CUDA-Accelerated SVM for Celestial Object Classification

Nanbo Peng, Yanxia Zhang, and Yongheng Zhao

Key Laboratory of Optical Astronomy, National Astronomical Observatories;
Chinese Academy of Sciences, 100012 Beijing, China

Table 1. Computation time of functions in SVM^{light}

Function name	Computation time [s]						
	q=2	q=4	q=8	q=16	q=32	q=64	q=128
select_working_set	29.58	24.11	15.75	12.32	8.20	5.61	2.44
*cache_multiple_kernel_row**	64.21	43.12	56.92	79.46	113.6	159.0	119.5
*optimize_svm**	7.45	5.28	2.68	3.88	6.32	19.61	73.44
*update_linear_component**	15.77	16.98	21.70	27.40	35.41	43.66	38.06
calculate_svm_mode	0.03	0.02	0.09	0.07	0.06	0.04	0.07
check_optimality	19.95	9.42	7.46	6.69	3.77	3.10	1.19
reactivate_inactive_examples	23.70	22.24	21.42	22.76	20.74	21.79	19.71
shrink_problem	0.50	0.23	0.22	0.14	0.12	0.06	0.06
kernel_cache_shrink	152.6	124.0	65.23	42.30	22.76	13.91	9.77

Table 2. Comparison of the SVM Models Derived from CPU and GPU

Data size	Iterations		Support vectors		Threshold *b*	
	CPU	GPU	CPU	GPU	CPU	GPU
500	13	25	202	202	0.6698	0.6698
5000	186	203	859	858	0.7582	0.7650
25000	1325	1323	2790	2788	0.7340	0.7400
50000	727	731	4736	4727	0.7481	0.7574

Table 3. Comparison of computation times on CPU and GPU

Data size	Training			Classifying		
	CPU [s]	GPU [s]	Speedup [x]	CPU [s]	GPU [s]	Speedup [x]
500	0.49	0.99	0.50	0.65	0.07	9.29
5000	9.61	7.66	1.25	6.55	0.11	59.54
25000	236.6	59.13	4.00	32.79	0.13	252.2
50000	398.6	40.01	9.96	65.64	0.18	364.6

Introduction

Quasars are the most luminous, powerful, and energetic celestial objects known in the universe. They show a very high redshift, which is an effect of the expansion of the universe. So, many astronomical survey projects make a schedule for finding new quasars as many as possible. Support Vector Machine (SVM) is one of the most promising tool to discovery new quasars and we have successfully used it to get a more than 95% performance both on precision and recall. Although SVM has a stable high performance, training a SVM model and using this model to predict class label are time-consuming, especially with massive data sets. In this work, we implement a parallelized SVM using CUDA to accelerate.

Result

We tested the computation time of each function which repeats many times in the optimization procedure of SVM-light. As shown in Table 1, all of them are affected by parameter "q" which means maximum size of working set in SVM model. The running time of the three functions cache_multiple_kernel_row*, optimize_svm* and update_linear_component* marked with asterisk increase as the size of "q" becomes larger, and they handle the most computation task and are suited to be accelerated by CUDA.

Table 2 gives the slight difference of training models produced by SVM-light and CUDA-accelerated SVM. Although our implementation of SVM is completely equivalent to SVM-light in logic, there is still a very small difference between the calculated results of CPU and GPU. This problem produced by the structure of GPU itself and we do not need to pay much attention to it. Table 3 shows the performance of CUDA-accelerated SVM in training and predicting. Because the predicting module is computing intensive and easily be parallelized, the speedup of predicting is much higher than that of training and can even reach 364.6x. Since he training strategy of SVM is rather complicated, the best speedup of training only reaches 9.96x in our data sets. However the exception for the small sample happens, for example, when the size of data set is just 500, the running time even becomes longer owning to the failure of hiding the I/O latency of GPU. Whereas for large sample, the acceleration based on GPU shows its superiority.

Conclusion

This work has demonstrated the utility of CUDA-accelerated SVM for celestial object classification. Training time can be reduced by1.25 - 9.96x and predicting time can be reduced by 9.29 - 364.4x compared to the original SVM-light program. GPU provides astronomers a powerful tool to solve the large-scale computation problems faced in astronomy, i.e., rapidly labeling billions of sources in the photometric data is possible in short time. In addition, GPU is a very low cost way to achieve high performance compared with clusters. Our future work will focus on improve the performance of our algorithm in training session and apply GPU to more astronomical problems.

E-mail: nbpeng@bao.ac.cn; zyx@bao.ac.cn; yzhao@lamost.org;
Phone: 8610-64877703-1320;
Fax: 8610-64878240;

Address: Datun Road 20A, Chao Yang District, National Astronomical Observatories, Chinese Academy of Sciences, China 100012