A Brief Reference of Reputable Articles in Relevance Vector Machines

Kexin Tian ktian8@wisc.edu 9080135420
Jack Yuan yuan88@wisc.edu 9079800406
Yiheng Wei wei73@wisc.edu 9080232094

Introduction: (Yiheng Wei)

The relevance vector machine is a recently proposed approach to machine learning, specifically in regression and classification problems. The model of RVM was first proposed by Micheal E. Tipping around 2000 and has raised great interest in the field. As its name suggests, RVM has a similar form as the support vector machine, which has been a popular machine learning algorithm for a long time. One of the notable characteristics that differ RVM from SVM is that the former provides reliable probabilistic predictions. To achieve this output, Tipping has applied Bayesian methods and kernel tricks to build the algorithm, and the resulting model has shown several benefits from calculations to model performance.

In decades, the RVM model has been developed and improved in several ways. As an efficient tool to tackle regression and classification tasks, RVM has been applied to various real-world applications, such as disease prediction as well as product quality modeling and prediction. Nevertheless, we will focus on the RVM method in this paper and introduce some academic journals that have significant contributions to the development of RVM.

Source 1: (Yiheng Wei) Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research, 1(Jun), 211-244

In the article, the author Tipping introduces the fundamental thesis of the relevance vector machine and demonstrates its potential. The first section of this paper explains how to develop the RVM based on the Bayesian model respectively for regression and classification problems. An important step in RVM is to define an explicit prior probability distribution on weights to avoid overfitting and controlling complexity. The learning algorithm will require an iterative re-estimation of the hyperparameters that are associated with each weight and the noise variance for maximum marginal likelihood. For each iteration step, it is then possible to figure out the posterior covariance and mean. When the result converges, predictions can be made based on the posterior statistics. For classification that has no closed-form solution, RVM instead makes use of Laplace approximation. Tipping notes that only a few posterior probabilities over the weights are concentrated at non-zeros, and this contributes to the sparsity of the method. To evaluate the accuracy and efficiency of RVM, Tipping compares the performance of RVM and SVM in various experiment settings and shows that RVM uses fewer relevance vectors but has an even smaller error compared to SVM. Moreover, SVM would need extra help from cross-validation to tune its parameters, which is not a requirement for RVM. In conclusion, RVM has six main advantages-its generalizations, highly sparse representation, as well as the ability of probabilistic classification, easy validation, and freedom in basis function choices as well as parameter scaling.

This paper is one of the fundamental works that construct the RVM model. As the main content of this paper is about the algorithm, it is difficult to evaluate the efficiency of the method or its logistics besides the evidence that the author provides. Nevertheless, there are some other perspectives I think can be considered here. In my opinion, the representation of the research question is clear and straightforward. Tipping provides a step-by-step demonstration and the implementation logic for his proposed algorithm, allowing readers like me to understand and follow his approach easily. Besides, the data analysis for the example experiments is comprehensive. By applying several different kernel functions with various function settings, Tipping shows that this new algorithm can achieve an impressive result in varying scenarios. However, there is only a small section in the paper that talks about the limitations of this new algorithm. The author seems to emphasize the results that meet his expectations more often but pay insufficient attention to the potential problems of the algorithm. Nevertheless, this paper presents a high-quality research project and has proved to have a long-term impact on the further study and applications of RVM.

Source 2: (Yiheng Wei) Tipping, M. E., & Faul, A. C. (2003, January). Fast marginal likelihood maximisation for sparse Bayesian models. In AISTATS.

This is another paper written by Tipping, in which he proposes a faster algorithm for marginal likelihood maximization. Even though the original RVM has some notable advantages compared to SVM, it is not suitable for training large data sets as its run time would increase rapidly as the input size increases. To solve this problem, Tipping focuses on one crucial step in RVM-marginal likelihood maximization. Instead of keeping the full basis sets in the model during the whole training procedure, this new strategy allows "deletion" and "addition" of basis functions for every iteration. In this way, even though the model starts with the full size of basis sets, this size will decrease when qualified basis vectors get deleted from the model. To realize this function, the algorithm adjusts hyperparameters accordingly based on the influence of each vector. This approach will eventually maximize the marginal likelihood and guarantee generalization. The author verifies the effectiveness of this new algorithm by comparing the performance of SVM-Light, the original RVM approach, and the new RVM equipped with this technique. The first improvement has been shown in the capacity for training sets. The new RVM is capable of calculating a total of training sets that is four times larger than the original RVM. Moreover, adjusted RVM still maintains a similar error rate as the original RVM, but at the same time, it further lowers down the model complexity with a sparser solution. The most significant improvement is the decrease in run time. The old RVM requires about 5 minutes when training 1000 data pairs while the adjusted RVM and SVM manage to do all calculations within 15 seconds. This method largely improves the practicability and efficiency of RVM.

