2012 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 23–26, 2012, SANTANDER, SPAIN

HAUSSDORFF AND HELLINGER FOR COLORIMETRIC SENSOR ARRAY CLASSIFICATION

Tommy S. Alstrøm[a], *Bjørn S. Jensen*[a], *Mikkel N. Schmidt*[a], *Natalie V. Kostesha*[b], *Jan Larsen*[a]

[a]Dept. of Informatics and Mathematical Modeling, Technical University of Denmark
Richard Petersens Plads 321, 2800 Kgs. Lyngby, Denmark
{tsal,bjje,mns,jl}@imm.dtu.dk

[b]Dept. of Micro- and Nanotechnology, Technical University of Denmark
ørsteds Plads 345 East, DK-2800, Kgs. Lyngby, Denmark
natalie.kostesha@nanotech.dtu.dk

ABSTRACT

Development of sensors and systems for detection of chemical compounds is an important challenge with applications in areas such as anti-terrorism, demining, and environmental monitoring. A newly developed colorimetric sensor array is able to detect explosives and volatile organic compounds; however, each sensor reading consists of hundreds of pixel values, and methods for combining these readings from multiple sensors must be developed to make a classification system. In this work we examine two distance based classification methods, K-Nearest Neighbor (KNN) and Gaussian process (GP) classification, which both rely on a suitable distance metric. We evaluate a range of different distance measures and propose a method for sensor fusion in the GP classifier. Our results indicate that the best choice of distance measure depends on the sensor and the chemical of interest.

Index Terms— Hausdorff distance, Hellinger distance, chemo–selective compounds, feature extraction, K–nearest neighbor classification, Gaussian Process Classification

1. INTRODUCTION

The development of rapid, reliable, and portable solutions for detection of chemical compounds is an important challenge with many possible applications including screening luggage and packages for explosives, detecting land mines, and monitoring the environment for hazardous compounds. In recent years a number of methods have been developed based on different technologies such as gas chromatography, Raman spectrometry, mass spectrometry, ion mobility spectrometry and colorimetric sensors.

Colorimetric sensors can be used to detect a wide range of organic compounds in gas as well liquid phase [1–6]. We have developed a novel colorimetric sensor array that is useful for identifying a wide range of explosives [7, 8] as well as other compounds such as amines, cyanides, alcohols, arenes, ketones, aldehydes and acids. A colorimetric sensor contains a chemo-selective compound: A dye which changes color when exposed to a target analyte. The sensor is read by capturing a digital image of the dye before and after exposure. Typically, the mean color change is used for detection; however, in earlier work we have shown that accuracy can be improved by considering the complete distribution of color change for different image pixels [9]. Furthermore, the sensor array comprises several colorimetric sensors, and detection could likely be improved further by combining information from multiple sensors.

In this paper, we present a new method for analyzing of the output of a colorimetric sensor array using the complete distribution of color changes. To classify a given analyte, we compare a K-Nearest Neighbor (KNN) approach with a Gaussian process (GP) classifier. Both of these approaches rely on a measure of similarity between sensor readings—the KNN through a suitable distance measure and the GP through a covariance function. We motivate and compare several measures of similarity, demonstrating that a proper choice of similarty measure can significantly improve classification accuracy using a single dye. Finally, we propose a method for fusing information from several sensors leading to superior performance compared to using single dyes.

We acknowledge the support from the Danish Agency for Science and Technology's, Program Commission on Nanoscience Biotechnology and IT (NABIIT). Case number: 2106-07-0031 - Miniaturized sensors for explosives detection in air. Further we acknowledge assistant professor Raviv Raich, Oregon State University, for lending his code for distance calculation based on Hellinger distances using Parzen windows.

2. METHODS

In this section we describe the colorimetric sensor data and the proposed classification approach.

