Relevance Vector Machines: An Introduction[1]

Koruk, Kasimcan (kasimcan.koruk@queensu.ca)

Abstract

Support Vector Machines (SVM) is a well-known machine learning method that has been available for many years. Although SVM offers a strong solution for machine learning problems combining generalization properties with sparse kernel technique, it normally does not provide posterior probabilities. As an alternative, Relevance Vector Machine (RVM) offers a Bayesian formulation to classification and regression problems. RVM is a promising machine learning method and open to new developments. This article reviews some basic principles of RVM, and it summarizes advantages and disadvantages of the method in comparison to SVM. The article also compares RVM and SVM according to the results of applications on a real drillhole dataset. The applications shows that the biggest challenge of RVM to be overcome is training time for huge datasets.

1. Introduction

Over-fitting is generally the main challenge for classification problems. SVM is one of the valid supervised learning methods which can handle over-fitting with minimum misclassification and maximum possible margins thanks to its generalization properties with sparse kernel technique (Tipping, 2001). SVM has become a popular method with several application examples in the literature (Tipping, 2000; Géron, 2017). With the help of support vectors utilized in decision functions, SVM provides sparsity to the solutions of machine learning problems.

Although SVM is a strong decision machine, it does not output posterior probabilities and the sparsity of SVM is limited, because the number of support vectors can increase linearly as the number of training data increases (Tipping 2000). As an alternative, Relevance Vector Machine (RVM) offers sparser solutions, and more importantly it offers a Bayesian formulation to classification problems (Bishop, 2006). RVM principally possesses the structure of SVM with some modifications. The article underlines the modifications and summarizes the differences between RVM and SVM in terms of formulation. The article also compares RVM and SVM according to the results of applications on a real drillhole dataset.

2. Relevance Vector Machines

2.1. Theoretical Review of RVM

Fundamentals of RVM are presented by Tipping (2000). RVM is fundamentally the same as SVM considering the functional form:

[1] Cite as: Koruk K (2021) Relevance Vector Machines: An Introduction, Predictive Geometallurgy and Geostatistics Lab, Queen's University, Annual Report 2021, paper 2021-08, 88-92.

$$y(x) = \sum_{n=1}^{N} w_n k(x, x_n) + b \tag{1}$$

The main difference is the introduction of a new hyper-parameter α which is assigned to each weight vector w_n. Posterior probability of the targets, t or y(x), is given by

$$p(t|X, w, \beta) = \prod_{n=1}^{N} p(t_n|x_n, w, \beta^{-1}) \tag{2}$$

where $\beta=\sigma^{-2}$. With these newly assigned hyper-parameters α, the posterior probability of the weight takes the form

$$p(w|\alpha) = \prod_{n=1}^{M} N(w_i|0, {\alpha_i}^{-1}) \tag{3}$$

After an initial value to hyper-parameters α and β, RVM predicts the probability iteratively. With each iteration, α and β are aimed to be maximized. As the hyper-parameters approximate to maximum values, the weights approximate to zero mean and covariance, and thus become redundant on the probability prediction. The rest of the vectors with non-zero weights control the model, and they are called 'relevance vector'. Unlike SVM, relevance vectors are not necessarily located on the boundary (Figure 1).

