# Best Clustering Configuration Metrics: Towards Multiagent Based Clustering

Santhana Chaimontree, Katie Atkinson and Frans Coenen

Department of Computer Science
University of Liverpool, UK.
Email: {S.Chaimontree,katie,coenen}@liverpool.ac.uk

**Abstract.** Multi-Agent Clustering (MAC) requires a mechanism for identifying the most appropriate cluster configuration. This paper reports on experiments conducted with respect to a number of validation metrics to identify the most effective metric with respect to this context. This paper also describes a process whereby such metrics can be used to determine the optimum parameters typically required by clustering algorithms, and a process for incorporating this into a MAC framework to generate best cluster configurations with minimum input from end users.

**Keywords**: Cluster Validity Metrics, Multi-Agent Clustering.

## 1 Introduction

Clustering is a core data mining task. It is the process whereby a set of objects, defined in terms of a global set of features, are categorised into a set of groups *(clusters)* according to some similarity measure or measures. Most clustering algorithms require user-supplied parameters, such as the desired number of clusters or a minimum cluster size. Identification of the most appropriate parameters to produce a "best" cluster configuration is difficult and is normally achieved through a "trial and error" process conducted by the user. The identification of the most appropriate parameters is further hampered by difficulties in defining what we mean by a cluster configuration that "best" fits the underlying data.

The need to be able to automatically identify best parameters with respect to a notion of a "best" cluster configuration is of particular relevance in the context of Multi Agent Clustering (MAC) as suggested in [6, 7]. The view of clustering presented in [6] is that of a "anarchic" collection of agents: some equipped with algorithms to conduct clustering operations or to validate the output from clustering algorithms, some holding data, and others performing house keeping and management tasks. The aim of this Multi-Agent System (MAS), although not fully realised, is to produce MAC solutions to clustering problems that require minimal input from the user and output a most appropriate cluster configuration. This objective is currently hampered by the lack of a clear understanding of what is meant by a best cluster configuration, and how this might be measured. Further, assuming we can define a best cluster configuration, how can

appropriate parameters be derived so that the desired best configuration can be realised?

In this paper a "best" cluster configuration, in the context of MAC, is defined in terms of some validity metric that must be optimised. The desired configuration is generated using a sequence of parameter values (the nature of which is dependent on the clustering algorithm adopted) to produce a collection of cluster configurations from which the most appropriate can be selected according to the adopted validity metric. Much depends on the accuracy of this metric. In this paper three such metrics are considered: the Silhouette coefficient, the Davies-Bouldin (DB) index and WGAD-BGD (this last derived by the authors).

The rest of the report is organised as follows: Section 2 surveys some relevant previous work. An overview of the metrics used to identify "best" cluster configurations, is given in Section 3. Section 4 describes the generation propose. Some evaluation and comparison is then presented in Section 5. Some conclusions are presented in Section 6.

## 2 Previous Work

This section provides a review of a number of established metrics used to evalualte clustering results, and current work on MAC within the context of the cluster configuration validity issue identified above. Different clustering algorithms provide different clustering results depending on the characteristics of the input data set and the input parameters used to define the nature of the desired clusters.

Cluster validity techniques are used to evaluate and assess the result of clustering algorithms, and may be used to select the best cluster configuration that best fits the underlying data. The available techniques can be typically classified into: (i) external criteria and (ii) internal criteria [10]. *External criteria* techniques use prelabelled datasets with "known" cluster configurations and measure how well clustering techniques perform with respect to these known clusters. *Internal criteria* techniques are used to evaluate the "goodness" of a cluster configuration without any priory knowledge of the nature of the clusters. This technique uses only the quantities and features inherent in the data set. Techniques based on external criteria require a pre-labelled *training* data set. Techniques based on internal criteria tend to be founded on statistical methods, A major drawback of which is that this often incurs high computational complexity.

