Dimension Reduction for Visual Data Mining

Edwige Fangseu Badjio and François Poulet

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

Abstract. We present a method for dimension reduction applied to visual data mining in order to reduce the user cognitive load due to the density of data to be visualized and mined. We use consensus theory to address this problem: the decision of a committee of experts (in our case existing attribute selection methods) is generally better than the decision of a single expert. We illustrate the choices operated for our algorithm and we explain the results. We compare successfully these results with those of two widely used methods in attribute selection, a filter based method (LVF) and a wrapper based method (Stepclass).

Keywords: visual data mining, dimension reduction, feature selection, filter, wrapper, consensus.

1 Introduction

The quantity of stored data doubles every 9 months, these data are not useful if at least a part of information they contain is not extracted. It is the goal of knowledge discovery in the databases (KDD) which can be defined as the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [Fayyad *et al.*, 1996]. In most of data mining (a step of KDD) approaches, the process of discovering correlations in data sets is performed in an automatic way. For users, understanding and explaining data with only automatic algorithms results can be difficult. Visual data mining is a new data mining approach using visualization as a communication channel for data mining. It lies in tightly coupling the visualizations and analytical process into one data mining tool that takes advantage of the strengths of all worlds [Wong, 1999]. Visualization is the process of transforming information into a graphical representation allowing the user to perceive and interact with the information.

Visual representation allows understanding data, determining what should be done about it. The human eye can capture complex patterns and relationships. Compared to data mining, the advantages of visual data mining are:

- the confidence in the results is improved, the KDD process is not just a "black box" giving more or less comprehensible results,

- the quality of the results is improved by the use of human pattern recognition capabilities,
- if the user is the data specialist, we can use the domain knowledge during the whole process (and not only for the interpretation of the results).

Computer devices can display vast amount of information with various techniques. This information must be appropriately communicated to us in order to make the best use of it. According to [Ware, 2000], in order to be visualized, data are passed through four basic stages : independently of any visualization technique, the first step of visualization is data collection and storage. Secondly, there is a data pre-processing which goal is to transform the data into a comprehensive form. At the third step, display hardware and software are used to produce a visual representation of the data. Lastly, the users perceive, interact with the visual representation and mine it. It is necessary to address the limits of human perception. When the collected data are multidimensional, there are some limits in the third and fourth steps.

For [Ferreira and Levkowitz, 2003], the conceptual boundary between low and high-dimensional data is round three to four data attributes. Their suggested guideline for characterizing dimensionality is the following: low: up to four attributes, medium: five to nine attributes and high: 10 or more. When the number of dimensions is over some dozen, the large number of axes needed to create these displays tends to overcrowd the figure, limiting the value of the plot for detecting patterns or other useful information.

We are interested in visual data mining methods performing supervised classification. Our objective is to select some dimension of a data set in order to create a visualization from which relevant information can be extracted. We want to identify attributes that are significant in order to reduce dimensionality. Dimension reduction can be used to improve the efficiency of visualization of large, multidimensional data sets and may be the accuracy of algorithms used for classification in visual data mining.

Knowing that:

1. an optimal subset of attributes is not necessarily unique,
2. the visualization of more than a dozen attributes is unusable for visual data mining,
3. without investigation, it is not possible to determine a dimension reduction method that can perfectly reduce the set of attributes (by taking account of different trade-offs between performance and complexity (tolerate lower performance in a model that also require less features)),
4. the decision of a committee of experts is generally better than the decision of a single expert,

we use a meta-analysis algorithm based on consensus theory for dimension reduction in visual data mining. The proposed algorithm combines decisions of several experts (in our case feature selection algorithms). More precisely, it maps a given set of dimension subsets to a single dimension subset.

The rest of this paper is organized as followed: section 2 explains the context of dimension reduction. In section 3, we present the visual data mining domain, the specificities related to this domain and the task analysis. Next, there is an explanation of the specificities of dimension reduction applied to visual data mining which allow us to design our dimension reduction method. Section 4 introduces this method before experiments, conclusion and future work.

