Attribute Selection for High Dimensional Data Clustering

Lydia Boudjeloud and François Poulet

ESIEA Recherche
38, rue des docteurs Calmette et Guérin,
Parc Universitaire de Laval-Changé,
53000 Laval-France
(e-mail: boudjeloud,poulet@esiea-ouest.fr)

Abstract. We present a new method to select an attribute subset (with few or no loss of information) for high dimensional data clustering. Most of existing clustering algorithms loose some of their efficiency in high dimensional data sets. One possible solution is to use only a subset of the whole set of dimensions. But the number of possible dimension subsets is too large to be fully parsed. We use a heuristic search for optimal attribute subset selection. For this purpose we use the best cluster validity index to first select the most appropriate cluster number and then to evaluate the clustering performed on the attribute subset. The performances of our new approach of attribute selection are evaluated on several high dimensional data sets. Furthermore, as the number of dimensions used is low, it is possible to display the data sets in order to visually evaluate and interpret the obtained results.
Keywords: Attribute Selection, Clustering, Genetic Algorithm, Visualization.

1 Introduction

Data collected in the world are so large that it becomes more and more difficult for the user to access them. Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [Fayyad *et al.*, 1996]. The KDD process is interactive and iterative, involving numerous steps. Data mining is one step of the Knowledge Discovery in Databases (KDD) process. This paper focus on clustering in high dimensional data sets, which is one of the most useful tasks in data mining for discovering groups and identifying interesting distributions and patterns in the underlying data. Thus, the goal of clustering is to partition a data set into subgroups such that objects in each particular group are similar and objects in different groups are dissimilar [Berkhin, 2002]. In real world clustering situations, with most of algorithms the user has first to choose the number of clusters. Once the algorithm has performed its computation the clustering method must be validated. To validate the clustering algorithm results we usually compare them with the results of other clustering algorithms or with the results obtained

by the same algorithm while varying its own parameters. We can also validate the obtained clustering algorithms results using some validity indexes described in [Milligan and Cooper, 1985]. Some of these indexes are based on the maximization of the sum of squared distances between the clusters and the minimization of the sum of squared distances within the clusters. The objective of all clustering algorithms is to maximize the distances between the clusters and minimize the distances between every object in the group, in other words, to determine the optimal distribution of the data set. The idea treated in this paper is to use the best index (according to Milligan and Cooper, it is the Calinski index), to first select the most appropriate number of clusters and then to validate the clustering performed on a subset of attributes. For this purpose we use attribute selection methods successfully used to improve cluster quality. These algorithms find a subset of dimensions to perform clustering by removing irrelevant or redundant dimensions. In section 2, we start with a brief description of the different attribute subsets search techniques and the clustering algorithm we have chosen (without forgetting that our objective is not to obtain a better clustering algorithm but to select a pertinent attribute subset with few or no loss of information for clustering). In section 3, we describe the methodology used to find the optimal number of clusters then we describe our search strategy and the method to qualify and select the subset of attributes. In section 5, we comment the obtained results and visualize the results to try to interpret them before the conclusion.

2 Attribute subset search and clustering

Attribute subset selection problem is mainly an optimization problem which involves searching the space of possible attribute subsets to identify one that is optimal or nearly optimal with respect to f (where $f(S)$ is a performance measure used to evaluate a subset S of attributes with respect to criteria of interest) [Yang and Honavar, 1998]. Several approaches of attribute selection have been proposed [Dash and Liu, 1997], [John *et al.*, 1994], [Liu and Motoda, 1998]. Most of these methods focus on supervised classification and evaluate potential solutions in terms of predictive accuracy. Few works [Dash and Liu, 2000], [Kim *et al.*, 2002] deal with unsupervised classification (clustering) where we do not have prior information to evaluate potential solution. Attribute selection algorithms can broadly be classified into categories based on whether or not attribute selection is done independently of the learning algorithm used to construct the classifier: filter and wrapper approaches. They can also be classified into three categories according to the search strategy used: exhaustive search, heuristic search, randomized search. Genetic algorithms [Goldberg, 1989] include a class-related randomized, population-based heuristics search techniques. They are inspired by biological evolution processes. Central to such evolutionary systems is the idea of a population

