Published in final edited form as:

Anal Chem. 2012 March 20; 84(6): 2686–2694. doi:10.1021/ac202755e.

Incorporation of support vector machines in the LIBS toolbox for sensitive and robust classification amidst unexpected sample and system variability

Narahara ChariDingari^{a,*}, Ishan Barman^{a,*}, Ashwin Kumar Myakalwar^b, Surya P. Tewari^b, and G. Manoj Kumar^{b,†}

^aLaser Biomedical Research Center, G. R. Harrison Spectroscopy Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

^bAdvanced Centre of Research in High Energy Materials (ACRHEM), South Campus, University of Hyderabad, Prof C R Rao Road, Central University Campus PO, Gachibowli, Hyderabad, 500046, India

Abstract

Despite the intrinsic elemental analysis capability and lack of sample preparation requirements, laser-induced breakdown spectroscopy (LIBS) has not been extensively used for real world applications, *e.g.* quality assurance and process monitoring. Specifically, variability in sample, system and experimental parameters in LIBS studies present a substantive hurdle for robust classification, even when standard multivariate chemometric techniques are used for analysis. Considering pharmaceutical sample investigation as an example, we propose the use of support vector machines (SVM) as a non-linear classification method over conventional linear techniques such as soft independent modeling of class analogy (SIMCA) and partial least-squares discriminant analysis (PLS-DA) for discrimination based on LIBS measurements. Using over-the-counter pharmaceutical samples, we demonstrate that application of SVM enables statistically significant improvements in prospective classification accuracy (sensitivity), due to its ability to address variability in LIBS sample ablation and plasma self-absorption behavior. Furthermore, our results reveal that SVM provides nearly 10% improvement in correct allocation rate and a concomitant reduction in misclassification rates of 75% (*cf.* PLS-DA) and 80% (*cf.* SIMCA)-when measurements from samples *not* included in the training set are incorporated in the test data – highlighting its robustness. While further studies on a wider matrix of sample types performed using different LIBS systems is needed to fully characterize the capability of SVM to provide superior predictions, we anticipate that the improved sensitivity and robustness observed here will facilitate application of the proposed LIBS-SVM toolbox for screening drugs and detecting counterfeit samples as well as in related areas of forensic and biological sample analysis.

1. Introduction

Laser-induced breakdown spectroscopy (LIBS) is an atomic emission spectroscopic (AES) technique that utilizes a laser to create a plasma source. As the plasma is created by focused optical radiation, the technique has many benefits compared to conventional AES techniques that use electrodes or coils to generate the vaporization source. Subsequent to the initial vaporization and ionization, the plasma plume undergoes relaxation and in the process emits radiation in the 200–1000 nm spectral region. The resultant optical emission spectrum can

[†]Address all correspondence to: G. Manoj Kumar; manojsp@uohyd.ernet.in.

*Authors have made equal contributions to this work

be used for identification and quantification of elements present in the ablated material. Importantly, this provides LIBS with the ability to probe samples *in situ* and remotely without necessitating any extraneous sample preparation [1, 2]. Exploiting these intrinsic characterization properties, LIBS has previously been applied in the fields of forensics, oceanography, planetary science, material science and biomedical sciences[3–7].

Nevertheless, one of the primary hurdles in the application of LIBS to real world situations is the development of appropriate analytical tools that can be used to quantitatively correlate the spectral responses with the sample properties under investigation. For example, one can formulate a quantitative calibration model between the measured spectra and the analyte concentrations in the training samples to predict the unknown concentration in test samples. Similarly, one can also employ these data interpretation and analysis tools to classify among various classes of samples. However, precise quantitative analysis of LIBS datasets using conventional univariate models (where the intensity of a specific peak is correlated with the elemental concentration) is very difficult due to the uncontrollable fluctuations of the experimental parameters and the physical and chemical matrix effects [8]. Such fluctuations are likely to distort the theoretical relationship between the intensity and concentration, presenting substantive challenges to the univariate approach. Further, such univariate modeling is extremely difficult to implement in multi-element samples, where inter-element interference due to overlap of the spectral signatures is unavoidable. Clearly, there is a need to develop and apply efficient multivariate analysis methods, which can use the collective information of a set of spectral lines acquired from the sample spectra.

To overcome this problem of analyzing large amounts of data to predict a specific property of interest, classical chemometric approaches, which have been extensively used in the fields of analytical chemistry and metabolomics for both descriptive and predictive problems, can be employed. Given the well-characterized ability of chemometric tools (such as principal component analysis (PCA)) to handle very large and highly complex datasets, their extension to LIBS spectral analysis is a natural one. However, it is worth mentioning that while considerable contemporary attention has been focused on developing multivariate chemometric methods to interpret spectroscopic data [9, 10], their application to LIBS studies has received relatively less attention. In the area of systematic chemometric analyses of LIBS spectra, De Lucia Jr. *et al.* [11] and Rehse *et al.* [12] have reported promising results in their investigations of high energy materials and biological samples, respectively.