2.1. Notation

$N = 253$	Number of observations
$L = 3$	Number of channels (RGB)
$M = 31$	Number of dyes
$N_{\text{pix}} = [49, 1009]$	Pixels per dye
$\boldsymbol{x}_{l,n}$	$1 \times N_{\text{pix}}$ vector of pixel values
$\boldsymbol{X}_n = [\boldsymbol{x}_{1,n}^\top \boldsymbol{x}_{2,n}^\top \boldsymbol{x}_{3,n}^\top]^\top$	$3 \times N_{\text{pix}}$ matrix with pixel values
$\mathcal{X} = \{x_n \| n = 1 : N\}$	a set of inputs
$\mathcal{Y} = \{y_n ; x_n \| n = 1 : N\}$	a set of labels for the given inputs
Q	the set of distance measures

2.2. Colorimetric sensor data

The colorimetric sensor array consists of a number of chemo–selective compounds immobilized onto silica gel resulting in circular spots (Fig. 1A). Each individual spot was approximately 3 mm in diameter with the total size of the sensor array of approximately 2.5 cm × 4.0 cm.

The dataset used in this paper has been discussed in detail in earlier work [10] but is summarized here for completeness. The sensor array has been exposed to analytes belonging to various chemical families – 9 families in total, making it a multi-class dataset. The chemical families are: acids (45), alcohols (27), amines (42), arenes (14), environment (28), explosives (56), inorganic explosives (14), ketones (13) and thiols (14). The number in the parenthesis denotes the number of examples measured for the class in question.

2.2.1. *Data acquisition*

The sensors were scanned using an ordinary flatbed scanner immediately after immobilization of dyes and then again after exposure of target analytes. The images are then aligned pixel, the dye locations are identified and finally pixel extraction is performed [10, 11]. The pixels are extracted by fitting a circular disc to the entire dye. Due to the nature of the sensor often a distinct ring near the perimeter of the dyes appears (the coffee stain effect) and this area of the dyes is deemed unreliable (Fig. 1B). In order to accommodate for this effect a smaller area of a dye is used for feature extraction, corresponding to 2/3 of radius of the fitted circular disc. Based on the pixel values features can be calculated, e.g. using the mean value, or distances between measurements.

2.3. Classification Methods

2.3.1. *K-Nearest Neighbor*

The KNN is a simple yet effective classification technique [12] which works as follows. When testing a data point belonging to an unknown class, the distances to all points with known class labels are calculated. The classes of the closest K points are then identified and the unknown point is classified using majority voting. In order to carry out both model selection and estimation of the generalization error, double-cross validation using *leave-one-out* is performed.

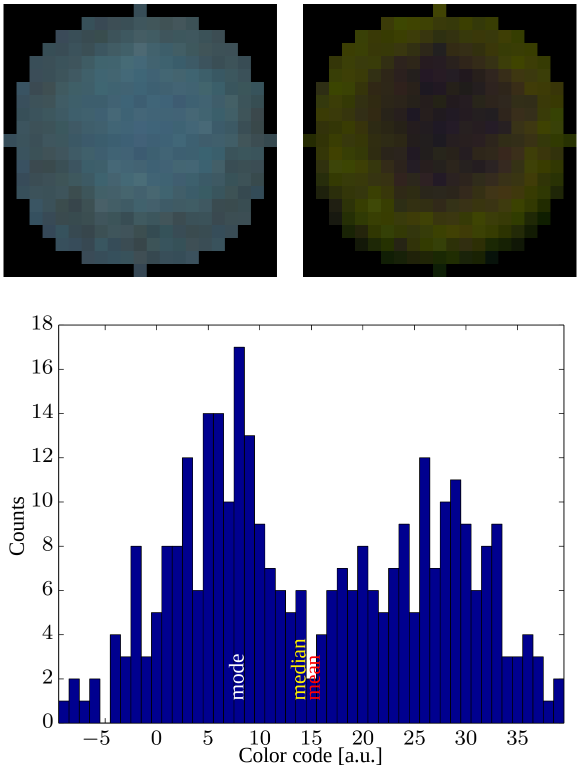

Fig. 1. An example of a specific dye of colorimetric sensor array exposed to the explosive analyte RDX. A: the sensor before exposure. B: the difference image. C: histogram of difference of the blue channel.

