I.J. Information Engineering and Electronic Business, **2013, 2, 22-27**
Published Online August 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2013.02.04

Optimal Clustering Algorithms for Data Mining

Omar Y. Alshamesti
Department of Computer science, Palestine Technical Colleges, Al-Aroub, Hebron, Palestine
oshamesti@ptca.edu.ps

Ismail M. Romi
College of Administrative sciences and Informatics, Palestine Polytechnic University, Hebron, Palestine
ismailr@ppu.edu

Abstract— Data mining is the process used to analyze a large quantity of heterogeneous data to extract useful information. Meanwhile, many data mining techniques are used; clustering classified to be an important technique, used to divide data into several groups called, clusters. Those clusters contain, objects that are homogeneous in one cluster, and different from other clusters. As a reason of the dependence of many applications on clustering techniques, while there is no combined method for clustering; this study compares k-mean, Fuzzy c-mean, self-organizing map (SOM), and support vector clustering (SVC); to show how those algorithms solve clustering problems, and then; compares the new methods of clustering (SVC) with the traditional clustering methods (K-mean, fuzzy c-mean and SOM). The main findings show that SVC is better than the k-mean, fuzzy c-mean and SOM, because; it doesn't depend on either number or shape of clusters, and it dealing with outlier and overlapping. Finally; this paper show that; the enhancement using the gradient decent, and the proximity graph, improves the support vector clustering time by decreasing its computational complexity to O(nlogn) instead of $O(n^2d)$, where; the practical total time for improvement support vector clustering (iSVC) labeling method is better than the other methods that improve SVC.

Index Terms **—** Data Mining, Clustering, Self-Organizing Map, Support Vector Clustering, Computational Complexity.

I. INTRODUCTION

In the early 1990's, the establishment of the internet made a huge quantity of data to be stored electronically; therefore, handling this quantity of data became to be necessary. Therefore, data mining emerged to extract useful information from a large quantity of heterogeneous data [1] using several techniques such as clustering. Where clustering divides data into several groups [2] depending on one of the proposed algorithms that have been developed by researchers [3, 4] such as K-mean, fuzzy c-mean, Self Organizing Map (SOM) and Support Vector Clustering (SVC). K-mean is a well-known partitioning method and one of the most popular clustering algorithms used in scientific and industrial applications [5]. Fuzzy c-mean [5, 6] is an iterative algorithm which is frequently used in pattern recognition, it allows one piece of data to be classified to more than one cluster. SOM algorithm can be classified as a powerful method for clustering high dimensional data [7]. SVC [8] is a nonparametric clustering process which depends on Support vector machine (SVM) concepts.

The fact that; there is no fixed method or technique, encourages researchers to keep developing algorithms and techniques to perform clustering in a variety of ways, where part of the studies focus on improving data clustering algorithms [3, 5, 9, 10, 11, 12], or develop new clustering methods [7, 8, 13], the other part focuses on comparing different data clustering algorithm using different factors [5, 12, 14, 15, 16, 17, 18].

This paper will focus on comparing k-mean, fuzzy C-mean, SOM and SVC algorithms to show how those algorithms solve clustering problems, and then compare those traditional methods with the new clustering method; mainly SVC, to find out the improvements and characteristics that reduce the computational complexity of this algorithm. Those comparisons will provide a tool for selecting the best clustering algorithm in specified area such as text mining, geographical information system, and information retrieval that depend on clustering.

This paper is organized as follow: A short description of data mining, k-mean clustering algorithm, fuzzy c-mean algorithm (FCM), self organizing map algorithms (SOM), and support vector clustering (SVC) are included in section 2. Section 3 includes the comparisons among the different data mining algorithms. Section 4, presents the conclusion of this paper, recommendations, and the required future researches.

II. BACKGROUND AND LITERATURE REVIEW

Data mining is the process of analyzing a large quantity of heterogeneous data to extract useful information [1]. This process could be performed using several techniques based on two types of learning paradigms [19]; mainly supervised and unsupervised learning. Clustering is one of those techniques which depend on unsupervised learning paradigm, and used to divide data into several groups; each of which called a cluster. Many algorithms are proposed for data clustering; where prior research's shows that the most used algorithms are k-mean, fuzzy c-mean, SOM and

SVC.