However, in a majority of these studies the prediction accuracy is used as the sole benchmark for establishment and optimization of the classification models. Unfortunately, conventional multivariate modeling often suffer from significant robustness issues, as pointed out in the relevant works of Somorjai and co-workers[13]. For example, many of the conventional classification methods used for LIBS analysis are constrained by the underlying linearity treatment. While such linear approximations may provide a reasonable first model for the LIBS data under controlled experimental conditions and for a limited range of samples, it does not address several of the real-world issues such as the saturation effect of the signal due to self-absorption in the plasma, fluctuations in the sample matrix and inconsistent ablation behavior. Furthermore, robust classification techniques should be insensitive to outliers, and should retain high generalization power by classifying unknown patterns correctly and reliably. For example, if the test set includes a sample that does not belong to any class of the training set, the sample should be correctly unallocated - as opposed to being misclassified into one of the existing classes.

In this article, we propose the application of a relatively new non-linear classification method, called least-squares support vector machines (LS-SVM) [14] (a variant of the 'original' support vector machines framework), to address the intrinsic curved effects in the

acquired LIBS data as well as to provide added robustness to outliers [15]. Using pharmaceutical sample analysis as a specific example, we demonstrate the effectiveness of this non-linear classification framework, with the understanding that this approach can be applied to similar LIBS applications with little or no alterations. Our choice of LIBS-based pharmaceutical sample analysis is motivated by the fact that the intricate chemical composition as well as the turbid media of the typical pharmaceutical samples substantially hinders the employment of other comparable optical and spectroscopic techniques. For example, most capsules and tablets consist of a complex mixture of organic active substances as active pharmaceutical ingredients (API) and excipients, in addition to possible impurities. Recently, we have initiated a research program exploring the possibility of LIBS application in the pharmaceutical industry, especially for quality assurance purposes [16]. This article provides a significant extension to our previous work by systematically studying the suitability of various classification approaches (including LS-SVM) for LIBS measurements on over-the-counter pharmaceutical drugs. Other contemporary studies that have preliminarily investigated the application of SVM for LIBS measurements include classification of pigments and inks [17] and quality control of welding process [18].

In this article, LIBS data acquired from over-the-counter pharmaceutical sample sets were used to quantify the performance of the linear and non-linear multivariate classification techniques. First, C4.5 algorithm was employed to visualize the decision tree for different classes based on their principal components. Subsequently, linear classification methods, namely soft independent modeling of class analogy (SIMCA) and partial least squares discriminant analysis (PLS-DA), and non-linear LS-SVM were used for prospective classification in order to assess the sensitivity and robustness of the respective models. Our analysis of the LIBS data of pharmaceutical samples reveals that LS-SVM substantially outperforms SIMCA and PLS-DA. In particular, the improvement in robustness provided by LS-SVM when an unknown class of samples is introduced in the test data is notable. Additionally, such methods may be misled by spurious correlations, especially those stemming from non-linear effects in the spectra-concentration relationship. We suspect that such curved effects may arise from small changes in the experimental conditions (*e.g.* laser power fluctuations) and variability in coating and matrix characteristics (even within the same class of drugs). Our results have substantive implications for LIBS-based process monitoring and quality control applications in the pharmaceutical industry. For example, the proposed LIBS-LS-SVM toolbox may provide an ideal platform for detection of counterfeit drugs (and identification of their sourcing) - a problem of considerable importance in the prevailing public health scenario for food and drug agencies worldwide. Furthermore, the proposed approach is inclusive and general enough to be extended to similar applications in process control and on-site reaction monitoring. Our future work will focus on generalizing the results of our current studies on different sample types acquired with different LIBS systems. In addition, we anticipate that combining with appropriate feature selection methods (given the sparsity of LIBS spectra) will enhance the robustness of the classification technique, heightening its efficacy for the industrial domain.

2. Materials and methods

Experimental studies are undertaken to accomplish the following objectives: (1) compare the classification performance of multivariate models developed using SIMCA, PLS-DA and SVM in discriminating over-the-counter pharmaceutical samples and (2) investigate the robustness of these models in classifying the samples. To accomplish objective (1), multivariate analysis is undertaken utilizing SIMCA (in which the variance of each class is separately modeled in a linear fashion using PCA), PLS-DA (where the linear PLS model fits all classes simultaneously) and SVM (where non-linear training and subsequent classification is performed). The rate of correct classification (and corresponding

misclassification and unclassification) is computed when all classes of pharmaceutical samples are included in the training data ("sensitivity test"). To achieve objective (2), the above-mentioned methods are evaluated to determine the rate of correct allocation (and corresponding unallocation and misclassification) when each class is alternately removed from the training set but is included in the test data ("robustness test"). In the context of pharmaceutical studies, this presents a more realistic scenario (than sensitivity analysis), which takes into account the fact that samples of an unknown class are likely to be encountered.

