Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004
June 9 - 11, 2004
Espoo, Finland

A New Approach to Robust Clustering by Density Estimation in an Autocorrelation Derived Feature Space

Dimitris Glotsos[1], Jussi Tohka[2†], Jori Soukka[3] and Ulla Ruotsalainen[2]

[1]Tampere University of Technology
Institute of Signal Processing,
FINLAND
University of Patras, Department of
Medical Physics, GREECE
E-mail: dimglo@yahoo.com

[2]Tampere University of Technology
Institute of Signal Processing,
FINLAND
[†]Tel. +358-3-31154508
[†]E-mail: jussi.tohka@ tut.fi
E-mail: ulla.ruotsalainen@tut.fi

[3]Arctic Diagnostics Oy
P.O.Box 51, 20521 Turku,
FINLAND

ABSTRACT

Robust clustering techniques aim to classify objects into partitions that have meaning for the particular problem, while dealing with outliers contaminating data. In this paper, we propose a new robust clustering method based on the concept of density estimation in an autocorrelation derived feature space. In that feature space, clusters are better separated than in the original data space, making clustering easier. The autocorrelation features comprise the input to a new Probabilistic Neural Network motivated clustering algorithm. We show that the method can also be applied for outlier detection when only one class of data exists. The proposed method was tested with simulated data and real data collected from bioaffinity assay applications. The algorithm was able to separate the clusters despite of the presence of outliers in every case. In addition, we demonstrate that combinations of traditional robust estimation techniques and clustering algorithms (k-means, fuzzy k-means) failed to detect different populations of data when applied to the same datasets due to existence of outliers.

1. INTRODUCTION

Traditional clustering algorithms (i.e. k-means) attempt to partition data points into distinct groups by optimizing a certain criterion function, which is usually derived from a distance metric [1]. One major limitation of these algorithms is their poor performance when there are outliers contaminating data; outliers are defined as points that severely deviate from the majority of the true data samples [2]. The presence of outliers jeopardizes clustering validity and even the existence of a single gross error may cause conventional clustering algorithms to fail [3]. As there usually exist outliers in data collected from measurements of biological phenomena, robust techniques for clustering are vital in these applications.

Most previous works in robust clustering [3-7] have explored the occurrence of clusters in the original data space. Moreover, robust clustering methods were focused mainly in partitioning data based on the known parametric form of the probability density function (PDF) for the data [6-7]. In this work we introduce a novel method to measure the degree of relation between data belonging to the same or different clusters: a) initially two features related to the autocorrelation function transform the input data into an autocorrelation feature space where the differences between data belonging to different clusters are more prominent. b) These features comprise the input to a new PNN (Probabilistic Neural Network)-motivated clustering algorithm that estimates non-parametrically the PDF of all data. Based on this estimation, clusters and outliers are defined as the peaks of the PDF using a k-nearest neighbour heuristic. The proposed algorithm is tested with simulated data and real data collected from bioaffinity assay applications.

2. METHODS

2.1 Autocorrelation feature space

The autocorrelation function is known to encode the degree of association between different data points and furthermore, in a broader sense, the degree of relation between different data clusters [8-9]. Motivated by this characteristic, we designed our clustering algorithm to evaluate not data in the original data space, but their transformation in an autocorrelation derived feature space. Assuming we have N d-component feature vectors $x_1,\ldots,x_N$, we derive two-component autocorrelation feature vectors $a_1\ldots,a_N$. We define for $i=1,\ldots,N$

$$B_i = \sum_{j=1,\, j\neq i} x_i^T x_j \qquad (1)$$

Autocorrelation features are now defined as

$$a_i = \left[(1-B_i)B_i, (1-B_i)^2 B_i\right] \qquad (2)$$

for $i=1,\ldots,N$.

These features are similar to those proposed for texture classification in [8]. As it can be observed from figure 1, in this autocorrelation feature space, individual clusters and outliers' differences are more pronounced. We used these autocorrelation features to cluster the data instead of using the original data.

