Quality Scheme Assessment in the Clustering Process

Conference Paper *in* Lecture Notes in Computer Science · January 2000

DOI: 10.1007/3-540-45372-5_26 · Source: dx.doi.org

3 authors, including:

Maria Halkidi
University of Piraeus
51 PUBLICATIONS **3,656** CITATIONS
SEE PROFILE

Michalis Vazirgiannis
École Polytechnique
284 PUBLICATIONS **9,117** CITATIONS
SEE PROFILE

Quality scheme assessment in the clustering process

M. Halkidi, M. Vazirgiannis, I. Batistakis
Dept. of Informatics, Athens University of Economics & Business
Patision 76, 10434, Athens, Greece (Hellas)
{mhalk, mvazirg, yannis}@aueb.gr

Abstract. Clustering is mostly an unsupervised procedure and most of the clustering algorithms depend on assumptions and initial guesses in order to define the subgroups presented in a data set. As a consequence, in most applications the final clusters require some sort of evaluation. The evaluation procedure has to tackle difficult problems, which can be qualitatively expressed as: i. quality of clusters, ii. the degree with which a clustering scheme fits a specific data set, iii. the optimal number of clusters in a partitioning. In this paper we present a scheme for finding the optimal partitioning of a data set during the clustering process regardless of the clustering algorithm used. More specifically, we present an approach for evaluation of clustering schemes (partitions) so as to find the best number of clusters, which occurs in a specific data set. A clustering algorithm produces different partitions for different values of the input parameters. The proposed approach selects the best clustering scheme (i.e., the scheme with the most compact and well-separated clusters), according to a quality index we define. We verified our approach using two popular clustering algorithms on synthetic and real data sets in order to evaluate its reliability. Moreover, we study the influence of different clustering parameters to the proposed quality index.

1 Introduction

Data Mining is a step in the KDD process that is mainly concerned with methodologies for knowledge extraction from large data repositories. There are many data mining algorithms that accomplish a limited set of tasks and produce a particular enumeration of patterns over data sets. These main tasks, according to well established data mining methods [5], are: the definition/extraction of clusters, the classification of database values into the categories defined, and the extraction of association rules or other knowledge artifacts.
More specifically regarding clustering, the outcome (from now on referred to as *clustering scheme*) is a set of clusters characterized by their center and their limits (i.e., the bounds within which the values of the cluster objects range). In the vast majority of KDD approaches, the clustering algorithms partition a data set in a number of groups based on some parameters, such as the desired number of clusters, the minimum number of objects in a cluster, the diameter of a cluster etc. For instance, assume the data set presented in Fig. 1a. It is obvious that we can discover three clusters in the given data set. However, if we consider a clustering algorithm (e.g. K-Means) with certain parameter values (in the case of K-means the number of clusters) so as to partition our data set in four clusters, the result of clustering process would be the clustering scheme presented in Fig. 1b. In our example the clustering algorithm (K-Means) found the best four clusters in which our data set could be partitioned. We define the term "optimal" clustering scheme as the outcome of running a clustering algorithm (i.e., a partitioning) that best fits (i.e., resembles) the real partitions of the data set. It is obvious from Fig. 1b that the clustering scheme presented in it does not fit well the data set. The optimal clustering for our data set will be a scheme with three clusters. As a consequence, if the clustering algorithm parameters are assigned an improper value, the clustering method may result in a partitioning scheme that is not optimal for the specific data set leading to wrong decisions. The problems of deciding the number of clusters better fitting a data set as well as the evaluation of the clustering results has been subject of several research efforts [14]. Although various validity indices have been introduced, in practice most of the clustering methods do not use any of them. Furthermore, formal methods in database and data mining applications for finding the best partitioning of a data set, are very few [12], [17].