2 Dimension reduction

Many techniques for the visualization of multidimensional data have been developed: pixel oriented techniques, parallel coordinates, survey plot, etc. With visualization techniques, large amount of data can be displayed on the screen, colors allow the users to instantly recognize similarities or difference of thousands of data items, the data items may be arranged to express some relationship. We try to solve the following problem: how can we select from a set of candidate dimensions, a subset that performs the best under visual tools and visual data mining and discard the others? We use visual data mining in order to find an accurate decision tree by using a visualization technique with interaction capabilities. The decision tree is interactively constructed by the user who uses his perception and data domain knowledge. This kind of interactive decision tree construction algorithm can only be used if the number of dimensions of the data is small enough (less than dozen).

Dimension reduction and attribute selection aim at choosing a small subset of attributes that is sufficient to describe the data set. It is the process of identifying and removing as much as possible the irrelevant and redundant information. Sophisticated attribute selection methods have been developed to tackle three problems: reduce classifier cost and complexity, improve model accuracy (attribute selection), improve the visualization and comprehensibility of induced concepts. There are two major components in a attribute selection/dimension reduction algorithm: the generation procedure and the evaluation function [Dash and Liu, 1997].

2.1 Generation procedure

Let N denote the number of varia in the original data set, attribute selection requires to test 2^N different subsets to find the optimal one. A solution in order to avoid this search is to proceed to random search or to use one of the following search strategies: backward, forward or both. After the generation of feature subsets, an evaluation function measures the goodness of the subset and this value is compared with the previous best subset of attributes [Dash and Liu, 1997]. The following section presents the available evaluation functions.

2.2 Evaluation functions

Two types of evaluation functions are used in attribute selection: in the first one, filter-based approach the dimensions are filtered independently of the induction algorithm. The relevance of each dimension is computed with some statistical information calculated from the training data set. Examples of statically measures used: information gain [Dumais *et al.*, 1998], [Quinlan, 1993], correlation [Hall, 2000], etc.

The other type is the wrapper approach [Kohavi and John, 1997]: a learning algorithm is used in order to select the subset of features, while discarding the rest. For any iteration of the wrapper algorithm, the quality of the feature subset is evaluated by an inductive learning algorithm.

Attribute wrappers often achieve better results than filters due to the fact they are tuned to the specific interaction between an induction algorithm and its training data. However, they tend to be much slower than attribute filters because they must repeatedly call the induction algorithm and must be run when a different induction algorithm is used [Kotsiantis and Pintelas, 2004].

2.3 Problems encountered in attribute selection

At the initialization step, the attribute selection algorithms require many parameters. In order to lead to best results, it is necessary to choose the most relevant parameters. Knowing that an attribute selection process may stop under one of the following reasonable criteria: a defined set of dimensions are selected, a defined number of iterations are reached, addition (or deletion) of any dimension does not produce a better result, an optimal subset according to the evaluation criteria is obtained.

3 Applying selection to visual data mining

As we said, if the data dimension is high (figure 1), the human cognitive task for detecting correlations or discover hidden patterns in data is very hard.

The figure presents a sequence of $\frac{n-1}{2}$ two-dimensional matrices (like scatter plot matrices [Chambers *et al.*, 1983]) generated by CIAD [Poulet, 2002], n represents the number of attributes. In order to deal with high dimensional data, the above approach of data exploration has been proposed, CIAD supports the user in selecting one representation which matches the best with his mining objective. The focus presents details of the most suitable view. Figure 1 does not allow distinguishing visually colors in order to mine the data set. This is because the number of attributes and the number of instances in the data set are too large. The following paragraph briefly presents the visual data mining task analysis.

3.1 Visual data mining task analysis

In order to mine a data set, the user interacts with a graphical representation (chart) of the data. The data model (knowledge) is built in an interactive and iterative way.

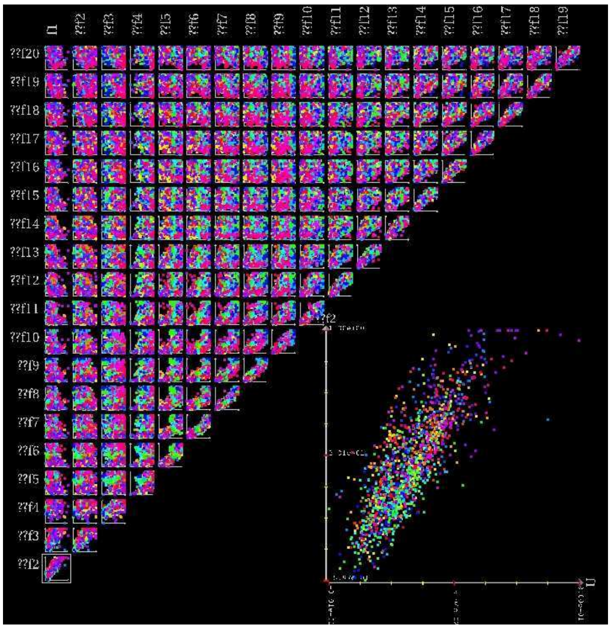

Fig. 1. Isolet data set visualization with CIAD

3.2 User categorization

A visual data mining environment can be used by several type of users:

