FLOating Window Projective Separator (FloWPS): a method of data transfer for expression-based features from cell lines to cancer patients during SVM-based prediction of drug efficiency in personalized medicine

Victor Tkachev, Anton A. Buzdin

First Oncology Research and Advisory Center, buzdin@ponkc.com

Nicolas M. Borisov

National Research Centre "Kurchatov Institute", nicolasborissoff@gmail.com

The sophisticated nature of intracellular processes and events, which lead to cell proliferation and cancer progression, gives a hint that several machine-learning methods may be applied for prediction of clinical efficiency of certain drugs and treatment methods for individual patients [1-3]. Normally, machine learning in personalized medicine uses a training (*T*-) dataset, e.g., the results of gene expression profiling (or any other expression-based values) taken from the set of patients with previously known clinical outcome (responder or non-responder to the therapy).

Support vector machines (SVM) are among the most advanced and powerful tools for machine-leaning-based classification [4-7]. In comparison with other classification algorithms, e.g., classical multi-layer perceptrons (MLP) that use the least square fitting procedure for training data [8], SVMs have proved to be more robust in terms of the changes in input data and, therefore, less demanding for the huge number of vectors in the training set. Unfortunately, for most anti-cancer drugs, it is still extremely difficult (if ever possible) to find hundreds of gene expression profiles that were obtained using the same investigation platform for the patients that were treated with the same drug *with known clinical outcome of the treatment*. From the other side, thousands of expression profiling cases exist for various cell lines that were used for testing the ability of hundreds of drugs to inhibit the cell proliferation.

Here we are proposing a novel method for the transfer of expression profiling results from the more numerous cell lines to less abundant cases of real patients for subsequent application of machine-learning techniques that predict the clinical efficiency of anti-cancer drugs.

The most complicated operation in construction of machine-learning drug efficiency

prediction is the transfer of data from the training (*T*-) dataset (expression-based data for cell lines) to the validation (*V*-) one (the same data for real cancer patients). The problem of extrapolation in SVM has been recognized previously in other fields of research rather than bioinformatics, such as quantum chemistry [9-10], analytical chemistry, material science [11] or environmental engineering [12]. If the ranges of expression-based values in the support vector space for the *T*- and *V*-datasets *do not overlap*, SVM is doomed *to extrapolate rather than interpolate* without additional tuning. As a result, the whole prediction may easily produce incorrect, if not meaningless, results.

To prevent SVM from meaningless extrapolation, we are proposing the *floating window* method for the SVM tuning.

1) First, for each point of the *V*-dataset, we keep only those dimensions in the support vector space (any dimension corresponds to a distinct expression-based feature) that contain at least *M* points of the *T*-dataset both below and above this point of the *V*-dataset. We call this number *the margin width*, expressed in the number of support vectors in the *T*-dataset. If a certain expression feature does not satisfy this criterion, the corresponding dimension in the support vector space should not be taken into account, and the whole space should be reduced using a rectangular geometric projection along this dimension.
2) Second, similarly to the *k nearest neighbor* (*kNN*) method [13], we should take into account only *K* support vectors in the *T*-dataset, which are the proximal to a point in the *V*-dataset, where the drug efficiency is predicted.

These two *parameters* (*M*, *K*), which define the *floating window for data transfer*, should be adjusted for each combination of the *T*- and *V*-dataset to provide successful predictions. The optimal parameters (*M*, *K*) should satisfy the following conditions. First, the area under the ROC curve (AUC) for SVM-based prognosis of drug response for the *V*-dataset should me higher than a certain quality threshold (θ), say 0.7 or 0.75. Second, we should select as an optimal solution (M_{opt}, K_{opt}) the "top" of the "biggest area" where AUC > θ in the two-dimensional (*M*, *K*) space.

We have tested our modification of the SVM method, termed *FLOating Window Projective Separator* (*FloWPS*), using as the *T*-dataset the results of the CancerRxGene study [14] with

227 cell lines that were treated with 22 different kinase inhibitors (nibs).