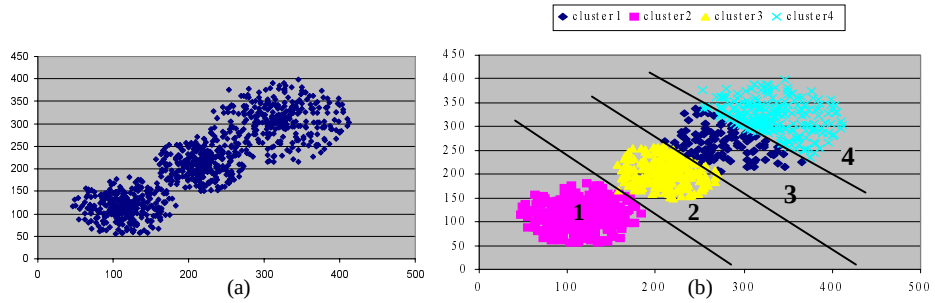

Fig. 1. a. A data set that consists of three clusters, b. The results from the application of K-means when we ask four clusters

In this paper, we present an approach for clustering scheme evaluation based on relative criteria. It aims at evaluating the schemes defined by a specific clustering algorithm, assuming different input parameter values. These schemes are evaluated using a clustering scheme quality index, which we define. Our goal it is not to propose a new clustering algorithm or to evaluate a variety of clustering algorithms, but to produce the clustering scheme with the most compact and well-separated clusters.
For the remainder of this paper, our approach can be applied and lead to good results, under the assumption that the data set can be partitioned to groups. This will prevent a misleading interpretation of the structure of a data set, when we apply a clustering algorithm. We also consider only the case of compact clusters, that is the clusters of our data set are non-ring shaped (Fig. 1). This is the most common structure of clusters presented in business databases. Moreover there are few clustering algorithms that discover successfully arbitrary shaped clusters.
The paper is organized as follows. Section two surveys related work in this area. Section three describes the proposed quality index for clustering scheme evaluation along with a qualitative and experimental evaluation of our approach. In section four, we present the algorithm for finding the best clustering scheme and we present the results of an evaluation study for our approach. We conclude in section five by summarizing and providing further research directions.

2 Related work

Clustering is one of the main tasks in the data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. The fundamental clustering problem is to partition a given data set into groups (clusters), such that the data points in a cluster are more similar to each other than points in different clusters [8]. In the clustering process, there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data [1]. This is what distinguishes clustering from classification. Classification is a procedure of assigning a data item to a predefined set of categories [6].
There is a multitude of clustering methods available in literature, which can be broadly classified into the following types [13]: i) *Partitional clustering*, ii) *Hierarchical clustering*, iii) *Density-Based clustering, iv) Grid-based.* For each of the types there exists a wealth of subtypes and different algorithms [1], [4], [9], [14], [18] for finding the clusters. In general terms, the clustering algorithms are based on a criterion for judging the quality of a given partitioning. More specifically, they take as input some parameters (e.g. number of clusters, density of clusters) and attempt to define the best partitioning of a data set for the given parameters. Thus, they define a partitioning of a data set based on certain assumptions and *not* the "best" one that fits the data set.
Since clustering algorithms discover clusters, which are not known a-priori, the final partitioning of a data set requires some sort of evaluation in most applications [11]. Another

important issue in clustering is to find out the number of clusters that give the optimal partitioning. A particularly difficult problem, which is often ignored in clustering algorithms is "how many clusters are there in the data set?". Though this is an important question that causes much discussion, the formal methods for finding the optimal number of clusters in a data set are few [12].

Previously described requirements for the evaluation of clustering results is well known in the research community and a number of efforts have been made especially in the area of pattern recognition [14]. However, the issue of cluster validity is rather under-addressed in the area of databases and data mining applications, even though recognized as important. In general terms, there are three approaches to investigate cluster validity [14]. The first is based on *external criteria*. This implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the data set. The second approach is based on *internal criteria*. We may evaluate the results of a clustering algorithm in terms of quantities that involve the vectors of the data set themselves (e.g proximity matrix). The third approach of clustering validity is based on *relative criteria*. Here the basic idea is the evaluation of a clustering structure by comparing it with other clustering schemes, resulting by the same algorithm but with different parameter values. The two first approaches are based on statistical tests and their major drawback is their high computational cost. Moreover, the indices related to these approaches aim at measuring the degree to which a data set confirms an a-priori specified scheme.

Our work puts emphasis to the third approach of clustering evaluation, based on relative criteria. In this approach, there are two criteria proposed for clustering evaluation and selection of an optimal clustering scheme [1]: i) *Compactness*, the members of each cluster should be as close to each other as possible, ii) *Separation*, the clusters themselves to be widely spaced.