of potential solutions that are members of a high dimensional search space. We have seen this decade, an increasing use of this kind of methods. Related works can be found in [Yang and Honavar, 1998]. However, all tests of the different authors are performed on data sets having less than one hundred attributes. The large number of dimensions of the data set is one of the major difficulties encountered in data mining. We are interested in high dimensional data sets, our objective is to determine pertinent attribute subsets in clustering, for this purpose we use genetic algorithm population-based heuristics search techniques using validity index as fitness function to validate optimal attribute subsets. furthermore, a problem we face in clustering is to decide the optimal number of clusters that fits a data set, that is why we first use the same validity index to choose the optimal number of clusters. We apply the wrapper approach to k-means clustering [McQueen, 1967], even if the framework presented in this paper can be applied to any clustering algorithm.

3 Finding the number of clusters

When we are searching for the best attribute subset, we must choose the same number of clusters than the one used when we run clustering in the whole data set, because we want to obtain a subset of attributes having same information (ideally) on the one obtained in the whole data set. [Milligan and Cooper, 1985] have compared thirty methods for estimating the number of clusters using four hierarchical clustering methods. The criteria that performed best in these simulation studies with a high degree of error in the data is a pseudo F-statistic developed by [Calinski and Harabasz, 1974]: it is a measure of the separation between clusters and is calculated by the formula: $\frac{S_b/(k-1)}{S_w/(n-k)}$, where S_b is the sum of squares between the clusters, S_w the sum of squares within the clusters, k is the number of clusters and n is the number of observations. The higher the value of this statistic, the greater the separation between groups. We first use the described statistic (Calinski index) to find the best number of clusters for the whole data set. The method is to study the maximum value max_k of i_k (where k is the number of clusters and i_k the Calinski index value for k clusters). For this purpose, we use the k-means algorithm [McQueen, 1967] on the Colon Tumor data set (2000 attributes, 62 points) from the Kent Ridge Biomedical Data set Repository [Jinyan and Huiqing, 2002], Segmentation (19 attributes, 2310 points) and Shuttle (9 attributes, 42500 points) data sets from the UCI Machine Learning Repository [Blake and Merz, 1998]. We compute all Calinski index values where k takes values in the set (2, 3,. . . , a maximum value fixed by the user) and select the maximum value max_k of the Calinski index and the corresponding value of k. The index evolution according to the different values of k for the Shuttle data set is shown in the figure 1 (we search the maximal value of the curve). We notice that the optimal value of Calinski index is obtained effectively for k=7. We obtain k=7 for Segmentation and

Shuttle data sets and k=2 for Colon Tumor data set. The optimal values found are similar to the original number of classes. Of course, these data sets are supervised classification data sets we have removed the class information. Now we try to find an optimal combination of attribute subset with a genetic algorithm having the Calinski index as fitness function. Our objective is to find a subset of attributes that best represent the configuration of the data set and discover the same configuration of the clustering (number, contained data, ...) for each cluster. The number of cluster is the value obtained for the whole data set and we search the attribute subset that has optimal value of Calinski index. The validity indexes give a measure of the quality of the resulting partition and thus usually can be considered as a tool for the experts in order to evaluate the clustering results. Using this approach of cluster validity our goal is to evaluate the clustering results in the attribute subset selected by the genetic algorithm.

4 Genetic algorithm for attribute search

Genetic algorithms (GAs) [Goldberg, 1989] are stochastic search techniques based on the mechanism of natural selection and reproduction. We use standard genetic algorithm with usual parameters (population, mutation probability), variation of these parameters have no effect for the convergence of our genetic algorithm. Our genetic algorithm starts with a population of 60 individuals (chromosomes) and a chromosome represents a combination (subset) of dimensions. The visualization of the data set is a crucial verification of the clustering results. With large multidimensional data sets (more than some hundred dimensions) effective visualization of the data set is difficult as shown in the figure 2.