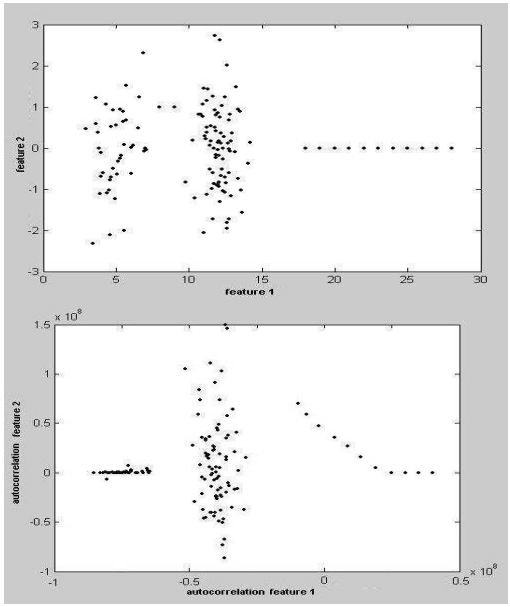

Fig. 1. Mapping the original data space (top panel) in the autocorrelation feature space (bottom panel). Differences between clusters and outliers are more prominent.

2.2 Non-parametric density estimation and Probabilistic Neural Networks

When the data generation function is unknown, non-parametric estimation techniques provide means for estimating the PDF without any prior assumption on its form. One of the most popular approaches was proposed by Parzen [10-11]. The Parzen estimate is given by

$$f_h(x) = \frac{1}{N}\sum_{i=1}^{N} k(x - x_i) \qquad (3)$$

where samples $x_1,\ldots,x_N$ are drawn from a population with the density function $f(x)$, $f_h(x)$ is the Parzen kernel density estimate and $k(\bullet)$ is the kernel function. A widely applied choice for the kernel function is the Gaussian:

$$k(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d}\exp\left(\frac{-\|x\|^2}{2\sigma^2}\right), \sigma = spread \qquad (4)$$

The adjustable parameter σ affects the estimate: Too small deviations cause very spiky approximation; too large deviations smooth out details. In this work we select $\sigma = \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\|x_i - x_j\|$.

The Parzen estimation approach can be implemented in a neural network configuration, by replacing the commonly used sigmoid activation function with the Gaussian function. In this way, the PNN for supervised classification is formed [12], which classifies data by comparing Parzen PDF estimates for each class.

The PNN architecture for supervised classification comprises 4 layers (Fig. 2). The input layer has a node for each feature of input data. The pattern layer has one pattern node for each training pattern. Each pattern node forms a product of the weight vector and the given example for classification, where the weights entering a node are from a particular example. After that, the product is passed through the Gaussian activation function (Eq. 4) to the summation layer which receives the outputs from pattern nodes associated with a given class. The output layer has as many nodes as existing classes. In this layer the classification decision is deduced by comparing the output of the PDF estimation for each class.

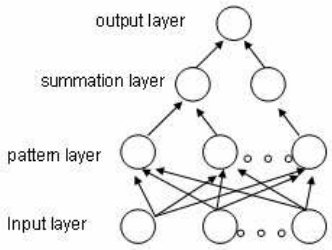

Fig. 2. PNN architecture for two category classification problems.

2.3 PNN-based robust clustering

In order to implement PNNs for robust clustering problems, the network architecture is alternatively constructed with 5 layers. 1) The input layer consists of as many nodes as the number of transformed data samples in the autocorrelation feature space. 2) In the pattern layer the PDF based on each data sample is calculated by using the Gaussian activation function. 3) In the summation layer the PDF of all data is computed by adding up the PDF estimates based on individual data samples. 4) In the clustering layer, clusters are defined as the peaks of the PDF. 5) Finally, in the output layer, the classifications assigned by the algorithm to each point are used to define clusters and outliers.

In the clustering layer, most prominent maxima (peaks) of the PDF are calculated by an iterative maximum likelihood estimation-based algorithm [13]. These maxima are considered as clusters centroids $c_1,\ldots,c_M$. In this paper we assume than M is fixed. A data point x_i is assigned to cluster c_j by using a k-nearest neighbor heuristic. Point x_i is classified to cluster j if the majority of k nearest neighbors is closer to c_j than any other cluster center. The k is defined as $k = 2 + \frac{d_{\max}}{\sqrt{2M}}$, where M is the number of clusters and d_{max} is maximum Euclidean distance between cluster centroids.

