. Given that the paper is meant to focus on its applications rather than theory, this is quite alright. The authors do a good job describing the modifications that they make to enable the use of SVMs as multi-class classifiers. To achieve this, they use a "1 versus rest" approach wherein the class labels for the i^{th} class are seen as being positive and the rest are considered negative. They also mention and provide references to the specific softwares and algorithms (SVM^{light}, LOQO) that were used in order to implement their SVMs. In addition, the authors, provide links to their own finished SVM software, 'SubLoc' and provide estimated run-times and computer specifications. This allows readers to reproduce their experiments if so desired. An important detail that the authors do not give enough attention to is the feature space used for their SVMs. They only state that it is the amino acid composition of proteins but are not more specific than that.

In order to validate their methods the authors run jackknife tests on the Reinhardt and Hubbard data set. This is a reliable and commonly used test wherein each protein is held out in turn as a test example while the others are used for training the SVM. They present their results in the form of tables, contrasting the choice of kernels for SVMs as well as the choice of learning algorithm, comparing SVMs to the ones mentioned before. The tables contain values of accuracy and the Matthew's correlation coefficient (MCC). Despite listing the formula for MCC, it might have been suitable for the author to give a brief description as to why it is important to use such a metric. Upon examining the results, one is able to conclude that it is due to the class sizes being very different from each other, with the cytoplasm sublocation being much larger than the rest. The MCC is able to provide a balanced measure of the quality of classification even in cases such as this one. The authors also provide a Reliability Index in their paper, which provides cut-offs above which a certain percentage of the sequences can be predicted with a certain accuracy. For example 75% of the sequences had an RI > 3 and of these sequences the SubLoc system classified 95.5% of them correctly.

The authors locate the kernels that best fit the prokaryote and eukaryote sequences of the Reinhardt and Hubbard data set but do not go into very much detail about tuning the cost

parameter of the SVM. They seem to have arrived at a fixed number of 500 but it is unclear whether other values were tested. Towards the end of the paper the authors do note their lack of testing with the cost parameter but dismiss it as having little to no influence on the results of their classifier. In terms of the results themselves, the authors are quick to claim that SVMs provide better accuracy than all other learning methods. While this is true for the total accuracy across all classes, it is not the case for a few of the individual classes and the authors make no mention of this. It also might have been useful for the authors to describe why every classifier performs poorly on the eukaryotic sequences when compared to the prokaryotic ones. My guess is that this is due to the addition of another class to the problem. Overall, I think the authors do a great job in the presentation of their results but do not deign to fully discuss them.

One of the authors' complaints about recognition of N-terminal sorting signals as a method of predicting PSLs was the lack of robustness against damaged genome sequences. In order to test for this characteristic in their own model, the authors cut out 10, 20, 30 and 40 length amino acid segments from the N-terminal of the genome sequences and compare the results. They conclude that the SVM SubLoc classifier is robust against such flaws in the genome sequence. Finally, the authors end by discussing the possibility of integrating their method with others such as Bayesian systems, and including more features in order to provide the best classifications possible.

Conclusion:

Listed as a top 100 SVM paper on some websites, I had high expectations going into this critical review. Some of these were met - in terms of methods used, relevance of the research, and overall presentation quality. Others were not - in terms of discussion and thorough tuning of all available parameters. Upon reading several other papers on the applications of SVMs however, I concluded that this one is probably one of the more detailed and well written ones available today and deserves its reputation as a top 100 SVM paper. Its goals are clearly stated, methods succinctly described and results well presented.