- data domain specialist: according to his knowledge about data, this type of user can select the best subset of attributes or request the support of an automatic tool for attribute selection.
- data analysis specialist: in this category, the user can be a statistician or a machine learning expert.
  - the statistician expert can adequately use filter approach and determine the appropriate parameters for the initialization of attribute selection algorithm.
  - the machine learning expert can perfectly initialize supervised classification algorithms used by wrapper approach. This type of user is able to choose a supervised classification algorithm to be used in order to evaluate the selected attributes in the attribute selection algorithm and to choose a best set of criteria for the evaluation of selected attributes.

  These users can also be interested in wrapper or filter based approach advantages and need to be supported by an automatic tool.

The automatic dimension reduction framework in all these cases will require a great accuracy of the results.

4 New dimension reduction algorithm

To obtain the best accuracy in attribute selection, the best is to operate an exhaustive search among the 2^N possible combinations of attribute subsets and to use a wrapper-based approach as evaluation function. For a large value of N, this approach is computationaly prohibitive. We propose to use random search and (backward, forward ((like sequential floating selection), knowing that the function used is non monotonic [Pudil *et al.*, 1994])). We believe that this procedure will allow us to treat a large number of attribute subsets.

The wrapper approach allows rising to interesting details for the data analysis specialist (data mining domain). Knowing that the classifier error rate capture two basic performance aspects: class separability ability and any structural error imposed by the form of the classifier. Other types of details, namely, properties that good dimension sets are presumed to have (class separability or a high correlation between the attributes) are more appropriate to statistician. These details could not be highlighted at all by the wrapper methods. In order to take this fact into consideration, we have added some filter-based criteria (consistency, entropy, distance) to our attribute subset selection method.

In input, there is a data set and the output is a subset of attributes of this data set. The generation procedure uses a combination of random search and sequential floating selection. Concerning the evaluation functions, we use a combination of filter (consistency, entropy, distance) and wrapper ((LDA, QDA, KNN) [Ripley, 1996]). LDA, QDA, KNN executions use ten fold cross validation. At each step of the execution of these algorithms, the following evaluation criteria are used: the correctness of the classification rule, the accuracy, the ability to separate classes, and the confidence. Next, we combine their selected attribute subset in order to derive a consensus of the most suitable subset of attributes. For this purpose, a learning step, based on the results of generation procedures evaluated by filter-based criteria and wrapper based approaches enables us to lead to final results.

More precisely, the domain we consider consist of a set of $N = 6$ experts (consistency, entropy, distance, LDA, QDA, KNN evaluation functions) $E = \{e_1, ..., e_N\}$, a set of dimension subsets $DS = \{D_1, ..., D_K\}$, where K is not a constant. Attribute subsets are available for expert/subset pairs $\{e, D\}$, where $e \in E$ and $D \in DS$. We define preference of a dimension d as the probability that the dimension appears in the experts feature subsets, $p(d) = \sum p_i(d)$. $p_i(d)$ represents the probability that expert i selects dimension d .

$p_i(d) = \frac{y}{Z}$ if expert i has selected feature$d$, 0 otherwise. y is the number of selected dimensions. Z represents the number of attributes in the original data set. The preference value of features is used in order to pool together the selected features and to rank them. Next, if the pool number of dimensions is greater than twenty (number of attributes which can be correctly display and visually mine), it is divided into relevant attributes (consensus) and less relevant attributes. At the cutting point, if some features have the same preference value (we consider these attributes as conflicting attributes), we use expert relevance score (ERS) in order to determine which features match the best. For each feature in the conflicting part, the decision to add it in consensus part of the pool or not is made according to the relevance score of the experts who choose the feature. The selected features are those with great expert relevance score computed as following: $ERS = \frac{g}{T}$, where g represents the number of attributes in the consensus part which have been selected by the expert and T the total number of features selected by that expert.

