RANSAC-BASED ENHANCEMENT IN DRUG CONCENTRATION PREDICTIONS USING SUPPORT VECTOR MACHINE

Wenqi You[(a)], Alena Simalatsar[(b)], Giovanni De Micheli[(c)]

[(a)(b)(c)]Ecole Polytechnique Federale de Lausanne

[(a)]wenqi.you@epfl.ch, [(b)]alena.simalatsar@epfl.ch, [(c)]giovanni.demicheli@epfl.ch

ABSTRACT

Training Support Vector Machines (SVMs) to predict drugs concentrations is often difficult because of the high level of noise in the training data, due to various kinds of measurement errors. We apply *RANdom SAmple Consensus* (RANSAC) algorithm in this paper to solve this problem, enhancing the prediction accuracy by more than 40% in our particular case study. A personalized sample selection method is proposed to further improve the prediction result in most cases.

Keywords: RANSAC, SVM, drug concentration predictions

1. INTRODUCTION

The decision-making regarding the drug dosage has been one of the main challenges of the pharmacological studies for decades. Population Analysis is a classical method to decide a dosage. It looks at a small number of data points per patient over many subjects [Bourne1995]. The models built by this method are applied to any new patient in clinical practice. However, due to both intra- and inter-differences of patients' characteristics, these models are not always accurate, therefore not applicable to some drugs whose therapeutic ranges are narrow. Furthermore, Population Analysis methods suffer from other limitations, such as not considering specific features, e.g. binary values, and moreover the number of features is limited. The Support Vector Machine (SVM) algorithm has been applied to address these problems [You2011]. Nevertheless, the performance of this algorithm highly relies on the quality of the training dataset. When building the predictor, SVM minimizes a cost function where a mean-squared error is very sensitive to noise in input data. Clinical measurements are particularly faced with the risk of measurement errors.

RANdom SAmple Consensus (RANSAC) is a general parameter estimation method proposed to filter out the outliers (errors) from input data [Fischler1981]. It resamples the input data and generates candidate solutions with respect to a minimum number of observations (data points) required to estimate the underlying model parameters. Depending on a threshold value, the input data are classified with different proportions into inliers (good data) and outliers. Only the inliers are considered to be useful to build the SVM model for drug concentration predictions. Unlike other sampling techniques that use as many data as possible to obtain an initial solution and then prune the outliers, RANSAC uses the smallest set possible and then enlarges this set with consistent data points [Fischler1981]. It has been applied to various domains such as sensor networks [Buttyan2006, Furukawa2006, Shafique2008], Integrated Chip (IC)'s three-dimensional information recognition [Liang2011] etc.

In this paper, we use the RANSAC algorithm to filter the datasets before running a Support Vector Machine (SVM) algorithm. Compared to an SVM-based algorithm in [You2011] and the Pharmacokinetic (PK) method [Widmer2006], it enhances the prediction by more than 40% in our experiments. Two scenarios for personalized predictions have also been tested to further improve the prediction accuracy.

The paper is organized as follows: Section 2 presents the methodology used in this paper. Section 3 shows the experimental results and comparisons with previous works. Finally, Section 4 draws a brief conclusion.

2. RELATED WORK

In the literature, predictions of drug concentrations are usually carried out using analytical models. These models are built based on some assumptions that the system of human body is one-, two- or three-compartment [Bourne1995, Bailey1991, Hahn2011]. This kind of assumption is widely applied to clinical practices nowadays, but suffers from some drawbacks such as it cannot take into account binary numbers and such as it is difficult to modify (add or remove) a parameter in the model. Thus, in [You2011], a Support Vector Machine (SVM)-based approach was proposed to overcome these drawbacks and tried to enhance the prediction accuracy.

SVM [Boser1992] was introduced by Boser, Guyon and Vapnik and became rather popular in several domains e.g. pattern recognition, computer vision, etc. It is a supervised learning model with associated learning algorithms that analyze data and recognize patterns. It has been successfully applied to human detection [Dalal2005], object recognition [Pontil1998],

