# Fuzzy k-NN applied to moulds detection.

Kuske Martyna, Rubio, Rubio Rafael, Nicolas Jacques, Marco Santiago, Romain Anne-Claude

Communication presented at ISOEN 2003 – RIGA- Latvia

## Introduction

Excessive humidity and subsequent fungal development are one of the most frequent problems in buildings. Moulds can provoke different symptoms: allergies, infections, irritations, general symptoms, and toxic effects. Some species are particularly dangerous because their toxins are involved in serious diseases leading even to death. Every source of fungal development should be diagnosed and removed as early as possible. However, classical methods evaluating the quantity of viable fungi don't give results before several days. Other, more rapid methods have been investigated recently; one of them is detection of microbial volatile organic compounds (MVOCs). An additional advantage of MVOCs detection is the possibility to detect hidden fungi, when the spores are not liberated in the air.
Several laboratories in the world use gas chromatography – mass spectrometry to detect volatile compounds produced by moulds. Electronic nose technology seems very interesting because, in difference with gas chromatography, the method is simpler, cheaper, and the results can be obtained in situ.

Concerning the signal processing, the K-nearest neighbours' decision rule is often implemented for this kind of applications. Its computational simplicity and the good results obtained in small problems still make this algorithm interesting, especially if some improvements can be done.

One of the problems encountered using the K-NN classifier is the lack of information about the "typicalness" of the samples used to label the input space. Due to the crisp character of the input membership given to make the classification, no information about how the data is distributed in the input space is provided. Another difficulty found is that the algorithm, once a input vector is classified, doesn't give information about the "strength" of membership to that class. These two problems were addressed by Keller, Gray and Givens in 1985 [1] incorporating fuzzy sets theory into the K-NN rule and developing a fuzzy K-NN algorithm.

The fuzzy set theory is introduced both, in the samples used to the classification and in the output given class. Concerning the samples, they're no longer described by a crisp label but with a fuzzy membership function for each class. On the other hand, another membership function is obtained as output for the input vector. This allows the user to make the last decision using this fuzzy information or to introduce thresholds of confidence level.

Although better performance, compared with classic K-NN, of this algorithm is expected, some a priori disadvantages are found. A membership function for the samples needs to be defined. This means that some information on how our samples are distributed has to be found. This can be a difficulty in cases with few training data. In addition, if this membership function is a hard partition, since the membership of one data for the different classes is normalized, some times samples are given non-zero membership to classes that originally they don't belong to. In last term, the fuzzy output makes necessary the introduction of a classification criterion.

**Goals**

The goal of the study is to develop an instrument that would be able to distinguish between "clean" materials and materials contaminated by moulds. In other words, the purpose is to detect the presence of moulds without identifying the kind of fungus.

To do so the two mentioned classification algorithms are going to be implemented. A further analysis of their performance, in combination with different dimensionality reduction techniques, is going to be compared.

**Measurement description**

Eight fungal species, *Aspergillus niger, Aspergillus versicolor, Cladosporium cladosporioides, Cladosporium sphaerospermum, Penicillium aurantiogriseum, Penicillium brevicompactum, Penicillium chrysogenum, Alternaria alternata,* all known to be often found in buildings, were cultivated on malt extract agar. Ten days later, the cultures were used to contaminate building materials. Typical building materials were selected for the study : gypsum board, particle board, oriented strand board (OSB), and wallpaper. Several combinations of these materials were examined.

Samples of materials were placed in jars (volume 500 ml), and incubated at room temperature. Forty millilitres of water was poured at the bottom of the jar, and the samples were placed in the way to avoid direct contact with water.

The e-nose was constructed of 12 metal oxide sensors (six Figaro and six Capteur) placed in two stainless steel boxes. Purified and humidified air is used as the reference air. During analyse, the air passes through the jars containing samples before entering to the boxes with sensors. Analyses start on the 21$^{st}$ day of incubation, and are made every 7 days, up to 3 months.

