Chemometrics and Intelligent Laboratory Systems 156 (2016) 174–180

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Machine learning for LC–MS medicinal plants identification

D.V. Nazarenko[a,*], P.V. Kharyuk[b], I.V. Oseledets[c,d], I.A. Rodin[a], O.A. Shpigun[a]

[a] Lomonosov Moscow State University, Faculty of Chemistry, Department of Analytical Chemistry, Moscow, Russia
[b] Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Department of Computational Technologies and Modeling, Moscow, Russia
[c] Skolkovo Institute of Science and Technology, Moscow Region, Russia
[d] Institute of Numerical Mathematics of Russian Academy of Sciences, Moscow, Russia

ARTICLE INFO

Article history:
Received 10 April 2016
Received in revised form 2 June 2016
Accepted 5 June 2016
Available online 8 June 2016

Keywords:
Plant species identification
Liquid chromatography–mass spectrometry
Machine learning
Multiclass classification

ABSTRACT

Herbal medicines are vigorously marketed, but poorly regulated. Analysis methodology for this field is still forming. One particular analytical task is confirmation of plant species identity for medicinal plants used as ingredients. In this work, machine learning approach has been implemented for LC–MS plant species identification. Samples for 36 plant species have been analyzed. Peak data (m/z, abundance) from respective samples have been used for development of classification algorithms. Namely, logistic regression (LR), support vector machine (SVM) and random forest (RF) techniques were used. For most of used machine learning algorithms, classification accuracy of 95% higher were obtained on cross-validation dataset. Now, massive training datasets are needed for full-scale application of this approach.

1. Introduction

Herbal remedies became popular alternative to modern health-care system. Global market is saturated with pills and decoctions based on various medicinal plants and amounts to billions of US dollars [1]. Some myths are persisting along with that trend [2]. The most popular one can be summarized briefly that chemically synthesized compounds are “unnatural” and, therefore, “unhealthy”, while naturally occurring substances are by the same logic all “harmless” regardless of their actual chemical composition. Surprisingly, FDA and EMEA hold to similar quality control strategies in this field and poorly regulate herbal remedies [3, 4]. Two main regulated characteristics are safety and effectiveness. But effectiveness is only assessed (if assessed at all) during registration procedures. Such drawbacks of quality control regulations caused among other things by complexity of analyzed substance.

Faced with herbal remedies which contain a large number of active compounds to be evaluated, analytical chemists developed a concept of “holistic analysis”. Main role in such strategies is given multivariate statistics. This approach can be described as:

1. Comprehensive chemical analysis (typically, LC-MS) of a batch of individual plant constituents and ready-to-use pills, if drug is a mix (e.g. TCM, going up to 10 and more plant decoctions in single formulae) [5].
2. Numerical analysis of obtained data in order to link a group of compounds with effectiveness and quality of respective drug. As a result, characteristic profile is generated, setting windows of acceptable concentrations for a number of quality markers. This profile is generally called “fingerprint” [6].

(U)HPLC coupled with mass-spectrometer is now standard instrumentation for such applications [7–12], with lesser number of works considering GC–MS [13, 14], NMR [15, 16], IR [17–19], and other analytical techniques (e.g. electronic nose [14], electrophoresis [12, 18, 20]).

As for multivariate statistics, researchers mostly use (O)PLS-DA, PCA and similar matrix factorization techniques (e.g. SIMCA [21], PARAFAC [12]), similarity analysis [7, 22, 23] etc. PLS-DA returns predictive model such as quality [8, 15] site of growth [17]/effectiveness assessment on base of constituents or for species discrimination [24]. PCA seeks for so called features, presented in case of LC/GC by peaks of components or their linear combinations, accounting for

Abbreviations: EMEA , European Medicines Agency; FDA , Food and Drug Administration; FN, false negatives; FP, false positives; (O)PLS-DA, (Orthogonal) Partial least squares-Discriminant Analysis; PARAFAC, parallel factor analysis; PCA, principal component analysis; RBF, radial basis function; SIL, standard isotope labeled; SIMCA, soft independent modeling by class analogy; SVM, support vector machine; TCM, Traditional Chinese Medicine; TN, true negatives; TP, true positives.

* Corresponding author.
E-mail address: dmitro.nazarenko@gmail.com (D. Nazarenko).

http://dx.doi.org/10.1016/j.chemolab.2016.06.003
0169-7439/

biggest differences in raw materials quality [7, 17, 25], plant material from different vendors [12]/growth sites [18, 26, 27] and differences between species [20, 24]. Also, PCA is used as a convenient tool for data dimensionality reduction.

General approach in fingerprinting is to employ technique to effectively (i.e. without significant loss of information) reduce size of dataset and obtain a reasonable number (up to tens) of quality control/identification markers.

Major flaws of such strategies are prices and availability of reference compounds and their SIL-analogues. Also, each species requires individual fingerprint development.

We started with the idea of "chemical image" of plant, i.e the ensemble of its major and/or characteristic compounds and their respective abundance ratios, uniquely defining plant species, part of the plant used, and, possibly, its quality. As most of plant extracts used in production of herbal medicines contain not more than about 50–100 compounds, the essence of "chemical image" should be the list of a few tens of compounds with normalized abundances.

