Dirk Habich, Peter B. Volk, Wolfgang Lehner, Ralf Dittmann, Clemens Utzny

Error-Aware Density-Based Clustering of Imprecise Measurement Values

Erstveröffentlichung in / First published in:

Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). Omaha, 28.-31.10.2007. IEEE, 2008, S. 471-477. ISBN 978-0-7695-3019-2

DOI: https://doi.org/10.1109/ICDMW.2007.88

Final edited form was published in "Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). Omaha, 2007". S. 417-477. ISBN 978-0-7695-3019-2.
https://doi.org/10.1109/ICDMW.2007.88

Error-Aware Density-Based Clustering of Imprecise Measurement Values

Dirk Habich, Peter B. Volk
Wolfgang Lehner
Dresden University of Technology
Database Technology Group
dbinfo@mail.inf.tu-dresden.de

Ralf Dittmann, Clemens Utzny
Advanced Mask Technology Center
Dresden, Germany
{ralf.dittmann, clemens.utzny}@amtc-dresden.com

Abstract

Manufacturing process development is under constant pressure to achieve a good yield for stable processes. The development of new technologies, especially in the field of photomask and semiconductor development, is at its physical limits. In this area, data, e.g. sensor data, has to be collected and analyzed for each process in order to ensure process quality. With increasing complexity of manufacturing processes, the volume of data that has to be evaluated rises accordingly. The complexity and data volume exceeds the possibility of a manual data analysis. At this point, data mining techniques become interesting. The application of current techniques is complex because most of the data is captured with sensor measurement tools. Therefore, every measured value contains a specific error. In this paper we propose an error-aware extension of the density-based algorithm DBSCAN. Furthermore, we present some quality measures which could be utilized for further interpretation of the determined clustering results. With this new cluster algorithm, we can ensure that masks are classified into the correct cluster with respect to the measurement errors, thus ensuring a more likely correlation between the masks.

1. Introduction

Current technology development is driving measurement technology to the border of physics and engineering. Especially in the field of semiconductors and its photolithographic mask production, the race for the smallest structures drives sensor technology to new dimensions. Photolithographic masks are used to imprint the structures on the wafer for chip production. Chips consist of multiple layers, and for each layer, a unique mask is needed. Ideally, masks are produced exactly once for every chip design. Wafer manufacturers are eager to receive a mask as perfect as possible, since every variance in the structures on the mask is reflected on every chip produced with the mask. Therefore, wafer manufacturers define the tolerance values of the target mask structures for the mask very tight. Current technology specifies a tolerance value of no more than 2.5% of the actual structure size. With tools being able to measure with an accuracy of 0.25% of the actual structure size, the difference between specified structure size and measured structure size may suffer from an uncertainty of up to 10% of the tolerance value. The difference of specified and measured size is a major attribute for the quality of a mask. Mask development always tries to optimize the processes against these process quality measures.

Almost every process step for mask manufacturing is influenced by the structures on the mask. Cleaning fluids, for example react differently for masks with only a small number of structures. For processes to settle, chip manufacturers can produce multiple wafers and tune process parameters with every new wafer until the correct setting has been found for this product. Ideally, a mask shop produces exactly one mask per order that can be shipped to the customer. Therefore, rules for the determination of manufacturing parameters must be built on the basis of historical masks. To be able to create rules for process parameters, masks with similar behavior must be found and grouped.

In general, clustering is defined as the problem of partitioning multiple data objects into groups. Objects in the same group have strong similarities with each other, while objects in different clusters bear weaker or no similaritiess at all. This definition requires a well-defined distance measure between data objects that captures intra-cluster similarity. Then, clustering becomes the problem of grouping data objects. With the help of such clustering algorithms, the necessary mask equivalence groups can be found. The selection of attributes used for clustering is one of the key factors for success of the clustering techniques. The number of dimensions for clustering used for mask production can fluctuate from only a few to several hundred, depending on the specific use case. Most of these values are measurements from sensors containing a specific uncertainty.