2.3.2. *Gaussian Process Classification*

From the instance based KNN classifier we now turn to a different but equally powerful classification framework based on a non-parametric Bayesian approach. The core of any probabilistic classification methods is a likelihood function modeling the likelihood of observing a specific outcome $y \in \{-1, +1\}$. Here we consider the cumulative Gaussian (known as a probit model) defined as

$$p\left(y_n | f\left(\boldsymbol{X}_n\right)\right) = \int\limits_{-\infty}^{y_n \cdot f(\boldsymbol{X}_n)} \mathcal{N}\left(t|0,1\right) dt = \Phi\left(y_n \cdot f\left(\boldsymbol{X}_n\right)\right)$$

parameterized by a given a functional value, $f(\boldsymbol{X}_n)$. Hence, the model or free parameter is the functional value, $f(\boldsymbol{X}_n)$, and by taking a Bayesian approach, we can directly consider the posterior over the function defined by the finite set of random variables, $\mathbf{f} = [f(\boldsymbol{X}_1), f(\boldsymbol{X}_2), ..., f(\boldsymbol{X}_N)]^\top$, i.e.,

$$p\left(\mathbf{f}|\mathcal{Y},\mathcal{X}\right) = \frac{p\left(\mathcal{Y}|\mathbf{f}\right)p\left(\mathbf{f}|\mathcal{X}\right)}{\int p\left(\mathcal{Y}|\mathbf{f}\right)p\left(\mathbf{f}|\mathcal{X}\right)d\mathbf{f}} = \frac{p\left(\mathcal{Y}|\mathbf{f}\right)p\left(\mathbf{f}|\mathcal{X}\right)}{p\left(\mathcal{Y}|\mathcal{X}\right)}$$

The natural prior for, $f(\cdot)$, is a Gaussian Process (GP) and we denote a function drawn from a GP as $f\left(\boldsymbol{X}\right) \sim \mathcal{GP}\left(\mathbf{0}, k(\cdot,\cdot)_{\boldsymbol{\theta}_c}\right)$ with a zero mean function, and $k(\cdot,\cdot)_{\boldsymbol{\theta}_c}$ referring to the covariance function with hyper-parameters $\boldsymbol{\theta}_c$, defining the covariance between the random variables $\boldsymbol{f}$. The GP can thus be considered a distribution over functions, i.e., $p\left(\mathbf{f}|\mathcal{X},\boldsymbol{\theta}_c\right)$.

Given the probit likelihood model, the posterior over $\boldsymbol{f}$ needs to be approximated and we resort to Expectation Propagation (EP) which provides a Gaussian approximation to the posterior [13]. The hyper parameters are approximated by point-estimates found by considering the marginal likelihood/evidence which can be optimized in regards to the hyper parameters by gradient methods. This possibility is a clear advantage over non-Bayesian methods.

The predictions for a new input $\boldsymbol{X}^*$ is obtained by first computing the predictive distribution, $p(f^*|\mathcal{Y},\mathcal{X},\boldsymbol{X}^*)$, which is Gaussian due to the EP approximation. The probability of a given outcome y^* is computed by $P\left(y|\mathcal{Y},\mathcal{X},\boldsymbol{X}^*\right) = \int p\left(y|f^*\right)p(f^*|\mathcal{Y},\mathcal{X},\boldsymbol{X}^*)df^*$.

2.4. Distances and Covariance Functions

The two classification methods outlined above both require some notion of similarity either in the form of a distance function/metric or in the form of a covariance function. In the following we present and motivate a number of distance measures that can be used for measuring similarites between dyes.

2.4.1. *Mean and "inner mean"*

The traditional approach when using colorimetric sensors is to calculate the mean value of the response by averaging the color change over all pixels.

Since we have observed that pixel values in the outer edge of the dyes are less reliable, we propose to estimate an optimal radius and compute the mean in the central region within this radius. Assuming that each dye should have a unique color change and such that any variation is due to noise, a reasonable assumption is to choose the radius that estimates the mean most accurately. This is done by choosing

$$\hat{r} = \arg\min_r \frac{1}{\sqrt{N_{\text{pix}}(r)}}\sigma(r),$$

where $N_{\text{pix}}(r)$ and $\sigma(r)$ denote the number of pixels and the standard deviation of pixel values within the central region with radius r. We denote this method "Inner Mean".

