Information Sciences Letters
An International Journal

http://dx.doi.org/10.12785/isl/040202

Comparison of Data Mining Methods on Different Applications: Clustering and Classification Methods

Richard Merrell[*] *and David Diaz*

Department of Computer Engineering, Yerevan University, Yerevan, Armenia.

Received: 12 Jan. 2015, Revised: 21 Apr. 2015, Accepted: 24 Apr. 2015
Published online: 1 May 2015

Abstract: Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. In this review we study different type if clustering methods.

Keywords: Clustering, Data Mining, Ensemble, Big data.

1 Introduction

According to Vladimir Estivill-Castro, the notion of a ”cluster” cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms [1,2,3,4]. There is a common denominator: a group of data objects. However, different researchers employ different cluster models, and for each of these cluster models again different algorithms can be given. The notion of a cluster, as found by different algorithms, varies significantly in its properties. Understanding these ”cluster models” is key to understanding the differences between the various algorithms. Typical cluster models include: Connectivity models: for example hierarchical clustering builds models based on distance connectivity. Centroid models: for example the k-means algorithm represents each cluster by a single mean vector. Distribution models: clusters are modeled using statistical distributions, such as multivariate normal distributions used by the Expectation-maximization algorithm. Density models: for example DBSCAN and OPTICS defines clusters as connected dense regions in the data space. Subspace models: in Biclustering (also known as Co-clustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes. Group models: some algorithms do not provide a refined model for their results and just provide the grouping information. Graph-based models: a clique, *i.e.*, a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Relaxations of the complete connectivity requirement (a fraction of the edges can be missing) are known as quasi-cliques. A ”clustering” is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each other, for example a hierarchy of clusters embedded in each other.

2 Clustering Methods

Clustering algorithms can be categorized based on their cluster model, as listed above. The following overview will only list the most prominent examples of clustering algorithms, as there are possibly over 100 published clustering algorithms. Not all provide models for their clusters and can thus not easily be categorized. An overview of algorithms explained in Wikipedia can be found in the list of statistics algorithms. There is no objectively ”correct” clustering algorithm, but as it was noted, ”clustering is in the eye of the beholder [4,5,6,7,8,9]. The most appropriate clustering algorithm for a particular problem often needs to be chosen experimentally, unless there is a mathematical reason to prefer one cluster model over another. It should be noted that an algorithm that is designed for one kind of model has no chance on a data set that contains a radically

[*] Corresponding author e-mail: R.Merrell@outlook.com

different kind of model. [4,5,7] For example, k-means cannot find non-convex clusters. [4,10,11,12,13].

2.1 Connectivity based clustering

Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect ”objects” to form ”clusters” based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name ”hierarchical clustering” comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don’t mix. These methods will not produce a unique partitioning of the data set, but a hierarchy from which the user still needs to choose appropriate clusters. They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge (known as ”chaining phenomenon”, in particular with single-linkage clustering). In the general case, the complexity is which makes them too slow for large data sets. For some special cases, optimal efficient methods (of complexity) are known: SLINK [5] for single-linkage and CLINK [6] for complete-linkage clustering. In the data mining community these methods are recognized as a theoretical foundation of cluster analysis, but often considered obsolete. They did however provide inspiration for many later methods such as density based clustering figure (1).

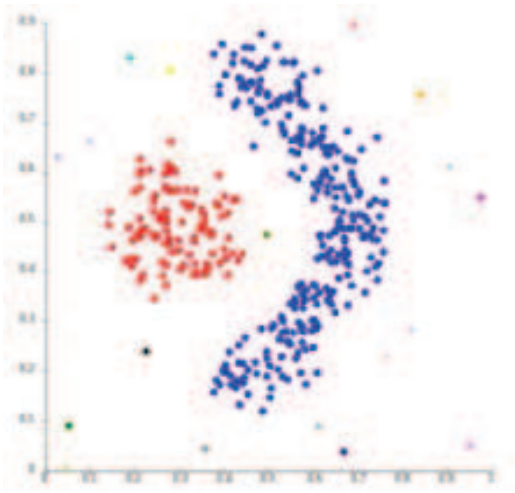

Fig. 1: Single-linkage on density-based clusters. 20 clusters extracted, most of which contain single elements, since linkage clustering does not have a notion of ”noise”. Many people used this method on the different application [15,16,17,18,19,20,21,22,23,24,25]