As we will see in the case study part of this paper, the main advantage of this approach is the combination of feature subsets from various feature selection algorithms.

5 Experiments

The purpose of this study was to see if the method would be able to effectively reflect the performance differences among experts.

In order to test proposed approach, we compare its results with the results of two widely used attribute selection methods. Namely, R language implementations of: Las Vegas Filter [Liu and Setiono, 1996] (package dprep) and a wrapper based feature selection algorithm (Stepclass, package klaR). Our consensus based algorithm is also implemented in R. We use a PC pentium IV, 1,7 GHz, Windows to perform these tests. The data sets (from the UCI [Blake and Merz, 1998] and the Kent Rigde Bio-Medical Data Set repositories [Jinyan and Huiqing, 2002] were chosen because of their large number of attributes (table 1).

Name	NbAt	NbInst	NbClass
Lung-Cancer	57	32	3
Promoter	59	106	2
Sonar	60	208	2
Arrhythmia	280	452	16
Isolet	618	1560	26
ColonTumor	2000	62	2
CentralNervSyst	7129	60	2

Table 1. Data set description

The final results of LVF, stepclass and consensus based algorithm were evaluated by IBk, a K nearest neighbor algorithm (KNN) found in WEKA, a free Java-based, open source, that provide a variety of machine learning algorithms.

Table 2 shows the difference (attribute size and KNN accuracy) between the original and the final data sets. The attribute subset selected by the consensus based approach (less or equal to 20) allows visualizing and mining the whole data sets. The changes in the accuracies of KNN classifier is minimal or there is no change. This is not the case of LVF or stepclass (table 3). The data set Arrhythmia for example has a subset with 109 attributes (LVF results) and for the data set Promoter, stepclass does not reduce the dimension.

Name	Initial NbAt	Final NbAt	Acc before	Acc after
Lung-Cancer	57	4	37.5%	75%
Promoter	59	9	85.84%	68.87%
Sonar	60	8	86.54%	71.15%
Arrhythmia	280	4	53.44%	59.96%
Isolet	618	14	85.57%	70.24%
ColonTumor	2000	19	77.42%	79.03%
CentralNervSyst	7129	20	56.67%	60%

Table 2. Comparison of number of attributes and accuracy with KNN algorithm before and after reduction

Feature selection frameworks as we said aim at reducing classifier cost and complexity, improving model accuracy. Our goal is firstly to reduce the number of dimensions in order that the data set could be visualized. Table 3 shows that we attend our principal goal and we obtain results that are comparable to those of the attribute selection algorithms which objective is to improve classifiers accuracy. Indeed, the consensus based approach allows obtaining the best result for data set Lung-Cancer and about the same accuracy rate for the data sets Sonar, Arrhythmia and colonTumor. It should be noted that two cases arise: either the attributes of the data set to be treated are redundant or irrelevant and then the results are comparable with those of filters or wrappers based approaches or it does not exist redundancy in the attributes and dimension reduction implies a loss of accuracy. The data sets in this category are: Isolet (best accuracy with LVF for 268 attributes) and Promoter (best accuracy with Stepclass for 59 attributes). For these data sets, the number of selected dimensions in spite of the best accuracy remains unusable for visual data mining.

Name	Final NbAt	Lvf NbAt	Wrap NbAt	Final Acc	Lvf Acc	Wrap Acc
Lung-Cancer	4	17	4	75%	62.5%	71.87%
Promoter	9	16	59	68.87%	80.19%	85.85%
Sonar	8	18	4	71.15%	82.21%	71.63%
Arrhythmia	4	109	4	59.95%	54.65%	60.84%
Isolet	14	268	8	70.24%	83%	57.98%
ColonTumor	19	918	5	79.03%	77.42%	79.03%
CentralNervSyst	20	3431	8	60%	58.33%	71.67%

Table 3. Comparison of our method with LVF and stepclass

6 Conclusion

The data visualization, the performance of classification algorithms are affected by attributes. When a data set has a large number of attributes, it is impossible to perform visual data mining. Irrelevant, redundant features have a negative effect on the accuracy of a classifier and on visual representations. We have defined a dimension reduction method for visual data mining. Then we have compared successfully the results of this framework to two widely used attribute selection algorithms. The data visualization (figure 1) which represents a visualization in which the relationships within the data are unclear is replaced by another visualization (figure 2) which is more usable and much more appropriate to visual data mining.