The main difficulty in that perspective is that depending on site of growth (i.e. type of soil, weather, altitude, etc.), developmental stage, storage conditions and other factors, ratios of components in plant may change significantly, even to a point of complete loss of secondary metabolite/volatile/unstable compounds thus complicating identification procedure. This could be overcome by classification techniques resistant to partial data loss/corruption. Therefore, machine learning is the obvious methodology to apply here. Firstly, due to previously mentioned resistance of a number of algorithms to data loss, secondly, machine learning algorithms are useful tools for chemical image construction, as they seek to capture internal structure of data.

To create a program tool for standard-free plant identification, certain conditions in chemical analysis procedure should be met:

1. Platform-independence, including mass-spectrometer, column, mobile phase, flow rate, gradient program, and injection volume
2. Easy extraction procedure
3. Consistency of results from different platforms
4. (optional) Ability to simultaneously recognize multiple plants in complex matrices.

In this particular study, efforts were made to get a rough estimation on possibility of creating such software. Consequently, collection of massive datasets will theoretically allow it to create an effective machine learning algorithm for plant species identification.

2. Materials and methods

2.1. Chemicals and plant material

Methanol, ethanol, acetonitrile and formic acid were purchased from Merck (Germany). Deoinized water was purified with Milli-Q water system (Millipore, Milford, MA, USA). Plant material was purchased in ethnopharmacological stores in Moscow and through internet-suppliers. Flavonoids were bought from Phytolab (Germany). The stock solutions of 17 flavonoids were prepared in methanol and stored at −70 °C. All solutions were diluted to the desired concentration with methanol prior to use.

2.2. Sample extraction

3 g of each plant was powdered in agate mortar and sonicated for 1 h in 30 mL of 70% EtOH. 2 mL of crude extract were centrifuged for 10 min (10,000g); supernatant was filtered through 0.45 μm membranes, diluted 1 : 10 with 0.1% FA and 10 μL were subjected to LC–MS analysis. Procedure were selected to recover diverse compound classes [28] with regard to US Pharmacopoeia procedures [29].

2.3. LC–MS analysis

Chromatography was performed on a Hypersil Gold aQ (Thermo scientific, USA) column (100 mm × 2.1 mm i.d.,1.9 μm) using an LC-20 Prominence system (Shimadzu Corp., Japan) equipped with a binary solvent delivery system, degasser, an auto-sampler, and column oven. A binary gradient elution system consisted of 0.1% aqueous formic acid (A) and acetonitrile containing 0.1% formic acid (B) was used for separation using the following gradient program: 0% to 95% B (0–12 min), 95% B (12–17 min), 95% to 0% B for 0.01 min and 0% B for 3 min. The flow rate was 0.3 $mL \cdot min^{-1}$.

The above UPLC system was coupled to an LCMS-IT-TOF (Shimaszu Corp, Japan) equipped with an electrospray ionization (ESI) source. The capillary voltage and cone voltage were set at −3500(+4500) V and −5(+5) V, for negative (positive) polarity. The nebulization gas was set to 15 $L \cdot h^{-1}$ at 200 ° C. The cone gas was 1.5 $L \cdot h^{-1}$, and the source temperature was 200 ° C. MS data was acquired from a mass-to-charge ratio (m/z) range of 100–900. Acquisition was performed in both positive and negative polarities during single run. Each sample was injected for 5–6 times in randomized order to ensure that the carry-over between injections was minimized and to acquire sufficient amount of replicates for training datasets. Mix of standard compounds (17 flavonoids) was injected per every 20 samples to check system performance.

2.4. Classification task

The main task of classification algorithms is to provide a specific rule to distinguish given samples among possible classes.

In case of multiclass classification one should select an appropriate strategy to work with. In our task we tested both "one-vs-all" type where algorithm build one classificator per each class and "one-vs-one" that requires $K(K-1)/2$ pairwise classificators, where K is a number of classes. Although latter variant can in some applications be more accurate as it constructs separating hyperplanes for each pair of classes, its effectiveness in our case is compromised by low cardinality of dataset and larger computational costs. That is the reason why we decided to use one-vs-all strategy in this work.

2.5. Numerical experiments

2.5.1. Preprocessing

After LC–MS experiments 870 samples have been generated. According to visual inspection, there was no carry-over detected in blank runs.

Run files were converted into mzXML format. Then, chromatogram files were subjected to peak picking procedure in Waters Progenesis QI. A table with peaks (each with m/z, t_R and abundance values) was obtained for each experiment. Low-resolution equipment has been simulated by rounding all m/z values to integers. Also, we have rejected retention times for overall simplicity.

Maximum value from all peak response candidates to be in cell with given polarity (i.e. positive or negative) and m/z had been selected to avoid ambiguities. Finally we have obtained 3-dimensional array $A(i,j,k)$ with $i = \overline{1,m}$, $j = \overline{1,n}$, $k = \{0,1\}$, - numeration of samples, m/z values from 100 to 900 in increment of 1, negative or positive polarity value consequently.

Data have been represented as a matrix with $m \times 2 \cdot n$ size. Because we rejected retention time, feature space consists of concatenated m/z with positive and negative polarity only. Simple feature selection has been preformed by selection column subspace

with non-zero means. Hence number of features had been reduced to 632.

2.5.2. Experiment description

In the beginning of experiment 20 samples per class have been randomly selected to make data presented in each class consistent. Experiments have been performed for each method separately.