2.1 K-mean clutsering algorithm:

K-mean was invented by Hartigan (1975) to represent each cluster by a mean called centroid [20, 21]. Prior researches find out that K-mean is a well-known partitioning method and the most popular clustering algorithm used in scientific and industrial applications [5]. K-mean aims to minimize the average squared distance of the object from their cluster center; where the cluster center is the mean of the objective in a cluster C as in (1). This algorithm is easy and fast to implement [5, 14], whereas this algorithm has no way to deal with outliers which not belonging to any cluster [5].

$$\mu(c) = \frac{\sum xi}{|c|} \tag{1}$$

c: number of clusters.

K-mean algorithm

1. Chooses the number of clusters, k.
2. Selects k points as an initial centroid of clusters.
3. Classifies each vector into the closest center by Euclidean distance measure.

$$\|xi - ci\| = \min \|\|xi - ci\|\| \tag{2}$$

4. Re-computes cluster center as in (3).

$$C(i) = \frac{\sum xi}{ni} \tag{3}$$

5. If no changes in step 4, stop; otherwise, repeat step 3.

2.2 Fuzzy c-mean algorithm (FCM):

Fuzzy c-mean [5, 6] is an iterative algorithm that can be used in pattern recognition, and allows one piece of data to belong to more than one cluster by a degree of membership by defining the percentage through which the data point belong to the cluster. FCM runs by finding the cluster center that minimizes the dissimilarity function as in (4).

$$jm = \sum \sum_{i=1}^{n} \sum_{j=1}^{n} \bigcup_{ij}^{m} \|xi - cj\| \tag{4}$$

m: a real number greater than 1.
Uij: the degree of membership of Xi in cluster J
Xi: the i^{th} of d-dimensional center of the cluster

$\|*\|$: used to express the similarity between any measured data and the cluster.

Fuzzy c-mean algorithm

1. Initialize $U = [Uij]$ matrix, $U(0)$. $U(0)$
2. Calculate the center vector for each step by computing:

$$Vij = \frac{\sum_{i=1}^{n} \bigcup_{ik}^{m} * Xkj}{\sum_{i=1}^{n} \bigcup_{ik}^{m}} \tag{5}$$

3. Calculate the distance matrix by computing:

$$Dij = \sqrt{\left[\sum_{j=1}^{m} Xkj - Vij\right]} \tag{6}$$

4. Update the membership matrix (U(k), U(k+1) by computing:

$$Uij = \frac{1}{\sum_{j-1}^{c} \left[\frac{xi - cj}{xi - ck}\right]^{\frac{2}{m-1}}} \tag{7}$$

5. If $\|U(k+1) - U(k) < \varepsilon\|$ then stop, otherwise repeat step 2.

2.3 Self organizing map algoritms(SOM):

Self-organizing map was proposed by Chokemen in 1982. It is a powerful method for clustering high dimensional data [7]. SOM algorithm is an artificial neural network used to map the high dimensional data into low dimensional space which is usually two dimensional space called; map (Figure1). This map consists of several neurons or units, each of which is represented by a weight vector [7]. The Chokemen neural network consists of an input and output layers; where, the input layer contains the dataset vectors, while the output layer forms a two dimensional array of nodes. SOM algorithm aims to put the sample unit in the map, and then close together the similar sample units. Where, the virtual units are modified iteratively through the artificial neural network (ANN) during the training process.

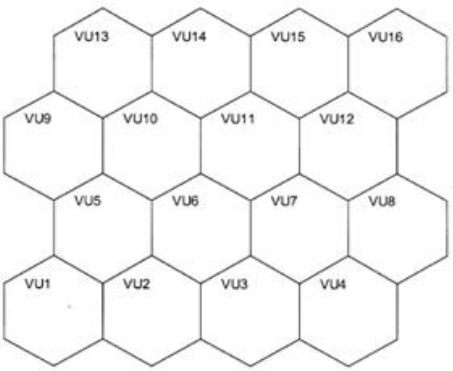

Figure1. A Self-Organizing Map formed by a rectangular Grid with a virtual unit V_{Uk} in each hexagon Source: (Asa et al, 2001) [8]