2.1 Experimental

For LIBS studies, laser pulses with energy of 25 mJ from a second harmonic of Nd:YAG laser are used at 532 nm with 7 ns pulse width and 10 Hz repetition rate. A lens system is used to collect and collimate the signal and is coupled to the Michelle ANDOR ME 5000 spectrograph equipped with an iSTAR DH734 ICCD. All spectra are recorded with an integration time of 1 μ s and a delay of 0.5 μ s. The resolving power of the LIBS spectrometer is 5000. A schematic of our LIBS system is shown in Figure S1 (Supporting Information).

For our present investigations, 85 sets of LIBS data acquired from 6 over-the-counter pharmaceutical tablets are used. Specifically, Brufen with coating (15 LIBS datasets), Brufen without coating (15 LIBS datasets), Glucosamine with coating (10 LIBS datasets), Glucosamine without coating (10 LIBS datasets), Paracetamol (20 LIBS datasets) and Vitamin C (15 LIBS datasets) were selected as representative over-the-counter drugs for further analysis. For Brufen and Glucosamine samples, spectra are obtained before and after removal of the coating using a standard protocol [16]. (Potential variations introduced into our acquired dataset by the preparation techniques are discussed in the Results and Discussion section.)

Twenty spectra from each drug type are acquired after averaging over ten consecutive pulses. It is worth noting that the acquired spectra are screened for further analysis by subjecting them to a minimum signal-to-noise ratio threshold and performing spectral outlier detection using student's *t*-test and the Mahalanobis distance function [20] resulting in the aforementioned number of datasets for each drug type. In order to prevent any spurious effects on our classification models, chemometric analysis is performed without the application of any additional pre-processing methods. Figure 1 shows a representative spectrum of each of the six drug classes measured on our LIBS system.

2.2 Data Analysis

A detailed description of the classification methods used in this article is provided in Supporting Information (Sec.-S1) (the interested reader is directed to the excellent references in the chemometrics literature for a more comprehensive understanding [21–36]). In this section, we emphasize the specific details of the application of the various classification schemes.

For the C4.5 decision tree analysis, the first six scores (corresponding to PCA decomposition of the entire dataset) are used as inputs. (While one may use specific spectral features based on characteristic wavelengths, the PC scores are typically more representative of the dataset in a complex multi-element dataset.) Here, J48, an open source implementation of C4.5 algorithm in the Wekadata-mining tool [37], is used for visualizing the importance of the specific PCs and their correlation to the actual elemental features. The entire dataset is used to generate the decision tree and a leave-one-out cross-validation routine is employed to compute its overall accuracy estimates.

Subsequently, the supervised classification techniques SIMCA, PLS-DA and LS-SVM are used to retrieve the class labels from the spectra. As stated earlier, two separate tests, one for sensitivity and one for robustness, are performed. MATLAB 7.12 is used as standard software for the ensuing classification analysis.

Sensitivity Test—The sensitivity investigations are performed for the three techniques by randomly splitting the 85 sample datasets into 55 for training and 30 for test, respectively. Specifically, the 30 test samples comprise of 5 randomly selected samples for each of the six drug classes. This size of the test set represents a compromise between opting for a test set size large enough to minimize the standard error in computing the misclassification rate, while maintaining the training set size large enough to construct an accurate classifier. Additionally, in order to obtain a representative estimate of the rates of correct classification, misclassification and unclassification (defined below), 100 independent iterations are performed by re-splitting the entire data into training and test sets. In the following, we describe the procedure for application of each classifier for one iteration where the dataset has already been randomly split into training and test sets.

For SIMCA model development, independent PC models are first created corresponding to each of the drug classes (i.e. six drug classes for sensitivity analysis). An important issue in SIMCA model development is to determine the number of the principal components to be retained in the subsequent analysis for each class. Here, we employed a standard leave-one-out cross-validation procedure to assess the number of PCs, which was observed to be in the range of 3–5 for each of the drug classes. Subsequently, for predicting the class of a spectrum, we used equally weighted scaled orthogonal and score distances. It is worth mentioning that an unclassification criterion is defined to prevent the misclassification of potential samples that are distant from the center of any of the PCA models. Specifically, we assumed that the distances of the training dataset to the center of the corresponding PCA model follows a normal distribution. This normal distribution is then used to calculate the probability of class membership of any test sample, given the distance of the corresponding spectrum to the center of the different classes. In the event that the membership probability for every class is observed to be less than 0.3% (i.e. the point lies beyond mean ± 3 times standard deviation), it is categorized as an "unclassified" sample. This method assists in decreasing the number of misclassifications. The decision threshold used here is arbitrary; any threshold could be used to analyze which SIMCA model is most effective for the specific application. A modified and extended version of the LIBRA toolbox originally developed by Verboven *et al* [38] is used for this analysis.