In the first part we seek for optimal set of parameters. For this purpose all input had been randomly divided into training set (70%) and validation set (30%). Separation had been performed inside each class in order to save data homogeneity. Then given method has been trained independently 100 times for the fixed parameters. Behavior of mean value of loss functional had been inspected for all settings, and those have been selected which had a minimal gap between training and validation curves.

In the second part the same randomized approach (100 starts with random permutations inside each class) have been used to plot learning curve dependent on the training and validation set size. Permuted input had been separated into equal sized sets due to the lack of experimental data. Training and validation sets have been incremented to show how it affects the loss functional. Because of very small number of samples, outstanding gap between both curves is typically observed. That means that algorithm tends to overfit the data. The problem had been particularly solved by regularization techniques but the explicit solution is to increase the number of samples and/or provide a more sophisticated feature selection/extraction method.

Final part included computation of common metrics for approach with selected parameters. These metrics are: accuracy, recall, precision, and F_1 value.

Let us consider a classification rule which decides either a given sample in $class_A$ or not. We can define following characteristics:

1. True positives (TP) as number of correctly classified samples within $class_A$;
2. True negatives (TN) as number of correctly classified samples without $class_A$;
3. False positives (FP) as number of misclassified samples within $class_A$;
4. False negatives (FN) as number of misclassified samples without $class_A$.

Now we can specify metrics considered above.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{F}_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision describes the relevance of items predicted as $class_A$ participants. Contrary, recall characterize how many relevant items are predicted. Accuracy indicates how well the classification algorithm classify input samples. F_1 score is another measure of accuracy defined as harmonic mean of precision and recall. It is usually applied to estimate comparative performance among different approaches.

Precision, recall and F_1-score have different extensions for multi-class tasks. We have chosen approach where metrics are calculated for each label followed by taking an unweighted mean.

2.5.3. Artificial data

The problem of low sample set can be attacked differently. One can increase the training set by adding similar data of other species. Due to the restrictions of chemical experiment (e.g., fixed extracting method) it is hard to collect appropriate samples from open databases. Extracting new samples manually presents severe difficulties (e.g., costs, species variety restriction).

In optical character recognition one way to go is to increase number of samples by applying several transformations like reflections and distortions (for instance, [30]). Similar idea may be applied to peak data leading to the artificial data generation.

In our simulations we disturbed from 15% up to 30% peak responses by scaling it with $\alpha \sim \text{Uniform}(-0.3, 0.3)$.

2.5.4. Tools

Computations were performed with scikit-learn package [31]. All experiments are implemented in Python programming language. Anaconda Python distribution [32] has been used as programming framework which includes various pre-built packages.

We provide open access to our source code and experiment data used in this study. All implemented algorithms and data are available by the following link: https://github.com/dmitro-nazarenko/chemfin-open.

3. Results and discussion

There are more than 400 species of medicinal plants in the biggest ethnopharmacology store in Moscow. Among them, 36 species were selected:

- Roots, seeds and leaves to include different parts of plants.
- All available species from Araliaceae (5 species). This family is interesting because it includes ginseng and other popular adaptogenes like *Eleutherococcus senticosus* and *Oplopanaxelatus* (Nakai).
- Umbelliferae (11 species). Umbelliferae is one of the most widely cultivated plant family.
- The rest 20 were chosen randomly to cover the assortment diversity (Table 1).

For most of plants we obtained samples from two consecutive harvests (2014 + 2015 or 2013 + 2015) and from different growth sites: Altai, Penza and Black Sea region. In one of the experiments, a batch of 36 extracts (all species) was dried in vacuum and heated at 60 °C for 5 days to simulate aging. Gradient elution program and MS analysis conditions were selected to be simple and reproducible. Our efforts were mainly focused on new methodology development and did not cover thorough species identification. On the other hand, species' identities were partly confirmed by LC–MS analysis results consistency for plant extracts from different suppliers. Typical chromatograms are presented in Fig. 1.

Fig. 2 demonstrates the overfitting trend for all tested approaches: the gap between a pair of learning curves of each

Table 1
Plant material used in experiment.

Part of the plant	Plant species
Roots	*Acanthopanax sessiliflorum*, *Eleutherococcus senticosus*, *Oplopanax elatus*, *Panax ginseng*, *Rhodiola rosea*, *Inula helenium*, *Helianthus tuberasus*, *Archangelica officinalis*, *Acorus calamus*, *Rosa majalis*, *Valeriana officinalis*, *Sambucus nigra*, *Glycyrrhiza glabra*, *Levisticum officinale*, *Ichorium intybus*, *Arctium lappa*, *Potenilla erecta*, *Dioscorea caucasica*, *Taraxacum officinale*, *Hedysarum neglectum*, *Aralia mandshurica*, *Astragalus membranaceus*
Seeds	*Coriandrum sativum*, *Daucus carota*, *Petroselinum crispum*, *Foeniculum vulgare*, *Anethum graveolens*, *Pimpinella anisum*, *Silybum marianum*, *Linum usitatissimum*
Leaves	*Bupleurum aureum*, *Pimpinella saxifraga*, *Heracleum sibiricum*, *Asarum europaeum*, *Aegopodium podagraria*

