Available online at www.sciencedirect.com
SCIENCE DIRECT
Sensors and Actuators B 117 (2006) 282–285
SENSORS AND ACTUATORS B CHEMICAL
www.elsevier.com/locate/snb

Decoding beads in a randomly assembled optical nose

Bahram G. Kermani, Igor Fomenko *, Theo Kotseroglou, Behrouz Forood, Lori Clark, David Barker, Michal Lebl

Illumina Inc., 9885 Towne Centre Drive, San Diego, CA 9212, United States

Received 6 September 2005; received in revised form 18 November 2005; accepted 18 November 2005
Available online 6 January 2006

Abstract

In Illumina's technology, the term bead is synonymous with microsensors used in optical arrays. Unlike orderly arranged microarrays, a randomly assembled array would need to be processed via a so-called decoding step, in order to identify the location of each beadtype. Illumina's O-nose technology is radically different from the electronic nose (E-nose) technologies by several factors, e.g., the number of sensors. In an O-nose application, one can easily obtain 2000 usable sensors. The quantity of sensors, however, does come at a price, i.e., the necessity for a decoding procedure. The decoding step plays a challenging role in the O-nose technology. A novel supervised learning technique of decoding randomly assembled arrays, based on subspace classifier method is proposed.

Keywords: Optical O-nose; Randomly assembled microarrays; Pattern recognition; Principal components

1. Introduction

In Illumina's technology, the term bead is synonymous with microsensors used in optical arrays. Unlike orderly arranged microarrays [1,2], a randomly assembled array would need to be processed via a so-called decoding step, in order to identify the location of each beadtype. In DNA-related applications, the decoding step is done via multi-stage hybridization to the complementary oligonucleotides (a.k.a., oligos) [3–6]. For optical nose (O-nose) chemical sensors, since the probes are not oligo-based, this method does not apply. Illumina's O-nose technology is radically different from the electronic nose (E-nose) technologies by several factors, e.g., the number of sensors. In an O-nose application, one can easily obtain 2000 usable sensors. The quantity of sensors, however, does come at a price, i.e., the necessity for a decoding procedure. Upon assembly, the beads (sensors) are randomly distributed on the array substrate. The process by which one would identify the location of each bead is referred to as decoding. The decoding step plays a challenging role in the O-nose technology.

In this paper, a novel method of decoding randomly assembled arrays is introduced. This is based on subspace classifier method [7].

The sensors are decoded by exposing the mixture of the sensor bead-types to a certain analyte or to a series of pre-selected analytes. More specifically, the time-course of the exposure of the sensors to nitrogen followed by the exposure to the specific analyte is obtained. By selecting an appropriate analyte, one can obtain different signatures from the different optical sensors. This idea is the main focus of the following study. In the more complex cases, the signature of the sensors may not be completely resolvable by a single analyte. In this case, the methods developed by this study are still applicable. However, one would need to perform a series of exposures to multiple analytes. At each stage of the series, the same procedure is executed. After the final stage, the individual results are pooled together, in order to make a composite decision. The major assumption for enabling the above claim is that even though each analyte, by itself, cannot decode the complexity of the mixture, the combination of the carefully selected analytes would enable one to do so.

An alternate method of decoding such a problem is based on unsupervised learning [8]. In this method, the time-course signal is first compressed, and then processed using a clustering algorithm, e.g., fuzzy C-means (FCM). In general, since the supervised learning method can make use of class labels,

* Corresponding author at: Systems Informatics Lead, Amgen, One Amgen Center Drive, MS 34-2-A, Thousand Oaks, CA 91320, United States.
Tel.: +1 8054479961/8052414130; fax: +1 8053758519.
E-mail address: ifomenko@amgen.com (I. Fomenko).

0925-4005/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.snb.2005.11.060

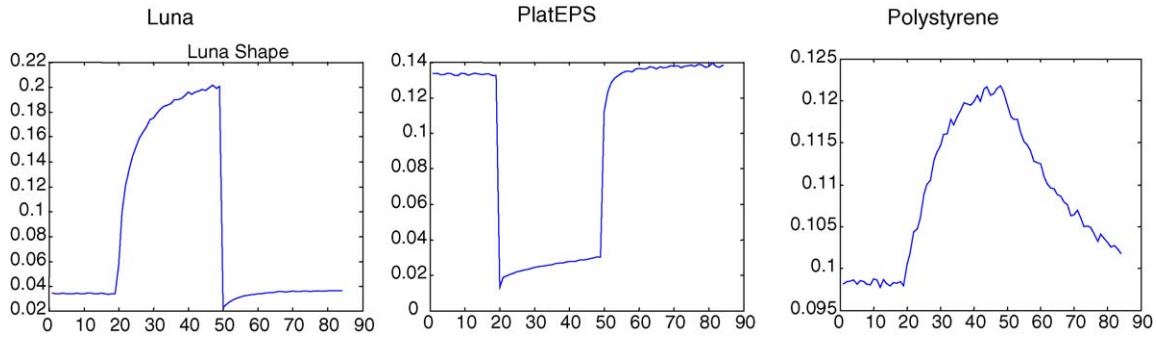

Fig. 1. Class prototypes, i.e., the projections onto the first principal component of the class luna (a), platEPS (b), and polystyrene (c). (Abscissa and ordinate represent the frame number (time) and bead intensity, respectively).

