Proceedings of the 5th WSEAS Int. Conf. on SIMULATION, MODELING AND OPTIMIZATION, Corfu, Greece, August 17-19, 2005 (pp373-378)

Exploration of very large data sets: The CiTree algorithm.

Manuel Castejon Limas; Joaquin B. Ordieres Mere; Antonio Ciampi; Fernando Alba Elias (*)
Escuela de Ingenierías Industrial e Informatica
Universidad de León
Campús de Vegazana. 24071. León. Castilla y León.
SPAIN
http://torio.unileon.es/~diemcl

Abstract: - Control variables and command history recorded during normal operation of industrial processes are routinely stored in databases for later analysis. These databases constitute a potentially precious source of information that can be extremely useful from a commercial and strategic perspective. However, the extraction of information from a database is often a non-trivial task, requiring the cooperation of various disciplines, in what has now become a field of research in itself, known as 'data mining'. One of the first tasks for the data miner is to summarize the complexity of the data into a number of distinct clusters, which represent 'interesting', often unexpected, behavior patterns of the process under analysis. Powerful computers and efficient clustering algorithms are now available; nonetheless limits are typically exceeded when mining massive databases such as those arising from industrial processes. Therefore, in order to uncover the potentially useful information held by large bodies of data, new clustering algorithms are needed that directly address the problem of size. We present here one such algorithm, which has been successfully applied to a variety of industrial processes. It yields a hierarchical structure of the clusters present in the process; thus providing a detailed representation of the relationships amongst sample units. We also show, as an example, the application of the algorithm to a real case study, which resulted in the extraction of useful information.

Key-Words: - Data Mining, Optimization, Model, Industrial, Clustering.

1 Introduction

This paper reports on the joint efforts of members of the EDMANS (Engineering Data Mining and Numerical Simulation) research group (Spain) and researchers from McGill University (Canada) and INRIA (**) (France). This joint work resulted in the development of a tool for revealing unexpected features in a manufacturing process, supposedly under control. Details can be found in [3])
The general aim of our research was to discover the latent models underlying an industrial process, so that its quality standards can be improved. Models can be inferred from process performance data, recorded and stored in huge databases, usually huge both in terms of number of variables and number of observational units. The use of computer databases allows us to numerically retrieve a picture of the real on going process during fairly long periods of manufacturing operation time. These records are often as detailed as possible, thus providing a unique opportunity to develop a good understanding of the process. Knowing a process from its recorded data allows us either to confirm previous assumptions or to reject hypotheses not supported by the data.

However, a direct glimpse at the contents of these databases is usually quite unenlightening and rarely suggests useful interpretations. Indeed, data should be considered as the raw material from which useful information must be distilled through complex analytic procedures. One of the main challenges is that there are no standard operating procedures to follow in this extraction process. Instead, one has to rely on a combination of intuition, understanding of the industrial context and knowledge of appropriate methods of data analysis. On the other hand, the results of this exercise may well be worth the effort. Uncovering key factors in the evolution of the industrial process of interest, would help quality control engineers make timely interventions aimed at controlling the levels of these key factors: thus, overall quality is improved, rejections are reduced, and both energy and money are saved.

There is a very rich literature describing methods that seem appropriate to the task. Most of them, however, fail when dealing with databases as large as those occurring in industrial problems. A methodology that is particularly appropriate for dealing with industrial databases is outlined in [1]. It consists of a logical sequence of actions to take in the process of information extraction, based on a variety of multidisciplinary techniques. Somewhere in the middle of this sequence, there are actions requiring

use of cluster analysis: in this paper we will focus on this aspect of the general methodology.

Cluster analysis aims at finding groups in a dataset. For instance in data from industrial processes one may wish to identify distinct patterns of process behavior by grouping together observational units with similar behavior. Focusing on the patterns rather than on the observational units, we loose fine-grained detail, but gain understanding of the broad features of the data. For example, in hierarchical clustering, a general structure emerges—a hierarchy, represented by a dendrogram—in which every unit has a specific place: this structure clearly exhibits the relationships amongst units through a powerful graphical representation. While clustering algorithms are quite useful in their own right for the role they play in direct data interpretation, their usefulness may be further enhanced if applied to data in combination with other techniques, which dig further in the structure of the data.