**Fuzzy K-NN algorithm description**

Both K-NN and fuzzy K-NN algorithms are used. The first is a well know algorithm that already has been implemented in mathematical tools like Matlab. For the implementation on the second one, the architecture proposed by Keller [1] is used. In this algorithm the membership function of a sample x to the cluster i is computed like:

$$u_i(x) = \frac{\sum_{j=1}^{K} u_{ij}\left(1/\left\|x - x_j\right\|^{2/(m-1)}\right)}{\sum_{j=1}^{K}\left(1/\left\|x - x_j\right\|^{2/(m-1)}\right)}. \quad (1)$$

As (1) shows the memberships of x is a weighted average of the memberships of the K closer neighbours .The weight is proportional to the inverse of the distance of the neighbour to x. The variable m determines how the distance is weighted. As m increases the effect of the relatives distances is less. The membership of the samples $u_{ij}$ to the class has to be given to algorithm as an input. The membership has to introduce information about how the data of the different classes are distributed. So, in order to build the membership function an assumption on how the data distribution data look like has to be done.

**Data processing**

The features used are the normalized resistance of the sensors excluding the humidity sensor. Two data sets with different normalizations are used. The first one is $(R-R_0)/R_0$ normalization

where $R_0$ is the resistance for the reference air. The second is the pattern normalization of the first one in order to eliminate the effect of intensity.

To build the membership function different approaches are taken. The first one is two suppose that the each class is composed by clusters with Gaussian distribution. A very simple first approach is just to suppose that the two classes are the same that the ones concerning the classification problem. This is moulds and no moulds. After this a more accurate model represent each of this classes formed by four different clusters corresponding to each material. Although four clusters are took into account, the classification problem still is a two-classes problem as the membership is only for these two classes (moulds and no moulds) and the four clusters are only considered to describe their inner distribution structure. To compute the membership of one element the normalized values of the distribution are taken.

The second approach tries to make a more accurate definition of the inner distribution of each cluster. Now each cluster is supposed to be and add of spherical Gaussians with different centres. The centres are calculated as prototypes of a clustering algorithm (fuzzy-C-means) applied to the single cluster.

These algorithms are applied on the raw space but also in the ones obtained by dimensional reduction using PCA and LDA. In these two last cases an outlier extraction is carried out by applying the residuals statistical confidence limit.

Determination of number of principal components is obtained by studying the residuals evolution. LDA using the two classes has been performed using the resulting first root and residuals for the further classification.

The cross-validation of the models proposed is performed using k-fold validation. In order to have more statistical information about the classification results, several executions of this cross validation is carried out to make a pro medium of the results. The best k neighbours parameters election is made looking the histogram of the k parameters obtained for the best classification rate each loop. This exploration was hold between k=1 and k=25. After that the classification rate assigned is the median of the obtained for that k during the different loops.

**Results**

A first PCA plot of the two sets of data gives an idea of the complexity of the problem faced.