SOM Algorithm

1. Initialize the virtual units using random sample drawn from the input dataset.
2. Choose a random sample unit as an input unit.
3. Compute the Euclidean distance between the sample unit and each virtual unit W_i.
4. Choose the closest virtual unit to the sample unit as a wining unit or neuron and it is called the best matching unit BMU.
5. Update the virtual unit using the following rule:

$$\omega_{ik}(t+1) = \omega_{ik}(t) + h_{ck}(t)\left[x_{ij}(t) - \omega_{ik}(t)\right] \quad (8)$$

t *: time*

h_{ck} *: is a neighborhood function which can be computed in several ways.*

The most common studies use the Gaussian function:

$$h_{ck}(t) = \exp\left(\frac{\left\|r_k - c_k\right\|^2}{2\sigma^{2(t)}}\right) \quad (9)$$

rk, ck *: the position of neuron t and c in the SOM grid.*

σ *: the learning factor-a decreasing function of the time, where σ converge to 0.*

1. $t = t + 1$.
2. If $t < t_{\max}$, repeat step 2, otherwise stop training.

2.4 Support vector clustering(SVC):

SVC clustering is a nonparametric clustering algorithm that is based on support vector machine (SVM) proposed by Cortes Vapnik in 1995 [15]. David et al [8] proposed this clustering method to search clustering solutions without any assumption of numbers or shapes.

2.4.1 SVC Algorithm

1. Mapping data points from a data space to a high dimensional space- called feature space.
2. Finding the smallest sphere that encloses the image of the data.
3. Mapping the sphere back to the data space.
4. The mapped sphere forms a set of contours that encloses the data points.
5. The set of contours are interpreted as cluster boundaries.
6. The number of clusters can be increased or decreased depending on the kernel width [8]. Where SVC algorithm can deal with outlier using soft margin constrains, and with overlapping cluster using a large value of kernel width.
7. Optimization stage which performed as follow:

   a. Looking for the smallest sphere that encloses a set of data points using (10), and then a soft constrains are employed to allow some data point to be enclosed in the sphere by adding a slack variable using (11).

$$\left\|\Phi(xj - a\right\|^2 \le R^2 \quad (10)$$

$$\left\|\Phi(xj - a\right\|^2 \le R^2 + \varepsilon_j \quad (11)$$

   b. Using lagrangian multiplier to perform optimization as in (12).

$$L = (R^2 + \varepsilon_{j-}\left\|\Phi(x_j) - a^2\right\|\beta_j\mu_j + C\sum\varepsilon_j \quad (12)$$

   c. Deriving (12) with respect to R, to produce the following results:

$$\sum\beta_j = 1 \quad (13)$$

$$a = \sum\beta_j\Phi(x_j) \quad (14)$$

$$\beta_j = C - \mu_j \quad (15)$$

   d. Applying Kuhn-Tucker complementary condition [8]; by using the equality constrains from (13), which will result in:

$$\varepsilon_j\mu_j = 0 \quad (16)$$

$$(R^2 + \mu_j - \left\|\Phi(x_j - a)\right\|^2)\beta_j = 0 \quad (17)$$

The above equations will produce three types of points [8]:

- Points with $\varepsilon_j > 0$ and $\beta_j > 0$ lie outside the hyper sphere in feature space, which is called Bounded Support vector or BSV.
- Points with $0 < \beta_j < C$ lie on the surface, which is called Support Vector or SV.
- The other points lie inside the sphere.

   e. Using the appropriate kernel function; such as Gaussian kernel [8], to represent the dot product:

$$K(x_i, x_j) = e^{-q\left\|x_{i-}x_j\right\|^2} \quad (18)$$

   f. The distance from the sphere center to each point is defined as [4]:

$$R^2(x) = \left\|\Phi(x) - a\right\|^2 \quad (19)$$

Therefore, equation (19), and the definition of the kernel can be concluded in:

$$R^2(x) = k(x,x) - 2\sum_j\beta_jK(x_i - x_j) + \sum_{ij}\beta_i\beta_jK(x_i, x_j) \quad (20)$$

2.4.2 Cluster assignment:

David et al [8] used a method called that complete graph (CG) to differentiate the data point that belongs to different clusters, considering that any path is connecting pairs of points which belongs to different clusters must exit from the sphere. Therefore, the authors used the definition of the adjacency matrix A_{ij} among pairs of points x_i and x_j as follow:

{1 if R(y) < R 0, otherwise} (20)

2.4.3 SVC complexity:

The time complexity for kernel evaluation according to testing benchmarks for SVC algorithm proposed by David et al [4] is $O(n^2)$, while the time complexity for clustering and labeling part is $O(N - n_{bsv})^2 n_{sv} d$; where n_{bsv} is the bounded support vector, n_{sv} is the

number of support vectors, and *d* is the dimensionality, therefore; the overall complexity is $O(n^2d)$.