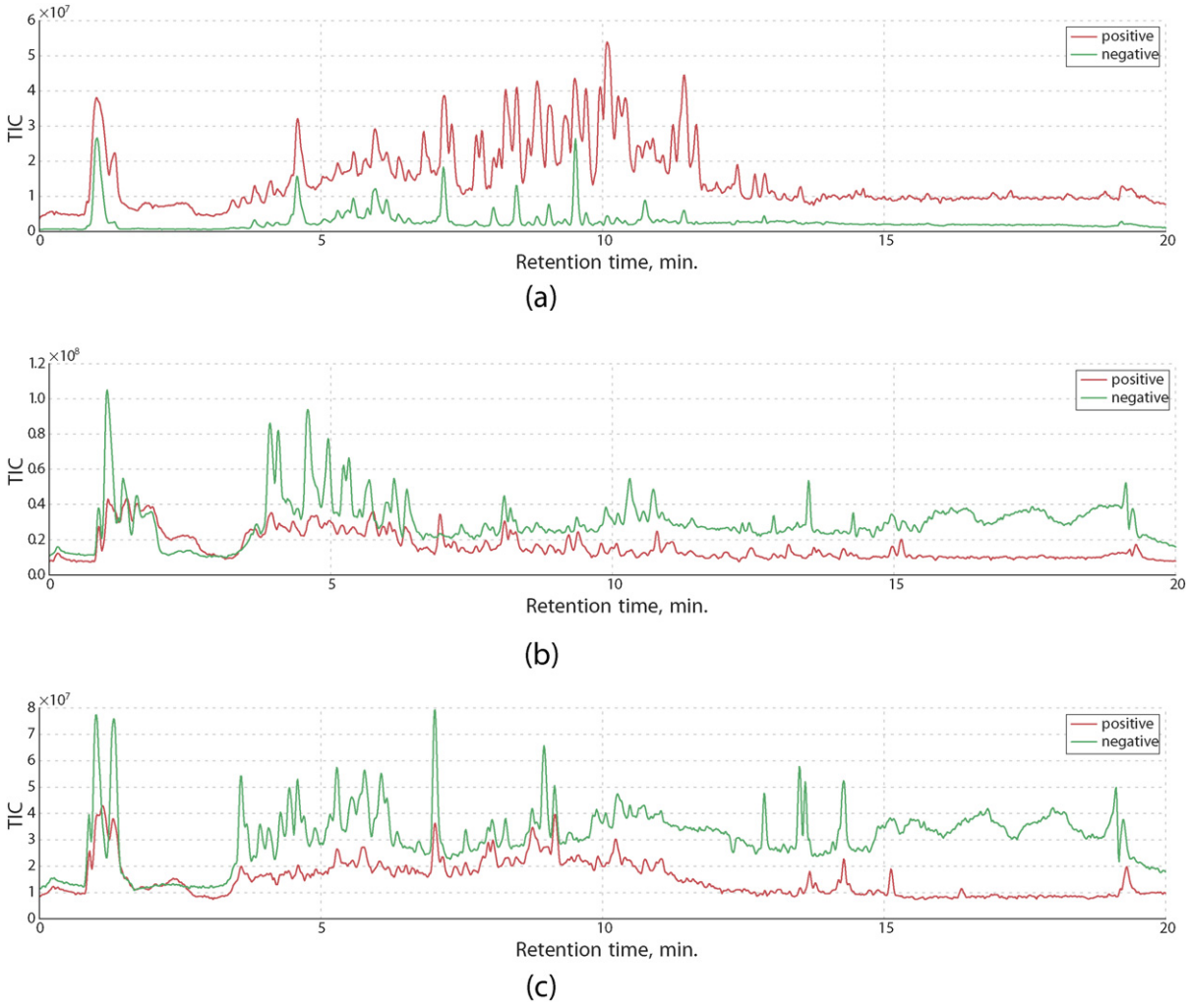

Fig. 1. Chromatograms of plant extracts: *Archangelica officinalis* (roots) (a), *Asarum europaeum* (leaves) (b) and *Pimpinella anisum* (seeds) (c) in postitive (red) and negative (green) ionization modes. TIC—Total ion current.

algorithm is preserved with set size moving higher. However, if we had larger cardinality of the training and validation sets, there is a strong possibility that this gap would be depleted. Even if this is not the case, more sophisticated approaches can be applied if only feasible dataset is acquired. Explanation of parameters and algorithms is given in appendix section.

Also, as mentioned above, we tested two multi-class classification strategies, “one-vs-all” and “one-vs-one” on the example of logistic regression approach. In Fig. 3 the comparison of both strategies is presented. Since results did not differ significantly and one-vs-one strategy requires more computational costs, one-vs-all strategy has been selected.

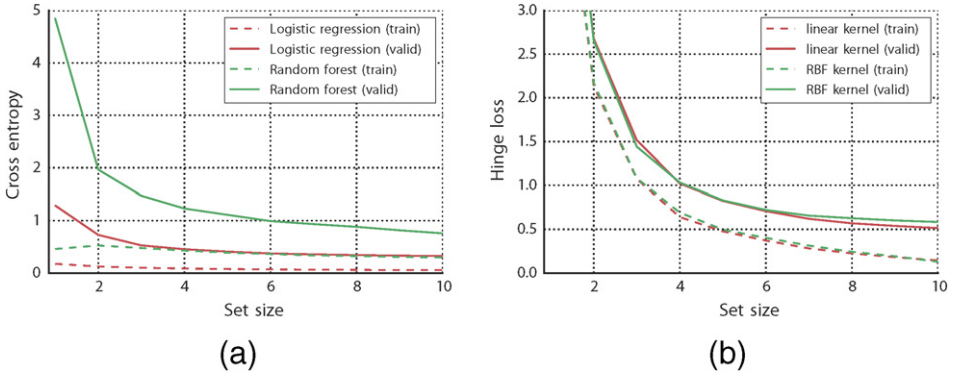

Fig. 2. Learning curves for tuned classification algorithms: dashed line indicates results for training set, solid line for validation, (a) in terms of cross entropy loss (logistic regression in red and random forest in green) and (b) hinge loss (SVM with linear kernel in red and RBF kernel in green).