A number of cluster validity indices are described in literature using the above criteria. A cluster validity index for crisp clustering proposed in [3], attempts to identify "compact and separated clusters". Other validity indices for crisp clustering have been proposed in [2] and [10]. The implementation of most of these measures is very computationally expensive, especially when the number of clusters and number of objects in the data set grows very large [17]. For fuzzy clustering, Bezdek proposed the partition coefficient (1974) and the classification entropy (1984). The limitations of these measures are: i) their monotonous dependency on the number of clusters and ii) the lack of direct connection to the geometry of the data [2]. Other fuzzy validity measures are proposed in [7], [11], [17]. We should mention that the evaluation of proposed measures and the analysis of their reliability are limited.

Another approach for finding the best number of cluster of a data set proposed in [12]. It introduces a practical clustering algorithm based on Monte Carlo cross-validation. This approach differs significantly from the one we propose. While we evaluate clustering schemes based on widely recognized quality criteria of clustering, the evaluation approach proposed in [12] is based on density functions considered for the data set. Thus, it uses concepts related to probabilistic models in order to estimate the number of clusters, better fitting a data set, while we use concepts directly related to the data. Our approach is based on the inter-cluster and intra-clusters distances.

In the following section we introduce the details of our approach concerning the evaluation of the clustering schemes defined in data mining process.

3 Our Approach of clustering scheme evaluation

The problem can be stated as follows: Given a data set of *n* objects containing non-categorical data, we aim at the definition of a *clustering scheme*, representing the best

partitioning of the specific data set based on a well defined quality index. In the following sections we elaborate on our approach.

3.1 Quality of Clustering Schemes

The objective of the clustering methods is to provide optimal partitions of a data set. In general, they should search for clusters whose members are close to each other (or have a high degree of similarity) and well separated. Another problem we face in clustering is to decide the optimal number of clusters that fits better a data set. The majority of clustering algorithms produce a partitioning based on the input parameters (e.g. number of clusters, clusters density etc) that finally lead to a finite number of clusters. Thus, the application of an algorithm assuming different input parameters values results in different partitions of a particular data set, which are not easily comparable. A solution to this problem is to run the algorithm repetitively with different input parameter values and compare the results against a well-defined validity index. The best partitioning of our data set will maximize (or minimize) the index. In this paper we aim at the evaluation of clustering schemes produced by applying repeatedly the same algorithm with different input parameters, and at the selection of the best clustering scheme. This process is based on an index that is better described by the term *clustering scheme quality index.*

A reliable quality index must take into account both the compactness and the separation of clusters in a partitioning [11]. Thus, the reliability of an index is based on the two fundamental criteria of clustering quality and more specifically to the degree an index combines both criteria. Many validation indices are proposed in the literature. However, the evaluation of the proposed indices and the analysis of their reliability are rather under-addressed in the literature [12] and especially in the area of databases and data mining.

In the sequel, we describe a clustering scheme quality index that can be used in crisp clustering and we analyse its reliability.

3.2 Quality index for clustering schemes

In this section, we define a quality index for evaluating clustering schemes based on concepts and the relative validity indices proposed in the literature.

Definitions[*]**.**

Variance of data set. The variance of a data set X, is called *σ(X)*. The value of the p_{th} dimension is defined as follows,

$$\sigma_x^p = \frac{1}{n}\sum_{k=1}^{n}\left(x_k^p - \bar{x}^p\right)^2 \quad (1)$$

where $\bar{x}^p$ is the p_{th} dimension of $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} x_k, \forall x_k \in X$ (2)

Variance of cluster i. The variance of cluster *i* is called $\sigma(v_i)$ and its p_{th} dimension is given by

$$\sigma_{v_i}^p = \sum_{k=1}^{n_i}\left(x_k^p - v_i^p\right)^2 \Big/ n_i \quad (3)$$

Quality index definition. The quality index of our approach is based on the validity index defined for FCM in [11]. We use the same concepts for validation, but transform them into equivalent ones for crisp clustering. Following, we give the fundamental definition for this index.

Average scattering for clusters. The average scattering for clusters is defined as

$$Scat(c) = \frac{1}{c}\sum_{i=1}^{c} \left\|\sigma(v_i)\right\| \Big/ \left\|\sigma(X)\right\| \quad (4)$$

[*] n=number of clusters, n_i = number of objects in cluster i, c = number of clusters, v_i = Center of cluster i, x_k^p= p_{th} dimension of x_k object, $||x_k|| = (x_k^T x_k)^{1/2}$

Total separation between clusters. The definition of total scattering (separation) between clusters is given by equation (5)

$$Dis(c) = \frac{D_{\max}}{D_{\min}} \sum_{k=1}^{c} \left(\sum_{z=1}^{c} \| v_k - v_z \| \right)^{-1} \quad (5)$$

where D_{max} = max($\|v_i - v_j\|$) $\forall i, j \in \{1, 2,3,\ldots,c\}$ is the maximum distance between cluster centers. The D_{min} = min($\|v_i - v_j\|$) $\forall i, j \in \{1, 2,3,\ldots,c\}$ is the minimum distance between cluster centers.