Due to the subsequent use of the clustering results in

Final edited form was published in "Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). Omaha, 2007". S. 417-477. ISBN 978-0-7695-3019-2.
https://doi.org/10.1109/ICDMW.2007.88

important processing steps, the uncertainty included in the data must be taken into consideration when applying clustering algorithms. For this reason, this paper proposes a seamless error-aware extension of the density-based clustering technique DBSCAN[3]. Fundamentally, an important aspect within this area is the definition of a distance measure between two uncertain data objects. This definition is usually of complex nature and in general, one single distance measure is difficult to determine. Therefore, our approach creates the opportunity to adjust the used similarity for the clustering in a well-defined way. The advantage of this approach is that each application is able to specify a desired similarity measure based on uncertainy. Aside from presenting this seamless extension, we propose further means of quality measures for clusters. These quality measures enable a more detailed insight into uncertain data and the computed clustering results. Based on one quality measure, we propose an extended clustering algorithm including a novel similarity measure.

The remainder of the paper is structured as follows: In the following section, we give a more detailed specification of the problem. In Section 3, we present our basic error-aware extension of the density-based clustering algorithm DBSCAN. In Section 4, we propose new quality measures and present an approach on how they could be efficiently used for further data analysis. The paper closes with some preliminary evaluation results in Section 5, an overview of related work in Section 6 and with a conclusion in Section 7.

2. Problem Specification

Our specific application scenario is characterized by the following aspect: Each data object O is represented by a number of n different captured sensor values S_i ($1 \leq i \leq n$). As a result, a data object is described not only by one single feature vector as an n-dimensional data point but by an n-dimensional region in which all points within this region equally likely represent the object (Figure 1). In the ongoing description, we denote the n-dimensional region as data region. More formally, a data object O is described by the following vector:

$$O = \{S_1, \ldots, S_n, E_1, \ldots, E_n\}, \quad (1)$$

where n repesents the number of captured sensor values. $S_1, \ldots, S_n$ are the measured sensor values and each E_i ($1 \leq i \leq n$) represents a complex function as error quantification for each sensor value. These error functions depend on several factors. Moreover, these error functions may be independent of or dependent on several measured sensor values. Figure 1(b) shows data objects where the error functions are independent from each other, and therefore, the shapes of the data regions are hypercubes. However,

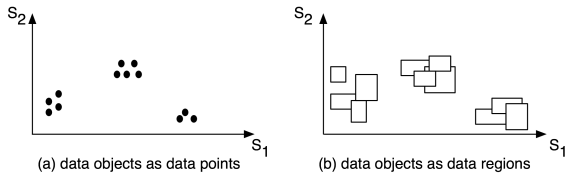


Figure 1. Representation of Data Objects

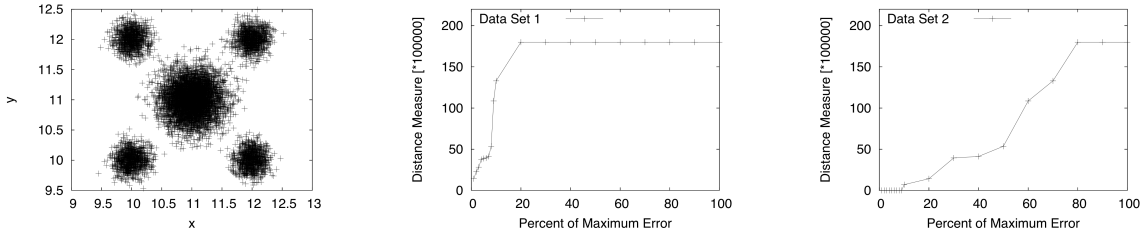
shapes of the data regions may be arbitrary. Within the data region, we assume that each possible data point represents the data object with the same probability.

The starting point is now a data set $D = \{O_1, \ldots, O_m\}$ containing a number of m different data objects O_j. The goal of the clustering task is to determine groups of similar data objects represented by data regions. However, all well-known distance functions, like the Euclidean or maximum distance functions, require n-dimensional data points. These distance functions idealize the data and do not take the uncertainty into account. To analyze the impact of imprecise measurement values on the clustering result with a traditional approach, we conducted a number of experiments. In these experiments, we observed the following effects: *clusters may split*, *clusters may merge*, *data points may jump from one cluster to another*, and *data points may fall out of or fall in to a cluster*.

For the conducted experiments, we generated a synthetic two-dimensional data set containing 5 clusters with a Gaussian distribution around the centroids and $100,000$ data points in total. As illustrated in Figure 2(a), four smaller clusters are arranged as satellites around a center cluster. Furthermore, we defined an uncertainty for each dimension (± 1 cm as maximum error). From this original data set, a number of different data sets are derived by changing the position of the data point within the data region specified by the uncertainy. To investigate the uncertainy in detail, each derived data set includes only points that are moved by a certain percentage smaller than 100 percent of the allowed maximum. Then, each data set was clustered with DBSCAN, whereas the values of $MinPts$ and ϵ are determined using the proposed heuristic [3].