Clustering methods are classified according to the principles on which they are based. Nearly all of the most popular algorithms fall into one of the following classes:

- Neural Classifiers: e.g., Self-Organizing Maps.
- Hierarchical Methods: e.g., Agglomerative and Divisive algorithms.
- Partitioning Methods: e.g. K-means and K-Nearest Neighboors.
- Model Based Algorithms
- ...

As mentioned above, hierarchical algorithms provide a very interesting representation of the data: they show the complete set of nested inclusion relationships that the observed cases support, thus yielding not just one but a whole family of nested classification schemes. By contrast, partitioning techniques just focus on finding and describing clusters, much as if, studying a dendrogram, one was only interested in the partition generated by a transversal cut at a specified level of similarity. Most people would easily recognize the advantages of hierarchical classifications. On the other hand, hierarchical clustering techniques quickly become inapplicable as the number of observational units in the database increases. Indeed, at the outset, one must calculate the similarities between each pair of units. Therefore, no matter the computer used, the procedure becomes sooner or later unfeasible, due to the heavy CPU usage requirement. Clearly, it would be unwise to approach a massive data set armed with just one hierarchical classification algorithm: the size of the former exceedingly surpasses the capabilities of the latter.

An alternative approach that circumvents the problem of size would be to concentrate on few of the top branches. Thus, instead of building the tree from individual units, we may start off by arranging these into more complex intermediate structures: as many as we could handle later on. The interested reader may find details in [3] and [4]).

2. An experience from a galvanizing line.

2.1 The study case.

In order to manufacture galvanized steel products efficiently and economically a strict control of the production process is needed. Unfortunately, however, current close loop techniques are unable to control the swift sheet of steel running at a rate of 30 m/s. Hydraulic systems are incapable of reacting as it would be needed at such speed. Furthermore, in this as in many other processes, it is impossible to feed some of the signals back online, since measuring it requires laboratory analyses that might take half an hour, to say the least. Under these circumstances, a different control strategy must be proposed, one based on the accurate forecast of some key variables. It is at this stage that data mining methodology can be advantageously introduced.

The application of data mining techniques to the manufacture of galvanized steel products was discussed in [5]. The aim of that work was to improve the quality of galvanized steel products by means of an optimized predictive model for some mechanical features of steel. The development of this predictive model requires data mining techniques. The size of the data set was 311,456 cases, a relatively small set but big enough to make traditional techniques inadequate. The chosen predictors were mainly the chemical composition and the furnace condition variables; the average furnace temperature; Carbon, Manganese, Silicon, Sulfur, Phosphorus, Aluminum, Copper, Nickel, Chromium, Niobium, Vanadium, Titanium, Boron and Nitrogen concentrations; the elastic limit, enlargement and ultimate strength.

2.2 Preclustering the data set.

Before performing hierarchical agglomerative cluster analysis, we should first arrange the individual cases in pre-clusters by means of a quicker method capable of handling bigger data sets than the hierarchical algorithms. We have obtained good pre-clusters using Self-Organizing Map. We show in Fig. 1 the resulting Kohonen Map with twenty cells.

