Clinica Chimica Acta 455 (2016) 118–127

Contents lists available at ScienceDirect

Clinica Chimica Acta

journal homepage: www.elsevier.com/locate/clinchim

A new self-partition clustering method for robust identification of subsets with heterogeneous size and density and its clinical application to leukocyte differential counting

Shouichi Sato [a,b], Kiyoshi Ichihara [a,*], Tohru Inaba [c]

[a] Department of Clinical Laboratory Science, Faculty of Health Sciences, Yamaguchi University Graduate School of Medicine, Ube, Japan
[b] Chiba Emergency Medical Center, Chiba, Japan
[c] Department of Infection Control and Laboratory Medicine, Kyoto Prefectural University of Medicine, Kyoto, Japan

ARTICLE INFO

Article history:
Received 16 November 2015
Received in revised form 4 January 2016
Accepted 21 January 2016
Available online 26 January 2016

Keywords:
Clustering
Mahalanobis distance
Image processing
Iterative truncation correction method
Non-hierarchical cluster method
EM algorithm

ABSTRACT

Background: Identification of clusters in 2-dimensional scatterplots generated by hematology analyzer is a classical challenge. Conventional clustering algorithms fail to process cases with complicated mixtures of overlapping clusters and noise.
Method: A new method was developed that features an image processing algorithm for rational identification of initial clusters and a self-partition clustering (SPC) algorithm with iterative truncation-correction (ITC) method to handle overlapping and noise. All clusters are assumed to follow bivariate Gaussian distributions with specified means, SDs, and correlation coefficient. While, each data point is assumed to belong to all clusters but with different proportions according to the likelihood of belonging to each cluster (computed by the Mahalanobis distance) and the data size of the cluster. Bivariate cluster statistics are computed in consideration of a weight factor determined cluster by cluster by each data point. In the computation, the ITC method minimizes the effect of overlapping and data.
Results: Performance of SPC/ITC method was evaluated by its application to differential leukocyte counting. It showed comparable performance with manual counting and much better performance than the commonly used expectation maximum algorithm.
Conclusion: The SPC/ITC method showed superior performance in situations with overlapping and low-density clusters such as leukopenia or leukocytosis.

1. Introduction

Identification of clusters in the multidimensional space is a classical problem in many fields of scientific research. Various statistical algorithms have been developed, but 2 basic algorithms commonly used are the K-means [1,2] and expectation maximum (EM) algorithms [3].

K-means clustering first assumes the number of possible clusters, k, and assigns initial centers for the k clusters randomly. Then, each of all data points in the multi-dimensional space is partitioned into one of the k clusters based on the closest distance to the centers. This approach has well known problems of (1) dependence on how the initial cluster centers are estimated and (2) difficulty in properly separating out mutually overlapping clusters. The EM algorithm was developed to overcome these problems. It assumes that the dataset consists of k clusters each following a multi-dimensional Gaussian distribution. The initial clusters are estimated by a trial-and-error procedure, but once the appropriate clusters are found, the EM algorithm is not affected by the overlapping between clusters. However, the following problems remain: (1) a large imbalance in data size between overlapping clusters still affects the result, (2) the method is too sensitive to identifying a

Abbreviations: 2D, 2-dimensional; a, number of rows and columns in the tracer matrix; Bas, basophils; c, column sequence number (c = 1,2, …, p); CE, confidence ellipse; D[r, c], density matrix for 2D dataset (r, c = 1,2,…, p); EM, expectation maximum algorithm; Eos, eosinophils; F[u, v], filter matrix (kernel) to generate a RD matrix from D matrix; g, cluster sequence number (g = 1, 2,…,k); i, data sequence number (i = 1, 2,…, n); IP, image processing for identifying initial clusters; ITC, iterative truncation-correction method; L[g]i, likelihood of the i-th data belonging to the g-th cluster; Lym, lymphocytes; MD[g]i, Mahalanobis distance of the i-th data to the g-th cluster; Mon, monocytes; Mx[g], mean of variable x for the g-th cluster; My[g], mean of variable y for the g-th cluster; n, data size of all points in the scatterplot; N[g], data size of the g-th cluster; Neu, neutrophils; p, number of rows and columns in D or RD matrix; P[g]i, probability of the i-th data belonging to the g-th cluster; r, row sequence number, correlation coefficient; RD[r,c], relative density matrix for the 2D dataset; SDx[g], SD of variable x for the g-th cluster; SDy[g], SD of variable y for the g-th cluster; SPC, self-partition clustering; SPC/ITC, the new clustering method by use of SPC and ITC method; sumLL[g], sum of log-likelihood for the g-th cluster; u, row sequence number relative to the filter matrix; v, column sequence number relative to the filter matrix; w[g]i, weight of the i-th data point belonging to the g-th cluster; xi, variable x of the i-th data (i = 1,2,…,n); yi, variable y of the i-th data (i = 1,2,…,n); zx[g]i, variable x of the i-th data standardized for the g-th cluster; zy[g]i, variable y of the i-th data standardized for the g-th cluster.