2.4.2. *Hellinger Distance*

As an alternative to computing the average color change, we can consider the distribution of color change over the dye. The Hellinger distance measures similarity between two probability measures $f_i(x)$ and $f_j(x)$ and is given by

$$d_{\text{He}}(f_i, f_j)^2 = \int \left(\sqrt{f_i(x)} - \sqrt{f_j(x)}\right)^2 dx$$

Using the Hellinger distance require us to choose how to represent the probability measures. We consider a nonparametric and a parametric approach: The first is a Parzen window Kernel density estimator

$$f_i(x) = \frac{1}{n_i}\sum_{k=1}^{n_i} K(x - x_{ik})$$

where $K(x) = 1/(2\pi\sigma^2)^{d/2}\exp(-||x||^2/2\sigma^2)$ and the kernel width is set to $\sigma = 1$. The scond approach is a multivariate normal distribution using a full covariance matrix

$$f_i(x) = \frac{1}{(2\pi)^{d/2}|\overline{\Sigma}_i|^{d/2}}\exp\left(-\frac{1}{2}(\overline{\mu}_i - x)^\top \overline{\Sigma}_i^{-1}(\overline{\mu}_i - x)\right)$$

where $\overline{\boldsymbol{\mu}}_i$ is the mean and $\overline{\Sigma}_i$ the covariance matrix which is estimated by maximum likelihood.

2.4.3. *The Hausdorff Distance*

The Hausdorff distance measures distance between two point sets, and it is small if all points in each set is close to some point in the other set. This could be useful for comparing dye color changes that are not uniform over the dye. First define the distance between two points a and b as the Euclidean distance $d(a,b) = \|a - b\|_2$. The distance between a point a and a set $\mathcal{B}$ is then $d(a,\mathcal{B}) = \min_{b\in\mathcal{B}} d(a,b)$. The Hausdorff distance is defined as

$$d_{\text{Ha}}(\mathcal{A},\mathcal{B}) = \max\left\{\max_{a\in\mathcal{A}} d(a,\mathcal{B}), \max_{b\in\mathcal{B}} d(b,\mathcal{A})\right\}$$

As an alternative approach one can use the modified Hausdorff distance which is more robust in the presence of noise and outliers

$$d_{\text{MH}}(\mathcal{A},\mathcal{B}) = \max\left\{\frac{1}{N_a}\sum_{a\in\mathcal{A}} d(a,\mathcal{B}), \frac{1}{N_a}\sum_{b\in\mathcal{B}} d(b,\mathcal{A})\right\}$$

It should be noted that this distance is not a metric as the triangle inequality is not fulfilled [14].

2.4.4. *From Distances to Covariance Functions*

Given a distance metric $d^2(i,j)$ we use the distance substitution approach [15] based on a squared exponential kernel,

$k\left(i,j\right) = {\sigma_f}^2 \exp\left(-\frac{1}{\sigma_l} d^2(i,j)\right)$. We note that a valid covariance function may be constructed directly for the Hellinger distance by considering the inner product given by the integral which is known as the Probability Product Kernel [16]; however, to make a fair comparisons we treat the Hellinger distance like the other distance measures. Note also, that since the modified Hellinger distance is not a metric, the distance substitution kernel is not positive definite [15].

The use of kernels provides a convenient way of integrating information from different sensors by combining different kernels in a weighted sum. Thus, different dyes can be combined by constructing the following kernel

$$k\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) = \sigma^2 \mathbf{I} + \sum_{m=1}^{M} \alpha_m k\left(\boldsymbol{X}_i^m, \boldsymbol{X}_j^m\right)$$

where each kernel function is the distance substitution kernel with one of the respective metrics. We generally consider the conic sum such that $\alpha_i \in \mathbb{R}_+$. We may also consider the combination of all dyes and all metrics for that dye, i.e.,

$$k\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) = \sigma^2 \mathbf{I} + \sum_{m=1}^{M} \sum_{q=1}^{Q} \alpha_{m,q} k_q\left(\boldsymbol{X}_i^m, \boldsymbol{X}_j^m\right)$$

The main issue is the estimation of the individual α's, however the Gaussian Process method has the option of learning the weights using evidence optimization. The combination of kernels is generally known as Multiple Kernel Learning (MKL) where one often places further constraints on the α's in order to obtain a convex optimization problem (see [17] for a recent review).