After assigning all points to clusters, a procedure to detect outlying points is initiated. First, we select points $O=\{o_1,\ldots,o_r\}$ that are possibly outliers. These are selected starting from point o_1 that has maximum Euclidean

distance from its closest cluster centroid c_j. Point o_2 is then the nearest neighbor of o_1 and the procedure continues this way until with find a point o_{r+1} that is closer to some cluster center than o_r. To decide if o_i is an outlier, we compute how many points in set O are closer to the point o_i than its distance to its closest cluster centroid. If this number is greater than $r/2$ then point o_i is considered to be an outlier. This process is effective in determining outlying points severely deviating from the rest of data. To determine outliers located in between clusters, the Mahalanobis distance (with covariance matrix determined for the outliers previously recognized) from all points to the closest cluster centroid is computed. This is compared to the smallest Mahalanobis distance that an outlier has to its closest cluster centroid. Those points exceeding this distance are also considered as outliers.

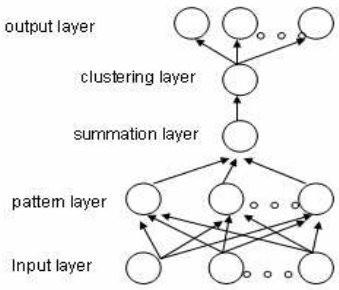

Fig. 3. PNN architecture for robust cluster analysis.

3. RESULTS

The proposed algorithm was tested with a simulated dataset of two dimensional data. It comprised two well defined clusters of 40 and 80 data points and a small set (12 points) of outliers, which consisted of outliers in between the two clusters and outliers severely deviating from the rest of the data. Clusters and outliers were successfully estimated (fig. 3), because the differences between clusters and outliers became more prominent in the autocorrelation feature space. Figures 4-5 depict the performance of k-means and fuzzy k-means in discriminating between the two clusters for the same dataset. Due to the existence of outliers, classification thoroughly failed. We performed the simulation also without outliers with the k-means and fuzzy k-means algorithm, and then the clustering result was excellent. Moreover, the m-type Huber robust estimator was applied for separating outliers (points having distance larger than three times standard deviation from the robust mean) [2] from the rest of the data. As it can be observed in fig. 6, the Huber estimator erroneously accounts as outliers a significant portion of points belonging to the proper clusters.

In order to test the performance of the proposed algorithm in outlier detection, the method was additionally applied to bioaffinity assay data. The assay principle is explained in references [14-16]. The sample contained a suspension of fluorescent microparticles which were excited and measured one at a time. The fluorescence signal was dependent on the analyte concentration in the sample. However, the physics behind the particle detection and variations in fluorescent labelling induced noise and outliers in the measurement. The goal was therefore to provide a robust estimation of the mean value of the fluorescence signal. By using the calibration curve of the instrument, the fluorescence values could be then transformed to concentration values. Results are illustrated in fig. 7. The real molecule concentration was 1nM. The estimate by the PNN-based algorithm was 0.93nM whereas the robust Huber estimate was 0.83nM.

4. DISCUSSION AND CONCLUSIONS

Several techniques have been proposed to deal with data outlier contamination [3-7]; the usefulness of parametric density estimation and distance based variant robust clustering methods [3-6] has been indicated. Among the most recent advances, neural network (NN) approaches have been suggested [6-7]. Sykacek et al [6] proposed a Bayes' inferred NN model that estimated the whole data PDF and asserted outliers as those points having PDF value below a certain threshold. The threshold was defined as a function of the distribution's standard deviation. Hawkins et al [7] employed a regression based NN model that was trained to minimize the mean square error of all training data. Based on this error they defined a threshold value to separate outliers from correct data values. However, these parameter estimation based studies [6-7] incorporated information from the whole data PDF making assumptions on its form, which is misleading for many practical applications. Furthermore, these techniques investigated the existence of clusters in the original raw data space.