A survey of well established cluster validity techniques, with respect to the above, can be found in [10]. Well known techniques include: Dunn indexing [9], Davies-Bouldin indexing[8], BR-index [20], SD [12], S_Dbw [11], the Silhouette coefficient [21], and SSB and SSE [23]. Reviews of a number of these techniques are presented in [16] and [20]. They can be categorised as measuring either: (i) the degree of intra-cluser cohesion, or (ii) the degree of inter-cluster separation; or (iii) both cohesion and separation (i.e. hybrid methods). Four of these techniques (Davies-Bouldin, Silhouette and SSB and SSE) are used in the study

described in this paper and are considered in further detail in Section 3. The four techniques were selected as they represent a mixture of approaches; SSB is used to measure inter-cluster separation and SSE is used to measure intra-cluster cohesion, while the Davies-Bouldin index and the Silhouette coefficient represent hybrid methods.

An essential feature of MAC is that the agents should determine the most appropriate number of clusters for a given data set, and the associated parameters required for cluster generation, without end user intervention. As far as the authors are aware there has been no reported work on this element of MAC. There are a number of proposed tools that present alternatives to end-users for selection. The presentation is usually in some graphical format, hence these tools may be labelled as "visual" cluster validation tools. One such tool is CVAP (Cluster Validity Analysis Platform) [24]. CVAP operates by applying a number of clustering algorithms, with a sequence of parameters, to a given data set. Each result is assessed using a "validity" index which is plotted on a graph which may then be inspected and a selection made. Another such tool is described in [17].

There has been some previous work on multi-agent clustering. The earliest reported systems are PADMA [13] and PAPYRUS [4]. The aim of these systems was to achieve the integration of knowledge discovered from different sites with a minimum amount of network communication and a maximum amount of local computation. PADMA is used to generate hierarchical clusters in the context of document categorisation. PADMA agents are employed for local data accessing and analysis. A facilitator (or coordinator) agent is responsible for interaction between the mining agents. All *local clusters* are collected at the central site to generate the *global clusters*. PADMA is therefore based on a centralised architecture, whereas PAPYRUS adopts a Peer-To-Peer model where both data and results can be moved between agents according to given MAS strategies. Another MAC based on the Peer-To-Peer model is proposed in [22] where a distributed density-based clustering algorithm, called KDEC [15], is used. Density estimation samples are transmitted, instead of data values, outside the site of origin in order to preserve data privacy. Reed et al. [19] proposed a MAS for distributed clustering for text documents which assigns new, incoming documents to clusters. The objective here was to improve the accuracy and the relevancy of information retrieval processes. Kiselev et al. [14] proposed a MAC dealing with data streams in distributed and dynamic environments whereby input data sets and decision criteria can be changed at runtime. Clustering results are available at anytime and are continuously revised to achieve the global clustering. (There are also some reported agent based systems for supervised learning, such as that reported in [3, 5, 1] from which some parallels may be drawn with respect to MAC.) What the above reported systems have in common, unlike the generic vision espoused in this paper, is that they are founded on a specific clustering algorithm that feature preset parameters.

## 3 Quality Measures

In this section four of the quality measures introduced in the previous section (Davies-Bouldin, Silhouette, SSB and SSE) are considered in further detail. Recall that to determine the effectiveness of a cluster configuration we can adopt three approaches: (i) we can measure the cohesion of the objects within clusters, (ii) the separation between clusters, or (iii) a combination of the two. Cohesion can be measured using the Sum Squared Error (SSE) measure and separation the Sum of Squares Between groups (SSB) measure. While the Davies-Bouldin index and the Silhouette coefficient combine the two. The basic notation used through out the discussion is given in Table 1. For the work described in this paper the maximum number of clusters, $maxPts$, is calculated as the square root of the number of objects (records), $N$, thus $maxPts = \lceil \sqrt{N} \rceil$. The intuition here is that the maximum number of clusters that the records in a data set can be categorised into is proportional to $N$ (the number of records). The minimum number of clusters ($minPts$) is typically set to 2.