Similarly, PLS-DA was applied to the same dataset. The total class memberships are stored in the form of a 85×6 matrix of dummy variables, where the columns are equal to the number of classes and rows are equal to the number of samples. Each column in this matrix consists of '1's for the corresponding class; otherwise it consists of '0's. To test the model for the unknown spectrum, it yields a 1×6 vector containing predicted membership probability values. It is worth emphasizing that for PLS-DA, only a single "global" model is developed for classification purposes, in contrast to the independent PCA submodels created for the aforementioned SIMCA analysis. Similar to the SIMCA analysis, we used probability of class membership for proper assignment. For PLS-DA, we have used the same 0.3% decision threshold to unclassify samples.

Finally, the LS-SVM computations are performed using a LS-SVM MATLAB toolbox [39]. Here, a radial basis function (RBF) kernel with a Gaussian profile $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2)$ is used for non-linear classification, where σ^2 is the RBF kernel parameter (precisely, the width of the RBF) (see also Ref. [40]). The optimal model parameters γ (cost parameter) and σ^2 that give minimum error in cross-validation in the training set are

determined by performing a grid search over the range of 1 to 10000 (γ) and $1-10^7$ (σ^2), respectively. (In principle, a relatively large value of the cost parameter (i.e. closer to 1000 than 1) bears the risk of overfitting, but in our study such complications are largely reduced by using cross validation for optimization and by invoking an independent test set to assess the prospective results.) Again, the 0.3% decision threshold is used on the LS-SVM predicted metrics (i.e. the latent variables of the classifier) to unclassify the test samples.

Robustness Test—The robustness investigations follow a very similar path to the sensitivity tests outlined in the above paragraphs with one key difference. In this set of investigations, one drug class is removed from the calibration set one at a time and this procedure is alternately repeated for each class. Hence, if the spectra of the unknown test class is unclassified by the developed models, it should be considered to be robust (and a positive identification for our statistical analysis). To account for this change from the previous sensitivity analysis, we have altered the classification categories to the following: "correct allocation", "misclassification" and "unallocation". We define "correct allocation" as all correctly classified spectra (from unknown drugs) as well as all correctly unclassified spectra (from known drugs), whereas "unallocation" refers to incorrect unclassification of the spectra from the known drugs. Incorrect classification (from known or unknown drugs) falls under the general category of "misclassification".

It is worth emphasizing that the size of the test set remains same as before, i.e. 30 samples with 5 samples per drug class (including the class omitted from the training set). For each removed drug class, 100 iterations are again performed to gain a representative value of the classification errors.

3. Results and discussion

Figure 2(A) shows the first six PCs corresponding to the entire 85-sample dataset (spatially offset for clarity). These six PCs explain 98.47% of the net variance and the subsequent PCs are observed to be noisier (not shown here). These components, while abstract in form as they are obtained from singular value decomposition, are valuable in assessing the key spectral markers. Comparing Fig. 1 to Fig. 2(A), we find that PC1 is primarily indicative of the Glucosamine coated sample spectrum. We suspect that PC1 may be largely based on the coating composition of the Glucosamine coated tablets. Similarly, PC2 and PC3 have important contributions from API elements, such as carbon, hydrogen, nitrogen and oxygen (as obtained from the NIST database [41]). In addition, PC 2 has substantive manganese contribution.

The decision tree resulting from the scores of these six PCs is given in Fig. 2(B). Although we have used first six PC scores in the analysis, we can see from the figure that algorithm only used scores corresponding to PC 1, 3, 4 and 6 to classify the samples. At each node of the tree, C4.5 chooses one PC score of the LIBS data that most effectively splits the set of pharmaceutical samples into subsets enriched in one drug class or the other. In our decision tree, PC score 3 lies on the top level, which then splits into two nodes representing score 1 and score 4. Score 1 separates Brufen and Paracetamol based on a decision threshold (0.040383). Score 4 was split into two nodes Score 1 and Score 6 which leads to the separation of Vitamin C and Glucosamine-coated based on the value of Score 1 and the separation of Brufen-coated and Glucosamine based on the value of Score 6. While the decision tree provides substantially more detailed information than PCA alone, one can find commonalities in the discrimination analysis from the more visual process (PCA) and the more mathematical one (C4.5). For example, it is obvious from the PC plots that PC1 is instrumental in the correct classification of glucosamine coated samples, whereas its impact in the separation of Brufen (uncoated) and Paracetamol samples is not visually evident

(which emphasizes the need for decision tree analysis). Further, from a leave-one-out cross-validation analysis, we observe that decision tree analysis gives a total of 78 correctly classified instances (i.e. approx. 91% classification accuracy). In particular, glucosamine coated samples were perfectly classified while all other classes had at least one misclassification. The coating removed Brufen and Glucosamine samples showed more than one classification errors each, which potentially be attributed to experimental errors in removal of the coating.

Sensitivity Test

Table 1(A) gives the results of sensitivity analysis for SIMCA-based classification of the six drug classes (i.e. the results of prospective prediction over 100 iterations when all the six classes are included in the training set). Here, the model is judged to be sensitive if the correct classification rate is high and the unclassification rate is low. First, it is noted that we obtain a high rate of correct classification for all classes with an average rate of correct classification of *ca.* 94%. In particular, for Paracetamol and Vitamin C no misclassifications are observed and the correct classification accuracy is fairly high ($>96\%$). However, it appears that Brufen, Brufen coated and Glucosamine are more difficult to analyze via SIMCA with substantial number of misclassifications observed in each case. Finally, from the unclassification perspective, the Glucosamine coated and Paracetamol samples perform relatively poorly, which implies the existence of spectral outliers in these cases.