In this work, we presented a new method for robust cluster analysis based on non-parametric density estimation using a PNN-based algorithm in an autocorrelation derived feature space. The method proved effective and performed better than k-means, fuzzy k-means and robust estimation techniques in identifying meaningful clusters when tested with a simulated and real bioaffinity assay datasets. The key features of the proposed algorithm are: a) It performs cluster analysis in a autocorrelation feature space where the differences between clusters and outliers are more prominent. b) It evaluates non-parametrically the PDF of data and determines clusters and outliers based on its peaks. c) Outliers either severely deviating or located in between clusters can be effectively detected based on a k-nearest neighbor heuristic.

ACKNOWLEDGMENTS

The authors wish to thank J. T. Soini (Arctic Diagnostics Oy) for providing us the bioaffinity assay data. The work by Dimitris Glotsos and Jussi Tohka has been funded by the EU Marie Curie Fellowship and the Academy of

Finland, respectively.

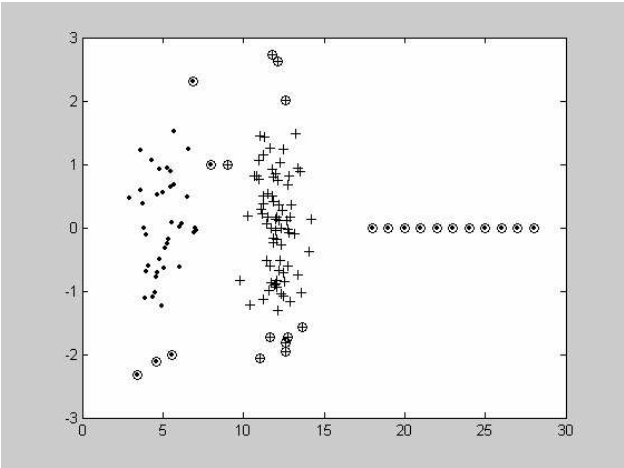

Fig. 4. Results of the PNN-based algorithm in the simulated data: with '.' are points belonging to cluster A, '+' to cluster B and 'o' outliers.

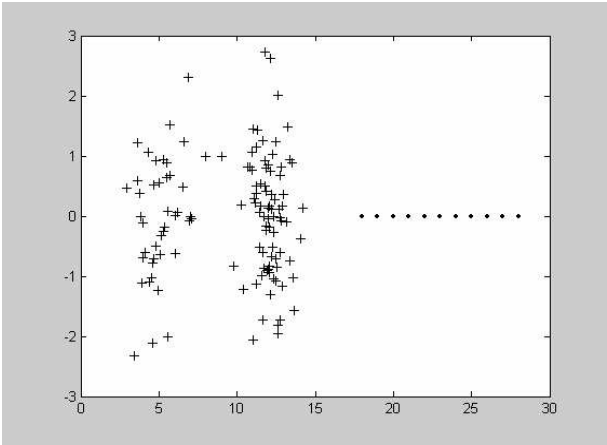

Fig. 5. Results of k-means and fuzzy k-means in the simulated data: with '+' points belonging to cluster A and '.' to cluster B.

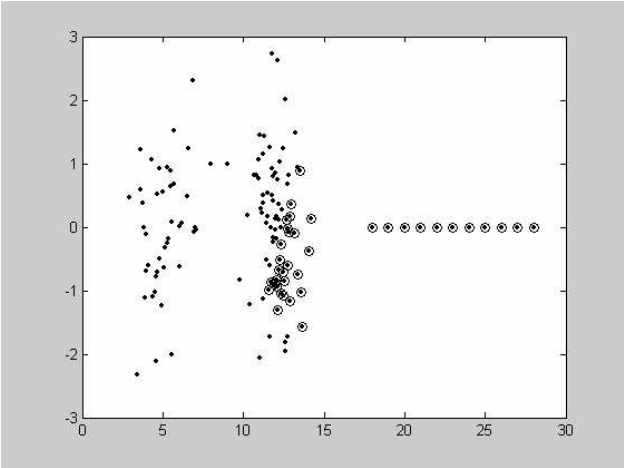

Fig. 6. Results of robust estimation (m-type Huber) in the simulated data in localizing outlying points: 'o' depicts outliers (points having distance more than three times standard deviation from the robust mean)

