RESEARCH ARTICLE OPEN ACCESS

Metacluster Model Using an Aggregate Function In The Presence Of Data Domain Membership

Zurini Madalina, Cioloca Cecilia, Georgescu Mihai
PhD Candidates, Department of Economy Informatics Bucharest Academy of Economic Studies, Romania

ABSTRACT
Once with the exponential increase of data generated by the business environment, the need of rapid, precise and robust algorithms appeared in data analysis. The improvements given by database technologies, computational performances and artificial intelligence have contributed to the development of intelligent data analysis. The principal clustering methods are presented in comparison. A technique for object grouping validation generated by the clustering methods is proposed and applied on a dataset a priori classified according to their domain membership. Metaclustering is introduced as an aggregation method for more clustering techniques in order to improve the performances of the clustering process in terms of results' correctness.
Keywords **–** Cluster evaluation, DBSCAN, k-Means, k-Medoids, Metacluster

I. INTRODUCTION

Found at the intersection of fundamental domains such as computer science, information technology, decision theory, geometry, probability theory and statistical mathematics, [12], pattern recognition knows in the present applications of which wide stretches from anthropology research to hardware and software projection, [1]. Pattern recognition theory is defined as representing the totality of rules, principles, methods and tools for analysis and decision used to identify the membership of objects, units, phenomena, events, actions, processes, to certain well defined sets of individuality. Pattern recognition theory is divided into two categories: supervised and unsupervised learning. Classifies are part of the supervised learning, using information about the membership of an object to a set in order to classify new objects in one of the defined sets, while unsupervised learning, represented by the clustering process, groups the set of objects according to the objects' characteristics in partitions of the initial set.

The need for classification, grouping and differentiation of objects into categories or classes appear in all human activities and in various fields of knowledge, such as: informatics, biology, medicine, physics, [13], financial analysis, political science or marketing. The grouping is made in categories that are clearly and natural defined, with a concrete meaning in the studied reality. Differentiation is based on the fundamental properties of the objects and the criteria of separation are given by the degree of similarity of the objects' analyzed properties.

The paper is composed as it follows. Chapter 2 contains a comparative analysis of main techniques in clustering. In chapter 3 it is proposed a validation technique of the clustering methods comparing the result given by the cluster membership of the objects with their initial membership to the a priori defined classes. A procedure for mapping of clusters to classes is implemented. Chapter 4 contains an aggregation function of the results of a minimum three clustering methods with the homogenization between the clusters and initial classes. The aggregation function proposes an improvement of the clustering process.

The paper finishes with the results and future work, in chapter 5, where a dataset formed out of 900 bi-dimensional objects distributed into four clusters are used as input data for running k-means, k-medoids and DBSCAN clustering algorithms using four types of distance functions, Euclidian, Canberra, Manhattan and Cosine. The results are compared according to the execution time and the level of correctness of the clustering process. Future work is related to the implementation of other clustering algorithms and aggregating them using a meta-cluster in order to improve the objective function defined by the clustering analysis.

II. COMPARATIVE ANALYSIS OF MAIN CLUSTERING TECHNIQUES

Data clustering, also known as cluster analysis is defined by Webster, Merriam-Webster Online Dictionary, as a statistical classification technique used to discover whether the individuals of a population can be separated into different groups by making quantitative comparisons of multiple characteristics.

The objectives of cluster analyses are given by:

- the understanding of structures, for generating hypothesis upon data or to detect abnormalities;
- the natural classification, in order to identify the degree of similarity among forms;
- the compression, as a method of data organization and resuming within structures such as clusters.

Clustering algorithms based on partitions start from an initial set of m objects and separate the set of data into k partitions, value k being given as an input to the algorithm. The principle of clustering is intuitive and is done with the scope of achieving optimum criteria within an iterative process. Partition methods are of two types: k-means, [2], in which each cluster is represented by its gravity center formed out of the totality of objects part of the cluster, and k-medoids, [3], context in which each cluster is represented by its closed object to its gravity center, called a medoid.