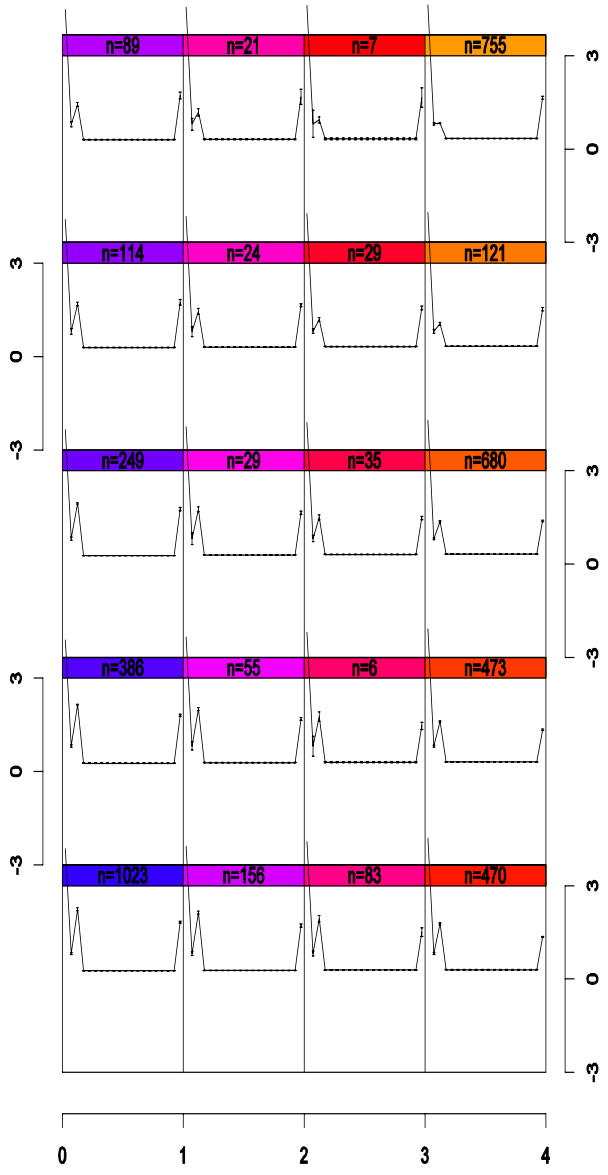

Fig. 1 – Self Organizing Map obtained from the Galvanizing Line.

2.3 Building the dendrogram from the Pre-clusters.

The cells of the Self Organizing Map constitute bins that contain the samples generated by the subpopulations present in the dataset. In order to construct a hierarchy with bins as atoms, we must define an appropriate dissimilarity measure for complex objects, such as bins or clusters of bins; then, any agglomerative algorithms using dissimilarities may be used to obtain a dendrogram. Our "CiTree" algorithm [4] is based on this approach. We show in Fig. 2 the resulting agglomerative dendrogram for our data.

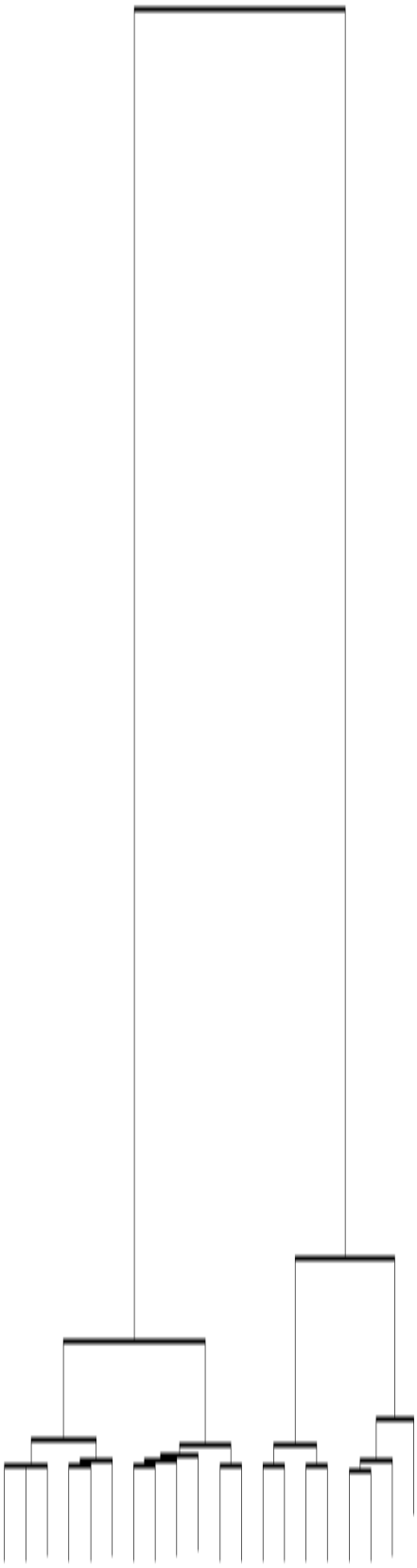

Fig. 2 - Top few branches of the galvanizing line CiTree dendrogram.