2.4.4 SVC enhancement:

Many techniques are proposed to improve the clustering labeling process for SCV; such as support vector graph, Proximity graph technique, Gradient decent technique (GD), and Improved support vector clustering.

Support vector graph:

This method [22] proposed as a modification of the method proposed by David et al [8]. Where, instead of checking the linkage among all pairs of data, a linkage only among points and support vectors were considered. This method takes $O(N-n_{bsv})^2 n_{sv}d$; where n_{bsv} the number of bounded support vector is, n_{sv} is the number support vectors and *d* is the dimensionality.

Proximity graph technique:

Despite, the ability of SVC algorithm which is proposed by David et al [8] to deal with outliers, and to make a cluster of arbitrary shape, it still suffering from two problems in the cluster labeling process; mainly, its low efficiency when the number of support vector increase, and producing a false negative. Therefore, Jianhua et al [10] presents a new clustering assignment method based on proximity graph [5, 9]. In this technique; instead of calculating the adjacency matrix coefficient *xi and xj* for each pair of points, it calculates A_{ij} only for pairs of points x_i and x_j; where x_i and x_j are a *SV* [22]. The A_{ij} coefficients are calculated for the point x_i and x_j; where those points are linked by an edge with time complexity *O (nlogn)*.

Gradient decent technique (GD):

The gradient decent method was proposed by Lee and Lee [11] to treat the problem of clusters labeling strategy in [8, 10]. Even; the method proposed by David et al [4] is easy to implement, its time complexity is $O(n^2d)$. Furthermore, despite the ability of the method discussed in [10] to reduce the time complexity of David et al [8] to $O(n\log n)$, it fails frequently in labeling the cluster correctly [11]. The gradient decent method solves the problem by decomposing data set into a small number of disjoint groups. Each group is represented by its candidate point, and all points that belong to the same cluster. The candidate points are labeled; which result in labeling the whole data points with $O(n\log n)$ time complexity.

Improved support vector clustering:

The previous methods for SVC [8, 10, 11] are still suffering from two important problems; which are computational cost, and poor labeling performance. Ling et al [12] proposed a new support vector clustering method to overcome these problems. This method performs a reduction strategy on the data set to extract the qualified subset of the data. This reduction strategy depends on Schrodinger [12] equation; by presenting a new labeling strategy to label the separate vector first, and then label the other data based on labeled SVs.

The optimization part of this approach is $_{O(M^3)}$, which is lower than the time taken by SVC which is $_{O(N^3)}$. Table 1 shows that; the time for SVC, and iSVC in real data set. The overall time taken by this strategy is $O(Nsv)^3 + N - Nsv$; where the time taken to decompose the eigen value is $_{O(Nsv)^3}$; *sv* is the number of support vector, and the time taken to label the other data is $_{O(N-N_{sv})}$. Table 2 compares the time taken by this approach with the other labeling techniques.

Table1: Time Comparison Between SVC and iSVC

	SVC		**iSVC**	
	Size	***Time***	***Subset size***	***Time***
Liver	354	115.1	100	0.661
Sonar	208	3.32	60	0.093
wine	178	2.32	52	0.087
Iris	150	9.09	46	0.138
Vote	435	126.6	125	0.811
Diabetes	768	261.3	219	5.687
Ionosphere	351	55.47	104	0.507

Source: (Sairam and Sowndary, 2011) [23]

Table2: Time Comparison for Labeling Approaches

	CG1	**SVG**	**PG**	**GD**
Liver	657	202	109	131
Vote	815	286	119	89
Ionosphere	1069	301	187	205

Source: Asa et al, 2001), [24]

III. COMPARISON

Pawan [5, 2] presents a comparative study that compares k-mean, and Fuzzy c-mean, in terms of time, and space complexity. This study was carried out on MATLAB, and shows that the time and space complexity for HCM are $O(ncdi)$, and $O(cd)$ respectively, time and space complexity for FCM are $O(ndc^2i)$ and $O(nd+nc)$ respectively; where n is the number of data point, c is the number of clusters, i is the number of iterations, and d is the number of dimensions. Tables 3, 4, and 5 show the results of this comparison.