Algorithm 1 RANSAC algorithm, where *data* is a set of observations, *model* is a model that can be fitted to data, K is the minimum number of data points required to fit the model parameters, N is the number of trials performed by the algorithm, T is a threshold determining if a data point fits a model, and *bestmodel* is the model fitting the highest number of data points.

```
Input: $data, model, K, N, T$
Output: $bestmodel$
  $bestinliers \leftarrow \emptyset$
  for $i = 1 \rightarrow N$ do
    $possibleinliers \leftarrow \text{SampleUniformly}(data, K)$
    $possiblemodel \leftarrow \text{Fit}(model, possibleinliers)$
    $inliers \leftarrow \emptyset$
    for all $point \in data$ do
      if $\text{Distance}(point, model) < T$ then
        $inliers \leftarrow inliers \cup \{point\}$
      end if
    end for
    if $|inliers| > |bestinliers|$ then
      $bestinliers \leftarrow inliers$
    end if
  end for
  return $bestmodel \leftarrow \text{Fit}(model, bestinliers)$
```

image classification [Chapelle1999], etc. It is simple in computation but also robust in data classification and regression, compared with other common machine learning methods, e.g. decision trees, neural networks, etc. However, according to our survey, SVM has not yet been applied to estimating drug concentrations.

To further enhance the performance of SVM, several previous works applied the *RANdom Sample Consensus* (RANSAC) algorithm. RANSAC, proposed by Fischler and Bolles [Fischler1981], is a general parameter estimation method used to deal with a large proportion of outliers in the input data. It was developed within the area of computer vision and applied to many other domains for data analysis. In [Nishida2008], the author claims a reduction of the computation requirement to about 1/170 compared with SVM libraries, and in [Kuo2007], RANSAC algorithm was used in a fine-selection stage for face recognition and achieved a lowest mean error rate.

3. METHODOLOGY

In this section, we will first introduce RANSAC algorithm that is used to improve drug concentration predictions together with Support Vector Machine (RANSAC-SVM). We will then propose a personalized drug concentration prediction scheme based on the RANSAC algorithm with two clinical scenarios.

3.1. RANSAC Algorithm

The RANSAC [Fischler1981] algorithm works as described in Algorithm 1. The number of trials N is set to be big enough to guarantee that at least one of the sets of possible inliers does not include any outlier with a high probability *p*. Usually *p* is set to 0.99. Let us assume, that *u* is the probability that any selected data point is an inlier, then *v=1-u* is the probability of selecting an outlier. *N* trials of sampling each *K* data points are required, where $1\text{-}p = (1\text{-}u^K)^N$. This way:

$$N = \frac{\log(1-p)}{\log(1-(1-u)^K)} \tag{1}$$

Algorithm 2 RANSAC-based personalized algorithm, where *training* is a set of *M* training samples, *newpatients* is an ordered set containing one sample per new patient , *Y* is the index of a particular feature, *model*, *K*, *N*, *T* are parameters of the RANSAC algorithm, and *bestmodels* is an ordered set containg the SVM model fitting the best each new patient.

```
Input: $training, newpatients, F, model, K, N, T$
Output: $bestmodels$
  $bestmodels \leftarrow \emptyset$
  for all $patient \in newpatients$ do
    $data \leftarrow \{patient, \ldots, patient\}$                  $\{|data| = M\}$
    $data \leftarrow data \cup training$
    $inliers \leftarrow \text{RANSAC}(data, model, K, N, T)$   $\{\text{predict } Y\}$
    $inliers \leftarrow inliers \setminus \{patient\}$
    $model \leftarrow \text{SVM}(inliers)$
    $bestmodels \leftarrow bestmodels \cup model$
  end for
  return $bestmodels$
```

The model of the RANSAC algorithm is a linear combination of several basis functions. The number of basis functions corresponds directly to the minimum number of points *K* required to fit the model. The parameters of the model are the weights of each basis function. In this paper, the drug concentration prediction method enhanced with filtering of the training dataset using RANSAC algorithm is called RANSAC-SVM method.

3.2. RANSAC-based Personalization

In Algorithm 1, inliers and outliers are separated without considering the information of a new patient. The SVM predictor for any new patient is estimated with the same inliers chosen as the training data. However, we believe that it might happen that the set of inliers for one patient is actually a set of outliers for another. Therefore, a predictor built out of the same set of inliers for a number of new patients might not be applicable for some others. Hence, it is important to find an individual set of inliers for each patient. So we propose to use RANSAC-based personalization method to solve this task. Previously [You2011], a 'closest point' strategy has been used which, despite using a much fewer number of training points (up to 30% of the total number), retains the initial performance of the original SVM (<3% degradation). However, that strategy needs a set of predefined weights for each feature in order to select the 'closest point'.