Clustering algorithms based on hierarchies use a method based on grouping objects according to a gradual aggregation and disaggregation of the objects, [4]. Disaggregation algorithms build the clusters in a downward manner, starting from a single cluster that contains all the initial objects, and forming, using a successive division, *m* clusters, and each cluster containing an object. Aggregation algorithms build a tree starting from the leaf level, the one with the m clusters each containing an object, and aggregates the clusters found at close distances until a single cluster is composed out of the m initial objects.

Depending on the similarity method of evaluation of two clusters in agglomerative hierarchic algorithms, five types of algorithms differentiate:

- *single link* uses as distance between two clusters the distance of its closest two objects;
- *complete link* uses the distance formed from its outermost two objects; *average link* evaluates the distance calculated as the average of the totality of distances between each two objects from two different clusters;
- *centroid link* uses the distance between two clusters equal to the distance between its centroids, the gravity center of the set of objects from which a cluster is formed out;
- *Ward* uses the consolidation methods of that clusters that minimizes the sum of squares of the deviations in the clusters.

The fundamental difference between cluster algorithms based on hierarchy and those based on the partitions is given by the fact that the hierarchic technology creates a decomposition or aggregation tree of the objects, without a priori knowledge of the number of clusters which decompose the original objects, while partitioning clustering needs an additional input data, along with the set of objects, parameter k, the number of clusters in which the object space is separated in. The evaluation of the clustering methods based on hierarchy is given by the high complexity level and its limitations involved by the stopping condition that needs additional knowledge of the domain from which the objects are part of.

Density based clustering algorithms use the agglomeration of objects through the calculation, in a maximal context, of the set of points connected at density level. Because of the identification model of the clusters through the direction given by the density gradient, density based algorithms bring major advantages in their implementation when it comes to multiple arbitrary shapes and noise objects, also called outliners. Among the algorithms used, the main ones are: DBSCAN, Density Based Spatial Clustering Algorithm with Noise, and OPTICS, Ordering Points To Identify the Clustering Structure.

Network based algorithms separate the object space into grids or cells, and the grouping technique is done upon those cells. The main algorithms from this class are: STING, Statistical Information Grid, Denclue, Density Clustering, CLIQUE, Clustering in Quest, MAFIA, Merging of Adaptive Intervals Approach to Spatial Data Mining and WaveCluster. The grid structure forms a finite number of cells that gives the great advantage of operating in a low time, independent of the number of objects from the initial space. The method is suitable to those sets of objects that their density and cardinality is high in a limited space.

Suffix Tree Clustering is a technique applied upon text documents that are represented using STDM representation, Suffix Tree Document Model. This method uses a tree representation of data, because of the need of introducing the dependency factor of the characteristics of each object.

Multiple studies are achieved upon cluster analysis lately, but the attention is concentrated on the new clustering techniques. Scalability and high dimensionality of the space aren't the only concerns of the current research. In [11], the main requirements of a clustering technique are describes through:

- *scalability*; cluster methods are applied upon large database and the performance laniary decreases along with the volume of data;
- *versatility*; the objects are of different types: numerical, binary or qualitative, so a cluster method needs to by applicable to all data representation;
- *the ability of discovering different forms of the clusters*; the majority of clustering techniques discovers sphere forms;
- *a minimum number of input parameters*; given the diverse context of applicability, a large number of input parameters slows the grouping process, a high level of knowledge of the domain being required;
- *robustness when it comes to noise*; a clustering algorithm shouldn't be affected by the noise present in data;
- *independence of the order of input data*; the order in which the input data is given shouldn't influence the result of the cluster;
- *scalability for high dimensions*; the capacity to operate upon multidimensional data represented in a high causal space.

The multiple forms that a cluster can take may lead to the need of generating a clustering algorithm that detects not only sphere forms.