The corresponding PLS-DA results for the sensitivity test are provided in Table 1(B). It is noted that five loading vectors were used in developing the global PLS-DA model (for perspective, as mentioned above, 3–5 PCs were used in constructing the PCA submodels for SIMCA). Here, we observe that the average rate of correct classification is even higher (*ca.* 95.6%) with the Paracetamol and Vitamin C samples being classified with very high accuracy ($\geq 98\%$). The improvement in average rate of correct classification from SIMCA to PLS-DA is extremely statistically significant ($p < 10^{-4}$). However, the average rate of unclassification for PLS-DA (*ca.* 4%) is significantly higher than in the case of SIMCA (*ca.* 2%) ($p < 10^{-4}$). We note that Glucosamine coated still has the highest rate of unclassification, similar to SIMCA analysis, reinforcing our previous suspicion of the existence of spectral outlier for these samples. This poses a difficult question on the issue of comparative sensitivity of SIMCA and PLS-DA, as one has higher correct classification rate and the other has lower unclassification rate. Apparently based on the specific application, one could choose either chemometric method depending on the relative importance of correct classification and unclassification accuracy.

Table 1(C) shows the sensitivity test results for LS-SVM classification analysis. Compared with SIMCA (Table 1(A)) and PLS-DA (Table 1(B)), we find that the correct rate of classification is higher and the rate of unclassification is also lower. The increase in the average rate of correct classification from SIMCA to LS-SVM as well as from PLS-DA to LS-SVM reaches statistical significance ($p < 10^{-4}$ in either case). Furthermore, the average rate of unclassification is significantly lower for LS-SVM as compared to PLS-DA ($p < 10^{-4}$) but the comparison with SIMCA does not attain statistical significance ($p > 0.05$). Looking deeper at the classification results of each drug, some important commonalities are observed between the classification analyses of the three chemometric methods. For example, the correct classification accuracy of Paracetamol and Vitamin C samples is high ($\geq 96\%$), comparable to what we detected in Table 1(A) and Table 1(B) for SIMCA and PLS-DA, respectively. Similarly, the rate of unclassification for Glucosamine coated samples is the highest among the six drug classes (5%), as also observed for SIMCA and PLS-DA.

For the sake of completeness, we also performed prospective prediction using decision tree analysis, where we observed an average correct classification rate of 90.6%. Clearly, the

consider only a single global model and illustrate the detailed results for each drug class, whereas in the robustness tables each row represents the performance of a different model, where the particular drug class has been removed from the training set. In essence, each row of a robustness table is a condensed form of a sensitivity table, *i.e.* it gives the allocation/classification errors averaged over all the classes for the particular model.

From Table-2(A), we note that our SIMCA models exhibit an average correct allocation percentage of *ca.* 84%. This is considerably poorer than the analogous correct classification performance observed in the sensitivity analysis (Table-1(A)) and, ominously, SIMCA also shows a fairly high rate of misclassification (*ca.* 15.5%). Further, we do not observe any strong dependence of the correct allocation (or the misclassification) rate on the type of drug omitted from the training set. Evidently, while SIMCA provided acceptable sensitivity performance (Table-1(A)), its robustness performance is not particularly desirable. In this context, PLS-DA shows significant improvement in lowering the average misclassification rate to *ca.* 12.7% (Table-2(B)) from the above SIMCA modeling ($p=0.006$). Its correct allocation rate also increases marginally over the corresponding SIMCA metric but the improvement is not statistically significant ($p>0.05$). Once again, the rate of correct allocation does not strongly depend on the omitted drug class.

Taken together, these results seem to indicate that PLS-DA is more robust than SIMCA, although even for PLS-DA the correct discrimination rates are still much lower than in sensitivity analysis. One may attribute this enhanced robustness to the fact that PLS-DA seeks to determine the maximum separation between the classes, which makes unallocation of unknown spectrum relatively easier. This is in contrast to class-specific PCA submodel-based SIMCA analysis, which explains the variance of each of the classes separately but does not necessarily address the class-to-class variability. As noted by De Lucia and co-workers, this may be present a problem where inter-class variability is of the same order as intra-class variability [28]. Previously, Sirven *et al.* [5] had also performed a comparative investigation of SIMCA and PLS-DA for LIBS-based rock identification. Our observations are in partial agreement with their results, where the point of commonality is the higher correct allocation accuracy of PLS-DA in each case. Nevertheless, their report showed that SIMCA had a lower average misclassification rate than PLS-DA (for an analogous robustness study), which is in contrast to the numbers obtained above. Based on this discrepancy in the two investigations, it is appropriate to say that the choice of the linear classification methodology (particularly with respect to robustness) depends heavily on the specific application and type of dataset.