To fuse the difference dyes we employ a forward selection method using the following steps. 1) For a given dye, perform a grid search of hyper parameters. 2) Optimize the evidence using the optimal point found in the grid as initial guess. 3) Perform *leave-one-out* cross validation to get classification error. 4) Choose the dye that yields the lowest classification error.

2.5. Evaluation

We apply nearest neighbor classifiers to each dye for each feature extraction technique in a one vs all setting. From earlier work [10] it was shown that the sensor is proficient in detecting among others acids, alcohols, amines and explosives so these are the classes we evaluate. The generalization error is estimiated using LOOCV [18]. This scheme result in a total of 124 classifiers of each type per feature extraction method (31 dyes × 4 classes).

To judge the differences in classification performance we use the McNemar significance test [19]. The McNemar is a paired test which uses the the number of cases where two classifiers disagree about a decision. From this test we calculate p–values for each comparison and use the multiple hypothesis framework proposed by Storey [20]. Based on the p–values we can calculate an expected false discovery rate (eFDR) for our significant differences, that is, the expected quantity of wrongly significant results amont all found significant results.

Chemical	M	P	IM	MH	G	eFDR	n
Acids	1	3	4	8	4	0.11	20
Alcohols	0	1	4	0	5	0.15	10
Amines	1	0	1	0	0	1.27	2
Explosives	0	4	0	13	3	0.09	20

Table 1. Number of instances the best performing method is significantly better than another method. The methods are Mean, Parzen, Inner Mean, Mod. Hausdorff and Gaussian.

3. RESULTS AND DISCUSSION

Initially we employ 1NN and KNN and rank the dyes according to their best performance. Fig. 2 top panels show the dye ranking for KNN. The modified Hausdorff is the best performer for dye 19 and 30 where it is significantly better than the single feature methods. Further the distribution methods are generally better or on par compared to the single feature methods.

Table 1 shows a summery of the McNemar significance tests. The overall best distance is the modified Hausdorff method although it is noteworthy that this method is not best even once for alcohols whereas is is best for noumerous occasions for acids and explosives. Looking into the dyes reveal that the top dyes for acids largely overlap with the top dyes for explosives whereas for alcohols one of the top dyes is dye number 5. For this dye the Inner Mean is significantly better than any of the other methods. It should be noted that the majority of the differences is between one of the distribution methods and the single feature method, i.e. of the 52 significant results, only on 3 instances was the parzen window worst and only 4 times where the Mod. Hausdorff method worst. Considering we have a total of 6 false discoveries we can only conclude that the distribution methods are significantly better than the single feature methods. To explore the effect of the methods deeper we fuse the dyes using GP.

Initially we want to establish the performance of GP in the same setting as KNN. Fig. 2 show the classification performance for GP when classifying explosives. Again the modified Hausdorff is the top performer for the first two dyes, which is again dye 19 and 30. However for dye 28 and 24 the modified Hausdorff coupled with GP yields an error of 21% and 30% respectively. This might very well be an effect of the fact that the modified Hausdorff is not a valid metric and as such the corresponding kernel might not be psd [14, 15]. But as the case of KNN, the figure does not show a clear indication of which method is superior although it seems that