2.2 Centroid-based clustering

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized [29,30,31,32,33]. The optimization problem itself is known to be NP-hard, and thus the common approach is to search only for approximate solutions. A particularly well known approximative method is Lloyd’s algorithm [7,24,25,26,27,28] often actually referred to as ”k-means algorithm”. It does however only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data set (k-medoids), choosing medians (k-medians clustering), choosing the initial centers less randomly (K-means++) or allowing a fuzzy cluster assignment (Fuzzy c-mean. Most k-means-type algorithms require the number of clusters - k - to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. Furthermore, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. This often leads to incorrectly cut borders in between of clusters (which is not surprising, as the algorithm optimized cluster centers, not cluster borders. K-means has a number of interesting theoretical properties. On the one hand, it partitions the data space into a structure known as a Voronoi diagram. On the other hand, it is conceptually close to nearest neighbor classification, and as such is popular in machine learning. Third, it can be seen as a variation of model based classification, and Lloyd’s algorithm as a variation of the Expectation-maximization algorithm for this model discussed below figure(2).

2.3 Distribution-based clustering

The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. A convenient property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution. While the theoretical foundation of these methods is excellent, they suffer from one key problem

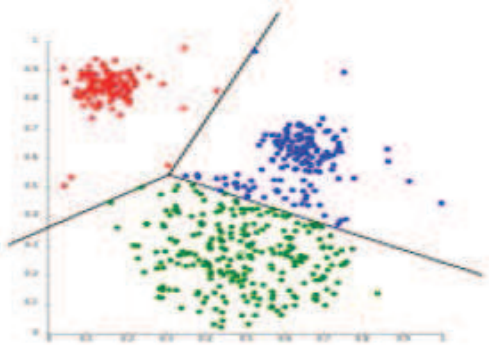

Fig. 2: Means separates data into Voronoi-cells, which assumes equal-sized clusters.

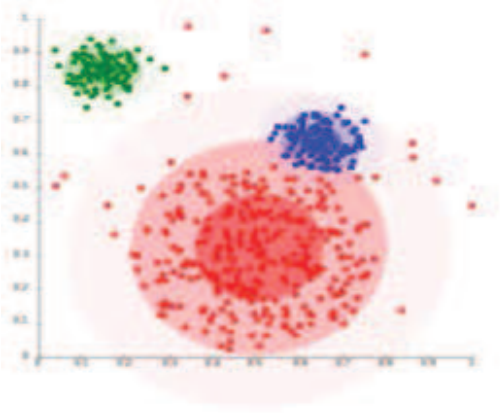

Fig. 3: On Gaussian-distributed data, EM works well, since it uses Gaussians for modelling clusters.

known as overfitting, unless constraints are put on the model complexity. A more complex model will usually be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult. One prominent method is known as Gaussian mixture models (using the expectation-maximization algorithm). Here, the data set is usually modelled with a fixed (to avoid overfitting) number of Gaussian distributions that are initialized randomly and whose parameters are iteratively optimized to fit better to the data set. This will converge to a local optimum, so multiple runs may produce different results. In order to obtain a hard clustering, objects are often then assigned to the Gaussian distribution they most likely belong to; for soft clusterings, this is not necessary. Distribution-based clustering produces complex models for clusters that can capture correlation and dependence between attributes. However, these algorithms put an extra burden on the user: for many real data sets, there may be no concisely defined mathematical model (figure 3). Many people used this method on the different application [34,25,36,37,38,39,40].