Assume that we already have *M* samples from previous patients (our training dataset), for each new patient we want to find a best subset of samples of *M* to train the SVM. To do so we treat them as if they were noisy samples of a new patient, and use a RANSAC to remove the outliers. The whole procedure is detailed in Algorithm 2. The new sample from each patient is first replicated *M* times to make sure that it will always be considered as an inlier. Then RANSAC is applied to those replicated samples plus to the original *M* training samples in order to predict the feature *Y* as a linear combination of basis functions of the remaining features *X*. The new patient sample is then removed from the inliers and finally an SVM is trained on the remaining original training samples to predict the drug concentration of this new patient.

Two clinical scenarios can be applied here: the target feature *Y* is set as:

1. Any feature other than the concentration value.
2. The measured drug concentration.

For scenario 1, no invasive blood test is required, while in 2 the drug concentration value should be measured after the first dosing.

4. EXPERIMENTS AND COMPARISONS

The experiments are conducted on a set of data collected during patients' treatment with *imatinib*, a drug designed to treat chronic myeloid leukemia and gastrointestinal stromal tumors [Widmer2006]. The training dataset consists of 54 patients and 252 samples, while the validation (testing) dataset contains the data of 65 patients and 209 samples.

To apply RANSAC, we first preset the basis using some typical functions: {x^{-2}, x^{-1}, x, x^2, x^3, log(x), cos(x), ($1-e^{-x}$), e^x}. This requires at least *K*=9 data points to estimate the parameters. However, not all the listed basis functions are useful to get the final model of drug concentration. Table 1 shows the experimental results on each basis function with respect to different thresholds (tolerable difference between the measured concentration and the predicted one). In practice, we set the threshold to be as small as possible to minimize the difference between the measured concentration values and the predicted ones. Hence, we combine the first two rows of the chosen basis functions (scored '1') in Table 1: f(x) = {x^{-2}, x, x^3, log(x), cos(x), ($1-e^{-x}$), e^x}. Figure 1 shows the AUC (Area Under the drug Concentration) curve estimated using RANSAC, the green points denote inliers and the blue represent outliers.

After determining the basis functions, the drug concentrations over the validated dataset are predicted via SVM algorithm. We evaluate the drug concentration prediction results of three algorithms (the traditional Pharmacokinetic (PK) [Widmer2006], SVM-based [You2011], and the proposed RANSAC-SVM) by computing an Absolute Difference between the Predicted concentration values and the Measured ones (ADPM). In practice, we expect ADPM values to be small. In our experiments, the RANSAC-SVM algorithm enhances the prediction performance by about 44.7% over the PK method and 42.6% over SVM-based method, respectively. Around 71% of mean ADPM values of RANSAC-SVM results are smaller than 500mg/L, while this number decreased to around 50% for the PK and SVM-based methods.

Table 1: RANSAC Basis Function Analysis With Respect To Different Thresholds. (T: Threshold with unit [mg/L]. '0' stands for 'unused' and '1' for 'in use'.)

T	x^{-2}	x^{-1}	x	x^2	x^3	$\log(x)$	$\cos(x)$	$1-\exp(-x)$	$\exp(x)$
250	1	0	1	0	1	1	1	1	0
500	0	0	1	0	0	1	1	0	1
1000	0	1	0	0	0	0	1	1	0
1500	0	1	0	0	0	0	0	0	0

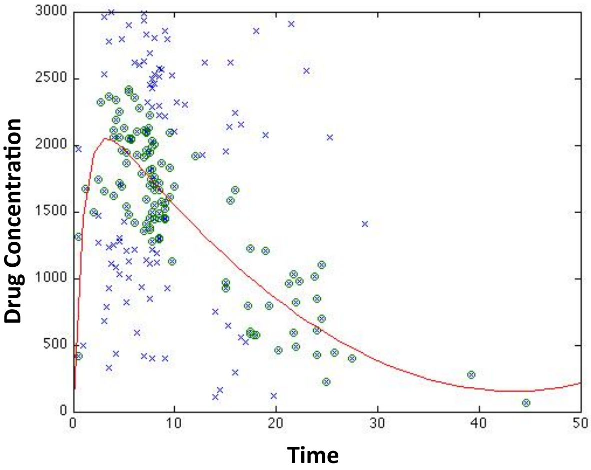

Figure 1: AUC (Area Under the plasma concentration time Curve) Using RANSAC Analysis. Green points are inliers and blue points are outliers.