Table 3: Time Comparison for FCM AND HCM

Number of Cluster	FCM Time Complexity	HCM Time Complexity
1	3000	3000
2	12000	6000
3	27000	9000
4	48000	12000
20	900	8

Source: (Abbas, 2008) [14]

Table 4: Space Comparison for FCM AND HCM

Number of Cluster	FCM Time Complexity	HCM Space Complexity
5	450	2
10	600	4
15	700	6

Source: (Abbas, 2008) [14]

Table 5: Time and Space Comparison for FCM AND HCM

Algorithm	Time Complexity	Space Complexity
HCM	*k*	*cd*
FCM	O(ndc^{2}i)	*O(nd+nc)*

Source: (Han and Kamber, 2006) [1]

Abbas [5] compares those algorithms; in term of size of dataset, number of clusters, type of dataset and type of software used. Table 6 and figure 2 show the results.

Table 6: The Relationship Between the Number of Clusters and Algorithm Performance

Performance		
Number of Cluster	***SOM***	***K-mean***
8	59	63
16	67	71
32	78	84
64	85	89

Source: (Anil, 2010) [4]

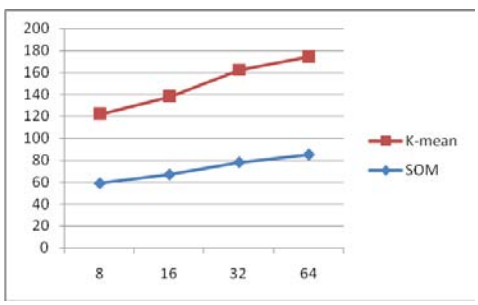

Figure2. The relationship between number of clusters, and algorithm performance. Source: (Anil, 2010) [2]

As a Comparison for different SVC enchantments, Table 7 shows that; the proximity graph and gradient decent method have the best time. But the practical total time for iSVC labeling method is the best among the other methods [20].

Table 7: Time and Complexity Analysis for SVC Improvements

Complete graph	Support vector graph	Proximity graph	Gradient decent	Improved SVC
$O(n^2d)$	$O(n - n_{bsv})n_{sv}^2$	$O(nlogn)$	$O(nlogn)$	$O(N_{sv})^3$

IV. CONCLUSION AND RECOMMENDATION:

The current study; compares two important groups of clustering algorithms; mainly, parametric and non-parametric clustering algorithm. K-mean and fuzzy c-mean are a parametric clustering algorithms that requires determining the number of clusters in a prior, where SOM and SVC are a nonparametric algorithms that don't require prior knowledge about the number of clusters and constructs. As a conclusion; fuzzy c-mean algorithm requires more time and space than k-mean, and SOM has a better performance over k-mean. Furthermore this study discusses the different enchantments for SVC; such as complete graph labeling strategy, support vector graph, proximity graph, gradient decent strategy, and improvement support vector clustering, where the comparisons show that SVC is better than other clustering methods; because it solves many problems which are not solved by the other clustering algorithms. Where from the SVC improvements, iSVC shows a better practical total time labeling than the other methods.

This study finds that; the iSVC method solves many problems, which are not solved by the other clustering algorithms, and deals with outlier, and overlapping; by controlling the kernel width and the soft margin constrains, besides to the better practical total time labeling than the other methods. Where; further researches are required to empirically test those findings.

REFERENCES

[1] Han. J., Kamber. M., (2006). Data mining: Concepts and technique, 2nd ed. *Morgan Kaufmann: USA.*

[2] Pawan, K. Pankaj, V and Rakesh, S., (2010). Comparative analysis of fuzzy c mean andhard c mean algorithm. *international journal of information technology and knowledge management*, 2(1): pp. 1-5.

[3] Jain, A.K., Murty, M.N., Flynn, P.J., (2000). Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3): pp. 264-323.