A mere visual inspection may suggest the presence of two to four main clusters. The numerical value of the Fowlkes-Mallows [6] index at each level of the dendrogram definitely suggests the presence of four main classes. The reader may have a look Fig. 3, which shows the data points projected onto a Linear Discriminant plane. It suggests that the clusters we obtained are nicely separated.

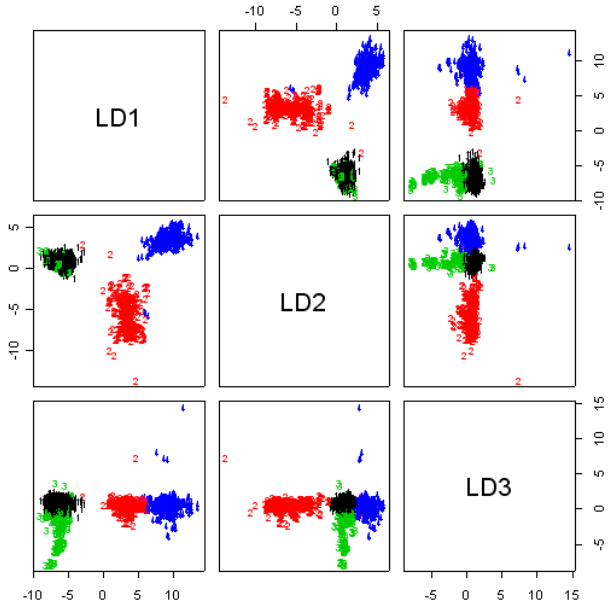

Fig. 3 - LDA projection of the galvanizing line according to the CiTree results.

2.4 Further processing of the clustering results.

As we mentioned in the introduction, the results of a clustering algorithm may provide already a satisfactory description of the various patterns of behavior present in the state space of our process. At the same time, they may also be taken as a starting point for a new appraisal of our assumptions and further analysis, taking now into account what is displayed in the dendrogram. As an example, we may consider testing for outliers. Do all the data points correctly represent the process in its normal functioning or have some of them suffered an unexpected perturbation? We may now attempt to answer this question. We can apply our PAELLA [2] algorithm –for outlier identification in non-normal samples – to the results just obtained from the CiTree algorithm. The results can be seen in Fig. 4.; it is clear from the picture that some of our data points,

those marked in red, may be considered outliers. This means that, in the course of the analysis, we should have taken appropriate precautions to reduce the effect of these outliers on the results. Now, we can repeat the analysis, but this time with due precaution.

In addition to what has already been shown, the PAELLA results may also help determine which were the most perturbed variables, or the ones that had the strongest influence in the exceptional character of the outliers. A detailed analysis of the components of the discriminant functions obtained in the linear discriminant analysis shows that the Carbon and Manganese concentration levels are responsible, with a 94.64 percent of influence, for the appearance of perturbations. This new information should warn the engineers to pay special attention to these variables and to the procedures for measuring them.

Perhaps the most important advantage of combining the Citree and the PAELLA algorithms is that a preliminary Citree analysis dramatically increases the speed of the PAELLA algorithm.

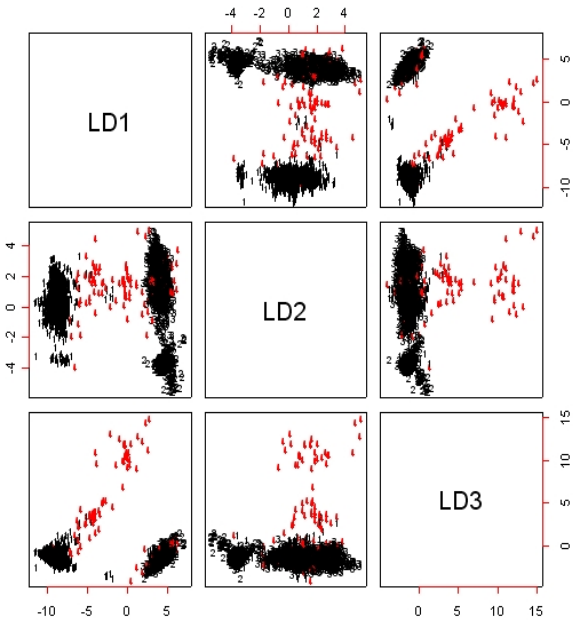

Fig. 4 – Outlier identification based on the CiTree results.