As the *V*-dataset, we have used the gene expression profiling results for following below-tabulated cases [15, 16]. As a set of expression-based features, we used the pathway activation score (PAS) values calculated according to our method OncoFinder [2] instead of expression levels for distinct genes, since pathway-based values are more robust than gene-based [17]. Only cell signaling pathways, which contain molecular targets of certain drugs, were taken into account to construct the support vector space.

For each cancer-like disease type, which we have used for optimized *floating window* parameters (M_{opt}, K_{opt}) allow separation of responders from non-responders in the *V*-dataset with AUC > 0.70. Optimal *floating window* parameters (M_{opt}, K_{opt}) appear to be stable and robust according to the leave-one-out quality assurance procedure for the *V*-dataset.

The quality of whole *FloWPS* procedure was evaluated as the total area in the two-dimensional (*M*, *K*)-space when AUC > θ. In the case of myeloma therapy with bortezomib [18], this quality value expresses essential correlation (cor = 0.5, p-value = 0.004) with the median cosine between the normal vectors to the separating surfaces for the *T*- and *V*-datasets.

To check if the *FloWPS* procedure is not overtrained, we have done the permutation test, in which the real responder/non-responder flags for the samples of the *V*-dataset were replaced with random values. For all three disease types, this random permutation leads considerable decrease of total area where AUC > θ. The tabulated p-value shows the probability for unperturbed quality value to belong to the distribution for the perturbed quality values.

Disease type, drug	Renal cancer, sorafenib (current study)	Lung cancer, sorafenib [15]	CML, imatinib [16]
Number of samples	28 (13 responders, 16 non-responders)	37 (23 responders, 14 non-responders)	28 (16 responders, 12 non-responders)
AUC for optimized (*M*, *K*) parameters	0.81	0.72	0.78
p-value for Gaussian test whether *FloWPS* works better with real *V*-dataset than with the permutated one	0.04	0.16	0.06

One should note that our new *FloWPS* method enjoys advantages of both *local* (like the *kNN* method) and *global* (like the SVM) machine-learning techniques. This *hybrid/compromise* nature of *FloWPS* allows it to act successfully when purely local and global approaches fail. The confirmed *FloWPS* ability for the data transfer from one sample type (cell lines) to another (cancer patients) makes the novel method promising, e.g., for more the less radical data transfer between different expression profiling platforms, since universal cross-platform harmonizers of expression data of has not been fully tested yet.

References

1. S. Aggarwal (2010). *Nat Rev Drug Discov,* **9:** 427–428.

2. A.A. Buzdin et al. (2014). *Front Genet,* **5:**55.

3. A. Artemov et al. (2016). *Oncotarget*, **6**: 29347–29356.

4. E. Osuna et al (1997). An improved training algorithm for support vector machines, In: *Proceedings of the 1997 IEEE Workshop*, 276–285.

5. P. Bartlett, J. Shawe-Taylor (1999). Generalization performance of support vector machines and other pattern classifiers, In: *Adv. Kernel Methods—Support Vector Learn*, 43–54.

6. V. Vapnik, O. Chapelle (2000). *Neural Comput*, **12:** 2013–2036.

7. X. Robin et al. (2009). *Expert Rev Proteomics*, **6:** 675–689.

8. M.L. Minsky, S.A. Papert (1987). Perceptrons: An Introduction to Computational Geometry, Expanded Edition (MIT Press).

9. R. Arimoto et al. (2005). *J Biomol Screen*, **10:** 197–205.

10. R.M. Balabin, E.I. Lomakina (2011). *Phys Chem Chem Phys*, ***13*:** 11710–11718.

11. R.M. Balabin, S.V. Smirnov (2012). *Analyst*, **137:** 1604–1610.

12. G.D. Betrie et al. (2013). *Environ Monit Assess*, **185:** 4171–4182.

13. N.S. Altman (1992). *The American Statistician*, **46:** 175–185.

14. W. Yang et al (2013). *Nucleic Acids Res*, **41:** D955-D961.

15. G.R. Blumenschein (2013). *Clin Cancer Res*, **19:** 6967–6975.

16. L. C. Crossman (2005). *Haematologica*, **90:** 459-464.

17. N.M. Borisov et al (2014). *Oncotarget*, **5:** 10198–10205.

18. G. Mulligan et al (2007). *Blood*, **109:** 3177-3188.