The general model of clustering uses as input the initial set of objects represented in a n-dimensional space and the parameters need for running multiple clustering algorithms. The result of the algorithm is given by object grouping in k number of cluster, a priori known or not, in the context of criteria optimization, with the minimization of inter-cluster variance and the maximization of intra-cluster variance.

Let x_i, $i = \overline{1,m}$, be the object *i* from the set of objects of the n-dimensional space, of the form:

$$x_i = (x_{i1}, x_{i2}, \dots, x_{in})^T, \forall i = \overline{1,m}$$

The set of x_i objects is noted with *X*, so that:

$$X = \{x_1, x_2, \dots, x_m\}$$

Each object, after running the clustering algorithm, is assigned to one of the k+1 formed clusters:

$$\forall i = \overline{1,m}, \exists! l \in \{0,1,2,\dots,k\}\ a.i.\ x_i \in C_l$$

where:

- *l* is the cluster number from which object x_i is part of;
- C_l is the l cluster;
- *k* is the number of clusters in which the objects are grouped in.

Cluster forming meets the condition system:

$$\begin{cases} \forall l, m \in \{0,1,2,\dots,k\}, l \neq m, C_l \cap C_m = \emptyset \\ \bigcup_{l=\overline{0,k}} C_l = X \end{cases}$$

The clusters $C_1, C_2, \dots, C_l, \dots, C_k$ represent the k groups through which the m objects from the causal space are characterized. The C_0 cluster is formed out of the totality of noise objects that are not integrated in the clusters *1,2,…,k*. According to the clustering algorithms implemented, the noise cluster may be generated or not.

The formulation of clustering analysis using an integer programming, IP, problem in [9] is done using an optimization problem by defining binary decision variables defined as:

$$x_{ij} = \begin{cases} 1, if\ x_i \in C_j \\ 0, otherwise \end{cases}$$

Given some performance measure $f: \{0,1\}^n \to \mathbb{R}$ of the quality of a partition, the IP problem is formulated as:

$$\begin{cases} \min f(x) \\ \sum_{j=1}^{k} x_{ij} = 1, i = 1,2,\dots,m \\ \sum_{i=1}^{m} x_{ij} \geq 1, j = 1,2,\dots,k \\ x_{ij} \in \{0,1\}, i = 1,2,\dots,m; j = 1,2,\dots,k, \end{cases}$$

III. CLUSTERING VALIDATION IN THE PRESENCE OF DATA DOMAIN MEMBERSHIP

The clustering analysis process is seen in [7] as a complex tool that unites the following components:

- data representation, representing the objects of a specific domain in order to use clustering algorithms;
- checking cluster tendency, the process of verification that a tendency for clustering exists in the data;
- using a clustering algorithm, the input for the algorithm is the proximity matrix and the output is a description of grouping the patterns into clusters;
- validation, cluster validation refers to the procedures that evaluate the result of the clustering algorithm.

The evaluation of clustering analysis is done at the level of how the clusters are formed, the intra-cluster homogeneity and at the level of distance between the clusters, taking over the principle of cluster forming, minimizing intra-cluster variance and maximizing inter-cluster variance, and at the level of cluster management, seen as a classification technique. This management achieves the verification of the degree of similarity between the proposed model for object grouping and the initial a priori known object grouping into classes.

The cluster method evaluation is done at the:

- local level, for the evaluation of the results obtained by a certain clustering method using different input parameters, the optimization function and the ending point, resulting a number of local optimum equal to the number of clustering methods applied;
- global level, for the evaluation of the local optimum and choosing the implementation that best matches the set of objects, resulting a global optimum that characterizes the set of objects.

In [5], an evaluation method is proposed without knowing a priori information concerning the objects' membership to certain classes and is based upon the calculation of the entropy for characterizing the structure of the information within the clusters.

In cluster analysis, the fundamental set of evaluation of the algorithm, described in [6], consists of:

- trend, demonstrates whether the structure formed is random or not;
- comparison, due to the stochastic and parameters' dependent character of clustering methods;
- stability, demonstrated though multiple implementations of clustering analysis and the comparison of their results;
- cohesion, depends on how much the cluster is compact or not.