For further prediction improvement, we apply two personalization scenarios with RANSAC algorithm (see Section 2). By choosing different features as *X* and *Y* to select the individual set of inliers for each new patient, we obtain the results shown in Table 2. In *imatinib* case study, the following features are available: {Measured Drug Concentration (MDC), Measuring Time (MT), Drug Dosage (DD), Age (A), Gender (G), and Body Weight (BW)}. Scenario 1 uses any feature other than MDC values while scenario 2 uses only MDC to be *Y*. We also compute the enhancement percentages with SVM [You2011] (shown as 'v.s. SVM' in the table) and Bayesian algorithm [Widmer2006] (shown as 'v.s. BAYE' in the table) with the Mean and STD results. The number of prediction samples whose predicted results are greater than 500mg/L from the measured values is denoted as '>500' in the table.

The Table 2 shows that the RANSAC-based personalization performs slightly better than RANSAC-SVM algorithm in many cases (1, 2, 3, 4, 5, 6), which results from a reduced number of predictions whose ADPM values are larger than 500mg/L. Both algorithms improve the prediction accuracy compared with SVM by around 40%. In scenario 2, BAYE outperforms the other two in the average prediction values in cases 7, 8, 9. However, in most cases, BAYE gives a larger STD value in that the predictions by BAYE deviate more from the measured values, while

Table 2: Comparisons Of The Drug Concentration Predictions Using RANSAC-based Personalization (RPER), RANSAC-SVM (RSVM), SVM [You2011], And Bayesian Estimation (BAYE) [Widmer2006]. ('>500': Number of prediction samples that are more than 500mg/L different from the measured values.)

Scenario 1: without blood measurement after first-dosing							
case	Features	Method	Mean	v.s. SVM	STD	v.s. SVM	> 500
1	Y =BW	RPER	258.60	42.10%	239.97	39.62%	6
	X =MT	RSVM	261.00	41.56%	240.64	39.45%	8
2	Y =A	RPER	222.39	50.21%	168.56	57.59%	3
	X =MT	RSVM	224.89	49.65%	168.40	57.63%	4
3	Y =G	RPER	282.21	36.81%	258.92	34.85%	7
	X =MT	RSVM	283.90	36.44%	260.56	34.44%	9
4	Y =DD	RPER	212.72	52.37%	170.93	56.99%	1
	X =MT	RSVM	213.35	52.46%	170.99	56.97%	1
5	Y =A	RPER	235.68	47.23%	181.60	54.31%	6
	X =BW	RSVM	243.79	45.42%	184.47	53.58%	6

Scenario 2: with blood measurement after first-dosing							
case	Features	Method	Mean	v.s. BAYE	STD	v.s. BAYE	> 500
6	Y =MDC	RPER	229.59	8.02%	211.34	43.65%	2
	X =MT	RSVM	239.58	4.01%	202.48	46.01%	3
7	Y =MDC	RPER	244.67	-17.74%	168.92	20.09%	10
	X =DD	RSVM	232.77	-12.02%	168.79	20.16%	7
8	Y =MDC	RPER	401.31	-103.42%	363.09	-89.80%	21
	X =G	RSVM	243.98	-23.67%	184.74	3.43%	7
9	Y =MDC	RPER	279.63	-15.30%	196.42	46.58%	8
	X =A	RSVM	247.46	-2.04%	199.13	45.84%	5
10	Y =MDC	RPER	219.33	8.18%	173.5	53.53%	3
	X =BW	RSVM	212.23	11.15%	155.22	58.43%	2

the other two algorithms estimate the concentration values without a large deviation. Hence, we can see that the proposed algorithms are robust to predict the concentrations more accurately for any individual patient, while Bayesian algorithm only predicts well for some patients and less accurate for the others.

5. CONCLUSIONS

This paper presents a RANSAC-based SVM algorithm to estimate drugs concentrations. RANSAC filters out the outliers in the input datasets to reduce the number of measurement errors. It exceeds the traditional Pharmacokinetic and SVM-based prediction methods by more than 40% in accuracy. The paper also introduces a way to personalize drug concentration prediction with the RANSAC algorithm. Experiments show that it enhances the proposed initial RANSAC-based SVM algorithm in many cases.

ACKNOWLEDGMENTS

Authors thank Carlotta Guiducci from EPFL for the help in manuscript revision and T. Buclin, N. Widmer and V. Gotta from CHUV Hospital of Lausanne for the precious suggestions on clinical data modeling and provision with sufficient data.

This work was supported by ISyPeM Project, with a grant from Swiss Nano-Tera.ch initiative, evaluated by Swiss National Science Foundation.