its performance is expected to be superior to that of the unsupervised learning.

2. Methods

A method of supervised learning was devised for this problem. More specifically, three sets of data were provided along with their corresponding class labels. A fourth set, which contained the combination of the three bead-types, was also provided. For this latter set of beads, no class label was provided. The objective of this study was to place labels (1, 2, or 3) on every bead of the fiber bundle containing the mixture of the three bead-types. Fig. 1 illustrates the prototype time-course of the three bead-types under test, along with the name of the compounds that the sensors are made of [8].

Fig. 2 shows the profile of 49 randomly selected bead-types from the multi-bead-type fiber bundle, i.e., the fiber bundle containing the mixture of the beads. It is notable that not all of the bead-types follow the above general patterns, to a great degree.

For every bead-type, three features were extracted, one feature per class. These features were based on the normalized cross-correlation between the pattern of a bead-type and the class prototypes. The representation of the bead-type in this N-dimensional space ($N=3$) is projected into the $N-2=1$-dimensional subspaces. In each subspace, a negative label is assigned to the bead-type, if it falls on the negative side of the subspace. For instance, consider the projection into the subspace spanned by X3. In this case, if a bead-type's projection falls on the negative side of the X3 axis, the bead-type is labeled as "not belonging to Class 3." If the projection falls into the positive side, no label is assigned. This process is repeated for all the other possible subspaces. At the end of this process, the partial (negative) calls are combined, and a composite call is deduced. For example, if the bead-type does not belong to X3 and does not belong to X2, then by deduction it has to belong to $\mathrm{UoD}-\{\mathrm{X2},\ \mathrm{X3}\}=\mathrm{X1}$, where UoD is the universe of discourse. Occasionally, a bead-type may not have enough negative calls to satisfy a unique deductive

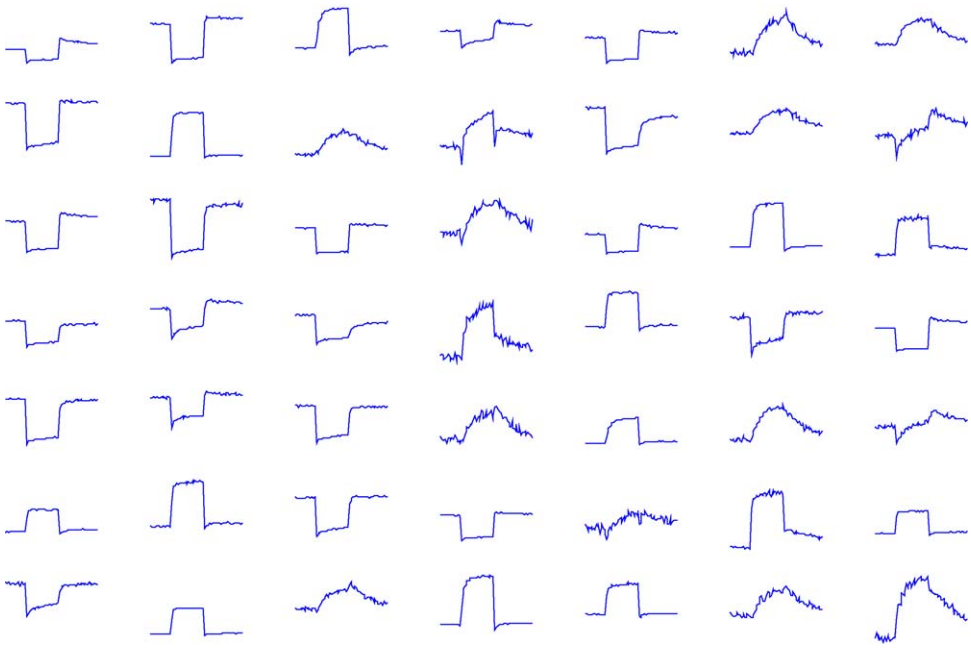

Fig. 2. A sample of the three bead-type mixture. Each time trace represents one bead-type in the mixture.