**Table 1:** Basic notation

| notation | Description |
|---|---|
| $K$ | The number of clusters |
| $N$ | The number of objects (records) in a data set |
| $\{C_1, ..., C_K\}$ | Set of $K$ clusters |
| $D = d(x_i, x_j)$ | Matrix of "distance" among objects |
| $x_1, ..., x_N$ | Set of objects to be clustered |
| $minPts$ | The minimum number of possible objects in a cluster |
| $maxPts$ | The maximum number of possible objects in a cluster |
| $\|C_i\|$ | The number of objects in a cluster $i$ (i.e. the size of a cluster) |

SSE [23] is the sum of the squared intra-cluster distance of each cluster centroid, $c_i$, to each point (object) $x_j$ in that cluster. More formaly the total SSE of a given cluster configuration is defined as:

$$Total\ SSE = \sum_{i=1}^{i=K} \sum_{j=1}^{j=|C_i|} dist(x_j, c_i)^2 \ . \tag{1}$$

The lower the total SSE value the greater the intra-cluster cohesion associated with the given cluster configuration.

SSB [23], in turn, is the sum of the squared distance of each cluster centroid ($c_i$) to the overall centroid of the cluster configuration ($c$) multiplied (in each case) by the size of the cluster. Formaly the total SSB of a given cluster configuration is defined as:

$$Total\ SSB = \sum_{i=1}^{i=K} |C_i|(dist(c_i, c))^2 \ . \tag{2}$$

The higher the total SSB value of a cluster configuration the greater the degree of separation. Note that the cluster centriod is used to represent all the points in a cluster, sometimes referred to in the literature as the *cluster prototype*, so as to reduce the overall computational complexity of the calculation (otherwise distance from every point to every other point would have to be calculated).

Using the above the calculated SSE and SSB values tend to be large numbers because of the squared operation. Therefore, in the context of the work described here, the authors use variations of SSE and SSB metrics. We refer to these metrics as the total Within Group Average Distance (WGAD) [18] and the total Between Group Distance (BGD). WGAD and BGD are derermined as follows:

$$Total\ WGAD = \sum_{i=1}^{i=K} \frac{\sum_{j=1}^{j=|C_i|}\ dist(x_j, c_i)}{|C_i|}\ . \tag{3}$$

$$Total\ BGD = \sum_{i=1}^{i=K} dist(c_i, c)\ . \tag{4}$$

Experiments conducted by the authors (not reported here) suggested that both separation and cohesion are significant and thus we find the difference between a pair of WGAD and BGD values to express the overall validity of a given cluster configuration. We refer to this measure as the WGAD-BGD measure. To obtain a best cluster configuration we must minimise the WGAD-BGD measure.

The Overall Silhouette Coefficient ($OverallSil$) of a cluster configuration is measure of both the cohesiveness and separation of the configuration [21]. It is determined by first calculating the *silhouette* ($Sil$) of each individual point $x_j$ within the configuration as follows:

$$Sil(x_j) = \frac{b(x_j) - a(x_j)}{max(a(x_j), b(x_j))}\ . \tag{5}$$

where $a(x_j)$ is the average intra-cluster distance of the point $x_j$ to all other points within its cluster, and $b(x_j)$ is the minimum of the average inter-cluster distances of $x_j$ to all points in each other cluster (see Figure 1 for further clarification). The overall silhouette coefficient ($OverallSil$) is then calculated as follows:

$$OverallSil = \frac{\sum_{i=1}^{i=K} \frac{\sum_{j=1}^{j=|C_i|} sil(x_j)}{|C_i|}}{K}\ . \tag{6}$$

The resulting overall silhouette coefficient is then a real number between $-1.0$ and 1.0. If the silhouette coefficient is close to $-1$, it means the cluster is undesirable because the average distance to points in the cluster, is greater than the minimum average distance to points in the other cluster(s). The overall silhouette coefficient can be used to measure the goodness of a cluster configuration. The larger the coefficient the better the cluster configuration.

The Davies-Bouldin (DB) *validity index* is the sum of the maximum ratios of the intra-cluster distances to the inter-cluster distances for each cluster $i$:

$$DB = \frac{1}{K} \sum_{i=1}^{i=K} R_i \ . \tag{7}$$

Where $R_i$ is the maximum of the ratios between cluster $i$ and each other cluster $j$ (where $1 \leq j \leq K$ and $j \neq i$). The lower the DB value the better the associated cluster configuration. The individual ratio of the intra-cluster distances to the inter-cluster distances for cluster $i$ with respect to cluster $j$ is given by:

$$R_{ij} = \frac{S_i - S_j}{d_{ij}} \ . \tag{8}$$

where $d_{ij}$ is a distance between the centroid of cluster $i$ and the centroid of cluster $j$, and $S_i$ ($S_j$) is the average distance between the points within cluster $i$ ($j$):

$$S_i = \frac{1}{|C_i|} \sum_{n=1}^{n=|C_i|} d(x_n, c_i) \ . \tag{9}$$

where $c_i$ is the centroid of cluster $i$, and $x$ is a point (object) within the cluster $i$. The derivation of DB is illustrated in Figure 2.