In the first experiment series, the four satellite clusters are very close to the center cluster in the original data set, so that the resulting data regions acoording to the specified error range functions highly overlap. As we can see in Figure 2(b), the higher the percentage of the allowed maximum by which the points are randomly moved, the more the clustering results are differ from the clustering result of the original data sets. The differences between two results are determined by measuring the disagreement between two clusterings C and C' as a pair of data points (v, u), such that C places them in the same cluster, while C' places them in a

Final edited form was published in "Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). Omaha, 2007". S. 417-477. ISBN 978-0-7695-3019-2.
https://doi.org/10.1109/ICDMW.2007.88



(a) Data Set (b) Satellites dense to Center Cluster (c) Satellites more remote from Center Cluster

Figure 2. Evaluation of Impact of Uncertainty

different cluster, or vice versa (according to [4]). The graph runs asymptotically against a fixed number because some pairs of data points do not change their behavior (always grouped together or never).

In the second experiment series, the four satellite clusters are more remote from the center cluster. As we can see in Figure 2(c), smaller uncertainty has less significant impact on the clustering results than in the first experiment series. The impact increases with increasing uncertainty (hihger percentage). Again, the graph runs asymptotically against a fixed number.

To summarize, the presented experiment indicates that the consideration of error information in the clustering task could lead to very different clustering results with traditional clustering algorithms. In the next section, we present our error-aware density-based clustering algorithm.

3. Error-Aware Extension of DBSCAN

In this section, we present our error-aware extension of the density-based clustering approach DBSCAN [3]. As already outlined, our starting data set D consists of m data objects $\{O_1, \ldots, O_m\}$. Rather than describing each object O_i $(1 \leq i \leq m)$ by one single vector as an n-dimensional data point, we use an n-dimensional region (*data region*) in which all points within this region are equally likely to represent the object.

In order to be able to apply clustering algorithms to such data regions, a distance model is required. In case of data regions, we are always able to determine two extreme distance values between two data regions P and Q: a *minimum distance* $d_{min}(P, Q)$ and a *maximum distance* $d_{max}(P, Q)$. Figure 3 illustrates the two distance measures with an example. These two measures are only extremes with regard to two data objects. If the two data regions intersect each other, the minimum distance is equal to zero.

Fundamentally, the computation of the maximum distance and the minimum distance between two data objects described by n-dimensional data regions of arbitrary shape

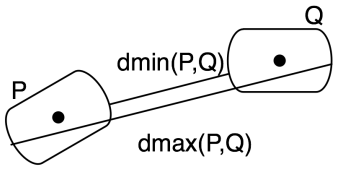


Figure 3. Min-Max Distances between P and Q (two-dimensional example)

is not trivial. In our approach, we assume that we can approximate these measures either with a minimal bounding rectangle (MBR) technique or via object sampling. Object sampling is a common method in the computer graphics field. A sample frequency defines the density for the surface of data regions. The knots of the resulting meshes on the objects are used to calculate the distances.

This distance model is now the foundation of our error-aware DBSCAN (DBSCAN^{EA}) clustering approach. Moreover, our DBSCAN^{EA} approach is a seamless extension of the original DBSCAN [3] approach to process uncertain data. In a first step, DBSCAN checks the ϵ-neighborhood of each data point in the database. If the ϵ-neighborhood $N_\epsilon(o)$ of a point o contains more than $MinPts$ elements, the data point o is a so-called *core data point*. In our approach, we define a (ϵ, λ)-neighborhood of a data region P as follows:

Definition 1 ((ϵ, λ)-neighborhood of a data region) The (ϵ, λ)-neighborhood of a data region P, denoted by $N_{(\epsilon,\lambda)}(P)$, is defined by

$$N_{(\epsilon,\lambda)} = \{Q \in D | dist(P, Q) \leq \epsilon\} \quad (2)$$

with

$$dist(P, Q) = (1 - \lambda) * d_{max}(P, Q) + \lambda * d_{min}(P, Q) \quad (3)$$

Final edited form was published in "Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). Omaha, 2007". S. 417-477. ISBN 978-0-7695-3019-2.
https://doi.org/10.1109/ICDMW.2007.88