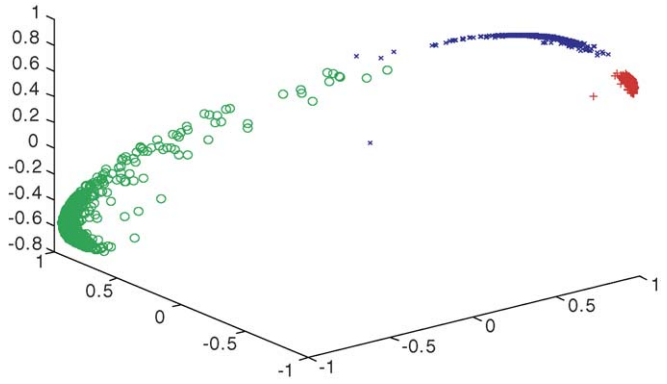

Fig. 3. Bead labeling, based on training separate classes. Each symbol/shade represents one labeled class. *X*, *Y*, and *Z* axes represent the projection of the bead's time-response onto the first principal component of Classes 1, 2, and 3, respectively.

solution. In that case, the bead-type would receive a no-call label.

In an attempt to rank the quality of the decoded beads, a score was assigned to each decoded bead-type. This score was a function of the Mahalanobis distance[9] of the bead to (the other members of) its assigned class. The raw Mahalanobis distances were processed via sigmoidal functions, in order to transform the distances to scores, bounded in [0,1]. The sigmoidal function was designed such that at Mahalanobis distances of 3 and 10, the scores were approximately 1 and 0, respectively. These numbers were selected based on the heuristic assumptions that for a normal distribution, a *Z*-distance of 3 contains more than 99% of the data, and data points with *Z*-distance of 10 or higher can be labeled as outliers [9].

3. Results

Fig. 3 shows the 3-D representation of the individual bead-types, i.e., three fiber bundles, each containing only one bead-type. Components of each dot (bead) on *X*, *Y*, and *Z* axes correspond to the projection of the bead time-response onto the first principal component of Class 1, Class 2 and Class 3, respectively.

Fig. 4 corresponds to the multi-bead bundle. This fiber bundle contains beads from all the three types. This is obvious from the span of the values.

Fig. 5 shows the results of the multi-bead-type fiber bundle after decoding. The resemblance of this figure to Fig. 3 provides a visual confirmation on the quality of the decoding. The classes, however, are not completely separated, i.e., there is no significant gap between the classes. This can be partially attributed to the fact that sensor responses are not always pure, i.e., they do not always belong to one of the three classes. Some sensors may fail to respond properly, as it is evident in Fig. 2.

Fig. 6 shows the beads of the three classes of the multi-bead fiber bundle, with scores greater than the arbitrary threshold of 0.7 (all shown in red). The rest of the beads are shown in green.

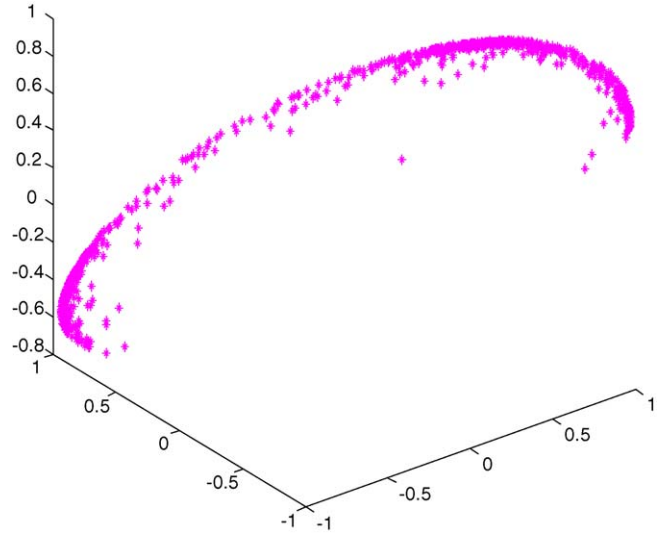

Fig. 4. The multi-bead bundle projections. The plot represents the mixture of three bead-types. *X*, *Y*, and *Z* axes represent the projection of the bead's time-response onto the first principal component of Classes 1, 2, and 3, respectively.