Now, we can define a quality index based on equations (4) and (5), as follows

$$SD(c) = a\, Scat(c) + Dis(c) \quad (6)$$

where α is a weighting factor equal to $Dis(c_{max})$ where c_{max} is the maximum number of input clusters. The first term (i.e., *Scat(c)* defined by (4)) indicates the average compactness of clusters (i.e., intra-cluster distance). A small value for this term indicates a compact cluster and as the scattering within clusters increases (i.e., they become less compact) the value of *Scat(c)* also increases. The second term *Dis(c)* indicates the total separation between the *c* clusters (i.e., an indication of inter-cluster distance). Contrary to the first term the second one, *Dis(c)*, is influenced by the geometry of the clusters centers and increase with the number of clusters. It is obvious for previous discussion that the two terms of *SD* are of the different range, thus a weighting factor is needed in order to incorporate both terms in a balanced way. The number of clusters, c, that minimizes the above index can be considered as an optimal value for the number of clusters present in the data set.

3.3 Evaluation study of clustering scheme quality index

In this section we present the evaluation of the quality index both qualitatively and experimentally.

Qualitative evaluation. Considering the SD index definition (eq. (6)), we can observe the following:

1. It uses cluster separation as an indication of the average scattering between clusters, minimizing thus the possibility to select a clustering scheme with significant differences in cluster distances. Also, we should mention that the total separation between clusters, *Dis(c)*, is influenced by the distribution of the cluster centers in space. As a consequence the quality index *SD* takes into account the geometry of the centers, which also influences the values of *SD*.
2. The maximum number of clusters influences the index *SD*. It is clear, from the index definition, that its values are equally affected by *Scat* (eq. (4)) and *Dis* (eq. (5)) due to the weighting factor definition, $\alpha = Dis(c_{max})$. Thus, different values of c_{max} (i.e., weighting factor) affect proportionally *SD* values. The influence of the weighting factor is an issue for further study as mentioned in [11]. However, to the best of our knowledge in recent literature there is no well-established study to analyze the effect of the weighting factor. In the sequel, we present the results of a study that we performed in order to evaluate the influence of the factor *a* to the *SD* values.

Experimental evaluation. In this study we have used real data sets to verify the results of the preceding theoretical study. More specifically, we use the Iris data set [19]. This is a biometric data set of measurements referring to flowers. The schema of this data set is R={sepal_length, sepal_width, petal_length, petal_width}. Also, we implemented and used for our study, the partitioning clustering algorithm K-Means [1] in order to define the clustering schemes. However, we should mention that any other clustering algorithm could be used. Our focus to the K-Means algorithm for our study is justified by the following: i) It is widely used in a variety of applications, ii) Our purpose is not to evaluate a variety of clustering algorithms but to illustrate a procedure for finding the optimal partitioning of an array of clustering schemes.

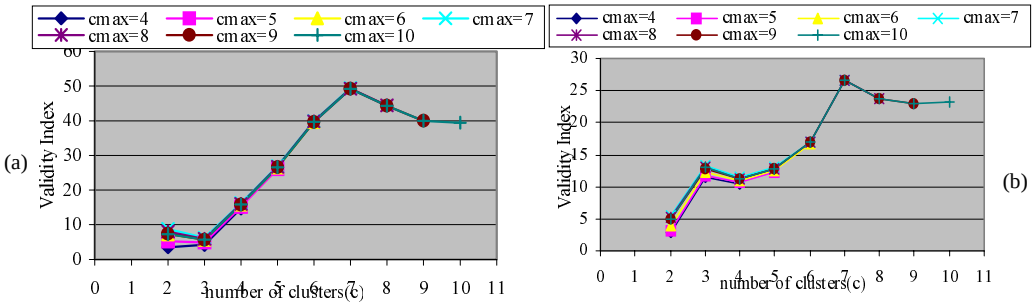

Fig. 2a: Quality index *SD* as a function of the number of clusters in :a) a two-dimensional data set (petal_length,petal_width), b) a four-dimensional data set (sepal_length, petal_length, petal_width, sepal_width).