Starting from the initial set of objects, *X*, each object is assigned to one of the existing states,

$T_1, T_2, \dots, T_r$, where r represents the number of classes in which the objects are assigned.

$$T_i = \{x_j | j \in \{1,2,\dots,m\}, t(x_j) = i\}$$

where $t(x_j)$ represents the class in which object x_j from the set of objects *X* is assigned.

Based on the partitions $T_1, T_2, \dots, T_r$ of the initial set *X*, the correlation between objects' matrix is formed, *MCP*, relating the interaction between the objects from the same partition, with $MCP \in \mathcal{M}_{m\times m}\{0,1\}$, so that:

$$mcp_{ij} = \begin{cases} 1, t(x_i) = t(x_j) \\ 0, t(x_i) \neq t(x_j) \end{cases}$$

After applying a method of clustering from the ones proposed, the result is given by the number of resulted clusters, k, and a set of partitions, $C_1, C_2, \dots, C_k$. Upon those partitions, the correlation operator implemented, resulting in the correlation matrix between the objects, *MCC*, with $MCC \in \mathcal{M}_{m\times m}\{0,1\}$, defined as:

$$mcc_{ij} = \begin{cases} 1, c(x_i) = c(x_j) \\ 0, c(x_i) \neq c(x_j) \end{cases}$$

where $c(x_j)$ represents the cluster in which object x_j is assigned to.

The matrix of clustering evaluation through the comparison between the initial r categories and the ones given by the clusters, *MEC*, is resulted after comparing each value from the MCP matrix with the ones from MCC matrix:

$$mec_{ij} = \begin{cases} 1, (mcp_{ij} + mcc_{ij})\%2 = 0 \\ 0, (mcp_{ij} + mcc_{ij})\%2 = 1 \end{cases}$$

where x*%y* represents the rest from the division between *x* and *y*.

Using the cluster evaluation matrix, Cluster Evaluation Indicator, IEC, is calculated, measuring the degree of correctness of clustering reported to the a priori grouping of the analyzed objects, so that:

$$IEC = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} mec_{ij}}{m^2} \times 100$$

The IEC indicator takes values between $[0\%; 100\%]$. *IEC=0%* if no cluster corresponds from the interaction between objects point of view from the same clusters with the initial grouping, and *IEC=100%* if all clusters are identical with the initial categories.

The mapping algorithm to the initial categories in which objects are classified uses a comparison of each category to each cluster resulted after applying clustering methods, resulting in *MAP_CT*, with $MAP\_CT \in \mathcal{M}_{2\times k}(\mathbb{N})$. Because the separation into *r* categories is a priori known, the number of clusters in which the objects are group in after clustering analysis is equal to r, *k=r*. The complexity level of the algorithm is $(k + k^2)$, figure 1.

```
Input:
t[], c[], k
Output:
MAP_CT[2][]
Algorithm MappingClustersIntoCategories
  For each j=1,k
       MAP_CT[1][j] := j
       use[j] := 0
  For each j=1,k
       max := 0
       For each i=1,k
         If (Card(Tj ∩ Ci) > max) AND (use[i]=0)
              max := Card(Tj ∩ Ci)
              MAP_CT[2][j] := i
              use[i] := 1
       If max = 0
         MPA_CT[2][j] := MIN(DIST(t[j],c[i]))
return MAP_CT[2][]
```

Fig. 1 Mapping clusters into categories pseudocode

The mapping function is used to bring to the same denominator the results of different clustering techniques in order to separate the causal space in subspaces of membership of objects to clusters.

IV. METACLUSTER AGGREGATION FUNCTION

The total number of clustering methods is noted with *nr_mtc*, methods that are applied upon an initial set of objects, for which the membership of objects to clusters' formed is returned. The number of clusters formed for each method is constant and equal to k.