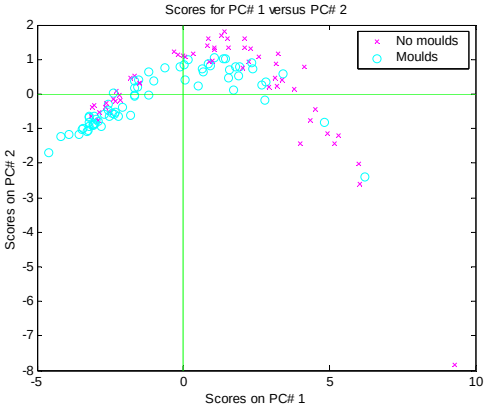


**Fig 1** PCA plot for $(R-R_0)/R_0$

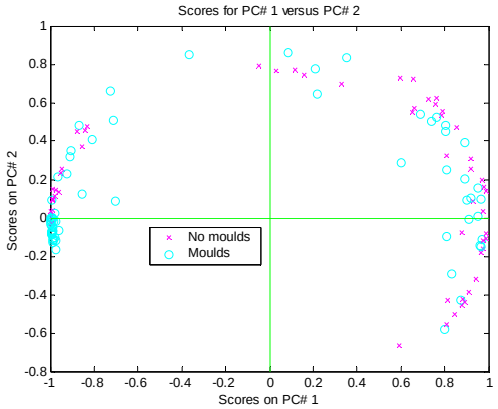


**Fig 2** PCA plot for pattern normalized $(R-R_0)/R_0$

*Fig 1* and *Fig 2* show that a simple classification in the two principal components space is impossible. In this case to perform a pattern normalization seems to spread more the data over

the two principal components space. Not information seems to be lost in this process of normalization.

The classification rates obtained for each classifier and dimensionality are summarized in *table 1*.

| **Model** | | | **Raw** | **5 pcs** | **LDA 2 classes** |
|---|---|---|---|---|---|
| **K-NN** | $(R-R_0)/R_0$ | | C.R:85±4 for k=1 | C.R:87±3 for k=1 | C.R:61±5 for k=1 |
| | Pattern normalized | | C.R:87±3 for k=1 | C.R:88±2 for k=1 | C.R:86±4 for k=5 |
| **Fuzzy** | Crisp | $(R-R_0)/R_0$ | C.R:87±2 for k=9 | C.R:88±1 for k=5 | - |
| **K-NN** | Membership | Pattern normalized | C.R:87±3 for k=1 | C.R:89±2 for k=11 | - |
| | 2 clusters | $(R-R_0)/R_0$ | C.R:92±3 for k=1 | C.R:97±1 for k=5 | - |
| | 1Gauss/cluster | Pattern normalized | C.R:91±2 for k=7 | C.R:96±1 for k=5 | - |
| | 8 clusters | $(R-R_0)/R_0$ | C.R:90±2 for k=5 | C.R:94±1 for k=5 | - |
| | 1Gauss/cluster | Pattern normalized | C.R:93±1 for k=9 | C.R:98±1 for k=1 | - |
| | 2 clusters | $(R-R_0)/R_0$ | - | C.R:98±1 for k=13 | C.R:58±7 for k=7 |
| | 4Gauss/cluster | Pattern normalized | - | C.R:96±2 for k=3 | C.R:84±3 for k=7 |
| | 8 clusters | $(R-R_0)/R_0$ | - | C.R:96±1 for k=3 | |
| | 2Gauss/cluster | Pattern normalized | - | C.R:97±1 for k=3 | |

**Table 1** Classification rate (%) obtained using K-NN and fuzzy K-NN with different samples membership functions. The classifiers are applied to different input spaces obtained by dimensional reduction.

A maximum membership criterion is used to determine the class of the input vectors. The number of principal components is optimized by the calculus of the residuals mean power. For both data sets an elbow is found for k=5.

**Conclusions**

In general a best performance of the fuzzy K-NN classifiers with fuzzy membership of the samples is observed in front of the traditional K-NN algorithm, specially for 5 PCA. Fuzzy K-NN with a crisp membership shows the same behaviour that K-NN. This is an indication on how important the fuzzyness of the membership function is. Further differences on the fuzzy membership function shape are not observed. The similarity of the models can be the reason

On the other hand not better results using fuzzy-K-NN are observed when applying the classifier with 2-classes LDA. In these last cases a better performance on the pattern normalized space is shown. But the for raw and the 5 principal components space there seems not to be loose of information by using the pattern normalized data set instead of the data set that still contains information about intensity. This can be indicative of the lack of importance of the intensity information for classification purposes.

Not a particular evolution of the number of neighbours is observed.

**Acknowledgements**

Generalitat de Catalunya.

**References.**

[1] J.M. Keller, M.R. Gray and J.A.Givens, JR, *A Fuzzy K-Nearest Neighbor Algorithm*, IEEE Trans. Syst, Man Cybern., vol. SMC-15, no. 4, pp 580-585, July/August 1985.