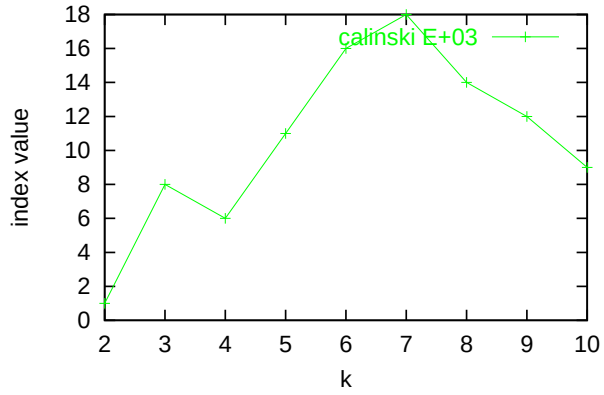

Fig. 1. Calinski index evolution for the Shuttle data set.

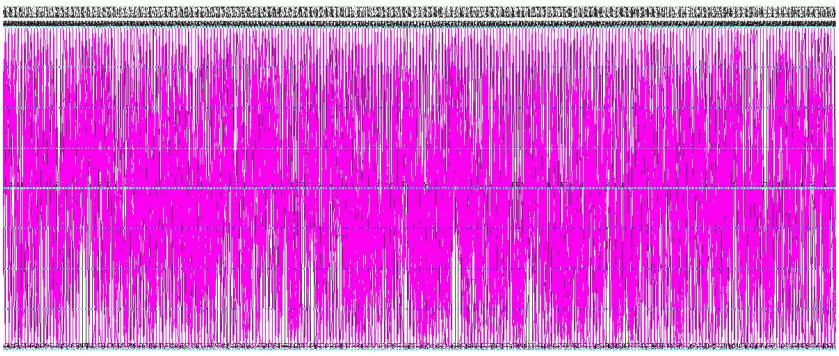

Fig. 2. Visualization of one hundred dimensions of Lung cancer data set.

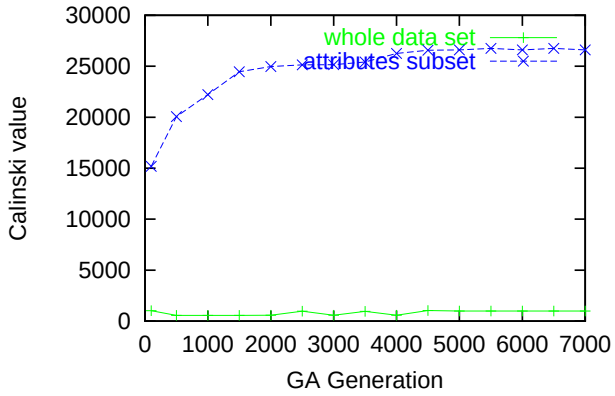

Fig. 3. Calinski index evolution for the Segmentation data set along genetic algorithm generations.

This is why the individuals (chromosomes) use only a small subset of the data set dimensions (3 or 4 attributes), we have used the same principle for outlier detection in [Boudjeloud and Poulet, 2004]. We evaluate each chromosome of the population with the Calinski index value. This procedure finds the combination of dimensions that best represents the data set with the same k as obtained for the whole data set and search attribute subset that have optimal Calinski index value. Once the whole population has been evaluated and sorted, we operate a crossover on two parents chosen randomly. Then, one of the children is muted with a probability of 0.1 and is substituted randomly for an individual of the second part of the population, under the median. The genetic algorithm ends after a maximum number of iterations. The best element will be considered as the best subset to describe the whole data, we will visualize the data set according to this most pertinent attribute subset.