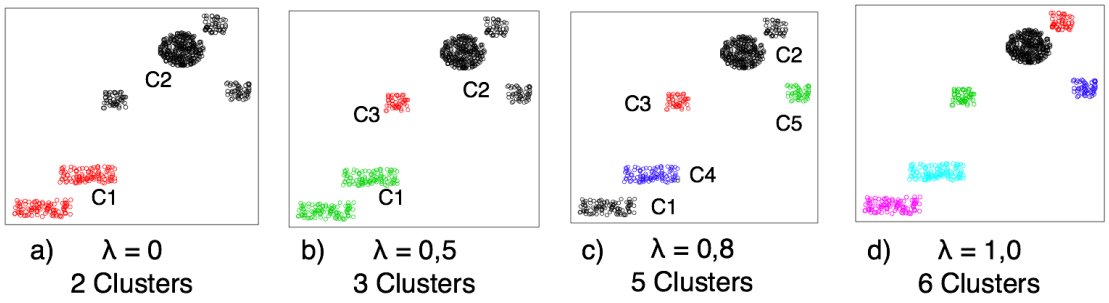


Figure 4. Clustering Results with Different Values for λ

This (ϵ, λ)-neighborhood definition is of complex nature and introduces a novel correlation parameter λ. Due to our distance model, we have two extreme distance measures $d_{max}(P,Q)$ and $d_{min}(P,Q)$ between two data regions P and Q. The first measure can be seen as a pessimistic distance, while the second measure is a very optimistic distance measure for the similarity of two considered data regions. In general, each value in the range of $[dist_{min}(P,Q), dist_{max}(P,Q)]$ is also a possible similarity measure. Therefore, our newly introduced parameter λ enables users to specify a correlation factor to be used in the determination of the similarity.

Definition 2 (correlation factor λ) The correlation factor, denoted by λ, is defined as $\{\lambda \in \Re | 0 < \lambda < 1\}$. In this way, a correlation factor of $\lambda = 1$ corresponds to a high (optimistic) correlation between two data regions, while a correlation factor of $\lambda = 0$ signifies a low (pessimistic) approach.

The idea behind this correlation factor is to give users a specialized factor they can easily use to adjust the analysis to their requirements. Furthermore, the complexity of DBSCAN^{EA} is similar to the orginal DBSCAN approach. The remaining definitions for the density-based clustering are as follows:

Definition 3 (directly density-reachable) A data region Q is *directly density-reachable* from a data region P with respect to ϵ, λ and $MinPts$ if

1. $Q \in N_{\epsilon,\lambda}(P)$, and
2. $|N_{\epsilon,\lambda}(P)| > MinPts$ (core data region condition).

Definition 4 (density-reachable) A data region Q is *density-reachable* from a data region P with respect to ϵ, λ and $MinPts$ if there is a chain of data regions $Q_1, \ldots, Q_k$ such that $Q_1 = P$ and $Q_k = Q$ and Q_{i+1} is directly density-reachable from Q_i for $1 \leq i \leq k$.

Definition 5 (density-connected) A data region Q is *density-connected* to a data region P with respect to ϵ, λ and $MinPts$ if there is a data region R such that both Q and P are density-reachable from R with respect to ϵ, λ and $MinPts$.

Definition 5 (density-based cluster region) Let D be a database of data regions. A cluster region CR in D with respect to ϵ, λ and $MinPts$ is a non-empty subset of D satisfying the following conditions:

1. $\forall P, Q \in D$: if $P \in CR$ and Q is density reachable from P with respect to ϵ, λ and $MinPts$, then $Q \in CR$
2. $\forall P, Q \in CR$: P is density-connected to Q with respect to ϵ, λ and $MinPts$.

Definition 6 (noise) Let $CR_1, \ldots, CR_k$ be the clusters of the data set D with respect to ϵ, $MinPts$, and λ. Then we define the noise as the set of data regions in the data set D not belonging to any cluster CR_i:

$$noise = \{P \in D | \forall i (1 \leq i \leq k) : P \notin CR_i\} \quad (4)$$

Figure 4 depicts example results of our clustering with four different values for the correlation factor λ. The error model was created as a hypercube (symmetric errors for all dimensions). With an increasing correlation factor (pessimistic to optimistic), the satellites around $C2$ (see Figure4a) in the investigated data set turn into their own clusters. Since the satellites represent major biasing effects during a measurement, the clusters make sense. On the other hand, a drift is recognised as a cluster $C1$ (see Figure 4(a,b)) with a correlation factor of $0 - 0.50$. Beyond this range, the drift splits into two clusters C_1 and C_4 (see Figure 4(c,d)). The parameters ϵ and $MinPts$ are determined according to the proposed heuristic in [3].