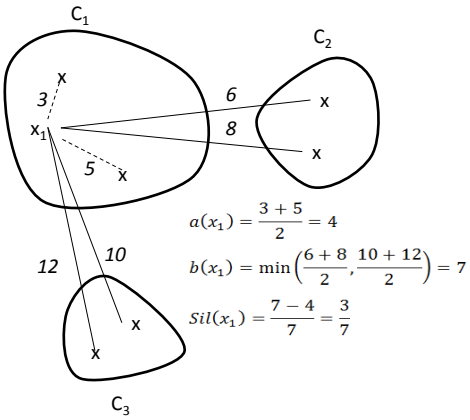


**Fig. 1:** Derivation of the Overall Silhouette Coefficient (*OverallSil*)

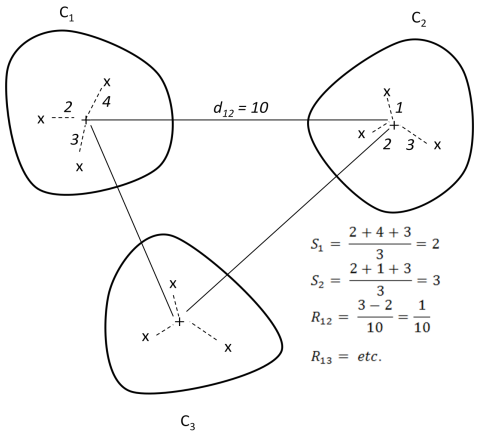


**Fig. 2:** Derivation of the Davies-Bouldin (DB) *validity index*

## 4 Parameter Identification for Clustering Algorithms

Most clustering algorithms require user-supplied parameters. For example: K-means requires the number of classes, $K$, as a parameter; and KNN requires a

threshold ($t$) to identify the "nearest neighbour". To obtain a best cluster configuration in terms of the metrics described above the most appropriate parameters are required. To act as a focus for this part of the research reported in this paper the well known K-means and KNN algorithms were adopted, although other clustering algorithms could equally well have been adopted.

The most obvious mechanism for indentifying appropriate parameters is to adopt some kind of *generate and test* loop whereby a cluster configuration is "generated" using a particular parameter value which is then "tested" (evaluated) using a validity metric (such as those discussed in the foregoing section) as a result of which the parameter value is adjusted. However, this did not prove successful. Figure 3 shows the validity measures obtained using the silhouette coefficient (Sil. Coef.) and the Davies-Bouldin index (DB) and with a range of $t$ values using the KNN clustering algorithm when applied to the Iris data set taken from the UCI data repository [2] (similar results were obtained using other data sets). From the figure it can be seen that there are local maxima and minima which means that a generate and test procedure is unlikely to prove successful. Figure 4 shows the WGAD-BGD measures obtained using the KNN algorithm and the Iris data set. Recall that, we wish to minimise this measure. From the figure it can be seen that a range of "best" $t$ values are produced. Thus a generate and test process would not find the most appropriate parameters. Instead, the process advocated here is to generate a sequence of cluster configurations for a range of parameter values.