A method of aggregation is proposed upon the results of each clustering method, respecting the principle of synergism. With the method of aggregation, the final result increases the degree of correct classification. Let FCA, the function of aggregated classification, $FCA: \mathbb{R}^n \rightarrow \{1,2,\dots,k\}$, that receives as input data a n-dimensional object and returns the class in which it is assigned using all the clustering methods. The method of aggregation uses the majority vote selection.

For applying the majority vote, a k-dimensional vector is formed, that retains the number of classification of the x_i object into each category by the *nr_mtc* clustering methods, so that:

$$FCA(x_i) = i, where\ t_i = \max_i \sum_{k=1}^{nr\_mtc} c_k(i)$$

where:

- t_i represents the aggregated result of the assignation of the object into category i, using *nr_mtc* clustering methods;
- $c_k(i)$ represents the result of the k clustering upon the object x_i.

The function returns the position of the maximum value from the vector, category in which the object x_i was mostly assigned to, following the pseudo code, figure 2.

```
Input:
x[][], nr_obj, k, nr_mtc, c[nr_mtc][k], category[]
Output:
cat
Algorithm MetaCluster
  max := 0
  cat := 0
  For each i=1,nr_mtc
       category[c[i][nr_obj]] := category[c[i][nr_obj]]+1
  For each i=1,nr_mtc
       If category[i] > max
         max := category[i]
         cat := i
return cat
```

Fig. 2 Metacluster pseudo code

With the applying of the majority vote, the level of correctness of the clustering methods isn't taken into account. For that, the weighted majority vote uses a weight for each clustering method equal to the value of IEC indicator. The aggregation techniques of meta-cluster type achieve an optimization of the process of classification, combining the results of all the clustering methods implemented and tested. The weighted majority vote uses the function *FCAP*:

$$FCAP(x_i) = i, where\ t_i = \max_i \sum_{k=1}^{nr\_mtc} IEC_k \times c_k(i)$$

where IEC_k represents the percentage of correct grouping given by the k clustering method.
The combination method of the results of a number of clustering methods proposed in [13] optimizes the process of grouping at the level of cluster mapping, based on the principle of partitioned graphs. The first method avoids the correspondence between the objects problem, with the identification of the pairs of similar objects. The improved method proposed by [13] analyzes the problem of hyper-graph cutting. Another method of correlation between different clustering methods applied upon the same set of objects uses as technique the clustering of the cluster for identifying the level of similarity among clusters.

V. CONCLUSION

For verifying the implementation of k-means algorithm with the four types of distances used, Euclidian distance, Canberra, Manhattan and Cosine, a set of data containing 900 objects from the bi-dimensional space. Table 1 contains the values of the centroids generated by each distance function used, for each four clusters.

Table 1. The centroids generated by k-means algorithm

Distance	Euclidian	Canberra	Manhattan	Cosine
Centroid 1	15 19	14 17	15 20	14 17
Centroid 2	31 5	30 5	31 5	30 5
Centroid 3	35 17	35 16	35 17	35 16
Centroid 4	41 32	38 32	41 32	38 32

The evaluation is done also at the time consuming, the sum of squares of the errors and the correctness of the clustering, using IEC indicator, table 2.

Table 2. K-means algorithm evaluation

Distance	Execution time	Sum of squares of the errors	Cluster's correctness
Euclidian	1 ms	6.996.971	90.83%
Canberra	2 ms	6.732.739	95.15%
Manhattan	2 ms	7.019.798	90.14%
Cosine	3 ms	6.732.739	95.15%

Figure 3 contains the visual results of the objects' membership to one of the four formed clusters after applying k-means clustering algorithm, 3.a) is the result using Euclidian distance, 3.b) implements Canberra distance, figure 3.c) implements Manhattan distance and 3.d) Cosine distance.