Finally, Table-2(C) exhibits the performance of LS-SVM when each drug is alternately removed from the training set. Clearly, LS-SVM shows significantly better performance than either SIMCA or PLS-DA, both in terms of a high correct allocation rate and a low misclassification rate. The improvement for correct allocation rate of LS-SVM over PLS-DA and SIMCA is approx. 10%, whereas the reduction in misclassification rate is 75% (compared to PLS-DA) and 80% (compared to SIMCA). In fact, the LS-SVM performance in this case approaches that of the three chemometric methods (SIMCA, PLS-DA and LS-SVM) in sensitivity analysis. Specifically, we note that the LS-SVM misclassification rate (*ca.* 3.1%) is lower than the misclassification rate of SIMCA even when all the drug classes are included in the training data (*ca.* 3.8%). This underlines the robustness of LS-SVM in dealing with samples of unknown classes. Previous investigators have also highlighted its robustness [33] and characterized its capability in dealing with sample outliers [24, 35]. Furthermore, the robustness is likely to make LS-SVM more efficient in handling real-world, noisy and variable spectra. Finally, other investigators have also noted the advantage of reduced computation time for SVM in comparison with SIMCA, thus making it valuable as a screening tool for large LIBS datasets [17, 44].

In the context of process monitoring of pharmaceutical tablet preparation and related quality assurance, one can come across the situation presented above when several drug types are produced and tested in one facility as well as when drugs are obtained from different manufacturing facilities even though they are of nearly identical composition. An even more important and critical case is that of detection of counterfeit samples, which is evidently difficult (if not impossible) to incorporate in the training dataset. As the sophistication of counterfeit products improves, it becomes more challenging to ascertain authenticity by testing the suspect product for the presence/absence or concentration of the API components - which necessitates the application of techniques that provide comprehensive elemental and molecular information. Our results here show coupling LIBS to a suitable non-linear classification algorithm such as LS-SVM may provide an important tool for obtaining the requisite information.

In summary, despite the limited number of drug classes used in this research as compared to what may be typically encountered in an industrial facility, the LS-SVM approach exhibited very good compliance and high prediction accuracy in classification of the blind cases. This work illustrates the power of LS-SVM as an enhanced spectral treatment tool for the determination of class labels and establishes the feasibility of its application as a potential screening tool in the pharmaceutical industry. Nevertheless, a more comprehensive characterization of SVM application for LIBS measurements is necessary in order to validate the generalizability of our current results. To this end, further investigations in a variety of sample types (*e.g.* pharmaceutical tablets from different vendors) and across multiple LIBS systems are currently underway in our laboratory (ACRHEM, University of Hyderabad).

4. Concluding Remarks

In this article, we have demonstrated that LIBS in combination with LS-SVM can provide a sensitive and robust tool for the analysis of pharmaceutical samples. Specifically, we have established the need for using support vector machines (or a similar classification method) that appears to be less susceptible to outliers in comparison to conventional linear least squares techniques. Our investigations systematically consider the real-world situation where a drug sample, which has not been included in the learning process, is tested by the proposed analytical method. Such considerations can be readily extended to identification and screening of counterfeit drugs, where an analytical method that is robust to spurious misclassifications is of utmost importance. Additionally, the possibility of sample and system variability in LIBS measurements may cause deviations from the idealized linear relationship (between the spectra and the concentration of the analyte of interest) necessitating the use of a non-linear method to provide adequate sensitivity.

Future investigations are needed to comprehensively characterize specific sources of non-linearity in LIBS-based monitoring of pharmaceutical samples and the robustness with which such effects can be modeled by our proposed scheme. In addition, given the high dimensionality of the LIBS dataset, it is worth exploring feature selection methods that can eliminate the excess (uninformative) regions of the spectrum, which are unrelated to the elemental composition, thereby establishing a classification model based solely on the physical features. In fact, utilizing full spectrum methods may worsen the chemometric model, as the redundancy is likely to incorporate spurious correlations and noise [45]. Importantly, such investigations can lead to the reduction of necessary hardware required for LIBS measurements by focusing on a select subset of wavelengths alone. One can also seek to improve the obtained classification accuracy by using active learning techniques such as boosting [46] and bagging [47]. The final outcome of LIBS application to the pharmaceutical sample studies will be a learning diagnostic tool that is based on a hybrid of

several chemometric techniques, which exploit the best features of each of the individual strategies. Taken together with the studies presented in this article, we believe that these future developments will establish LIBS as a powerful diagnostics component in the pharmaceutical industry and, more generally, in the field of product quality assurance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

GMK acknowledges the financial support of DRDO, India. IB and NCD acknowledge the support of NIH National Center for Research Resources (Grant No. P41-RR02594). The authors acknowledge Dr. Ramachandra Rao Dasari for valuable discussions.