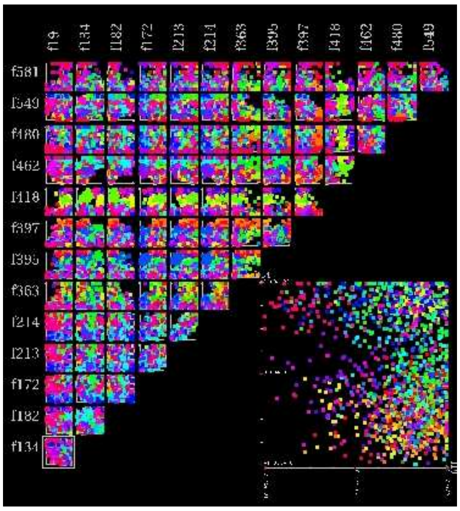

Fig. 2. Isolet Reduced data set visualization with CIAD

Our dimension reduction framework reduces the number of attributes. However, we remark that with a low number of attributes and a high number of instances, it is not easy to represent and mine data perfectly. We plan to develop some methods for reducing the number of instances in a data set to be treated.

References

[Blake and Merz, 1998]C. Blake and C. Merz. *UCI Repository of machine learning databases.* Irvine, University of California, Department of Information and Computer Science, from www.ics.uci.edu/~ mlearn/MLRepository.html, 1998.

[Chambers *et al.*, 1983]J. Chambers, W. Cleveland, and P. Turkey. *Graphical Methods for Data Analysis.* Wadsworth, 1983.

[Dash and Liu, 1997]M. Dash and H. Liu. Feature selection methods for classification. In *Intelligent Data Analysis: An International Journal*, pages 1–2, 1997.

[Dumais *et al.*, 1998]S. Dumais, J. Platt, D. Heckerman, and M. Shahami. Inductive learning algorithms and representation for text categorisation. In *Proc, The International Conference on Information and Knowledge management*, pages 148–155, 1998.

[Fayyad *et al.*, 1996]U. M. Fayyad, G. Piatetsky-Shapiro, and G. Smyth. *Advances in Knowledge Discovery and Data Mining.* AAAI Press / MIT Press, Menlo Park, CA, 1996.

[Ferreira and Levkowitz, 2003]d.O. Ferreira and Levkowitz. From visual data exploration to visual data mining: a survey, visualization and computer graphics. *IEEE Transactions*, pages 378–394, 2003.

[Hall, 2000]M. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proc, International Conference on Machine Learning*, pages 359–366, 2000.

[Jinyan and Huiqing, 2002]L. Jinyan and L. Huiqing. *Kent Ridge Bio-medical Data Set Repository.* http://sdmc.lit.org.sg/GEDatasets, 2002.

[Kohavi and John, 1997]R. Kohavi and G. H. John. Wrappers for feature subset selection. In *Artificial Intelligence*, pages 273–324, 1997.

[Kotsiantis and Pintelas, 2004]S. B. Kotsiantis and P. E. Pintelas. Hybrid feature selection instead of ensembles of classifiers in medical decision support. In *Proc, IPMU*, 2004.

[Liu and Setiono, 1996]H. Liu and R. Setiono. A probabilistic approach to feature selection: a filter solution. In *Proc, The 13th International Conference on Machine Learning*, pages 319–327, 1996.

[Poulet, 2002]F. Poulet. Cooperation between automatic algorithms, interactive algorithms and visualization tools for visual data mining. In *Proc, Visual Data Mining workshop, PKDD2002*, 2002.

[Pudil *et al.*, 1994]P. Pudil, J. Novovicova, and J. Kittler. Floating search meathods in feature selection. *Pattern Recognition Letters*, pages 1119–1125, 1994.

[Quinlan, 1993]J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan-Kaufman, San Mateo, CA, 1993.

[Ripley, 1996]B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[Ware, 2000]C. Ware. *Information Visualization, Perception for design.* Morgan-Kaufman Publishers, San Diego, USA, 2000.

[Wong, 1999]P.C. Wong. Visual data mining. *IEEE Computer Graphics and Applications*, pages 20–21, 1999.