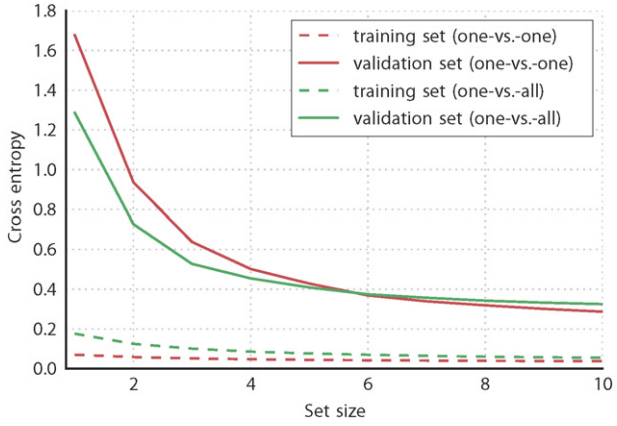

Fig. 3. Comparison of "one-vs-all" and "one-vs-one" strategies for tuned logistic regression classification.

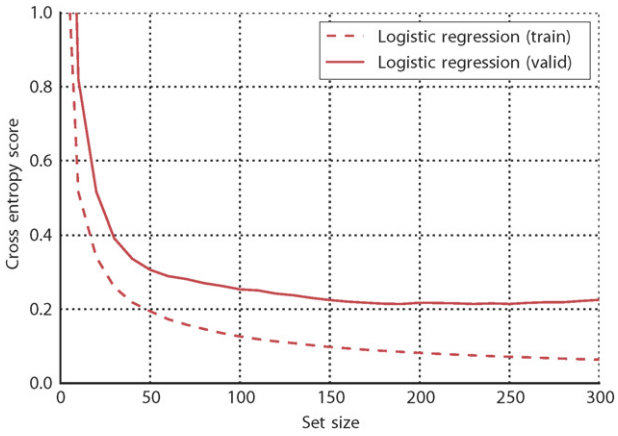

Fig. 5. Learning curves for logistic regression on artificially generated data, dashed line indicates results for training set, solid line for validation set, in terms of cross entropy loss.

It is worth nothing that parameters of each algorithm had been tuned with respect to available data. Parameter values had been increased exponentially without hard evaluation because it was not an objective of this work. Example for random forest is demonstrated in Fig. 4. Parameters must be agreed with appropriate loss functional to minimize it on validation and training sets simultaneously. In order to this requirement, depth of each decision tree varies from 15 to 25 and number of elementary decision trees should be greater than 20.

In Fig. 5 learning curves for artificial data are presented. It is apparent that naive approach considered above does not work and more sophisticated modeling is required.

Table 2 demonstrates performance of selected algorithms in terms of characteristics specified in one of the previous sections. There are two values in cells of the table, mean percentage for training and validation set consequently. All approaches showed ¿95% accuracy on validation set; however, eliminated behavior of learning curves indicates high variance trend and the problem of generalization appears.

Using retention time in classification task can greatly facilitate algorithm specificity, but at the same time presents a whole new level of difficulty. Retention times of compounds greatly depend on type of column, solvent, gradient elution program, mobile phase modifiers etc. Therefore, it could be efficient to introduce some categorization of retention time values. Since C18 columns are dominant in this field, retention time values strongly correlate with polarity of eluted compounds. Accordingly, categories can be introduced, for example, as: very polar, polar, intermediate, low-polar and non-polar. Hence, retention time values present additional feature axis for large scale classification purposes.

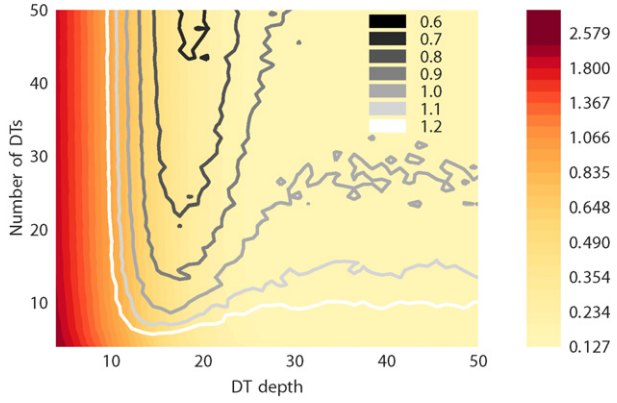

Fig. 4. Logistic loss values for random forest according to number of decision trees and depth of each tree. Heat map corresponds to losses on training dataset; contour lines coincide with losses on validation data.

4. Conclusion

So far, no standard-free strategy for LC–MS plant species identification has been developed. Machine learning approaches have been implemented to broaden existing plant analysis methodology. Practically all contemporary mass-spectrometers can be used for MS1 spectra collection in scanning mode both for positive and negative polarity during single run. This data was used for recognition algorithms development (i.e. plant species identification). Alternatively, GC–MS instruments allow individual component identification and in theory facilitate species identification. But to the best of our knowledge, no general strategy has been proposed in this field yet.