2.4 Density-based clustering

In density-based clustering [8,41,42,43] clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. The most popular [9,44,45,46, 47,48,49] density based clustering method is DBSCAN [10]. In contrast to many newer methods, it features a well-defined cluster model called ”density-reachability”. Similar to linkage based clustering, it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects’ range. Another interesting property of DBSCAN is that its complexity is fairly low - it requires a linear number of range queries on the database - and that it will discover essentially the same results (it is deterministic for core and noise points, but not for border points) in each run, therefore there is no need to run it multiple times. OPTICS [11,50,51] is a generalization of DBSCAN that removes the need to choose an appropriate value for the range parameter ε, and produces a hierarchical result related to that of linkage clustering. DeLi-Clu [12,52,53]. Density-Link-Clustering combines ideas from single-linkage clustering and OPTICS, eliminating the varepsilon parameter entirely and offering performance improvements over OPTICS by using an R-tree index. The key drawback of DBSCAN and OPTICS is that they expect some kind of density drop to detect cluster borders. Moreover, they cannot detect intrinsic cluster structures which are prevalent in the majority of real life data. A variation of DBSCAN, EnDBSCAN [13,54,55] efficiently detects such kinds of structures. On data sets with, for example, overlapping Gaussian distributions - a common use case in artificial data - the cluster borders produced by these algorithms will often look arbitrary, because the cluster density decreases continuously. On a data set consisting of mixtures of Gaussians, these algorithms are nearly always outperformed by methods such as EM clustering that are able to precisely model this kind of data. Mean-shift is a clustering approach where each object is moved to the densest area in its vicinity, based on kernel density estimation. Eventually, objects converge to local maxima of density. Similar to k-means clustering, these ”density attractors” can serve as representatives for the data set, but mean-shift can detect arbitrary-shaped clusters similar to DBSCAN. Due to the expensive iterative procedure and density estimation, mean-shift is usually slower than DBSCAN or k-Means (figure 4).

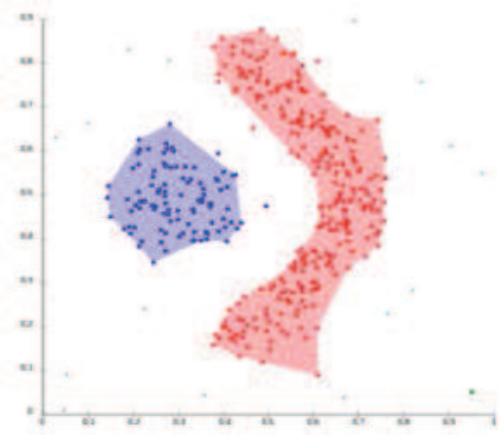

Fig. 4: Density-based clustering with DBSCAN.

3 Recent developments

In recent years considerable effort has been put into improving algorithm performance of the existing algorithms [14,15] Among them are CLARANS (Ng and Han, 1994) [16] and BIRCH (Zhang et al., 1996) [17] With the recent need to process larger and larger data sets (also known as big data), the willingness to trade semantic meaning of the generated clusters for performance has been increasing. This led to the development of pre-clustering methods such as canopy clustering, which can process huge data sets efficiently, but the resulting ”clusters” are merely a rough pre-partitioning of the data set to then analyze the partitions with existing slower methods such as k-means clustering. Various other approaches to clustering have been tried such as seed based clustering [18,56]. For high-dimensional data, many of the existing methods fail due to the curse of dimensionality, which renders particular distance functions problematic in high-dimensional spaces. This led to new clustering algorithms for high-dimensional data that focus on subspace clustering (where only some attributes are used, and cluster models include the relevant attributes for the cluster) and correlation clustering that also looks for arbitrary rotated (”correlated”) subspace clusters that can be modeled by giving a correlation of their attributes. Examples for such clustering algorithms are CLIQUE [19] and SUBCLU [20]. Ideas from density-based clustering methods (in particular the DBSCAN/OPTICS family of algorithms) have been adopted to subspace clustering HiSC [21] hierarchical subspace clustering and DiSH [22] and correlation clustering HiCO [23] hierarchical correlation clustering, 4C [24] using ”correlation connectivity” and ERiC [25] exploring hierarchical density-based correlation clusters. Several different clustering systems based on mutual information have been proposed. One is Marina Meil?’s variation of information metric [26]; another provides hierarchical clustering [27]. Using genetic algorithms, a wide range of different fit-functions can be optimized, including mutual information [28]. Also message passing algorithms, a recent development in Computer Science and Statistical Physics, has led to the creation of new types of clustering algorithms [29].

4 Conclusion

Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, botryology and typological analysis. The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goals. Here we studied different type of clustering methods.

References

[1] Ken Bailey, ”Numerical Taxonomy and Cluster Analysis”. Typologies and Taxonomies. p. 34. ISBN 9780803952591, (1994).