The results of our experiments can be summarized as follows:

- *The index SD evaluates clustering schemes considering scattering within and between clusters.* Fig. 2a, b show the values of *SD* as a function of the number of clusters, considering two and four-dimensional data sets respectively. From these figures it is clear that the number of clusters does not influence the SD index. So it does not decrease monotonously as the number of clusters increases. Indeed the *SD* index behaves in a homogeneous and expected way almost independently of the c_{max}, c_{min} values for a particular data set. The study of *SD* in different data sets and with a variety of attributes shows that the values of *SD* are influenced by the geometry of the centers of clusters. The different shape of graphs in Fig. 2a, b is due to the difference in the geometry of the clusters' centers.
- *The index SD is slightly influenced by the maximum value of clusters' number.* As Figures 2a, b depict, the distributions of the index values for different values of c_{max} is similar. It is also noteworthy that there is no significant influence of c_{max} value on the value of SD, corresponding to the optimal clustering scheme (minimum value of SD). Hence SD results in the optimal clustering scheme almost irrespectively of the maximum number of clusters. We can only observe a slight decrease to the minimum value of the SD index when the value c_{max} becomes very small. For instance (Fig.2a) the best number of clusters is three when the c_{max} ranges from 10 to 5 while it is two when c_{max} = 4.

Our experiments verify that aforementioned results are valid in data sets with higher dimensionality.

4 Finding the "best" clustering scheme

4.1 An algorithm for clustering scheme evaluation

In previous sections we have defined a quality index for clustering scheme evaluation. We exploit this index during the clustering process in order to define the optimal number of clusters, c_{opt}, and as a consequence, the optimal clustering scheme for our dataset. More specifically, we define the range of input parameters of a clustering algorithm. Different cluster algorithms require different input parameters (e.g. number of clusters, number of objects in each cluster, least number of points in a cluster etc). Let us call *p* this set of parameters associated with a clustering algorithm. The range of values for *p* is defined by an expert, so that the clustering schemes produced are compatible with expected data set partitions. For example, if we consider the Iris data set described in previous sections, it would be meaningless to search for clustering schemes where the number of clusters is too small (i.e., one) or too big (i.e., twenty). A suitable range of the number of clusters for the given example will be [2, 10]. Then, a clustering algorithm is performed for each value *p* and the results of clustering are evaluated using the *SD* quality index (Sect. 3.2). The number of

clusters at which *SD* reaches its minimum value and also a significant local change (i.e., decrease) in its value occurs, can be considered as the best one. The basic steps of the algorithm that defines the optimal clustering scheme using *SD* as quality index is described as follows:

```
1.  Define the value ranges of the clustering algorithm input parameters.
    Let p_max and p_min are the parameters, resulting in schemes with the max
    and min number of clusters respectively.
2.  Initialize p ← p_max ;
3.  Run the clustering algorithm using p. The output is a clustering scheme
    of c clusters for the data set.
   if (p=p_max)
        a ←Disc(c_max);
        indexValue←SD(c);
       dist_between_indexValues=0;
       c_opt←c }
   else  if (SD(c)<indexValue)
    if (dist_between_indexValues/abs(SD(c) –SD(c-1)))<3
     c_opt←c;
     indexValue ← SD(c);
     dist_between_indexValues= SD(c) –SD(c-1);
  5.  p ← p-1,
     if (p=p_max-1)
      stop
    else   goto 3.
```

The previously described procedure aims at choosing the optimal clustering scheme for a specific data set, which can be partitioned into compact clusters. It is based on: i) the iterative application of a clustering algorithm assuming different parameter values as input, and ii) the calculation of the SD quality index (see Sect. 3.2) for each clustering scheme that discovered by an algorithm. Let $[c_{min}, c_{max}]$ the range of the number of clusters, *c*. The clustering algorithm is repeated for successive values of *c* starting from c_{max} and decreasing to c_{min}. The choice of the optimal cluster number is the one for which: i) SD reaches a local minimum (there number of clusters *c*), ii) the slope of the linear segment (c_{i-1}, c_i) is greater than the inclination of any other linear segment (c_k, c_{k+1}), $k \in [c_{min}, c_{max}]$. We have considered for our algorithm implementation that the best quality index value corresponds to the point at which the index reaches its minimum value and at the same time the change in its value is not less than 1/3 of the previous significant change in the index value. For instance, as Fig. 4b depicts the value of SD is smaller for two than for three clusters. However, the change of the index value between three and two clusters is not significant comparing it with the previous significant change between four and three clusters (it is less than 1/3 of the index value change occurring between four and three). Thus, we conclude that the best number of clusters for Iris data set is three.