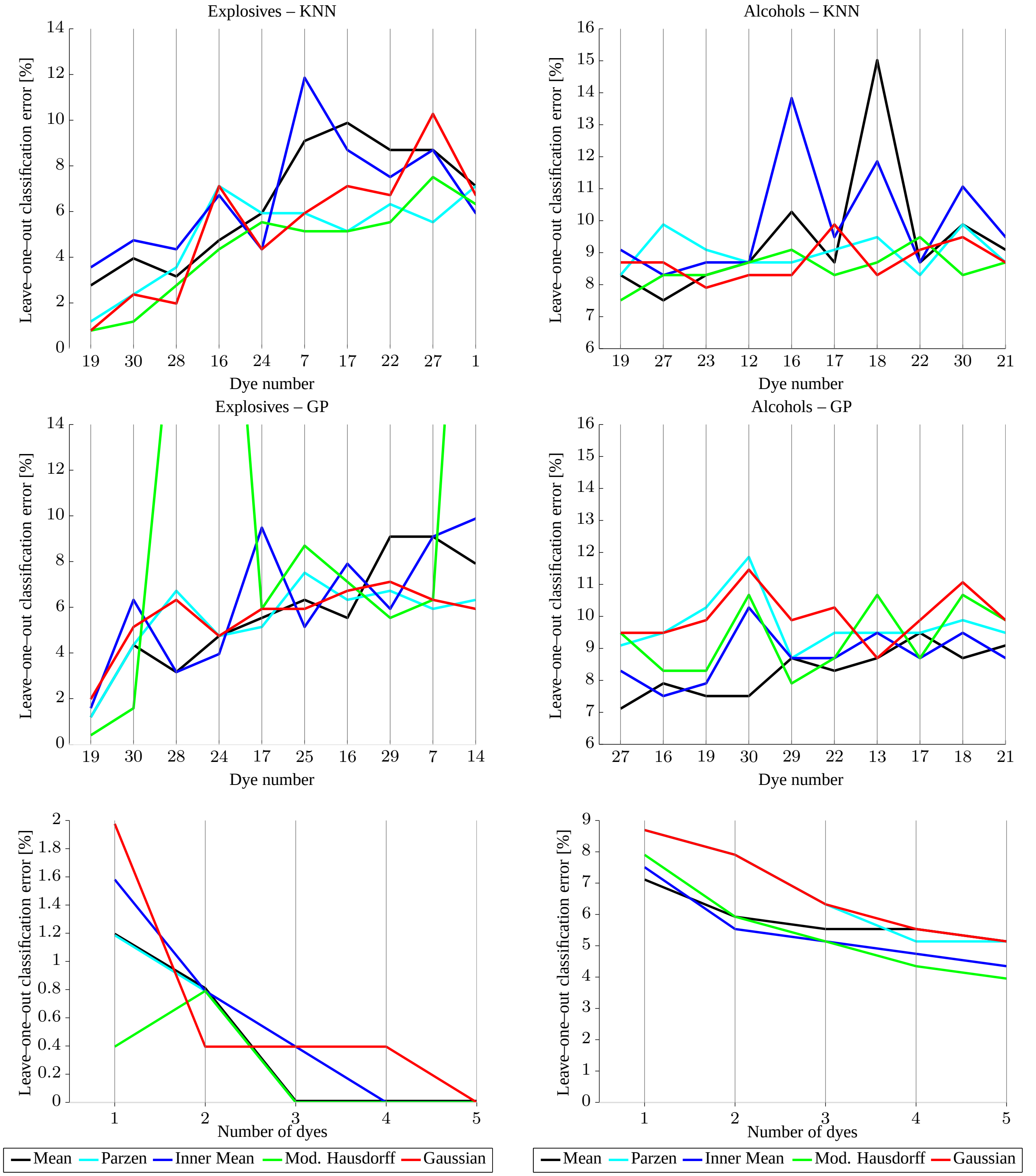

Fig. 2. Classification error for the different measures of similarity for GP and KNN. The top panels show the best performing dyes for explosives and alchohols using KNN. The middle panels show the best performing dyes using GP. The bottom panels shows the result of fusing the dyes together using GP.

modified Hausdorff is able to capture more information about dye 19 and 30 than the the methods.

When fusing the dyes one by one using forward selection we find no significant results between the methods (Fig. 2). To get a significant result the performance must at least differ 2.4 percent points but already with one dye we are below that margin.

4. CONCLUSIONS

We have proposed three new methods for representation of sensory data in colorimetric sensor arrays, namely the Inner Mean, Hellinger distance using a Gaussian distribution and the Modified Hausdorff. Each method have it's merits. The Inner Mean seem to be particularly proficient for dye number 5 and 22 whereas the modified Hausdorff is especially strong for dye number 19 and 30. This could indicate that the best way to represent dye measurements depends on the dye in question.

The modified Hausdorff method is strong for explosives, and since the distance calculation is based on sets of pixels and is parameterfree it can potentially work a lot better for high dimensional data where the Hellinger distance method might be insufficient.