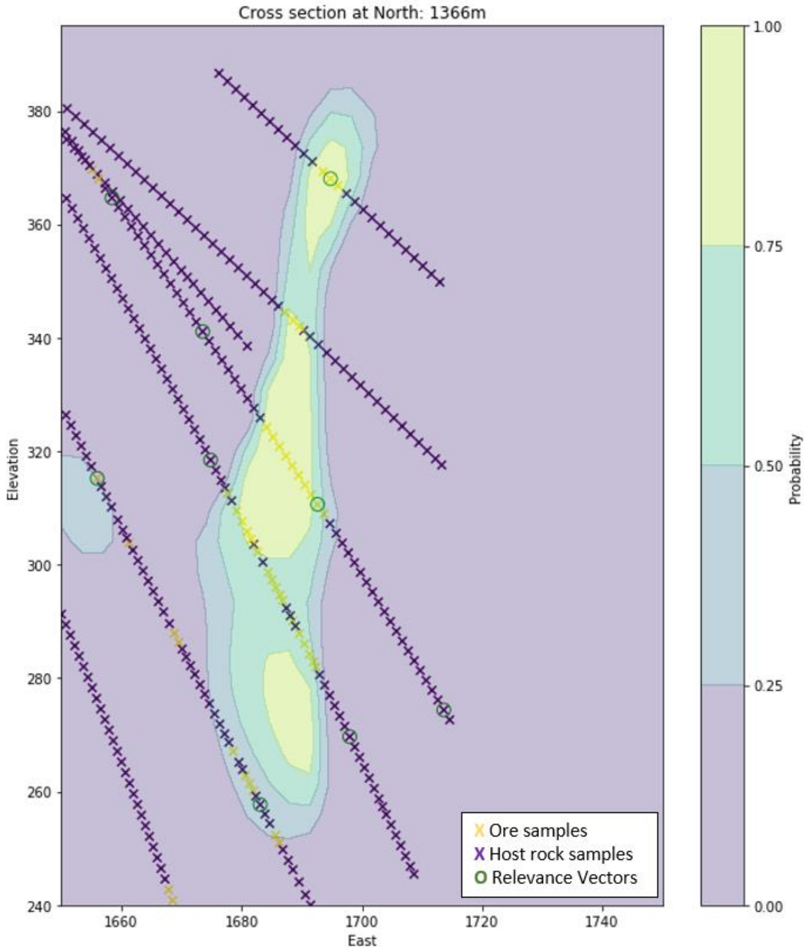

Figure 1 A cross-section of RVM model showing Bayesian probability overlain by training samples (composite drillhole samples)

While RVM looks more promising compared to SVM, there are some theoretical drawbacks of RVM worth to mention. All advantages and disadvantages of the RVM are listed in table below.

Table 1 Comparison of RVM and SVM

Advantages
RVM can make probability prediction.
Better than SVM when number of classes is more than two.
No need for cross-validation because there is no regularization parameter C.
Fewer decision functions because most of the weights of samples approximate to zero.
Disadvantages
More time for the training step.
Computation cost increases exponentially as number of classes increases.

2.2. Application of RVM and SVM

RVM and SVM are applied on a real drillhole data to make a tangible comparison. The comparison was based on the time spent on the training and accuracy of prediction results. Accuracy of the probabilities are left aside to further studies. The information about drillhole data is kept simple for the sake of the privacy of the project. The data is huge with more than 10000 binary-class samples. Ore samples are classified as 1, and the rest of the samples are classified as -1. When RVM is applied on the data directly, the training step lasts for hours and even days. To decrease the time spent on the training step, RVM and SVM are applied on a specific zone of the data narrowed to 1616 samples. The ratio of ore samples over the total samples is 12.19%. The data are split into training and test groups to assess the accuracy with 25% ratio.

SVM is applied on the data using SVC module of Scikit-learn (Pedregosa et al., 2011). The critical parameter shaping a model is gamma and regularization parameter C. Gamma defines how far a sample can have influence on the model. A low gamma makes the model general, and a large gamma can cause individuality of samples. The parameter C with low values makes the model smooth, and C with high values may cause overfitting. To apply SVC, cross validation is applied first to determine ideal parameters. Cross validation is performed to determine optimum parameters, then C and gamma are determined as 1000 and 5, respectively. Moreover, imbalance between the classes is taken into account using an option embedded in the module. RVM trials are done using the module EMRVC of the library sklearn-rvm 0.1.1, compiled by a university research group of King's College London. The project group claims that the library is compiled according to the implementation of Tipping (2000), and they adapted the API of scikit-learn to the module. As stated in the previous section, there is no parameter C in RVM. Gamma is determined equal to 5 as the final decision.