The result of the experiment is certainly encouraging. However, as Tipping also points out in the article, the applications that run these two models are different, and it may affect the performance of RVM. At this stage, there is no final implementation for this adjusted RVM but only "prototype-level Matlab code". It means the final result may not be finalized until the algorithm can be fully implemented into a program. Another limitation of the research may be associated with memory, which this study barely mentions. I think it would be more convincing if comparative experiments concerning the matrix size limit are provided. Nevertheless, the primary purpose of this paper is to improve the training speed, and thus the memory limitation may not be a focus here. Moreover, as this new approach is purely based on algorithm improvement, the author did not provide any experiments in realistic problems nor state its potential usage in different applications. If such information was provided, the value of this research would further increase, in my opinion.

Source 3: (Jack Yuan) B. Demir and S. Erturk, "Hyperspectral Image Classification Using Relevance Vector Machines," in IEEE Geoscience and Remote Sensing Letters, vol. 4, no. 4, pp. 586-590, Oct. 2007, doi: 10.1109/LGRS.2007.903069.

In order to check whether RVM actually results in better performance than SVM does, they should be compared when being applied on the exact same classification problem. In this paper written by Demir and Erturk, the performances of RVM and SVM are examined as both classification methods were applied to a sample hyperspectral image with a large number of training and test samples. The aspects that are being compared are the number of SVs (support vectors) and RVs (relevant vectors), classification accuracy, the training and testing time duration. For the number of SVs or RVs, the goal is to minimize this number so that the computational complexity and cost is minimized. The experimental results show that the theoretical analysis of the performances of SVM and RVM is correct: the number of RVs when using RVM classification is significantly less than that of SVs when using SVM classification. The accuracy of RVM classification is indeed lower than that of SVM classification, but the difference is not large, particularly when compared to the difference between the number of SVs and RVs, which is about a factor of 6. In the paper, the authors also mention that despite that RVM is definitely preferred when considering the time performance in the classification phase, RVM does have a much longer training time.

Based on the information provided in the paper, the authors tested every aspect of the possible factors that would influence the performance of RVM and SVM classification thoroughly. They even tried to use different kernels: linear, polynomial, and radial basis function kernels, all of which confirm their theoretical hypothesis: RVM classifier is more suitable for real-time processing, and it notably reduces the computational cost whilst keeping the accuracy in an acceptable margin. However, as mentioned at the end of the above paragraph, there is a disadvantage when using RVM, which is that it requires more time to train the classifiers. This becomes even more critical since RVM classification usually needs much more training samples than SVM classification does. While I think the paper is quite believing, it would be more convincing if more real life classification comparisons are examined.

Source 4: (Jack Yuan) I. Psorakis, T. Damoulas and M. A. Girolami, "Multiclass Relevance Vector Machines: Sparsity and Accuracy," in IEEE Transactions on Neural Networks, vol. 21, no. 10, pp. 1588-1598, Oct. 2010, doi: 10.1109/TNN.2010.2064787.

Although RVM is generally considered more effective than SVM, RVM encounters obstacles when adapting to multiclass settings, namely classifying input into three or more classes instead of two (binary classification). In this paper, Psorakis, Damoulas and Girolami attempt to solve this problem that RVM is facing by expanding the original RVM to multiclass multi-kernel relevance vector machines (mRVMs). The authors provided theoretical insights on mRVMs and their convergence properties, analyzed sparsity vs. accuracy trade-off as mentioned in the previous section, introduced a new methodology called “informative sample selection” for mRVMs in order to reduce the computational complexity, and evaluate the performance of mRVMs and other classification models. This paper contains extensive explanations about two mRVMs, including algorithms and mathematical derivations that lead to those algorithms, and a lot of graphs which present the experimental results the authors collected. The conclusion they arrive at is that mRVMs are definitely effective ways to solve multiclass classification problems, especially with the methodologies they proposed to boost the performance in terms of both computational cost and multiclass discrimination. The authors developed two mRMVs, one of which ($mRVM_1$) leads to more confident predictions as it has better sample identification properties, and the other one has more predictive power and better outlier detection capabilities.