3 CONCLUSIONS

In this paper we have discussed a new approach, to the exploration of very large data sets. In particular, we have shown how to apply hierarchical agglomerative clustering dendrogram when data size largely surpasses the capabilities of traditional algorithms. Our proposal is to obtain a manageable data set by pre-clustering the original data points into a much smaller number of composite elements, on which a natural dissimilarity measure can be defined: then, using these elements as 'atoms', a hierarchical agglomerative algorithm may reveal interesting structures.

This approach, named CiTree, is a hierarchical agglomerative algorithm in which the construction of the lower branches of the hierarchical tree is replaced by a basic and fast non-hierarchical algorithm. Indeed, in most cases the bottom branches of a hierarchical tree are not as useful as the top branches in providing sense and interpretation. Therefore it seems reasonable to short-circuit the early and most computationally heavy phase of dendrogram construction, sacrificing some less significant details; instead, complex computations should be concentrated on the later phase, where important information is more likely to be revealed.

We applied the proposed strategy to a real data from a galvanizing line, where the use of the CiTree algorithm discovered four clusters, which appeared clearly separated in the LDA graphical output. The suggestions gleaned from visual interpretation of the dendrogram were confirmed by numerical analysis based on the Fowlkes-Mallows index, thus showing the power of the visual representation provided by the dendrogram.

These results, useful in themselves, can also provide a deeper insight into the structure of the data if used in synergy with other data mining techniques. Finally, we showed an example of outlier identification where the CiTree algorithm accelerated the bottleneck already found in non-normal outlier identification.

4 REFERENCES

1. Castejón Limas, M; Ordieres Meré, J.B.; de Cos Juez, F.J.; Martínez de Pisón Ascacibar, F.J. "Control de Calidad. Metodología para el análisis previo de los datos en procesos industriales. Fundamentos teóricos y aplicaciones en R." Servicio de Publicaciones de la Universidad de La Rioja. 2001. ISBN: 84-95301-48-2

2. Castejón Limas, M.; Ordieres Meré, J.B.; Martínez de Pisón Ascacibar, F.J.; Vergara González , E.P.; "Outlier detection and data cleaning in multivariate non-normal simples. The PAELLA algorithm." Data Mining and Knowledge Discovery, Vol. 9, 2004, pp. 171-187.

3. Castejón Limas, M., "Desarrollo de estrategias basadas en técnicas de inteligencia artificial para la mejora de la calidad en procesos industrials". PhD Thesis. Universidad de La Rioja, 2004.

4. Ciampi, A; Lechevallier, Y.; Castejón Limas, M.; González Marcos, A. "Hierarchical Clustering of Sub-Populations with a dissimilarity based on the likelihood ratio statistic: Application to Clustering Massive Data Sets." Under revision.

5. Ordieres Meré, J.B, González Marcos, A., González, J.A., Lobato Rubio, V., "Estimation of mechanical properties of steel strip in hot dip galvanizing lines". Ironmaking & Steelmaking, Vol. 31, nº 1, 2004, pp. 43-50.

6. Fowlkes, E., Mallows, C., "A new method for comparing two hierarchical clusterings." Journal of the American Statistical Association, Vol. 78, 1983, pp. 553-569.

NOTES:

(*) Joaquín B. Ordieres Meré and Fernando Alba Elías work at the Universidad de La Rioja, Departamento de Ingeniería Mecánica, Área de Proyectos de Ingeniería. Antonio Ciampi works at Mcgill University, Department of Epidemiology and Biostatistics.

(**) Antonio Ciampi – McGill University – and Yves Lechevallier – INRIA – in collaboration with their Spanish colleagues developed the CiTree algorithm used in this paper about which we refer to in [4].

Aknowledgments:

We gratefully acknowledge support from the Ministerio de Educación y Ciencia de España, Dirección General de Investigación, by means of the DPI2004-07264-C02-01 research contract; from the "II Plan Riojano de I+D+i"; from the European Union by means of the CEUTIC INTERREG IIIA Spain / France trans-border cooperation project; and from the RFCS program by means of the RFS-CR-03012, RFS-CR-04023 and RFS-CR-04043.