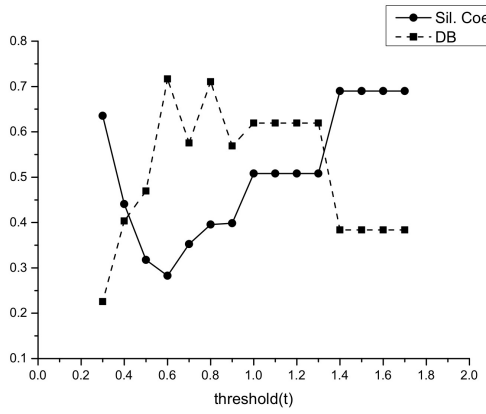


**Fig. 3:** Validity values using Sil.Coef. and DB index for KNN algorithm using the Iris data set.

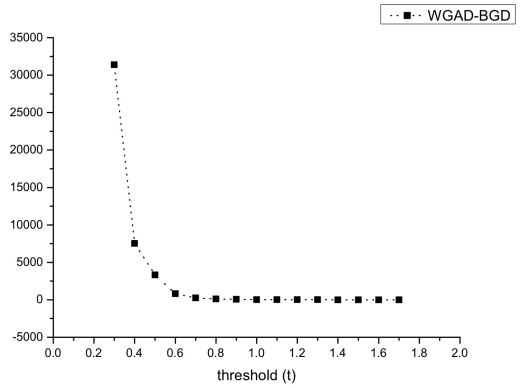


**Fig. 4:** Validity values using WGAD-BGD for KNN algorithm using the Iris data set.

Thus, for K-means, to identify the most appropriate number of clusters the algorithm is run multiple times with a sequence of values for $K$ ranging from 2 to $\lceil\sqrt{N}\rceil$, where $N$ is the number of records in the given data set. The "goodness" of each generated cluster configuration was tested using the identified validity metrics. The generation algorithm is presented in Table 2.

In the case of KNN the algorithm is run multiple times with a range of different values for the nearest neighbour threshold ($t$). Note that any $t$ value that does not generate a number of clusters of between 2 and $\lceil\sqrt{N}\rceil$ is ignored. The algorithm is presented in Table 3.

The process was incorporated into a MAC framework founded on earlier work by the authors and reported in [6, 7]. An issue with K-means is that the initial points (records/objects) used to define the initial centroids of the clusters are randomly selected. Experiments indicated that the selection of start points can greatly influence the operation of K-means, to the extent that different best values of $K$ can be produced depending on where in the data set the algorithm starts. Therefore use of K-means to identify a "proper" number of clusters does not represent a consistent approach. However, the KNN parameter selection process can be used to indirectly determine the most appropriate value for $K$ to be used in the K-means approach.

## 5 Experimental Evaluation

In Section 3 a number of metrics for identifying a best cluster configuration were identified. These in turn were incorporated into the parameter identifictaion mechanism identified in the foregoing section. The evaluation of the proposed approach is presented in this section. The evaluation was conducted using ten data sets taken from the UCI repository [2]. Table 4 gives some statistical information concerning these data sets. Note that the data sets display a variety of features.

**Table 2:** Algorithm to identify the most appropriat number of clusters using K-means.

*Algorithm: KmeansIdentifiesK*

Input: $x_1, ..., x_N$
Output: $K$, *the overall validity value*

1. For $K = 2$ to $K = \lceil\sqrt{N}\rceil$ do
2. Do K-means clustering
3. Evaluate the clustering result (a set of clusters) by using a chosen metric
4. Keep $K$ and *the overall cluster validity*
5. Select $K$ which provide the good cluster configuration

Table 5 shows a comparison of the operation of the above process using: K-means; and the identified cluster validity techniques, namely silhouette coefficient (Sil. Coef.), Davies-Bouldin index (DB index), and the combination between WGAD and BGD (WGAD-BGD); when applied to the data sets listed in Table 4. The comparison is conducted by considering the number of clusters ($K$) associated with the best cluster configuration and the known value for $K$. Table 6 presents a similar comparison using the KNN algorithm.

**Table 3:** Algorithm to identify the most apppropriate number of clusters using KNN.

*Algorithm KNNIdentifiesK*

Input: $x_1, ..., x_N$
Output: $t$, $K$, *the overall validity value*

1. Calculate distance between objects in a data set
2. Ascending order all of objects in the data set
3. Choose one point in which its position is at the middle
4. Choose distance between the object from step 3 and other objects in the data set
5. Set $t = distance$
6. Ascending order distances and select distinct distances
7. For each $t$ do
8. Do KNN clustering
9. If the result generated from step (8) is different from the last result
10. Evaluate a clustering result.
11. Keep $t$, $K$ and *the overall validity value*
12. Select $t$ providing the "best" cluster configuration

In both Table 5 and Table 6 the values in bold indicate where the identified number of clusters ($K$) exactly matches the "known" value of $K$. (In Table 5 it can also be argued that the $K$ value of 8 produced using the DB index metric is significantly close to the "required" value of 7.)