5 Tests and results

We have tested GA with size 4 for the subset of attributes for the Segmentation and the Colon tumor data sets and size 3 for the Segmentation data set. Figure 3 shows the evolution of the Calinski index for all generations of the genetic algorithm for the Segmentation data set. We can see a large gap between the indexes computed with the whole data set and the indexes calculated with a subset of attributes. Our objective was to try to find the same index value for a subset of attributes as the one obtained with the whole data set. The obtained results show that the values of the indexes with the subset of attributes are better than those obtained with the whole data set. One can explain this by the fact that the data set can be noisy according to some attributes and when we select some other attributes we can get rid of the noise and therefore we obtain better results. To confirm the obtained results, we have performed tests to verify the clustering result in the different subsets of attributes that are supposed to be optimal and compared these results with the clustering obtained in the whole data set. We have used the Calinski index as reference because it is classified as the best index by Milligan and Cooper. The results with the colon Tumor data set are shown in table 1. This table describes different values obtained when we change

	Whole data set 2000 att.	Whole data set 2000 att.	Data set 20 att.	Data set 20 att. **GA opt.**	Data set 4 att.	Data set 4 att. **GA opt.**
Nbr. clusters (k)	2	3	2	2	2	2
Nbr. elemt./Cluster	18/44	10/30/22	11/51	48/14	11/51	**18/44**
Calinski	**28.91**	21.88	41.66	**56.06**	79.84	**88.50**

Table 1. GA optimization results.

the value of k (cluster number), we illustrate the obtained index values when k=2 and k=3, the optimal value is obtained for k=2 with 18 objects in the cluster number 1 and 44 objects in the cluster number 2. We have tested the program for a subset of 20 attributes, we describe in the third column the results obtained when we compute different index values for a subset of 20 randomly chosen attributes, after this we apply the GA to optimize the result of the index. We obtain a better Calinski index with object affectation not very different from the whole data set. We also tested our program for a subset of 4 attributes and we have obtained the optimal values described in the table (last 2 columns) for the subset of attributes: 1089, 890, 1506, 1989. We note that the cluster content for this optimal subset is similar to the cluster content in the whole data set. We presented the optimal solution of GA i.e. the subset of attributes, which has obtained the optimal values of

all indexes. Then we visualize these results using both parallel-coordinates [Inselberg, 1985] and 2D scatter-plot matrices [Carr *et al.*, 1987], to try to explain why these attribute subsets are different from the other ones. These kinds of visualization tools allow the user to see how the data are presented in this projection. For example, figure 4 shows the visualization of clustering, with the optimal subset of attributes obtained by the GA and we can see a separation between the two clusters.

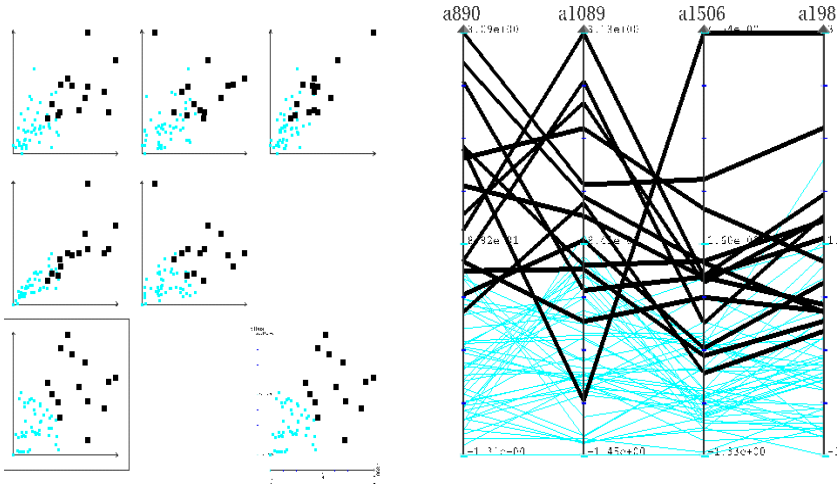

Fig. 4. Optimal subset visualization for the Colon data set.