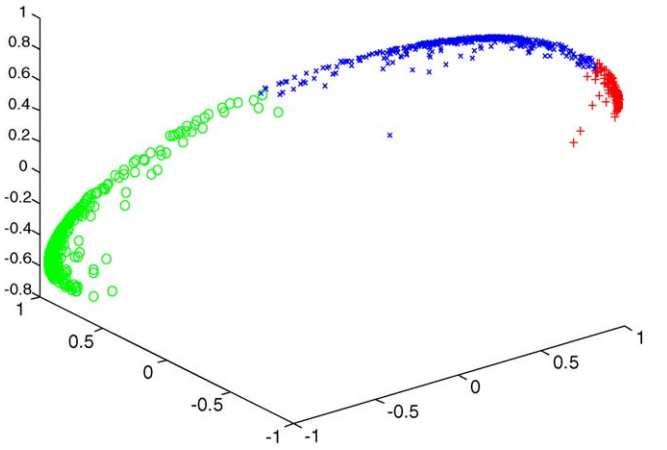

Fig. 5. Bead labeling using supervised learning. *X*, *Y*, and *Z* axes represent the projection of the bead's time-response onto the first principal component of Classes 1, 2, and 3, respectively.

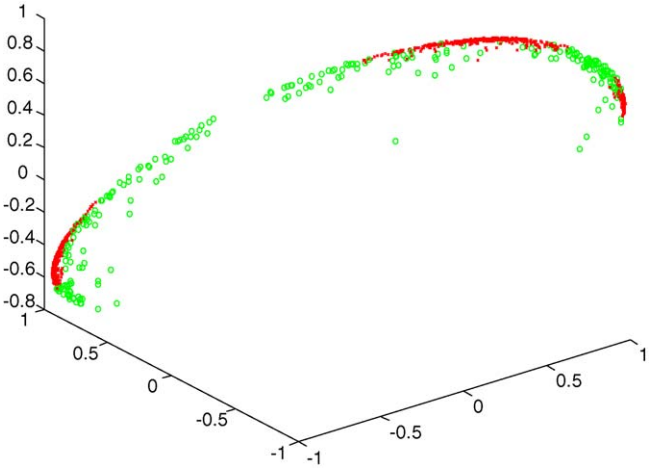

Fig. 6. Points with score greater than 0.7. *X*, *Y*, and *Z* axes represent the projection of the bead's time-response onto the first principal component of Classes 1, 2, and 3, respectively.

Table 1
The number of high-quality beads (score >0.7) in different classes

Sensor	Total beads decoded	Acceptable beads
Luna	212	168
PlatEPS	557	453
Polystyrene	473	367
Total	1240	988

Given the arbitrary threshold of 0.7 on the scores, one can bin the high-quality beads into three classes, as shown in Table 1.

According to Table 1, approximately 80% of the beads were decoded (988/1240 = 0.8) with score >0.70. The resultant decoded beads were visually confirmed for accuracy of the calls. The decode efficiency (DE) of 80% is to contrast with the 10% DE achieved by an unsupervised learning method[8].

4. Conclusion

Supervised learning results in higher decoding efficiency (80%) in the prediction of the unknown classes, as compared with an unsupervised learning method (which resulted in 10% decoding efficiency). The supervised learning is also less sensitive to the number of elements in each class. In particular, it is less sensitive than the unsupervised learning to the imbalance in the number of items in clusters. This is mainly due to the fact that in supervised learning, one can exploit the domain-specific prior knowledge.

References

[1] M.J. Heller, DNA microarray technology: devices, systems and applications, Annu. Rev. Biomed. Eng. 4 (2002) 129–153.
[2] D. Shalon, S.J. Smith, P.O. Brown, A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization, Genome Res. 6 (7) (1996) 639–645.
[3] T.A. Dickinson, K.L. Michael, J.S. Kauer, D.R. Walt, Convergent, self-encoded bead sensor arrays in the design of an artificial nose, Anal. Chem. 71 (1999) 2192–2198.
[4] J. White, J.S. Kauer, T.A. Dickinson, D.R. Walt, Rapid analyte recognition in a device based on optical sensors and the olfactory system, Anal. Chem. 68 (1996) 2191–2202.
[5] T.A. Dickinson, J. White, J.S. Kauer, D.R. Walt, A chemical-detecting system based on cross-reactive optical array sensors and temporal responses, Nature 382 (1996) 697–699.
[6] K.S. Lam, M. Lebl, V. Krchnak, The “one-bead one-compound” combinatorial library method, Chem. Rev. 6980 (1997) 411–448.
[7] J. Laaksonen, Proceedings of the ICANN’97 on Local Subspace Classifier, Stockholm, Sweden, 1997, pp. 37–40.
[8] B. Forood, T. Kotseroglou, L. Clark, M. Lebl, M. Lieu, B.G. Kermani, D. Barker, T. Dickinson, Chemical detection using the optical nose system, in: Ninth International Symposium on Olfaction and Electronic Nose ISOEN’02, Rome, Italy, 2002.
[9] MATLAB Statistics Toolbox, Math Works, Inc. Natick, MA.