4.2 Time Complexity

It is obvious from above description that the time complexity for the definition of the optimal clustering scheme depends on the complexity of clustering algorithm we consider (*O(f(n,c,d))*) and on the time complexity for the quality index calculation. The complexity for computing the quality index *SD*, is: $O(ndc + c^2d)$ where *d* is the number of attributes (data set dimension), *c* is the number of clusters, *n* is the number of database tuples. According to our approach, the clustering algorithm and the computation of the quality index are performed for each value of the parameter p in $[p_{min}, p_{max}]$. Thus, the overall complexity for finding the optimal number of clusters is $O((p_{max}-p_{min}+1)(f(c,n,d) + ncd + c^2d))$.

For example, in the case we use K-Means (time complexity is *O(tcnd)*) as a clustering algorithm, the overall complexity for finding the optimal number of clusters is $O((c_{max}-c_{min}+1)(tncd+c^2d))$. The graphs in Fig. 3a, b show the results of an experimental study referring to the execution time of our approach using K-Means as the clustering algorithm.

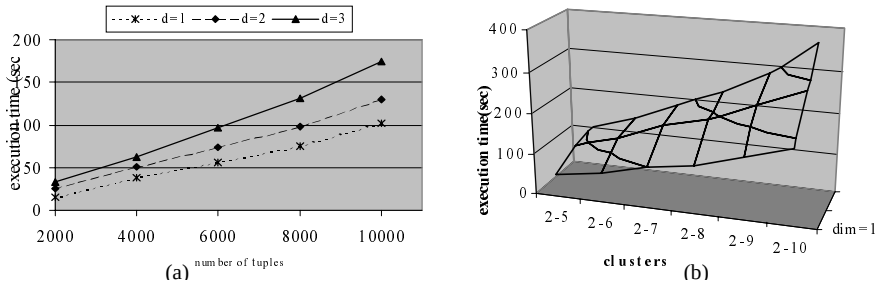

Fig. 3. a. Execution time as function of number of tuples, **b.**Execution time in seconds as function of the number of tuples

4.3 Evaluation study of our approach - Experimental Results

Based on the algorithm described in Section 4.1, we implemented a clustering system for evaluating the clustering schemes and selecting the scheme that fits a data set based on a specific clustering algorithm. It is a system, which is implemented in Java and uses ODBC to connect to a database. At the current stage of development, we have implemented and applied the K-Means algorithm in order to define the clustering schemes. Nevertheless, our approach can also be applied by considering any other clustering algorithm on our data set. This claim was verified by using a hierarchical algorithm (single link algorithm)[14] on our data set. Using our system, we experimented with synthetic and real data sets so as to evaluate our approach as regards the selection of the best clustering scheme for a data set that a specific clustering algorithm defines.

Data Sets description. We evaluate the proposed approach for finding the optimal clustering scheme using two data sets: a real one based on Iris Data Set (see Sect. 3..3) and a synthetic one. A good partitioning of the data set in three groups of flowers can be obtained if we use only two features and more specifically the attributes petal_length, and petal_width of the Iris Data Set [11]. The second data set is a synthetic one and includes 10000 two- dimensional data points (Fig 5a). We produce the data set using normal distribution.

Discussion on experimental results. The goal of this experiment is to evaluate our approach regarding the selection of the optimal clustering scheme that a specific clustering algorithm can define. More specifically, we apply two-dimensional clustering to Iris Data Set considering the attributes petal_length, and petal_width. The clustering schemes are discovered using the K-means algorithm while its input parameters (number of clusters) take values between 2 and 10. Applying the steps of the algorithm (see Sect. 4.1) our system results in a clustering scheme of three clusters. We can verify this result based on Fig. 2a, which presents the quality index values as a function of the number of clusters. We observe there, that the clustering scheme with three clusters corresponds to the minimum value of quality index and also is the point at which a significant local decrease in its value occurs. Thus, according to our system three clusters is the optimal c (Fig. 4a) for our data set, which is also the number of clusters proposed by experts.