The results are summarised in Table 7 (again values in bold indicate best results). From Table 7 it can be observed that the DB-index does not perform well, particularly when used in conjunction with K-means. In this latter case the use of DB-index over specified the number of clusters that define a best cluster configuration, in all cases. Both the Silhouette Coefficient and WGAD-RGD produced better results, with the best result generated using Silhouette in conjunction with KNN. Further experiments using K-means demonstrated that the results obtained were very inconsistent in that they were very dependent on the nature of the selected start locations (centroids). In this respect KNN produced more consistent results. It is also worth nothing that using some data sets (Heart, Pima Indians and Breast Cancer) consistent results were obtained, with respect to other data sets (Iris, Zoo, Ecoli, Yeast and Car) poorer results were obtained. Inspection of the nature of these data sets (Table 4) indicats that the proposed technique operates best where data sets feature a small number of clusters (classes).

**Table 4:** Statistical information for the datasets used in the evaluation.

| No. | Data Set | Num Records ($N$) | Num Attr | Num Classes | Attribute Description |
|---|---|---|---|---|---|
| 1 | Iris | 150 | 4 | 3 | 4 Numeric |
| 2 | Zoo | 101 | 16 | 7 | 15 Boolean, 1 Numeric |
| 3 | Wine | 178 | 13 | 3 | 13 Numeric |
| 4 | Heart | 270 | 13 | 2 | 6 Real, 1 Ordered, 3 Binary, 3 Nominal |
| 5 | Ecoli | 336 | 7 | 8 | 7 Real |
| 6 | Blood Transfusion | 748 | 4 | 2 | 4 Integer |
| 7 | Pima Indians | 768 | 8 | 2 | 8 Numeric |
| 8 | Yeast | 1484 | 8 | 10 | 8 Numeric |
| 9 | Red wine quality | 1599 | 11 | 6 | 11 Numeric |
| 10 | Breast Cancer | 569 | 21 | 2 | 21 Real |

**Table 5:** Results using K-means to generate a best set of clusters

| No. | Data Set | Known $K$ | $K$ | Sil. Coef. | $K$ | DB Index | $K$ | WGAD-BGD |
|---|---|---|---|---|---|---|---|---|
| 1 | Iris | 3 | 2 | 0.68 | 4 | 0.31 | 2 | 3.93 |
| 2 | Zoo | 7 | 4 | 0.47 | 8 | 0.68 | 2 | 3.66 |
| 3 | Wine | 3 | 2 | 0.66 | 12 | 0.21 | 2 | 583.02 |
| 4 | Heart | 2 | **2** | 0.38 | 12 | 0.67 | **2** | 81.75 |
| 5 | Ecoli | 8 | 4 | 0.43 | 12 | 0.67 | 2 | 0.57 |
| 6 | Blood Transfusion | 2 | **2** | 0.70 | 15 | 0.20 | **2** | 3044.25 |
| 7 | Pima Indians | 2 | **2** | 0.57 | 3 | 0.51 | **2** | 223.53 |
| 8 | Yeast | 10 | 3 | 0.27 | 26 | 0.78 | 2 | 0.29 |
| 9 | Red Wine | 6 | 2 | 0.60 | 34 | 0.87 | 2 | 3.00 |
| 10 | Breast cancer | 2 | **2** | 0.70 | 20 | 0.29 | **2** | 1331.33 |

## 6 Conclusions

In this paper a set of experiments directed at identifying the most appropriate metrics to determine the validity of a cluster configulation have been reported. Three validity metrics were considered, the silhouette coeficient, the DB index and WGAD-BGD. The last being a combination of the SSB and SSE metrics. Experiments were conducted using K-means and KNN, and a collection of data sets taken from the UCI repository. The context of the work described was Multi-Agent Clustering (MAC) directed at generating a cluster configuration that best fits the input data using a variety of clustering mechanisms. This in turn required a mechanism for identifying best cluster configurations. The main findings of the work are as follows: (i) the overall best cluster configuration validation technique is the Silhouette coefficient, (ii) KNN is a much more reliable technique to find the best value for $K$ than K-means (but can be used to discover the $K$ value required by K-means), and (iii) a *generate and test* process does not necessarily achieve the desired result.