4. Quality Measures

Our presented *DBSCAN^{EA}* algorithm has three input parameters: ϵ, $MinPts$, and λ. That means the clustering re-

Final edited form was published in "Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). Omaha, 2007". S. 417-477. ISBN 978-0-7695-3019-2.
https://doi.org/10.1109/ICDMW.2007.88

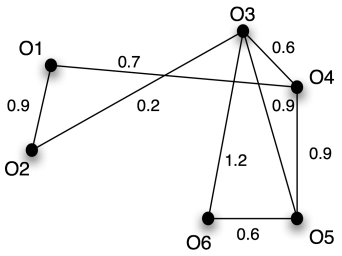


Figure 5. Similarity Graph

sult depends on those parameters. In this section, we want to explore the λ parameter in more detail.

Using the λ parameter, users are able to define the similarity correlation between data regions used in the clustering. Therefore, this parameter has a strong impact on the clustering result. On the other hand, the λ parameter allows to derive further means of quality measures of clusters: the *cluster stability* and the *similarity measure*.

Cluster Stability

In Section 2, we showed, that data points change their cluster association when we directly add error values to the points within the error range. As a coarse-grained effect, merging and splitting of clusters occur. The same effects are observable in the clustering of data regions using our $DBSCAN^{EA}$ approach with different values of λ.

The *cluster stability* (CS) is a quality measure for clusters determining the range of correlation factors in which the cluster is stable. To calculate CS, our clustering algorithm is applied multiple times on the data set. Every time the algorithm is applied, the correlation factor λ is increased by a constant factor $\triangle\lambda$ until the maximal correlation factor $\lambda = 1$ is reached. For the first application of the algorithm, the correlation factor is either set to an initial value specified by the user or to zero. After each iteration, the resulting cluster association is compared to the association of the previous run with a lower correlation factor. Two clusters are equal if the difference in the associated data regions is not larger than $\delta\lambda$ with respect to the total number of data regions in the larger cluster.

In general, the interpretation of the cluster stability measure specifies the degree of robustness of the extracted clusters with respect to the maximum error of the data.

Similarity Measure

Aside from cluster stability, we are able to determine a specific similarity measure between data regions. This quality measure determines how close two data regions are associated with each other in the clustering. To compute this

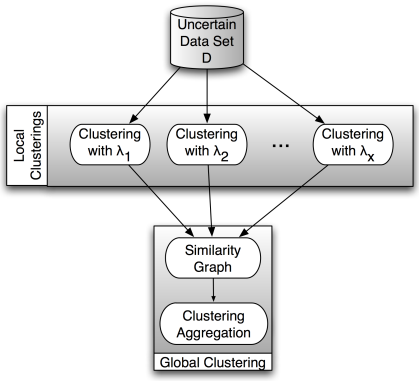


Figure 6. Extended Error-Aware Clustering Approach

specific similarity measure, our proposed $DBSCAN^{EA}$ method is applied several times. With every new application, the correlation factor λ is decreased by a constant factor $\triangle\lambda$ from the maximum value $\lambda = 1$ to the mimimum value $\lambda = 0$. The similarity measure between two data regions P and Q is computed as follows:

1. During each application of the clustering algorithm, we compute a local similarity measure. If both data regions, P and Q, are in the same cluster for the currently considered correlation factor λ, the local similarity LS is equal to $(\triangle\lambda)^{(1-\lambda)}$. The smaller the correlation factor λ (more optimistically), the higher is the local similarity. If the two data regions do not belong to the same cluster, the local similarity is set to zero.

2. The global similarity between P and Q is the sum of all local similarity values with respect to P and Q. The higher the sum of the local similarity values, the higher is the similarity of the two data regions. The advantage of this global similarity measure is that this measure is computed across several clustering results considering the uncertainty in different ways.