REFERENCES

Bailey, J. M. and Shafer, S. L., 1991. A Simple Analytical Solution to the Three-Compartment Pharmacokinetic Model Suitable for Computer-Controlled Infusion Pumps. *IEEE Transactions on Biomedical Engineering*, vol. 38, no. 6, pp. 522-525.

Bourne, D. W. A., 1995. *Mathematical Modeling of Pharmacokinetic Data*. 2nd ed. Technomic Publishing Company, Inc.

Buttyan, L., Schaffer, P. and Vajda, I., 2006. Ranbar: Ransac-based Resilient Aggregation in Sensor Networks. *Proceedings of the fourth ACM workshop on Security of ad hoc and sensor networks*. Oct. 30.

Chapelle, O., Haffner, P. and Vapnik, V. N., 1999. Support Vector Machines for Histogram-Based Image Classification. *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1055-1064.

Dalal, N. and Triggs, B., 2005. Histograms of Oriented Gradients for Human Detection. *Proceeding of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 886-893.

Fischler, M. and Bolles, R., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, vol. 24, no. 6, pp. 381-395.

Furukawa, Y., Sethi, A., Ponce, J. and Kriegman, D. J., 2006. Robust Structure and Motion From Outlines of Smooth Curved Surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 302-315.

Hahn, J. O., Dumont, G. A. and Ansermino, J. M., 2011. Closed-Loop Anesthetic Drug Concentration Estimation Using Clinical-Effect Feedback. *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 3-6.

Kuo, C. H. and Lee, J. D., 2007. A Two-Stage Classifier Using SVM and RANSAC for Face Recognition. *TENCON 2007*, pp. 1-4.

Liang, Z., Ye, B. and Xiao, Z., 2011. Vector Optimization of Integrated Chip Micro-image Chromatic Characteristic. *International Journal of Advancements in Computing Technology*, vol. 3, no. 11, pp. 170-177.

Nishida, K. and Kurita, T., 2008. RANSAC-SVM for Large-Scale Datasets. *Proceeding of the 19th International Conference on Pattern Recognition*, pp. 1-4.

Pontil, M. and Verri, A., 1998. Support Vector Machines for 3D Object Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 637-646.

Shafique, K., Hakeem, A., Javed, O. and Haering, N., 2008. Self Calibrating Visual Sensor Networks. *IEEE Workshop on Applications of Computer Vision*, pp. 1-6.

Widmer, N., Decosterd, L., Csajka, C., Leyvraz, S., Duchosal, M. A., Rosselet, A., Rochat, B., Eap, C. B., Henry, H., Biollaz, J. and Buclin, T., 2006. Population Pharmacokinetics of Imatinib and the Role of a_1-acid Glycoprotein. *British Journal of Clinical Pharmacology*, vol. 62, no. 1, pp. 97-112.

You, W., Widmer, N. and De Micheli, G., 2011. Example-based Support Vector Machine for Drug Concentration analysis. *Engineering in Medicine and Biology Society, EMBC*, pp. 153-157. Aug. 30 – Sept. 3, Boston, USA.

AUTHORS BIOGRAPHY

Ms. Wenqi You is currently a Ph.D. candidate in School of Computer and Communication Sciences at Ecole Polytechnique Federale de Lausanne, Switzerland. Her thesis is Mathematical Modeling for Decision Support System of Personalized Medicine. She holds a Bachelor degree from School of Electronic, Information and Electrical Engineering at Shanghai Jiao Tong University, China and a Master degree from Graduate School of Information, Production and Systems at Waseda University, Japan.

Dr. Alena Simalatsar is a postdoctoral researcher at Ecole Polytechnique Federale de Lausanne, Switzerland since February 2011. She holds a Bachelor and a Master degrees from the Faculty of Radiophysics and Computer Technologies of Belarusian State University in 2005. She received the Ph.D. degree (2009) in Computer Science and Telecommunication Technologies from University of Trento, during which she also spent six months as a Visiting Scholar at Electrical Engineering and Computer Science Department in University of California at Berkeley (2007).

Dr. Giovanni De Micheli is Professor and Director of the Institute of Electrical Engineering and of the Integrated Systems Centre at Ecole Polytechnique Federale de Lausanne, Switzerland. He is program leader of the Nano-Tera.ch program. Previously, he was Professor of Electrical Engineering at Stanford University. He holds a Nuclear Engineer degree (Politecnico di Milano, 1979), a M.S. and a Ph.D. degree in Electrical Engineering and Computer Science (University of California at Berkeley, 1980 and 1983). Prof. De Micheli is also a Fellow of ACM and IEEE and a member of the Academia Europaea.