[4] Anil, K.J., (2010). Data Clustering: 50 Years Beyond K-Means. *Journal of Pattern Recognition Letters*, 31(8).
[5] Estivill, V.C and Lee, I. A., (2000). Hierarchical clustering based on spatial proximity using Delaunay diagram. *In Proc. of the 9th Int. Symposium on Spatial Data Handling*, pp 26 – 41.
[6] Karthikeyani, V., Suguna, J., (2009). K-Means Clustering using Max-min Distance Measure. *The 28th North American Fuzzy Information Processing Society Annual Conference (NAFIPS2009)*.
[7] Fedja, H and Tharam,S.D., (2005). CSOM: Self-Organizing Map for Continuous Data. *3rd IEEE international Conference on Industrial Informatics (INDIN)*.
[8] Asa, B.H., David, H., Hava T. S and Vladimir, V., (2001). Support Vector Clustering. *Journal of Machine Learning Research*, 2(1): pp. 125-137.
[9] Rajashree, D., Debahuti, M., Amiya, K.R., Milu.A., (2010). A hybridized K-means clustering approach for high dimensional dataset. *International Journal of Engineering, Science and Technology*. 2(2): pp. 59-66.
[10] Vladimir, Jianhua, Y., E.C, and Stephan, K.C., (2003). Support Vector clustering Through Proximity Graph Modeling. *IEEE*, 2: pp. 898 – 903.
[11] LEE, D., LEE, J., (2007). DOMAIN DESCRIBED SUPPORT VECTOR CLASSIFIER FOR MULTI-CLASSIFICATION PROBLEMS. PATTERN RECOGNITION, 40(10: PP. 41-51.
[12] Ling, P, Zhou, C.G and Zhou, X., (2010). Improved support vector clustering. Engineering Applications of Artificial Intelligence. *Elsevier*, 23(4): pp. 552-559.
[13] Hsiang, C.L., Jenz, M.Y.,Wen,C.L., Tung, S.Liu., (2009). Fuzzy c-means algorithm based on PSO and mahalanobis disyance. *International Journal of Innovative Computing, Information and Control*, 5(12): pp. 5033–5040.
[14] Abbas, O., (2008). Comparison between Data clustering algorithms. *The international journal of information technology*, 5(3): pp.320-325.
[15] Hanafi, G., Abdulkader, S., (2006). Comparison of clustering algorithm for analog modulation classification. *Expert Systems with Applications: An International Journal*, 30(4): pp. 642-649.
[16] Satchidanandan, D. Chinmay, M. Ashish, G and Rajib, M., (2006). A comparative study of clustering algorithms. *Information technology journal*, 5(3): pp 551-559.
[17] Velmurugan, T and Santhanam, T., (2010). Computational complexity between k-mean and k-medoids clustering algorithm for normal and uniform distribution of data points. *Journal of computer science*, 6(3): pp. 363-368.
[18] kumar, P., Siri, K.W., (2010). Comparative analysis of k-mean based algorithms. *International journal of computer science and network security*, 10(4): pp.314-318.
[19] Scholkopf. B., Smola. A.J., (2002). Learning with Kernels. *London. MIT press*.
[20] Hartigan, J. A. and M. A. Wong (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 28.1: pp. 100-108.
[21] Hartigan, J. A. (1975). Clustering Algorithms (Probability & Mathematical Statistics). *John Wiley & Sons Inc*.
[22] Hammouda, K.M, (2008). A comparative study of data clustering techniques. *International journal of computer science and information technology*, 5(2), pp. 220-231.
[23] Sairam, Manikandan and Sowndary., (2011). Performance analysis of clustering algorithms in detecting outliers. *International journal of computer science and information technology*, 2(1): pp.486-488.
[24] Asa, B.H, David, H and Vapnik. V., (2001). A support vector method for hierarchical clustering. *MIT press*.

Authors' Profiles:

Omar Y. Alshamesti is a lecturer at Palestine Technical Colleges-AL-Aroub, and Al-Quds Open University. He received his master degree in informatics in 2011 from Palestine Polytechnic University. He taught topics in programming languages and network management. He worked as a programmer and Network Administrator in several organizations. His area of interest is machine learning, cloud computing, and information management.

Ismail M. Romi is an assistant professor of information systems at Palestine Polytechnic University. He received his BSc, MBA, and PhD in Business & information systems. He has more than 15 years experience in information systems, and is the author of over 8 peer-reviewed scientific publications and conference papers, and responsible of conducting many exhibitions, workshops and Symposiums, and scientific days. His current research interest include information management, information systems modeling and assessment, managing and planning information systems and technology, statistical databases, human computer interaction, and business strategic planning & alignment with information systems.