Generally speaking, this paper extends the capabilities of RVM, which has only one kernel and is not suitable when it comes to multiclass classification by using multiple kernels in RVM along with other methodologies. The experimental results are very convincing: the authors successfully balanced the classification accuracy and computational complexity (number of RVs used). The differences between the performance of mRVMs and other competing classification models such as k-NN and E-M models are significant. There is no doubt that the performance of mRVMs are efficient and accurate. However, this paper did not mention the training stage in the whole process of classification which is equally important as the testing/predicting stage. As the previous source suggests, the training time for RVM is very long compared with SVM and other classification models. In this case, it is very likely that there is a huge tradeoff between the training time and the testing performance. How to maximize performance in both the training stage and the predicting stage considering this tradeoff might be a further research topic. In addition to that, the authors also suggested that it would be interesting to further analyze the qualitative properties of mRVM results with certain problem contexts.

Source 5: (Kexin Tian) T. Wang, H. Xu, J. Han, E. Elbouchikhi and M. E. H. Benbouzid, "Cascaded H-Bridge Multilevel Inverter System Fault Diagnosis Using a PCA and Multiclass Relevance Vector Machine Approach," in IEEE Transactions on Power Electronics, vol. 30, no. 12, pp. 7006-7018, Dec. 2015, doi: 10.1109/TPEL.2015.2393373.

Following the previous section, this paper also uses mRVM. It is used with principal component analysis (PCA) to get a fault diagnosis strategy for a cascaded H-bridge multilevel inverter system (CHMLIS). RVM, like SVM, is sparse, meaning that some samples have weights 0 so that they will not interfere with the decision boundary. In this paper, the authors took advantage of this property of RVM to achieve higher sparsity and shorter diagnosis time. And again, they used mRVM instead of RVM to classify multiple classes. In this paper, various kernels are used to check which one works the best. The conclusion is that linear kernel function is better for simple fault diagnosis, whereas Gaussian kernel function is more suitable when the fault diagnosis is more complicated. The experimental results shown in this paper confirms the theory, which is that using PCA-mRVM strategy during the training process could achieve higher sparsity and shorter diagnosis time.

This paper provides a new application for RVMs: fault diagnosis. Compared to SVMs, RVMs and mRVMs can provide probabilistic outputs for different classes, whereas SVMs can only make binary decisions. With all the discussions in this paper, it should be confident to say that RVM is a better classification model compared to SVM since RVM's kernel can be extended for multiple classes and RVM could provide more insight to the data sample and provide better and more accurate diagnosis. In this paper, the authors did not only compared RVM with SVM, it also compared RVM with another traditional method, backpropagation (BP) neural network. Similarly, mRVM has better performance. In the context described in this paper, PCA-mRVM fault diagnosis method is validated, but what about in other scenarios? For example, would this method still be useful when the sampling data does not shape like the sampling data for cascaded H-bridge multilevel inverter system? I think it would require further analysis and experimentations to determine the adaptability of this newly introduced method in this paper.

Source 6: (Kexin Tian) O. Williams, A. Blake and R. Cipolla, "Sparse Bayesian learning for efficient visual tracking," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1292-1304, Aug. 2005, doi: 10.1109/TPAMI.2005.167.

In this paper written by Wiiliams, Blake, and Cipolla, they first introduced kernel based SVMs that can be used to solve object recognition problems “elegantly”. However, this method is adapted to localization of spatial perturbation, the authors would like to achieve the temporal fusion of data, which cannot be achieved using SVM. the authors then used a fully probabilistic RVM in order to get observations using Gaussian function kernel that can be fused as time goes on. This paper focuses on a face tracking software that utilizes RVM, which is trained from a single image called seed image to generate a training set. The training process does not take a lot of time, hence making the RVM model efficient in terms of time duration.

This paper introduces yet another possible application of RVM. The paper is very comprehensive, and it includes experiment setups, results, statistics, and graphs, making it very encouraging.

Conclusion: (Jack Yuan and Kexin Tian)

Both SVM and RVM are state of the art classification models that are widely used after artificial neural networks. As discussed in this paper, these methods can be applied to physics, biology, computer vision, and many more fields. They can also be used with other methods, for instance PCA, to obtain more insightful results. In conclusion, RVM and SVM have comparable prediction accuracy, but RVM has a much higher sparsity, making the number of RVs significantly less that of SVs, saving a huge amount of computational complexity and time. RVM can be used, modified according to different problem context, and combined with other classification methods to maximize the computational efficiency. It is more flexible and more capable than SVM, hence it has a lot more potential to be discovered in many different fields in the future.