[*] Corresponding author at: Department of Clinical Laboratory Science, Faculty of Health Sciences, Yamaguchi University Graduate School of Medicine, Minami-Kogushi 1-1-1, Ube 755-8505, Japan.
E-mail address: ichihara@yamaguchi-u.ac.jp (K. Ichihara).

http://dx.doi.org/10.1016/j.cca.2016.01.018
0009-8981/© 2016 Elsevier B.V. All rights reserved.

set of scattered or outlying points, with resultant appearance of a large spread-out cluster [4], and (3) its optimization step often takes a large amount of time [5].

In the field of laboratory medicine, the most familiar situation that requires such clustering techniques is the differential counting of leukocytes or white blood cells (WBCs) [6]. Although there might have been many attempts to use the classical approaches, no reports on such use exist. This appears to be due to the difficulty in ensuring successful classification with the existence of a vast variety of pathological patterns. For example, dense overlapping of clusters occurs when there is a large increase in WBCs in the presence of bacteremia or leukemia, whereas sparsely scattered clusters are seen when WBCs are reduced to a very low level, and bizarre uncommon clusters emerge in hematological malignancy. Therefore, the clustering algorithm has to cope with those situations by: (1) rational identification of the initial location of the clusters, (2) separation of densely overlapping clusters even with unbalanced data sizes, (3) identification of clusters with sparse, widely spread-out clusters, and (4) ignorance of noise/debris points. Manufacturers of hematology analyzers appear to have developed a variety of algorithms that use available information to handle the multitude of situations but which have not been disclosed openly.

We had an opportunity to develop a new clustering algorithm to the feasibility of upgrading the software currently in use in a hematology analyzer, the Horiba Pentra MS CRP [7], so that it can cope with the above situations, which are often processed as difficult cases with a warning from the analyzer requesting the special attention of a hematologist. This new clustering method consists of a series of algorithms to handle various situations. Although the new method can be generalized into multi-dimensional cases, in this study, we limited its application to the 2-dimensional (2D) case only.

First, to ensure rational estimation of initial clusters, we devised an image processing (IP) algorithm that features mapping of the relative density of the scatterplot and empirical scanning of density peaks to ensure identification of clusters, even those with low density.

Second, to solve the problem of overlapping of clusters especially of unbalanced data size, a new self-partitioning clustering (SPC) algorithm was developed. In the conventional clustering algorithm, the membership of data points is determined on the side of clusters according to their closeness. Conversely, in the SPC algorithm, each data point calculates its proportions of membership to each of k clusters based on its closeness (likelihood) to cluster g (g = 1,2, …, k) L[g], which is determined by a measure of distance (such as the Mahalanobis distance) and the data size N[g] of the cluster. By use of an iterative process, bivariate statistics (means, SD, and correlation coefficient) of each cluster are updated based on the proportional weight w[g] provided by each data point to cluster g, $\mathrm{w[g]} = (\mathrm{L[g]} \times \mathrm{N[g]})/(\sum_{j=1}^{k} \mathrm{L[j]} \times \mathrm{N[j]})$.

Third, an iterative truncation-correction (ITC) algorithm [8] was applied to minimize the influence of adjacent clusters or noise in the surroundings. It features truncation of data points outside a confidence ellipse at a certain level (e.g. 80%) of each cluster, followed by correction of the shortened SDs by multiplication with a correction factor (1.29 for truncation at 80%). The performance of the ITC algorithm is well documented for the 1-parameter case [9], but we extended the algorithm to the bivariate case in this study.

The performance of the new clustering method was evaluated in comparison with the conventional EM algorithm by randomly generating clusters with variable density and dispersion. Its clinical utility was then evaluated by applying the algorithm to the dataset obtained from the hematology analyzer for differential counting of leukocytes.