6 Conclusion and future work

We have presented a way to select the cluster number and to evaluate a relevant subset of attributes in clustering. We used validity index of clustering algorithm not to compare clustering algorithms, but to evaluate a subset of attributes as a representative one or pertinent one for clustering results. We have used the k-means clustering algorithm, the best validity index (Calinski index) described by [Milligan and Cooper, 1985] and a genetic algorithm for the attribute selection, having the value of the validity index as fitness function. We introduced a new representation of genetic algorithm individual, our choice is fixed on small sizes of attribute subsets to facilitate visual interpretation of the results and then show the relevance of the attributes for clustering application. Nevertheless, the user is free to set up the size

of the attribute subset and there is no complexity problem with the size of the population of genetic algorithm. Our first objective is to obtain subsets of attributes that best represent the configuration of the data set (number, contained data). When we tested our method by verifying clustering results we notice that the optimal subset obtained has optimal value for the index with a number of elements in the clusters similar to the ones in the whole data set and they have the same elements. Furthermore, as the number of dimensions is low, it is possible to visually evaluate and interpret the obtained results using scatter-plot matrices or/and parallel coordinates. We must keep in mind that we work with high dimensional data sets. This step is only possible because we use a subset of dimensions of the original data. This interpretation of the results would be absolutely impossible if considering all the set of dimensions (figure 2). We think to follow our objective that is to find the best attribute combination to reduce the research space without any loss in result quality. We must find a factor or a fitness function for the genetic algorithm qualifying attribute combination to optimize the algorithm and improve execution time. We think also to involve more intensively the user in the process of cluster search in data subspace [Boudjeloud and Poulet, 2005].

References

[Berkhin, 2002]P. Berkhin. Accrue software: Survey of clustering data mining techniques. In *Working paper*, 2002.

[Blake and Merz, 1998]C.L. Blake and C.J. Merz. Uci repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[Boudjeloud and Poulet, 2004]L. Boudjeloud and F. Poulet. A genetic approach for outlier detection in high dimensional data sets. In *Modelling, Computation and Optimization in Information Systems and Management Sciences, MCO'04*, pages 543–550. Le Thi H.A., Pham D.T. Hermes Sciences Publishing, 2004.

[Boudjeloud and Poulet, 2005]L. Boudjeloud and F. Poulet. Visual interactive evolutionary algorithm for high dimensional data clustering and outlier detection. In *to appear in proc. of The Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining*. PAKDD'05, 2005.

[Calinski and Harabasz, 1974]R.B. Calinski and J. Harabasz. A dendrite method for cluster analysis. In *Communication in statistics*, volume 3, pages 1–27, 1974.

[Carr *et al.*, 1987]D. B. Carr, R. J. Littlefield, and W. L. Nicholson. Scatterplot matrix techniques for large n. *Journal of the American Statistical Association*, 82:424–436, 1987.

[Dash and Liu, 1997]M. Dash and H. Liu. Feature selection for classification. In *Intelligent Data Analysis*, volume 1, 1997.

[Dash and Liu, 2000]M. Dash and H. Liu. Feature selection for clustering. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 110–121, 2000.

[Fayyad *et al.*, 1996]U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. In *AI Magazine*, volume 17, pages 37–54, 1996.

[Goldberg, 1989]D.E. Goldberg. *Genetic Algorithms in Search: Optimization and Machine Learning*. Addison-Wesley, 1989.

[Inselberg, 1985]A. Inselberg. The plane with parallel coordinates. In *Special Issue on Computational Geometry*, volume 1, pages 69–97, 1985.

[Jinyan and Huiqing, 2002]L. Jinyan and L. Huiqing. Kent ridge bio-medical data set repository. 2002. http://sdmc.-lit.org.sg/GEDatasets.

[John *et al.*, 1994]G. John, R. Kohavi, and K. Pfleger. Irrelevant features and subset selection problem. In Morgan Kaufmann New Brunswick, NJ, editor, *the eleventh International Conference on Machine Learning*, pages 121–129, 1994.

[Kim *et al.*, 2002]Y. Kim, W. N. Street, and F. Menczer. Evolutionary model selection in unsupervised learning. volume 6, pages 531–556. IOS Press, 2002.

[Liu and Motoda, 1998]H. Liu and H. Motoda. Feature selection for knowledge discovery and data mining. In *Kluwer International Series in Engineering and Computer Science, Secs*, 1998.

[McQueen, 1967]J. McQueen. Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[Milligan and Cooper, 1985]G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. volume 50, pages 159–179, 1985.

[Yang and Honavar, 1998]J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. In *IEEE Intelligent Systems*, volume 13, pages 44–49, 1998.