References

1. Singh, JP.; Thakur, SN. Laser-induced Breakdown Spectroscopy. UK: Elsevier Oxford; 2007.
2. Miziolek, AW.; Palleschi, V.; Schechter, I. Laser-induced Breakdown Spectroscopy (LIBS): Fundamentals and Applications. Cambridge UK: Cambridge University Press; 2006.
3. Jantzi S, Almirall J. Anal. Bioanal. Chem. 2011; 400:3341–3351. [PubMed: 21461623]
4. Michel APM, Chave AD. Appl. Opt. 2008; 47:G122–G130. [PubMed: 19122693]
5. Sirven J-B, Salle B, Mauchien P, Lacour J-L, Maurice S, Manhes GJ. Anal. At. Spectrom. 2007; 22:1471–1480.
6. Harmon RS, Shughrue KM, Remus JJ, Wise MA, East LJ, Hark RR. Anal. Bioanal. Chem. 2011; 400:3377–3382. [PubMed: 21537914]
7. Marquardt BJ, Goode SR, Michael Angel S. Anal. Chem. 1996; 68:977–981.
8. Clegg SM, Sklute E, Dyar MD, Barefield JE, Wiens RC. Spectrochim. Acta, Part B. 2009; 64:79–88.
9. Martens, H.; Naes, T. Multivariate calibration. Chichester: John Wiley & Sons; 1989.
10. Lavine B, Workman B. J. Anal. Chem. 2008; 80:4519–4531.
11. De Lucia FC, Gottfried JL. Propellants, Explos., Pyrotech. 2010; 35:268–277.
12. Rehse SJ, Mohaidat QI, Palchaudhuri S. Appl. Opt. 2010; 49:C27–C35.
13. Somorjai RL, Dolenko B, Baumgartner R. Bioinformatics. 2003; 19:1484–1491. [PubMed: 12912828]
14. Suykens, JAK.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. Least Squares Support Vector Machines. Singapore: World Scientific; 2002.
15. Barman I, Dingari NC, Rajaram N, Tunnell JW, Dasari RR, Feld MS. Biomed. Opt. Exp. 2011; 2:592–599.
16. Myakalwar AK, Sreedhar S, Barman I, Dingari NC, Venugopal Rao S, Prem Kiran P, Tewari SP, Manoj Kumar G. Talanta. 2011; 87:53–59. [PubMed: 22099648]
17. Hoehse M, Paul A, Gornushkin I, Panne U. Anal. Bioanal. Chem. 2011
18. Anabitarte F, Mirapeix J, CondePortilla OM, Lopez-Higuera JM, Cobo A. IEEE Sensors Journal. 2012:12–1.
19. Missaghi S, Fassihi R. AAPS PharmSciTech. 2004; 5:32–39.
20. Dingari NC, Barman I, Kang JW, Kong C, Dasari RR, Feld MS. J. Biomed. Opt. 2011; 16 087009.
21. Quinlan, JR. C4 5: Programs for Machine Learning. San Mateo CA: Morgan Kaufmann; 1993.
22. Berman ESF, Wu L, Fortson SL, Kulp KS, Nelson DO, Wu KJ. Surf. Interface Anal. 2009; 41:97–104.
23. Brereton, RG. Applied Chemometrics for Scientists. England: John Wiley & Sons Ltd; 2007.
24. Brereton RG, Lloyd GR. Analyst. 2010; 135:230–267. [PubMed: 20098757]
25. Wold S, Esbensen K, Geladi P. Chemom. Intell. Lab. Syst. 1987; 2:37–52.

26. Blomquist G, Johansson E, Söderström B, Wold S. J. Chromatogr. A. 1979; 173:19–32.
27. Berrueta LA, Alonso-Salces RM, Héberger K. J. Chromatogr. A. 2007; 1158:196–214. [PubMed: 17540392]
28. De Lucia JFC, Gottfried JL, Munson CA, Miziolek AW. Appl. Opt. 2008; 47:G112–G121. [PubMed: 19122692]
29. De Lucia FC Jr, Gottfried JL. Spectrochim. Acta, Part B. 2011; 66:122–128.
30. Amador-Hernandez J, Garcia-Ayuso LE, Fernandez-Romero JM, Luque de Castro MD. J. Anal. At. Spectrom. 2000; 15:587–593.
31. Beale, R.; Jackson, T. Neural Computing: an introduction. Adam Hilger Bristol; 1991.
32. Vapnik, V. The Nature of Statistical Learning Theory. New York: Springer-Verlag; 1995.
33. Scholkopf, B.; Smola, AJ. Learning with Kernels. Cambridge MA: MIT Press; 2002.
34. Suykens JAK, Vandewalle J. Neural Processing Letters. 1999; 9:293–300.
35. Shao X, Bian X, Liu J, Zhang M, Cai W. Anal. Methods. 2010; 2:1662–1666.
36. Balabin RM, Lomakina EI. Analyst. 2011; 136:1703–1712. [PubMed: 21350755]
37. Witten, IH.; Witten, E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed.. San Francisco CA: Morgan Kaufmann; 2005.
38. Verboven S, Hubert M. Chemom. Intell. Lab. Syst. 2005; 75:127–136.
39. Suykens, JKT.; Van Gestel; De Brabanter, J.; De Moor, B.; Vandewalle, J. Least Squares Support Vector Machines. Singapore: World Scientific; 2002.
40. Barman I, Kong C-R, Dingari NC, Dasari RR, Feld MS. Anal. Chem. 2010; 82:9719–9726. [PubMed: 21050004]
41. NIST database of atomic spectral data. <http://physicsnistgov/PhysRefData/ASD>
42. Sirven JB, Bousquet B, Canioni L, Sarger L. Anal. Chem. 2006; 78:1462–1469. [PubMed: 16503595]
43. Sirven JB, Bousquet B, Canioni L, Sarger L, Tellier S, Potin-Gautier M, Hecho IL. Anal. Bioanal. Chem. 2006; 385:256–262. [PubMed: 16538460]
44. Balabin MB, Safieva RZ, Lomakina EI. Microchem J. 2011; 98:121–128.
45. Spiegelman CH, McShane MJ, Goetz MJ, Motamedi M, Yue QL, Coté GL. Anal. Chem. 1998; 70:35–44. [PubMed: 21644597]
46. Freund, Y.; Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting Computational Learning Theory. Vitányi, P., editor. Vol. Vol. 904. Heidelberg: Springer Berlin; 1995. p. 23-37.
47. Breiman L. Machine Learning. 1996; 24:123–140.