Finally, we have demonstrated an methodology for fusing the measurements from the different dyes. GP classification effectively identifies which dyes should be included and generally including more dyes reduces the LOOCV error rate.

5. REFERENCES

[1] C. Zhang and K. S. Suslick, "A colorimetric sensor array for organics in water.," *Journal of the American Chemical Society*, vol. 127, no. 33, pp. 11548–9, Aug. 2005.

[2] Kenneth S. Suslick, Neal A. Rakow, and Avijit Sen, "Colorimetric sensor arrays for molecular recognition," *Tetrahedron*, vol. 60, no. 49, pp. 11133–11138, Nov. 2004.

[3] N. Rakow, A. Sen, M. C. Janzen, J. B. Ponder, and K. S. Suslick, "Molecular recognition and discrimination of amines with a colorimetric array.," *Angewandte Chemie (International ed. in English)*, vol. 44, no. 29, pp. 4528–32, July 2005.

[4] C. Zhang, D. P. Bailey, and K. S. Suslick, "Colorimetric sensor arrays for the analysis of beers: a feasibility study.," *Journal of agricultural and food chemistry*, vol. 54, no. 14, pp. 4925–31, July 2006.

[5] S. H. Lim, L. Feng, J. W. Kemling, C. J. Musto, and K. S. Suslick, "An Optoelectronic Nose for Detection of Toxic Gases.," *Nature chemistry*, vol. 1, pp. 562–567, Sept. 2009.

[6] C. Zhang and K. S. Suslick, "Colorimetric sensor array for soft drink analysis.," *Journal of agricultural and food chemistry*, vol. 55, no. 2, pp. 237–42, Jan. 2007.

[7] N. V. Kostesha, T. S. Alstrøm, C. Johnsen, K. A. Nilesen, J. O. Jeppesen, J. Larsen, M. H. Jakobsen, and A. Boisen, "Development of the colorimetric sensor array for detection of explosives and volatile organic compounds in air," in *Proceedings of SPIE*, Apr. 2010, vol. 7673, pp. 76730I–76730I–9.

[8] N. V. Kostesha, T. S. Alstrøm, C. Johnsen, K. A. Nielsen, J. O. Jeppesen, J. Larsen, A. Boisen, and M. H. Jakobsen, "Multi-colorimetric sensor array for detection of explosives in gas and liquid phase," in *Proceedings of SPIE*, 2011, pp. 80181H–80181H–12.

[9] Tommy S. Alstrøm, Raviv Raich, Natalie V. Kostesha, and Jan Larsen, "Feature extraction using distribution representation for colorimetric sensor arrays used as explosives detectors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[10] Tommy S. Alstrøm, Jan Larsen, Natalie V. Kostesha, Mogens H. Jakobsen, and Anja Boisen, "Data representation and feature selection for colorimetric sensor arrays used as explosives detectors," *IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2011.

[11] T. S. Alstrøm and J. Larsen, "Feature Extraction and Signal Representation for Colorimetric Sensor Arrays," Tech. Rep., DTU Informatics, 2011.

[12] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Secaucus, NJ, USA, 2006.

[13] Carl Edward Rasmussen and Christopher K. I. Williams, *Gaussian processes for machine learning.*, Apr. 2006.

[14] M. P. Dubuisson and A. K. Jain, "A modified hausdorff distance for object matching," in *12th International Conference on Computer Vision Image Processing*, oct 1994, vol. 1, pp. 566–568.

[15] Bernard Haasdonk and Claus Bahlmann, "Learning with distance substitution kernels," in *in Pattern Rcognition - Proc. of the 26th DAGM Symposium*, 2004.

[16] T. Jebara and A. Howard, "Probability Product Kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.

[17] Mehmet Gönen and Ethem Alpaydin, "Multiple kernel learning algorithms.," *Journal of Machine Learning Research*, vol. 12, no. 7, pp. 2211, 2011.

[18] Lars K. Hansen and Jan Larsen, "Linear unlearning for cross-validation," *Advances in Computational Mathematics*, vol. 5, pp. 269–280, 1996.

[19] Quinn McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[20] John D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 479–498, 2002.