[2] Robert C. Tryon, Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers, (1939).

[3] R. B. Cattell, ”The description of personality: Basic traits resolved into clusters”. Journal of Abnormal and Social Psychology 38: 476?506. doi:10.1037/h0054116, (1943).

[4] Estivill-Castro, Vladimir (20 June 2002). ”Why so many clustering algorithms, A Position Paper”. ACM SIGKDD Explorations Newsletter **4 (1)**, 65, 75. doi:10.1145/568574.568575.

[5] R. Sibson, ”SLINK: an optimally efficient algorithm for the single-link cluster method”. *The Computer Journal (British Computer Society)* **16 (1)**, 30:34. doi:10.1093/comjnl/16.1.30, (1973).

[6] D. Defays, ”An efficient algorithm for a complete link method”. The Computer Journal (British Computer Society) **20 (4)**, 364?366. doi:10.1093/comjnl/20.4.364, (1977).

[7] H. Parvin, H. Alinejad-Rokny, S. Parvin, Divide and Conquer Classification, *Australian Journal of Basic & Applied Sciences*, **5(12)**: 2446-2452 (2011).

[8] H. Parvin, B. Minaei-Bidgoli, H. Alinejad-Rokny, A New Imbalanced Learning and Dictions Tree Method for Breast Cancer Diagnosis, *Journal of Bionanoscience*, **7(6)**, 673-678 (2013).

[9] H. Parvin, H. Alinejad-Rokny, M. Asadi, An Ensemble Based Approach for Feature Selection, *Journal of Applied Sciences Research*, **7(9)**: 33-43 (2011).

[10] N. Seyedaghaee, S. Rahati, H. Alinejad-Rokny, F. Rouhi, An Optimized Model for the University Strategic Planning, International Journal of Basic Sciences and Applied Research, **2(5)**: 500-505 (2013).