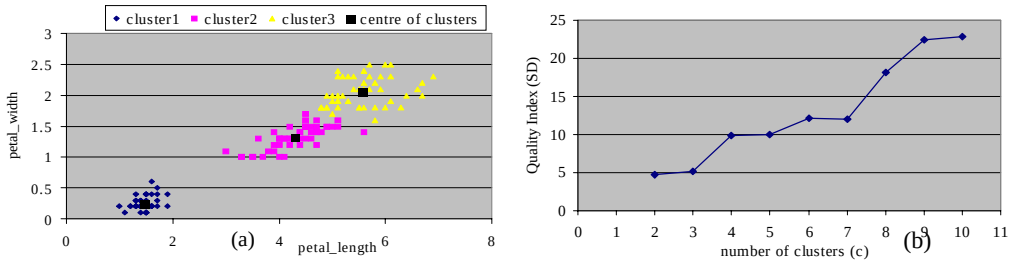

Fig 4. a .The best clustering scheme of Iris Data set, **b.** The graph of SD versus the number of clusters that corresponds to a clustering scheme of Iris Data set defined by applying an hierarchical clustering algorithm.

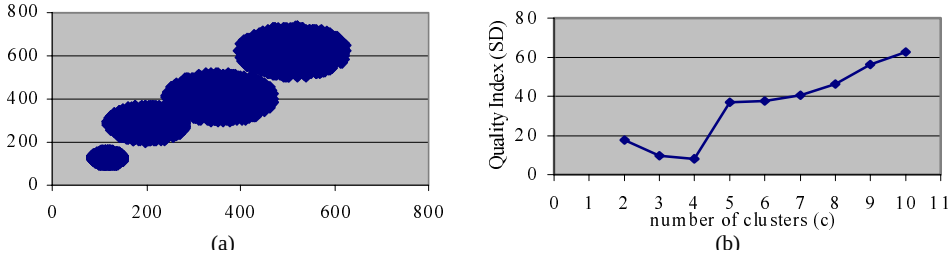

Fig 5. a.A data set that contains four clusters. **b**. The graph of SD versus the number of clusters (using K-Means).

We carried out a similar study using the synthetic data set described above. It is clear from Fig 5a that our data set consists of four clusters. This was verified by our system as the clustering algorithm (K-Means) parameter values (i.e., number of clusters) range from 2 to 10. According to the graph presented in Fig. 5b the number of clusters in which SD reaches its minimum value and also a significant decrease of its value occurs, is four.
In the sequel, we demonstrate that our approach is independent of the clustering algorithm used.
We consider the Iris Data Set (Fig. 4a) and we apply the single link algorithm [14] in order to discover clustering schemes. The algorithm will produce various schemes based on the range of values for the number of clusters that we define (e.g. [2, 10]). The result is a tree (called dendrogram) where each level corresponds to a specific number of clusters. Then, we evaluate each of the clustering schemes based on the *SD* quality index. The graph of *SD* versus the number of clusters (c) is presented in Fig. 4b. According to this graph, the optimal clustering scheme contains three clusters as in the case of K-Means. The above proves that our approach can be easily applied to any data set with compact clusters irrespective of cluster algorithm. We have to stress that our approach does not evaluate schemes discovered by different algorithms. It aims at selecting the optimal clustering scheme among the schemes that a specific clustering algorithm defines considering different input parameter values.

5 Conclusions and Further Work

In this paper, we presented an approach for the extraction of optimal clustering schemes (i.e., schemes that correspond and best fit the real data set partitioning). More specifically:

- We defined the quality index *SD*, for clustering scheme evaluation, based on the fundamental quality criteria for clustering (compactness and separation). The quality index is a variant of the indices defined for the FCM algorithm in [11], adapted to crisp clustering algorithms. Its definition is based on intra-cluster and inter-cluster distances.
- We evaluated the behavior of *SD* considering different values for the clustering algorithm input parameters (i.e., number of clusters, max number of clusters). The results of the experimental study are summarized as follows: i) the *SD* index reaches a local minimum as the number of clusters ranges in $[c_{min}, c_{max}]$. This happens because *SD* takes into account both of the quality criteria (compactness and separation), ii) the value of *c* at which *SD* reaches its minimum and also a significant local decrease of *SD* occurs, is the optimal number of clusters, iii) *SD* proposes an optimal number of clusters almost irrespectively of the maximum number of clusters, c_{max}.
- We evaluated the results of our approach using real and synthetic data sets with known partitioning structure. The experimental study demonstrated that the quality index *SD* can identify the best clustering scheme defined by an algorithm while we assume different input parameter values. Our approach is independent on the clustering algorithm used to partition the data set.