**Table 6:** Results using KNN to generate a best set of clusters

| No. | Data Set | Known $K$ | $t$ | $K$ | Sil. Coef. | $t$ | $K$ | DB Index | $t$ | $K$ | WGAD-BGD |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Iris | 3 | 1.49 | 2 | 0.69 | 1.49 | 2 | 0.38 | 1.49 | 2 | 3.97 |
| 2 | Zoo | 7 | 2.83 | 2 | 0.40 | 2.65 | 4 | 0.71 | 2.83 | 2 | 3.94 |
| 3 | Wine | 3 | 195.07 | **3** | 0.61 | 75.78 | 14 | 0.33 | 206.37 | 2 | 810.82 |
| 4 | Heart | 2 | 81.63 | **2** | 0.76 | 81.63 | **2** | 0.16 | 81.63 | **2** | 316.67 |
| 5 | Ecoli | 8 | 0.53 | 2 | 0.39 | 0.28 | 10 | 0.39 | 0.53 | 2 | 0.79 |
| 6 | Blood Transfusion | 2 | 2000.02 | **2** | 0,83 | 250.27 | 18 | 0.12 | 2000.02 | **2** | 7404.08 |
| 7 | Pima Indians | 2 | 193.36 | **2** | 0.79 | 193.36 | **2** | 0.23 | 193.36 | **2** | 682.60 |
| 8 | Yeast | 10 | 0.60 | 2 | 0.74 | 0.26 | 21 | 0.31 | 0.60 | 2 | 0.74 |
| 9 | Red Wine | 6 | 1.42 | 3 | 0.17 | 1.42 | 3 | 2.21 | 1.42 | 3 | 8.47 |
| 10 | Breast cancer | 2 | 992.39 | **2** | 0.80 | 992.39 | **2** | 0.13 | 992.39 | **2** | 3888.91 |

**Table 7:** Summary of results

| No. | Data Set | $K$ UCI | KNN | | | K-Means | | |
|---|---|---|---|---|---|---|---|---|
| | | | $K$ Sil. Coef. | $K$ DB index | $K$ WGAD-BGD | $K$ Sil. Coef. | $K$ DB index | $K$ WGAD-BGD |
| 1 | Iris Plants | 3 | 2 | 2 | 2 | 2 | 4 | 2 |
| 2 | Zoo | 7 | 2 | 4 | 2 | 4 | 8 | 2 |
| 3 | Red Wine | 3 | **3** | 14 | 2 | 2 | 12 | 2 |
| 4 | Heart | 2 | **2** | **2** | **2** | **2** | 12 | **2** |
| 5 | Ecoli | 8 | 2 | 10 | 2 | 4 | 12 | 2 |
| 6 | Blood Transfusion | 2 | **2** | 18 | **2** | **2** | 15 | **2** |
| 7 | Pima Indians | 2 | **2** | **2** | **2** | **2** | 3 | **2** |
| 8 | Yeast | 10 | 2 | 21 | 2 | 3 | 26 | 2 |
| 9 | Car | 6 | 3 | 3 | 3 | 2 | 34 | 2 |
| 10 | Breast Cancer | 2 | **2** | **2** | **2** | **2** | 20 | **2** |
| | Totals | | 5 | 3 | 4 | 4 | 0 | 4 |

These findings are currently being incorporated into a MAS framework [6, 7]. The techniques investigated sofar, and reported here, do not serve to find the best results in all cases and further investigation is therefore required, however the authors are greatly encouraged by the result reported in this paper.