Figure 1. Representative LIBS spectra of the over-the-counter pharmaceutical samples (from left to right) Brufen coated, Brufen, Glucosamine coated, Glucosamine, Vitamin C, and Paracetamol. The spectra shown are normalized.

2(A)

2(B)

Figure 2.

(A): The first six principal components corresponding to the 85 pharmaceutical sample dataset (PCs are vertically offset for clarity).

(B): Decision tree generated using C4.5 algorithm on the PC scores obtained from the 85 sample dataset.

Table 1

Sensitivity Test Results

(A) SIMCA predictions for each sample when all samples are included in the training set

Average rate of ...	Correct classification	Misclassification	Unclassification
Brufen	0.908	0.076	0.016
Brufen coated	0.934	0.062	0.004
Glucosamine	0.904	0.088	0.008
Glucosamine coated	0.940	0.002	0.058
Paracetamol	0.968	0	0.032
Vitamin C	0.988	0	0.012
Average	0.9403	0.038	0.0217

(B) PLS-DA predictions for each sample when all samples are included in the training set

Average rate of ...	Correct classification	Misclassification	Unclassification
Brufen	0.980	0	0.020
Brufen coated	0.916	0.03	0.054
Glucosamine	0.978	0	0.022
Glucosamine coated	0.888	0.004	0.108
Paracetamol	0.980	0.008	0.012
Vitamin C	0.998	0	0.002
Average	0.9567	0.007	0.0363

(C) LS-SVM sensitivity test: predictions for each sample when all samples are included in the training set

Average rate of ...	Correct classification	Misclassification	Unclassification
Brufen	1.0000	0.0000	0.0000
Brufen coated	0.9200	0.0600	0.0200
Glucosamine	0.9800	0.0000	0.0200
Glucosamine coated	0.9000	0.0500	0.0500
Paracetamol	0.9600	0.0100	0.0300
Vitamin C	1.0000	0.0000	0.0000
Average	0.9600	0.0200	0.0200

Table 2**Robustness Test Results****(A) SIMCA predictions when samples of a specific drug class are alternately omitted from the training set.**

Omitted drug	Correct allocation	Misclassification	Unallocation
Brufen	0.8070	0.1860	0.0070
Brufen coated	0.8157	0.1733	0.0110
Glucosamine	0.8323	0.1520	0.0157
Glucosamine coated	0.8673	0.1267	0.0060
Paracetamol	0.8747	0.1133	0.0120
Vitamin C	0.8147	0.1777	0.0077
Average	0.8353	0.1548	0.0099

(B) PLS-DA predictions when samples of a specific drug class are alternately omitted from the training set.

Omitted drug	Correct allocation	Misclassification	Unallocation
Brufen	0.8640	0.1017	0.0343
Brufen coated	0.9217	0.0407	0.0377
Glucosamine	0.9193	0.0557	0.0250
Glucosamine coated	0.8010	0.1807	0.0183
Paracetamol	0.7657	0.1797	0.0547
Vitamin C	0.7653	0.2040	0.0307
Average	0.8395	0.1271	0.0335

(C) LS-SVM predictions when samples of a specific drug class are alternately omitted from the training set.

Omitted drug	Correct allocation	Misclassification	Unallocation
Brufen	0.9300	0.0267	0.0400
Brufen coated	0.9550	0.0117	0.0333
Glucosamine	0.9117	0.0467	0.0417
Glucosamine coated	0.9400	0.0033	0.0567
Paracetamol	0.9017	0.0483	0.0500
Vitamin C	0.9100	0.0500	0.0400
Average	0.9247	0.0311	0.0436