Further work will be concentrated in the following issues:

- *Comparison of the clustering scheme quality index* proposed in this paper with a number of known validity indices. We will choose and implement some known validity indices presented in relevant literature. Then, we will carry out an evaluation study considering specific data sets to compare the performance of our approach regarding the selection of the best clustering scheme with the other indices.
- *Revealing and handling of the uncertainty in data mining process*. The system presented in this paper is part of a larger scale effort that aims at extracting useful knowledge and handling uncertainty in all stages of the KDD process. This effort has resulted in a classification scheme that handles uncertainty and supports decision-making [15]. We aim at enhancing our clustering evaluation scheme in order to deal with uncertainty. Then the results of the clustering process will directly be used as a classification scheme described in [15][16].

Acknowledgements

We would like to thank C. Amanatidis for his help in the experimental study and N. Berdenis for his comments on the paper style.

References

1. Michael J. A. Berry, Gordon Linoff. Data Mining Techniques For marketing, Sales and Customer Support. John Willey & Sons, Inc, 1996.
2. Rajesh N. Dave. "Validating fuzzy partitions obtained through c-shells clustering", *Pattern Recognition Letters*, Vol .17, pp613-623, 1996
3. J. C. Dunn. "Well separated clusters and optimal fuzzy partitions", *J. Cybern.* Vol.4, pp. 95-104, 1974.
4. Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proceedings of 2nd Int. Conf. On Knowledge Discovery and Data Mining*, Portland, OR, pp. 226-231, 1996.
5. Usama Fayyad, Ramasamy Uthurusamy. "Data Mining and Knowledge Discovery in Databases", *Communications of the ACM*. Vol.39, No11, November 1996.
6. Usama M. Fayyad, Gregory Piatesky-Shapiro, Padhraic Smuth and Ramasamy Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI Press 1996
7. Gath, B. Geva. "Unsupervised Optimal Fuzzy Clustering". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 11, No7, July 1989.
8. Sudipto Guha, Rajeev Rastogi, Kyueseok Shim. "CURE: An Efficient Clustering Algorithm for Large Databases", *Published in the Proceedings of the ACM SIGMOD Conference*, 1998.
9. Alexander Hinneburg, Daniel Keim. "An Efficient Approach to Clustering in Large Multimedia Databases with Noise". *Proceeding of KDD '98*, 1998.
10. Zhexue Huang. "A Fast Clustering Algorithm to Cluster very Large Categorical Data Sets in Data Mining", *DMKD*, 1997
11. Ramze Rezaee, B.P.F. Lelieveldt, J.H.C Reiber. "A new cluster validity index for the fuzzy c-mean", *Pattern Recognition Letters*, 19, pp237-246, 1998.
12. Padhraic Smyth. "Clustering using Monte Carlo Cross-Validation". KDD 1996, 126-133.
13. C. Sheikholeslami, S. Chatterjee, A. Zhang. "WaveCluster: A-MultiResolution Clustering Approach for Very Large Spatial Database". *Proceedings of 24th VLDB Conference, New York*, USA, 1998.
14. S. Theodoridis, K. Koutroubas. *Pattern recognition*, Academic Press, 1999
15. M. Vazirgiannis, "A classification and relationship extraction scheme for relational databases based on fuzzy logic", in the proceedings of the *Pacific-Asian Knowledge Discovery & Data Mining '98 Conference*, Melbourne, Australia.
16. M. Vazirgiannis, M. Halkidi. "Uncertainty handling in the datamining process with fuzzy logic", in the proceedings of the IEEE-FUZZ conference, San Antonio, Texas, May, 2000.
17. Xunali Lisa Xie, Genardo Beni. "A Validity measure for Fuzzy Clustering", *IEEE Transactions on Pattern Analysis and machine Intelligence*, Vol13, No4, August 1991.
18. A.K Jain, M.N. Murty, P.J. Flyn. "Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No3, September 1999.
19. Fisher,R.A. Machine readable .names file for MLC++ library. July, 1988