## References


1. Albashiri, K., Coenen, F.: Agent-enriched data mining using an extendable framework. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5680 LNAI, 53–68 (2009)
2. Asuncion, A., Newman, D.: UCI machine learning repository (2007), http://www.ics.uci.edu/∼mlearn/MLRepository.html
3. Baik, S., Bala, J., Cho, J.: Agent based distributed data mining. In: Liew, K.M., Shen, H., See, S., Cai, W. (eds.) Parallel and Distributed Computing: Applications and Technologies, Lecture Notes in Computer Science, vol. 3320, pp. 185–199. Springer Berlin / Heidelberg (2005)
4. Bailey, S., Grossman, R., Sivakumar, H., Turinsky, A.: Papyrus: A system for data mining over local and wide area clusters and super-clusters. IEEE Supercomputing (1999)
5. Canuto, A.M.P., Campos, A.M.C., Bezerra, V.M.S., Abreu, M.C.d.C.: Investigating the use of a multi-agent system for knowledge discovery in databases. International Journal of Hybrid Intelligent Systems 4(1), 27–38 (2007)
6. Chaimontree, S., Atkinson, K., Coenen, F.: Clustering in a multi-agent data mining environment. In: Proc. Int. Workshop on Agents and Data Mining Interaction (ADMI'10). pp. 103–114. Springer LNAI 5980 (2010)
7. Chaimontree, S., Atkinson, K., Coenen, F.: Multi-agent based clustering: Towards generic multi-agent data mining. In: Proc. 10th Industrial Conf. on Data Mining (ICDM'10). pp. 115–127. Springer LNAI 6171 (2010)

8. Davies, D.L., Bouldin, D.W.: A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-1(2), 224–227 (April 1979)
9. Dunn, J.C.: Well separated clusters and optimal fuzzy-partitions. Journal of Cybernetics 4, 95–104 (1974)
10. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Cluster validity methods: part I. SIGMOD Record 31(2), 40–45 (2002)
11. Halkidi, M., Vazirgiannis, M.: Clustering validity assessment: Finding the optimal partitioning of a data set. In: ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining. pp. 187–194. IEEE Computer Society, Washington, DC, USA (2001)
12. Halkidi, M., Vazirgiannis, M., Batistakis, Y.: Quality scheme assessment in the clustering process. In: PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery. pp. 265–276. Springer-Verlag, London, UK (2000)
13. Kargupta, H., Hamzaoglu, I., Stafford, B.: Scalable, distributed data mining using an agent based architecture. In: Proceedings the Third International Conference on the Knowledge Discovery and Data Mining. pp. 211–214. AAAI Press (1997)
14. Kiselev, I., Alhajj, R.: A self-organizing multi-agent system for online unsupervised learning in complex dynamic environments. In: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. pp. 1808–1809. AAAI Press (2008)
15. Klusch, M., Lodi, S., Moro, G.: Agent-based distributed data mining: The KDEC scheme. In: Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science). vol. 2586, pp. 104–122 (2003)
16. Legány, C., Juhász, S., Babos, A.: Cluster validity measurement techniques. In: AIKED'06: Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases. pp. 388–393. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA (2006)
17. Qiao, H., Edwards, B.: A data clustering tool with cluster validity indices. ICC2009 - International Conference of Computing in Engineering, Science and Information pp. 303–309 (2009)
18. Rao, M.: Clustering analysis and mathematical programming. Journal of the American statistical association 66(345), 622–626 (1971)
19. Reed, J.W., Potok, T.E., Patton, R.M.: A multi-agent system for distributed cluster analysis. In: Proceedings of Third International Workshop on Software Engineering for Large-Scale Multi-Agent Systems (SELMAS'04) W16L Workshop - 26th International Conference on Software Engineering. pp. 152–155. IEE, Edinburgh,Scotland, UK (2004)
20. Ristevski, B., Loshkovska, S., Dzeroski, S., Slavkov, I.: A comparison of validation indices for evaluation of clustering results of DNA microarray data. In: 2nd International Conference on Bioinformatics and Biomedical Engineering, iCBBE 2008. pp. 587–591 (2008)
21. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20(1), 53–65 (November 1987)
22. da Silva, J., Klusch, M., Lodi, S., Moro, G.: Privacy-preserving agent-based distributed data clustering. Web Intelligence and Agent Systems 4(2), 221–238 (2006)
23. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley (2005)
24. Wang, K.and Wang, B., Peng, L.: CVAP: Validation for cluster analyses. Data Science Journal 8, 88–93 (2009)