Now, we are able to assign each pair of data regions a global similarity measure. The higher the value, the more similar the data regions are. At this point, we can construct a graph as illustrated in Figure 5. Each vertex of the graph represents a data object (data region) of the data set. The weight of the edge (O_1, O_2) is the specific similarity between objects O_1 and O_2. Using algorithms from the correlation clustering [1] or clustering aggregation area [4, 5], a global clustering result based on this similarity graph can be computed.

The advantage of the overall extended error-aware approach (see Figure 6) is that each uncertain data set is in-

Final edited form was published in "Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). Omaha, 2007". S. 417-477. ISBN 978-0-7695-3019-2.
https://doi.org/10.1109/ICDMW.2007.88

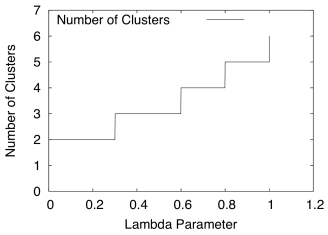


Figure 7. Influence of the λ Parameter on the Number of Clusters

vestigated with several local *DBSCAN*^{EA} algorithms considering the uncertainty in different ways. These local information are aggregated together to determine a global result. Our next research tasks focus on the refinement of this extended error-aware clustering process, including the following steps:

- investigating the parameterization of the local $DBSCAN^{EA}$ algorithms in the context of our whole approach,
- development and evaluation of several local similarity functions,
- evaluation of the influence of the value $\triangle\lambda$, and
- integration of arbitrary probalitity-density functions (further specification of data regions) in the determination process of $\triangle\lambda$

5. Evaluation

In this section, we present some evaluation results. Due to the current early phase of our overall *Extended Error-Aware Clustering Approach* (see Figure 6), we can only show some preliminary evaluation results regarding DBSCAN^{EA}. Figure 7 illustrates the influence of the λ parameter on the number of detected clusters within the data set. In this experiment, we used the illustrated data set of Figure 4(a). As depicted, the number of clusters increases if the value of the λ parameter increases. Due to our seamless extension to DBSCAN, our DBSCAN^{EA} shows the same efficiency as the orginal approach. In general, the performance depends on the approximation of the mimimum and maximum distance.

6. Related Work

An overview of state-of-the-art data clustering approaches is given by Jain et al. in [6]. Well-known algorithms are k-means [2] and DBSCAN [3]. K-means belongs to the class of partitioning algorithms, while DBSCAN is an example of a density-based clustering approach. However, most of the published data clustering techniques do not consider data objects with imprecise values. Kriegel and Pfeile presented a density-based [7] and a hierarchical density-based clustering approach [8] for uncertain data. They proposed a fuzzy distance function to measure the similarity between fuzzy objects. This function does not express the similarity by a single numerical value. Instead, the similarity is expressed by means of probability functions, which assign a numerical value to each distance value. Based on this fuzzy distance function, we proposed an extension of the DBSCAN algorithm. Ngai et al. [9] presented an extension of the k-means algorithm for uncertain data.

7. Conclusion

In this paper, we presented our basic error-aware extension to the density-based clustering algorithm *DBSCAN* to cluster uncertain data. Moreover, we proposed a novel quality measure to interpret the clustering results. Furthermore, we presented a method on how to use a resulting quality measure to determine a global clustering result. All these presented concepts are implemented and tested in the field of mask manufacturing. The first application on real data showed a significant improvement for the manufacturing yield in comparison to applications using the classical DBSCAN algorithm. Our next steps include (i) the exhaustive comparison with other algorithms, like [7, 9], and (ii) the refinement of our Extended Error-Aware Clustering algorithm.

References

[1] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. In *Proc. of FOCS*, 2002.
[2] E. Forgy. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifcations. *Biometrics*, 21:768, 1965.
[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of KDK*, 1996.
[4] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. In *Proc. of ICDM*, 2005.
[5] D. Habich, T. Wächter, W. Lehner, and C. Pilarsky. Two-phase clustering strategy for gene expression data sets. In *Proc. of SAC*, 2006.
[6] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3):264–323, 1999.
[7] H.-P. Kriegel and M. Pfeifle. Density-based clustering of uncertain data. In *Proc. of KDD*, New York, NY, USA, 2005. ACM Press.
[8] H.-P. Kriegel and M. Pfeifle. Hierarchical density-based clustering of uncertain data. In *Proc. of ICDM*, 2005.
[9] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip. Efficient clustering of uncertain data. In *Proc. of ICDM*, 2006.