The results are illustrated in confusion matrices (Figure 2). The result of SVC (Figure 2b) is relatively better than that of EMRVC (Figure 2a), considering accuracy of ore samples. Predictions on ore samples increases up to 76% (Figure 2c) when the imbalance is imposed on the SVC model, at the cost of slightly losing accuracy of host rock prediction. On the other hand, SVC showed clearly better performance compared to EMRVC in terms of training time. Even when the cross-validation is considered, the time spent on the training was only 5 seconds for the case of SVC. However, training took 6 minutes to finalize the training using EMRVC.

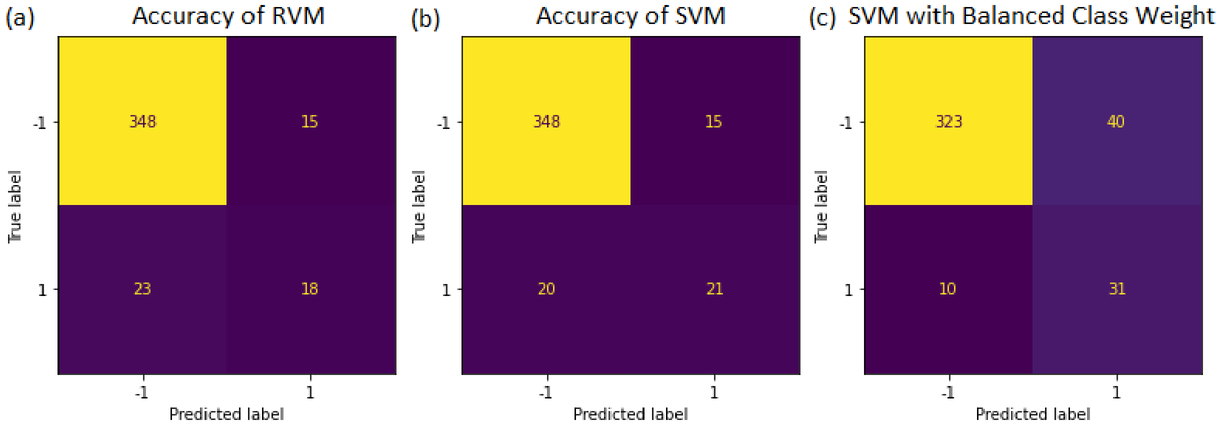

Figure 2 Accuracy of the results of (a) RVM, (b) SVM and (c) SVM with balanced class weight shown on confusion matrices

3. Discussion and Conclusion

RVM is theoretically a promising technique since it offers substantial developments to SVM. The sparsity of RVM is noteworthy considering the sparsity of SVM is limited. The absence of regularization parameter frees RVM from cross-validation step, which can be a time-consuming step for machine learning methods. Probabilistic outputs make RVM again valuable in the world of machine learning. Although RVM offers the advantages specified above, RVM has some practical obstacles to overcome against SVM. The application showed that training time spent by RVM is roughly 50 times more than the time spent by SVM. This problem is one of the most important obstacles to overcome for RVM. While the regularization parameter C requires cross-validation, it also offers higher control to the SVM which is not the case for RVM. The necessary changes on C made SVM model slightly more accurate compared to RVM. Lastly, while the SVC of Sci-kit learn is a fully developed module with detailed documentation, EMRVC of sklearn-rvc is a module developed by a university project group and requires some improvements. Because SVC offers solution for imposing the imbalance of classes, much higher accuracy is obtained in terms of predicting ore samples. Developing the option of balanced class prediction is especially necessary for the real-world datasets.

In conclusion, RVM has serious obstacles to overcome, and it requires some serious improvements before it displaces the position of SVM in the world of machine learning.

4. References

Bishop, CM, 2006. Kernel Methods & Sparse Kernel Machines. Pattern Recognition and Machine Learning. Springer Science+Business Media, LLC, Singapore.

Géron, A., 2017. Support Vector Machines. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow., 2nd Edition, O'Reilly Media, Inc.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, vol. 12, pp. 2825-2830.

Platt, JC., 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. MIT Press.

Tipping, Michael E., 2000. The relevance vector machine. Advances in neural information processing systems, pp. 652 – 658.

Tipping, Michael E., 2001. Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning Research 1, pp. 211-244.