[11] S. Lloyd, ”Least squares quantization in PCM”. IEEE Transactions on Information Theory 28 (2): 129-137. doi:10.1109/TIT.1982.1056489, (1982).
[12] Hans-Peter Kriegel, Peer Krger, Jrg Sander, Arthur Zimek, ”Density-based Clustering”. WIREs Data Mining and Knowledge Discovery **1 (3)**: 231?240. doi:10.1002/widm.30, (2011).
[13] Microsoft academic search: most cited data mining articles: DBSCAN is on rank 24, when accessed on: 4/18/2010
[14] Martin Ester, Hans-Peter Kriegel, Jrg Sander, Xiaowei Xu, Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. ”Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)”. AAAI Press. pp. 226-231. ISBN 1-57735-004-9. CiteSeerX: 10.1.1.71.1980, (1996).
[15] Mihael Ankerst, Markus M. Breunig, Kriegel, Hans-Peter; Jrg Sander, ”ACM SIGMOD international conference on Management of data”. ACM Press. pp. 49-60. CiteSeerX: 10.1.1.129.6542, (1999).
[16] H. Parvin, H. Helmi, B. Minaie-Bidgoli, H. Alinejad-Rokny, H. Shirgahi, Linkage learning based on differences in local optimums of building blocks with one optima, *International Journal of Physical Sciences*, **6(14)**: 3419-3425 (2011).
[17] H. Motameni, H. Alinejad-Rokny, M.M. Pedram, Using Sequential Pattern Mining in Discovery DNA Sequences Contain Gap, *American Journal of Scientific Research*, **14**: 72-78 (2011).
[18] H. Parvin, B. Minaei-Bidgoli, H. Alinejad-Rokny, S. Ghatei, An innovative combination of particle swarm optimization, learning automaton and great deluge algorithms for dynamic environments, *International Journal of Physical Sciences*, **6(22)**: 5121-5127 (2011).
[19] H. Parvin, H. Alinejad-Rokny, S. Parvin, A Classifier Ensemble of Binary Classifier Ensembles, *International Journal of Learning Management Systems*, **1(2)**: 37-47 (2013).
[20] E. Achtert, C. Bhm, P. Krger ”DeLi-Clu: Boosting Robustness, Completeness, Usability, and Efficiency of Hierarchical Clustering by a Closest Pair Ranking”. LNCS: Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science 3918: 119?128. doi:10.1007/11731139-16. ISBN 978-3-540-33206-0, (2006)
[21] S. Roy, D. K. Bhattacharyya, ”LNCS Vol.3816”. Springer Verlag. pp. 523-535, (2005).
[22] D. Sculley, ”Web-scale k-means clustering”. Proc. 19th WWW, (2010).
[23] Z. Huang, ”Extensions to the k-means algorithm for clustering large data sets with categorical values”. *Data Mining and Knowledge Discovery* **2**: 283?304, (1998).
[24] R. Ng and J. Han. ”Efficient and effective clustering method for spatial data mining”. In: Proceedings of the 20th VLDB Conference, pages 144-155, Santiago, Chile, (1994).
[25] Tian Zhang, Raghu Ramakrishnan, Miron Livny. ”An Efficient Data Clustering Method for Very Large Databases.” *In: Proc. Int’l Conf. on Management of Data, ACM SIGMOD*, pp. 103-114.
[26] H. Alinejad Rokny, M.M. Pedram, H. Shirgahi, Discovered motifs with using parallel Mprefixspan method, *Scientific Research and Essays*, **6(20)**: 4220-4226 (2011).
[27] H. Parvin, B. Minaei-Bidgoli, H. Alinejad-Rokny, W.F. Punch, Data weighing mechanisms for clustering ensembles, *Computers & Electrical Engineering*, **39(5)**: 1433-1450 (2013).
[28] N. Eftekhari, M. Poyan, H. Alinejad-Rokny, Evaluation and Classifying Software Architecture Styles Due to Quality Attributes, *Australian Journal of Basic and Applied Sciences*, **5(11)**: 1251-1256 (2011).
[29] H. Parvin, H. Alinejad-Rokny, B. Minaei-Bidgoli, S. Parvin, A new classifier ensemble methodology based on subspace learning, *Journal of Experimental & Theoretical Artificial Intelligence*, **25(2)**: 227-250 (2013).
[30] F. Can, E. A. Ozkarahan, ”Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases”. *ACM Transactions on Database Systems* **15 (4)**: 483. doi:10.1145/99935.99938, (1990).
[31] R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan, ”Automatic Subspace Clustering of High Dimensional Data”. *Data Mining and Knowledge Discovery* **11**: 5. doi:10.1007/s10618-005-1396-1, (2005).
[32] Karin Kailing, Hans-Peter Kriegel and Peer Krger. Density-Connected Subspace Clustering for High-Dimensional Data. In: Proc. SIAM Int. Conf. on Data Mining (SDM’04), pp. 246-257, 2004.
[33] E. Achtert, C. Bhm, H. P. Kriegel, P. Krger, I. Mller-Gorman, A. Zimek, ”Finding Hierarchies of Subspace Clusters”. LNCS: Knowledge Discovery in Databases: PKDD 2006. Lecture Notes in Computer Science 4213: 446?453. doi:10.1007/11871637-42. ISBN 978-3-540-45374-1, (2006).
[34] E. Achtert, C. Bhm, H. P. Kriegel, P. Krger, I. Mller-Gorman, A. Zimek, ”Detection and Visualization of Subspace Cluster Hierarchies”. LNCS: Advances in Databases: Concepts, Systems and Applications. Lecture Notes in Computer Science 4443: 152?163. doi:10.1007/978-3-540-71703-4-15. ISBN 978-3-540-71702-7, (2007).
[35] E. Achtert, C. Bhm, P. Krger, A. Zimek, ”Mining Hierarchies of Correlation Clusters”. Proc. 18th International Conference on Scientific and Statistical Database Management (SSDBM): 119-128, (2006).
[36] R. Javanmard, K. Jeddisaravi, H. Alinejad-Rokny, Proposed a New Method for Rules Extraction Using Artificial Neural Network and Artificial Immune System in Cancer Diagnosis, *Journal of Bionanoscience*, **7(6)**: 665-672 (2013).
[37] E. Hasanzadeh, M. Poyan and H. Alinejad-Rokny, Text Clustering on Latent Semantic Indexing with Particle Swarm Optimization (PSO) Algorithm, *International Journal of the Physical Sciences*, **7(1)**, 116-120 (2012).
[38] H. Parvin, H. Alinejad-Rokny, N. Seyedaghaee, S. Parvin, A Heuristic Scalable Classifier Ensemble of Binary Classifier Ensembles, *Journal of Bioinformatics and Intelligent Control*, **1(2)**: 163-170 (2013).
[39] M.H. Fouladgar, B. Minaei-Bidgoli, H. Parvin, H. Alinejad-Rokny, Extension in The Case of Arrays in Daikon like Tools, *Advanced Engineering Technology and Application*, **2(1)**: 5-10 (2013).
[40] H. Parvin, M. MirnabiBaboli, H. Alinejad-Rokny, Proposing a Classifier Ensemble Framework Based on Classifier Selection and Decision Tree, *Engineering Applications of Artificial Intelligence*, **37**: 34-42 (2015).
[41] C. Bhm, K. Kailing, P. Krger, A. Zimek, ”Computing Clusters of Correlation Connected objects”. Proceedings of the 2004 ACM SIGMOD international conference