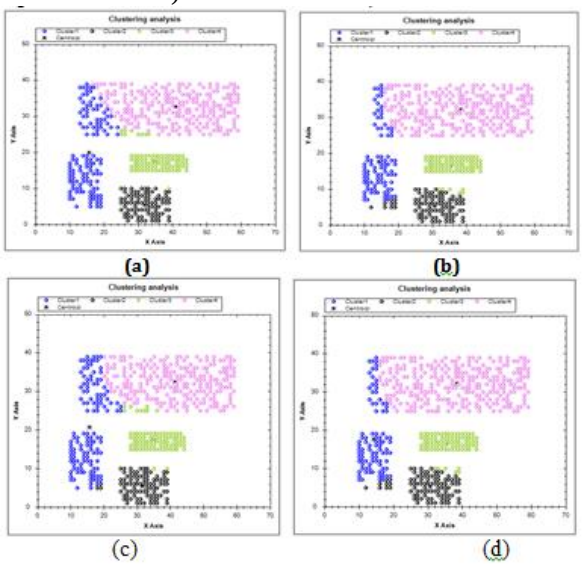

Fig. 3 K-means clustering results

The comparative analysis reveals that the local maximum of the k-means clustering algorithm is achieved with the Canberra distance function, with a level of 95.15% for IEC indicator, having a time consuming of 2 milliseconds and a sum of squares of the errors of 6.732.739.
For the evaluation of the k-medoids clustering algorithm, the same set of data formed out of m=900 objects from the bi-dimensional space is used. Table 3 contains the positions of the centroids resulted after running k-medoids algorithm for each four distance functions implemented.

Table 3. The centroids generated by k-medoids algorithm

Distance	Euclidian	Canberra	Manhattan	Cosine
Centroid 1	15 20	14 17	15 20	14 17
Centroid 2	31 5	30 5	31 5	30 5
Centroid 3	35 17	35 16	35 17	35 16
Centroid 4	41 32	38 32	41 32	38 32

The evaluation of the k-medoids clustering results concerning time execution, sum of squares of the errors and the level of correctness given by ICE indicator is done in table 4.

Table 4. K-medoids algorithm evaluation

Distance	Execution time	Sum of squares of the errors	Cluster's correctness
Euclidian	1 ms	7.039.978	89.91%
Canberra	2 ms	6.705.360	95.11%
Manhattan	2 ms	7.021.290	90.38%
Cosine	3 ms	6.705.360	95.11%

Figure 4 contains the results of objects' membership to one of the four existing clusters, using Euclidian distance (a), Canberra (b), Manhattan (c) and Cosine (d).

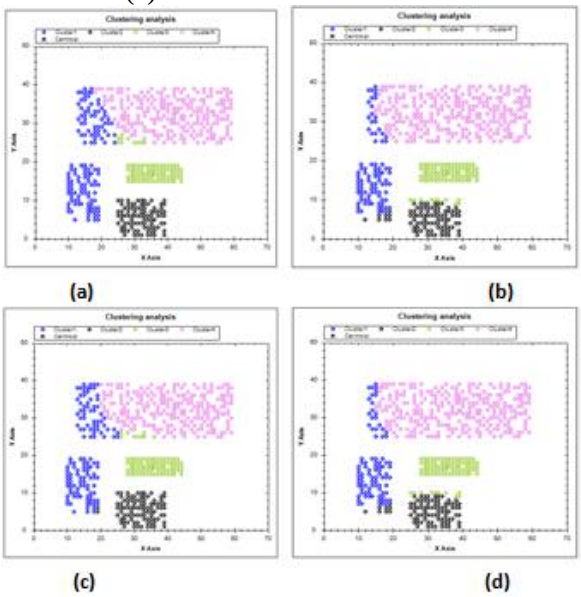

Fig. 4 K-medoids clustering results

The local optimum achieved in k-medoids implementation is given by the combination using Canberra distance function, having a time consuming of 2 milliseconds, the sum of squares of the errors of 6.705.360 and the level of correctness of 95.11%.
Comparing the two local optimum given by k-means and k-medoids, the best results are achieved using the combination of k-means with Canberra distance function, reaching a level of correctness of 95.15%.
For the evaluation of DBSCAN clustering algorithm, the same set of bi-dimensional 900 objects is used. Table 5 contains the result obtained using four types of different distance functions, Euclidian, Canberra, Manhattan and Cosine.