Logistic regression, SVM, and random forest were implemented to classify 36 species of medicinal plants with sufficient (¿95%) accuracy. It is essential now to test applicability of this approach by training algorithms for classification of a much larger species number. However, this step is limited by a significant lack of experimental data.

When available dataset is relatively small, machine learning algorithms tend to overfit dataset and thus fail to effectively operate on new samples. Experiments showed that generated dataset resulted in models with so called "high variance" problem, i.e. good modeling of training data and considerably worse perfomance on cross-validation data. This drawback can be minimized by regularization technique or by narrowing feature space. Following this course, retention time axis was rejected to result in 1-dimensional data vectors and regularization was applied. Nevertheless, the best way to capture internal structure of data is to use as large training dataset as possible. Also, more sophisticated feature selection approaches are likely to make a contribution.

Other practical solution — generation of artificial data. But this approach is likewise restricted: in order to correctly generate artificial data we need compound identification as concentration fluctuations depend on classes of phytochemicals. Furthermore, information about fluctuation patterns for all considered species and metabolite classes also needed.

To make an easy and reliable plant identification tool it will require to collect a huge amount of data about thousands of species. So it will most likely require collaborative efforts of analysts in this

Table 2
Comparative characteristics of implemented approaches on original data.

Method	Accuracy, %	Recall, %	Precision, %	F_1, %
Logistic regression ($\lambda = 0.6$)	99.74/96.05	99.74/96.05	99.75/96.72	99.74/96.02
SVM (linear; $C = 10^{-2}$)	98.11/96.45	98.11/96.45	98.18/96.98	98.10/96.41
SVM (RBF; $C = 10^2$, $\gamma = 10^{-4}$)	98.11/96.61	98.11/96.61	98.16/97.12	98.10/96.58
Random forest ($d = 20$, $n = 45$)	99.82/96.56	99.82/96.56	99.83/97.06	99.82/96.52

field, interested in proposed identification strategy as it was with many databases for recognition/identification algorithms.

Acknowledgments

Computational part of this work was supported by Russian Science Foundation Grant 14-11-00659.

Appendix A. Classification algorithms

There exists a great number of good practical and theoretical guides to machine learning including approaches used in this study, for instance [33–35], etc. Here our objective is to briefly describe those which used in this work.

Having multi-class problem where each sample must be distinguished among several classes. Let us consider two-class classification for simplicity.

There are different algorithms applicable in classification problems. The objective may be considered as obtaining a posterior probabilities of an input to be transformed to one over available labels, as in *logistic regression* approach.

Logistic regression is a special case of generalized linear model, where the dependent variables are assumed to be binary categorical. It uses the following hypothesis on the dependent variables:

$$h(x) = \theta\left(\sum_{i=1}^{n} \omega_i x_i\right) = \theta\left(w^T x\right), \tag{1}$$

where $\theta(z) = \frac{1}{1+e^{-z}}$ is a sigmoid function, $w \in \mathbb{R}^n$— parameters, $x \in \mathbb{R}^n$— input sample.

In logistic regression model we make an attempt to estimate posterior probabilities using a considered hypothesis. Parameters w are computed in order to minimize cross-entropy which is widely used to measure error between predicted output and real values:

$$w = \arg\min_{w} \left[\frac{1}{m}\sum_{k=1}^{m} \ln\left(1 + e^{-y_k \cdot w_k^T x_k}\right) + \lambda \|w\|_2^2\right], \tag{2}$$

(x_k, y_k)—k-th sample with class label, $x_k \in \mathbb{R}^n$, $y_k \in \{-1, +1\}$, $w \in \mathbb{R}^n$— parameters to estimate. The second additive component is regularization term: it penalizes high values of w with weight λ.

In contrary to logistic regression, there is no natural probabilities in *support vector machine (SVM)* approach. The objective of SVM method is to evaluate parameters w so that it has following properties:

1. $w = [w_0, w_1, \ldots, w_n] \in \mathbb{R}^{n+1}$ is a vector of minimum possible norm, $w = \arg\min_{w \in \mathbb{R}^{n+1}} \|w\|_2^2$;
2. w defines a hyperplane which separates representatives of differenet classes.

Here we considered two-categorical case for simplicity. Mathematically the task can be expressed as constrained quadratic optimization problem:

$$\begin{cases} w = \arg\min\limits_{w \in \mathbb{R}^{n+1}, \eta \in \mathbb{R}^n} \left[\frac{1}{2}\|w\|_2^2 + C\sum\limits_{k=1}^{m} \eta_k\right] \\ y_k\left[w_0 + (\hat{w}, x_k)\right] \geq 1 - \eta_k, \quad \forall k = \overline{1,m} \\ \eta_k \geq 0, \quad k = \overline{1,m} \end{cases}, \tag{3}$$

where $w \in \mathbb{R}^{n+1}$ is a weight vector to be tuned, $C > 0$ is a regularization parameter, $\eta_k \geq 0$ is an error on k-th sample. In other words, SVM naturally constructs the separating hyperplane with the largest margin (or gap) between classes.