2. Methods

2.1. Algorithm for robust non-hierarchical clustering

Assuming the scatterplot consists of n data points designated as (x_i, y_i) [i = 1, 2, …, n], the 2D scatterplot is partitioned into p segments (p = 30 used as a default) horizontally and vertically to form a p × p density matrix, and the number of points (density) in each cell is recorded as D[r, c] (r, c = 1,2,…, p), where r and c respectively represent a location of row and column within the matrix.

2.1.1. Estimation of initial clusters based on the image processing algorithm

2.1.1.1. Computation of a relative density matrix. To identify a cluster with small density, we first applied the following 7 × 7 filter matrix F (a filter kernel) to the density matrix D. As shown in Fig. 1, the filter matrix was designed to map relative local density using the density matrix. The filter matrix is moved from left to right, and from top to bottom to scan relative density matrix RD[r, c] (r, c = 1,2,…, p).

$$\mathrm{RD[r,c]} = \frac{\sum_{u=-3}^{3}\sum_{v=-3}^{3} \mathrm{D[r-u,c-v]} * \mathrm{F[r+3,c+3]}}{\sum_{u=-3}^{3}\sum_{v=-3}^{3} \mathrm{D[r-u,c-v]}},$$

where u and v point to a location in row and column relative to the filter matrix.

2.1.1.2. An image scan algorithm for appropriate identification of initial clusters. To identify the initial set of clusters in the relative density matrix RD, a small tracer matrix consisting of a × a cells (a = 4 or 5) is applied to the RD to scan for the location of the highest relative density. Once found, the tracer matrix is expanded 1 by 1 in the vertical direction either upward or downward until the relative density within the tracer matrix does not decrease to less than 80% of the density for the original a × a area. Then, the tracer matrix is expanded 1 by 1 to the horizontal direction either leftward or rightward until the relative density does not decrease to the same level. After completion of expansion in both directions, then the second largest spot of density is scanned using the original a × a tracer matrix. In the search, the area already occupied by the cluster identified in the previous scan is not allowed to be included. This process of detecting the cluster with maximum density will be continued for the area which remained in the previous scan. The minimum density to be identified as detecting an independent cluster was set to 3% of the maximum value of the density matrix.

2.1.2. The self-partition clustering algorithm

Based on the initial clusters identified from the above image processing step, the statistics (center and spread) for each cluster are computed. For that purpose, in this study, we assumed that data points in each cluster followed a bivariate Gaussian distribution and thus the distance of a given data point to the cluster is expressed as the Mahalanobis distance [10].

The initial estimate of bivariate statistics (means, SDs, and correlation coefficient) for each cluster (designated as Mx, My, SDx, SDy, and r, respectively) is computed from the data limited to those enclosed by the scan matrix.

By expressing the number of data points belonging to cluster g as N[g] [g = 1,2,…, k], the probability $\mathrm{P[g]_i}$ of each data point (x_i, y_i) [i = 1, 2,., n] belonging to the cluster g is then computed based on Mahalanobis distance $\mathrm{D[g]_i}$ to the cluster g as follows:

$$\mathrm{MD[g]_i} = \frac{\mathrm{zx[g]_i^2} + \mathrm{zy[g]_i^2} - 2 \times \mathrm{r[g]} \times \mathrm{zx[g]_i} \times \mathrm{zy[g]_i}}{1 - \mathrm{r[g]}^2}$$

where $\mathrm{zx[g]_i} = \frac{x_i - \mathrm{Mx[g]}}{\mathrm{SDx[g]}}$, $\mathrm{zy[g]_i} = \frac{y_i - \mathrm{My[g]}}{\mathrm{SDy[g]}}$; $\mathrm{MD[g]_i}$ corresponds to the χ^2 statistic of 2 degrees of freedom, and thus $\mathrm{P[g]_i}$ can be obtained as the probability of the upper tail of the χ^2 distribution. Therefore, the relative weight of the i^{th} point belonging to the g^{th} cluster $\mathrm{w[g]_i}$ is computed as

$$\mathrm{w[g]_i} = \frac{\mathrm{P[g]_i} \times \mathrm{N[g]}}{\sum_{j=1}^{k} \mathrm{P}[j]_i \times \mathrm{N}[j]}.$$