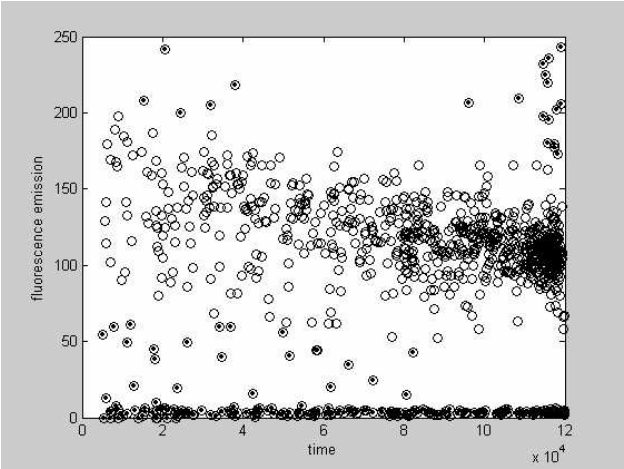

Fig. 7. Data described the fluorescence emission of each particle with respect to measurement time. With '.' are points identified as outliers by the proposed algorithm.

REFERENCES

[1] A. Jain and R. Dubes, *Algorithms for clustering data*, Prentice Hall 1988.

[2] F. Hampel, E. Ronchetti, P. Rousseeuw and W. Stahel, *Robust Statistics*, John Willey & Sons, 1985.

[3] R. N. Davé and R. Krishnapuram, "Robust clustering methods: A unified view," in *IEEE Transactions on Fuzzy Systems,* vol. 5(2), pp. 270-293, 1997.

[4] O. Nasraoui and R. Krishnapuram, "A robust estimator based on density and scale optimization and its application to clustering," In Proceedings of the Fifth IEEE International Conference on Fuzzy Systems, pp. 1031-1035. 1996.

[5] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," *IEEE Trans Fuzzy Systems*, vol. 1(2), pp.98-110, 1993.

[6] P. Sykacek, "Outliers and Bayesian Inference," in *Proceedings of NC 98,* Vienna, Austria, 1998, pp. 973-978.

[7] S. Hawkins, H. He, G. Williams and R. Baxter, "Outlier Detection Using Replicator Neural Networks," in Neural Networks, pp.170-180, 2002.

[8] *O. Faugeras, and W. Pratt ,* "Decorrelation Methods of Texture Feature extraction," in *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 14, pp. 323-332, 1980.

[9] P. Spyridonos et al, "Neural Network based segmentation and classification system for the automatic grading of histological sections of urinary bladder carcinoma," in *Analytical and Quantitative Cytology and Histology,* vol.26, pp. 317-324, 2002.

[10] E. Parzen, "On the estimation of a probability density function and the mode," in *Ann. Inst. Stat. Math.*, vol. 33, pp. 1065-1076, 1962.

[11] T. Cacoullos, "Estimation of a multivariate density," in *Ann. Inst. Stat. Math.*, vol. 18, pp. 179-189, 1966.

[12] D. Specht, "Probabilistic neural networks," in *Neural Networks*, vol. 3(1), pp. 109-118, 1990.

[13] W. Edmonshon, W. Lee and J. Anderson, "Maximum likelihood estimation of sinusoidal parameters using a global optimization algorithm," in Ninth Asilomar Conference on Signals, Systems and Computers, October 1995, vol.2, pp. 1167 – 1171.

[14] P. Hänninen, A. Soini, N. Meltola, J. Soini, J. Soukka, E. Soini, "A new microvolume technique for bioaffinity assays using two-photon excitation," *Nat. Biot.*, vol. 18, pp. 548–550, 2000.

[15] J. Soini, J. Soukka, A. Soini, N. Meltola, E. Soini and P. Hänninen, "Ultra sensitive bioaffinity assay for micro volumes," *Single Molecules*, vol. 1, pp. 203-206, 2000.

[16] J. Soini, J. Soukka, E. Soini, P. Hänninen, "Two-photon excitation microfluorometer for multiplexed single-step bioaffinity assays," *Review of Scientific Instruments*, vol. 73, pp. 2680-2685, 2002.