However, there is no guarantee for this method to perform well on the data which is not linearly separable. To deal with this problem either soft-margin approach or generalization of SVM with so-called kernel functions (or kernels) are commonly used. The idea of kernels is to transform feature set into new one by special function which must agree the conditions of Mercer theorem. For instance, radial basis function $K(u, v) = \exp\left(-\gamma \|u - v\|_2^2\right)$ (Gaussian kernel with parameter γ) is commonly used as similarity measure between pairs of samples, and instead of solving the task for samples directly, we make an attempt to classify it by its similarities.

Decision tree is a non-parametric supervised learning method applied in classification. It infers simple decision rules from the samples and builds hierarchical partitioning of feature space.

However, deeply built decision trees tend to overfit the training set leading to bad ability to generalization. In order to prevent it, a common practice is to use ensemble methods. One of them is *random forest* method. A random forest is an ensemble method that uses several simple estimators (decision trees) on various subsets of source dataset. Also a random selection of feature subspace can be provided for each outstanding classifier. Final decision is obtained by averaging technique which improves reliability and accuracy. Detailed explanation of random forests may be found in [36].

Loss function observed in this study vary according to the method:

- cross-entropy (logistic regression, decision tree);
- hinge loss (support vector machine).

Recall cross-entropy (or logistic loss):

$$J_{\text{CE}} = \frac{1}{m}\sum_{k=1}^{m} \ln\left(1 + e^{-y_k \cdot w_k^T x_k}\right), \tag{4}$$

(x_k, y_k)—k-th sample with class label, $x_k \in \mathbb{R}^n$, $y_k \in \{-1, +1\}$, $w \in \mathbb{R}^n$— estimated parameters.

The cumulated hinge loss is an upper bound of the number of mistakes made by the classifier and defined as

$$J_{\text{hinge}} = \frac{1}{m}\sum_{k=1}^{m} \max\left(0,\ 1 - y_{\text{out}}^k \cdot y_{\text{true}}^k\right), \tag{5}$$

where $y^k_{\text{out}} \in \mathbb{R}$, $y^k_{\text{true}} = \pm 1-$ predicted output and real label for k-th sample.

References

[1] R.R. Chaudhury, Herbal remedies and traditional medicines in reproductive health care practices and their clinical evaluation, J. Reprod. Health Med. 1 (1) (2015) 44–46.

[2] Geoffrey A. Cordell, Phytochemistry and traditional medicine the revolution continues, Phytoch. Lett. 10 (2014) xxviii–xl.

[3] European Parliament and of the Council, Directive 2004/24/EC, 2004, URL http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32004L0024&qid=1451884773824.

[4] Food and Drug Administration, Dietary Supplements, URL http://www.fda.gov/Food/DietarySupplements/.

[5] C. Tistaert, B. Dejaegher, Y.V. Heyden, Chromatographic separation techniques and data handling methods for herbal fingerprints: a review, Anal. Chim. Acta 690 (2) (2011) 148–161.

[6] J. Riedl, S. Esslinger, C. Fauhl-Hassek, Review of validation and reporting of non–targeted fingerprinting approaches for food authentication, Anal. Chim. Acta 885 (2015) 17–32.

[7] C. Fan, J. Deng, Y. Yang, J. Liu, Y. Wang, X. Zhang, K. Fai, Q. Zhang, W. Ye, Multi-ingredients determination and fingerprint analysis of leaves from *Ilex latifolia* using ultra-performance liquid chromatography coupled with quadrupole time-of-flight mass spectrometry, J. Pharm. Biomed. Anal. 84 (2013) 20–29.

[8] J.B. Wan, X. Bai, X.J. Cai, Y. Rao, Y.S. Wang, Y.T. Wang, Chemical differentiation of Da-Cheng-Qi-Tang, a Chinese medicine formula, prepared by traditional and modern decoction methods using UPLC/Q-TOFMS-based metabolomics approach, J. Pharm. Biomed. Anal. 83 (2013) 34–42.

[9] H. Sheridan, L. Krenn, R. Jiang, I. Sutherland, S. Ignatova, A. Marmann, X. Liang, J. Sendker, The potential of metabolic fingerprinting as a tool for the modernisation of TCM preparations, J. Ethnopharmacol. 140 (3) (2012) 482–491.

[10] J. Wang, H. Kong, Z. Yuan, P. Gao, W. Dai, C. Hu, X. Lu, G. Xu, A novel strategy to evaluate the quality of traditional Chinese medicine based on the correlation analysis of chemical fingerprint and biological effect, J. Pharm. Biomed. Anal. 83 (2013) 57–64.

[11] Y.X. Chang, A.H. Ge, S. Donnapee, J. Li, Y. Bai, J. Liu, J. He, X. Yang, L.J. Song, B.L. Zhang, X.M. Gao, The multi-targets integrated fingerprinting for screening anti-diabetic compounds from a Chinese medicine Jinqi Jiangtang Tablet, J. Ethnopharmacol. 164 (2015) 210–222.

[12] J. Mazina, M. Vaher, M. Kuhtinskaja, L. Poryvkina, M. Kaljurand, Fluorescence, electrophoretic and chromatographic fingerprints of herbal medicines and their comparative chemometric analysis, Talanta 139 (2015) 233–246.

[13] C. Cordero, P. Rubiolo, L. Cobelli, G. Stani, A. Miliazza, M. Giardina, R. Firor, C. Bicchi, Potential of the reversed-inject differential flow modulator for comprehensive two-dimensional gas chromatography in the quantitative profiling and fingerprinting of essential oils of different complexity, J. Chromatogr. A 1417 (2015) 79–95.