on Management of data - SIGMOD '04. p. 455. doi:10.1145/1007568.1007620. ISBN 1581138598, (2004).
[42] E. Achtert, C. Bohm, H. P. Kriegel, P. Krger, A. Zimek, "On Exploring Complex Relationships of Correlation Clusters". 19th International Conference on Scientific and Statistical Database Management (SSDBM 2007). p. 7. doi:10.1109/SSDBM.2007.21. ISBN 0-7695-2868-6, (2007).
[43] Meil, Marina. "Comparing Clusterings by the Variation of Information". Learning Theory and Kernel Machines. Lecture Notes in Computer Science 2777: 173?187. doi:10.1007/978-3-540-45167-9-14. ISBN 978-3-540-40720-1, (2003).
[44] Alexander Kraskov, Harald Stgbauer, Ralph G. Andrzejak, Peter Grassberger, (1 December 2003) [28 November 2003]. "Hierarchical Clustering Based on Mutual Information". arXiv:q-bio/0311039.
[45] B. Auffarth, (July 18-23, 2010). "Clustering by a Genetic Algorithm with Biased Mutation Operator". WCCI CEC (IEEE). CiteSeerX: 10.1.1.170.869.
[46] B. J. Frey, D. Dueck, "Clustering by Passing Messages Between Data Points". Science 315 (5814): 972?976. doi:10.1126/science.1136800. PMID 17218491, (2007).
[47] D. Christopher, Prabhakar Raghavan, Hinrich Schtze. Introduction to Information Retrieval. Cambridge University Press. ISBN 978-0-521-86571-5.
[48] J. Dunn, "Well separated clusters and optimal fuzzy partitions". Journal of Cybernetics 4: 95?104. doi:10.1080/01969727408546059, (1974).
[49] Ines Frber, Stephan Gnnemann, Hans-Peter Kriegel, Peer Krger, Emmanuel Mller, Erich Schubert, Thomas Seidl, Arthur Zimek, Xiaoli Z. Fern, Ian Davidson, Jennifer Dy, eds. "MultiClust: Discovering, Summarizing, and Using Multiple Clusterings". ACM SIGKDD, (2010).
[50] W. M. Rand "Objective criteria for the evaluation of clustering methods". Journal of the American Statistical Association (American Statistical Association) **66** (336): 846?850. doi:10.2307/2284239. JSTOR 2284239, (1971).
[51] E. B. Fowlkes & C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings", *Journal of the American Statistical Association* **78**, 553?569, (1983).
[52] L. Hubert et P. Arabie. Comparing partitions. *J. of Classification*, **2(1)**, (1985).
[53] D. L. Wallace. *Journal of the American Statistical Association*, **78**, 569-579, (1983).
[54] A. Bewley, et al. "Real-time volume estimation of a dragline payload". IEEE International Conference on Robotics and Automation 2011: 1571?1576.
[55] S.C Basak, V.R. Magnuson, C.J. Niemi, R.R. Regal, "Determining Structural Similarity of Chemicals Using Graph Theoretic Indices". *Discr. Appl. Math.*, **19**, 1988: 17-44.
[56] R. Huth, et al. "Classifications of Atmospheric Circulation Patterns: Recent Advances and Applications". Ann. N.Y. Acad. Sci. 1146: 105?152, (2008).