Table 5. DBSCAN algorithm evaluation

Distance	Execution time	Sum of squares of the errors	Cluster's correctness
Euclidian	4 ms	6.544.447	99.90%

In figure 5, the visual membership of the objects to the clusters formed is presented using the Euclidian distance.

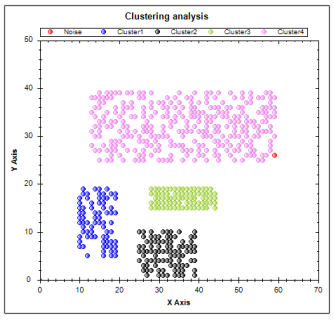

Figure 5. DBSCAN clustering results

DBSCAN clustering algorithm achieves an optimum of 99.90% of correctness of classification, returning only one noise point. In comparison with k-means and k-medoids, DBSCAN performs the best for the initial set of 900 bi-dimensional objects.
Future work is related to the implementation of meta-clusters, aggregate functions that optimize the correctness of clustering. Also, other clustering algorithms will be implemented and aggregated to meet all the requirements of a clustering algorithm.

ACKNOWLEDGEMENTS

This work was cofinaced from the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/107/1.5/S/77213 „Ph.D. for a career in interdisciplinary economic research at the European standards".

REFERENCES

[1] Ruxanda, Gh. „Analiza multidimensională a datelor", *Doctoral course Academy of Economic Studies*, 2011, 133 pp.
[2] MacQueen, J., „Some methods for classification and analysis of multivariante observations", *Proceeding of the 5th Berkeley Symp., Mathematics, Statistics and Probabilities*
[3] Kaufman, L., Rousseew, P.J., „Finding Groups in Data: an Introduction to Cluster Analysis", *John Wiley & Sons*
[4] Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X., „Exploiting Wikipedia as External Knowledge for Document Clustering", 2009
[5] Gokcay, E., Principe, J.C. „A New Clustering Evaluation Function Using Renyi's Information Potential", *International Conference on Acoustics, Speech and Signal Processing*, 2000
[6] Turkay, C., Parulek, J., Reuter, N., Hauser, H. „Integrating Cluster Formation and Cluster Evaluation in Interactive Visual Analysis", *Proceeding Spring Conference on Computer Graphics*, 2011
[7] Harever, M., Brailovsky, V.L. "Probabilistic validation approach for clustering", *Pattern Recognition Letters, Vol. 16*, 1995, pg. 1189-1196, ISSN 0167-8655

[8] Al-Sultan, K.S., Marrof Khan, M. "Computational experience on four algorithms for the hard clustering problem", *Pattern Recognition Letters, Vol. 17*, 1996, pg. 295-308, ISSN 0167-8655

[9] Lee, J.S., Olafsoon, S. "Data clustering by minimizing disconnectivity", *Information Sciences, 2011*, pg. 732-746

[10] Yousri, N.A., Kamel, M.S., Ismail M.A., "A distance-relatedness dynamic model for clustering high dimensional data of arbitrary shapes and density", *Pattern Recognition, Vol. 42*, 2009, pg. 1193-1209

[11] Zalik K.R. "An efficient k-means clustering algorithm", *Pattern Recognition Letters, Vol. 29*, 2008, pg. 1385-1391

[12] Duin, R., Pekalska, E. "The dissimilarity space: Bridging structural and statistical pattern recognition", *Pattern Recognition Letters, Vol. 33*, 2012, pg. 826-832

[13] Sause, M.G.R., Gribov, A., Unwin, A.R., Horn, S. "Pattern recognition approach to identify natural clusters of acoustic emission signals", *Pattern Recognition Letters, Vol. 33*, 2012, pg. 17-23

[14] Leisch, F., "A toolbox for k-centroids cluster analysis", *Computational Statistics & Data Analysis, Vol. 51*, 2006, pg. 526-544, ISSN 0167-9473