[14] H. Cheng, Z.H. Qin, X.F. Guo, X.S. Hu, J.H. Wu, Geographical origin identification of propolis using GC–MS and electronic nose combined with principal component analysis, Food Res. Int. 51 (2) (2013) 813–822.

[15] L. Tarachiwin, A. Katoh, K. Ute, E. Fukusaki, Quality evaluation of *Angelica acutiloba* Kitagawa roots by 1H NMR-based metabolic fingerprinting, J. Pharm. Biomed. Anal. 48 (1) (2008) 42–48.

[16] M.A. Farag, A. Porzel, L.A. Wessjohann, Unraveling the active hypoglycemic agent trigonelline in *Balanites aegyptiaca* date fruit using metabolite fingerprinting by NMR, J. Pharm. Biomed. Anal. 115 (2015) 383–387.

[17] Q. Fan, C. Chen, Y. Lin, C. Zhang, B. Liu, S. Zhao, Fourier Transform Infrared (FT-IR) Spectroscopy for discrimination of Rhizoma gastrodiae (Tianma) from different producing areas, J. Mol. Struct. 1051 (2013) 66–71.

[18] W. Liu, J. Xu, R. Zhu, Y. Zhu, Y. Zhao, P. Chen, C. Pan, W. Yao, X. Gao, Fingerprinting profile of polysaccharides from *Lycium barbarum* using multiplex approaches and chemometrics, Int. J. Biol. Macromol. 78 (2015) 230–237.

[19] M. Sandasi, G.P. Kamatou, S. Combrinck, A.M. Viljoen, A chemotaxonomic assessment of four indigenous South African *Lippia* species using GCMS and vibrational spectroscopy of the essential oils, Biochem. Syst. Ecol. 51 (2013) 142–152.

[20] H. Huang, J. Sun, J.-A. McCoy, H. Zhong, E.J. Fletcher, J. Harnly, P. Chen, Use of flow injection mass spectrometric fingerprinting and chemometrics for differentiation of three black cohosh species, Spectrochim. Acta, Part B 105 (2015) 121–129.

[21] J. Viaene, M. Goodarzi, B. Dejaegher, C. Tistaert, A.H.o.a.n.g.L.e. Tuan, N.N.g.u.y.e.n. Hoai, M.C.h.a.u. Van, J. Quetin-Leclercq, Y.V.a.n.d.e.r. Heyden, Discrimination and classification techniques applied on *Mallotus* and *Phyllanthus* high performance liquid chromatography fingerprints, Anal. Chim. Acta 877 (2015) 41–50.

[22] M. Goodarzi, P.J. Russell, Y.V. Heyden, Similarity analyses of chromatographic herbal fingerprints: a review, Anal. Chim. Acta 804 (2013) 16–28.

[23] G. Alaerts, J.V.a.n. Erps, S. Pieters, M. Dumarey, A.M. van Nederkassel, M. Goodarzi, J. Smeyers-Verbeke, Y.V.a.n.d.e.r. Heyden, Similarity analyses of chromatographic fingerprints as tools for identification and quality control of green tea, J. Chromatogr. B 910 (2012) 61–70.

[24] K.H. Wong, V. Razmovski-Naumovski, K.M. Li, G.Q. Li, K. Chan, Differentiation of *Pueraria lobata* and *Pueraria thomsonii* using partial least square discriminant analysis (PLS-DA), J. Pharm. Biomed. Anal. 84 (2013) 5–13.

[25] M. Yuan, R.F. Wang, L.J. Liu, X. Yang, Y.S. Peng, Z.X. Sun, Contribution evaluation of the floral parts to orientin and vitexin concentrations in the flowers of *Trollius chinensis*, Chin. J. Nat. Med. 11 (6) (2013) 699–704.

[26] F. dos Santos Grasel, M.F. Ferro, C.R. Wolf, Development of methodology for identification the nature of the polyphenolic extracts by FTIR associated with multivariate analysis, Spectrochim. Acta A Mol. Biomol. Spectrosc. 153 (2016) 94–101.

[27] B.-Y. Li, Y. Hu, Y.-Z. Liang, P.-S. Xie, Y.-P. Du, Quality evaluation of fingerprints of herbal medicine with chromatographic data, Anal. Chim. Acta 514 (1) (2004) 69–77.

[28] J. Azmir, I.S.M. Zaidul, M.M. Rahman, K.M. Sharif, A. Mohamed, F. Sahena, M.H.A. Jahurul, K. Ghafoor, N.A.N. Norulaini, A.K.M. Omar, Techniques for extraction of bioactive compounds from plant materials: a review, J. Food Eng. 117 (4) (2013) 426–436.

[29] USP36 NF31, 2013: U.S. Pharmacopoeia National Formulary, United States Pharmacopeial, Secaucus, NJ, USA, 2012.

[30] Y. Netzer, T. Wang, A. Coates, A. Bissacco, Bo. Wu, Andrew Y. Ng, Reading digits in natural images with unsupervised feature learning, NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[32] C. Analytics, Anaconda Software Distribution, Computer software. Vers. 2-2.4.0, 2015, URL https://continuum.io.

[33] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[34] A. Ng, Machine Learning. Standford Course, 2015. URL http://cs229.stanford.edu/materials.html.

[35] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, second ed., Springer, New York, 2009.

[36] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32. URL http://dx.doi.org/10.1023/A:1010933404324. http://dx.doi.org/10.1023/A:1010933404324.