DISCOVERING UNKNOWN PATTERNS IN MASSIVE DATABASES.

Castejón Limas, M.; Ordieres Meré, J.B.; Vergara González, E.P.; Martinez de Pisón Ascacibar, F.J. (*)
Área: Proyectos de Ingeniería.
Dpto.: Ingeniería Eléctrica y Electrónica.
Universidad de León.
Campus de Vegazana s/n (24071) León.
Tfn. : + 34 987291000-Ext.5382 Fax: + 34 987291790 e-mail: manuel.castejon@unileon.es

RESUMEN

El almacenamiento de los diferentes estados de un proceso industrial durante su normal funcionamiento preserva para posteriores análisis una preciosísima fuente de información. Sin embargo, esta información rara vez se encuentra representada de forma directamente interpretable. Más al contrario, su aprovechamiento, desde el punto de vista de estratégico y comercial de la industria, involucra la participación de tareas procedente de artes y ciencias muy diversas, en lo que se ha venido a denominar en su conjunto como "minería de datos". Estas técnicas representan un interesante campo en el que el experto debe conjugar sus habilidades prácticas y el valor de su experiencia, con un profundo conocimiento de las bases en las que se fundamentan sus herramientas de trabajo. Entre sus primeras ocupaciones, el experto deberá descubrir la presencia de distintas pautas de comportamiento del proceso en estudio. Sin embargo, y aun a pesar de la incorporación de los computadores a esta arena, la cantidad de información que somos capaces de almacenar supera habitualmente la cantidad de información que podemos procesar e interpretar. La numerosidad de casos habitualmente proporciona en una mano una rica fuente de información y en la otra una nutrida fuente de problemas para su obtención. Es por esto que son necesarios algoritmos como el aquí presentado para hacer posible la determinación de patrones en situaciones como ésta en la que las técnicas habituales pierden operatividad y sentido. El algoritmo ha sido utilizado con éxito en procesos industriales de muy diversa índole, proporcionando una jerarquía de las clases presentes en el proceso, representación completa de las distintas relaciones entre los diversos casos registrados. Se mostrará a modo de ejemplo uno de los casos reales en los que el método ha sido aplicado con éxito, constatando su utilidad y destacando la relevancia de la interpretación de sus resultados.

Palabras clave: Minería de Datos, Optimización, Modelo, Industrial, Clasificación.

ABSTRACT

The storage in databases of the different values of the control variables and commanding actions in an industrial process during its normal operation preserves for a later analysis the source of information of the utmost relevance. Nevertheless, this information is rarely found to be represented in a crystal clear manner. On the contrary, to be useful from the commercial and strategic perspective of the industry, very varied disciplines must cooperate, in what has come to be named ‘data mining". These techniques are an active and interesting field in which the practitioner must combine practical skills and experience with a profound knowledge of the principles in which the tools are based. Among the first tasks to perform, the practitioner must reveal the presence of different clusters and unknown patterns and behaviors of the process under analysis. Nonetheless and in spite of the appearance of computers in this arena, the available amount of information usually exceeds the limits beyond which no available algorithm can trespass. These massive data sets most frequently provide a rich source of information in one hand, while holding in the other an unlimited source of troubles in its retrieval. That’s why algorithms like the one here presented are needed. They make possible the identification of clusters in situations like this where traditional techniques lack both sense and effectiveness. The algorithm we have used has been successfully applied in varied industrial processes, providing a hierarchical structure of the clusters present within the process; a complete representation of the relationships amongst the registered samples. We will show in this paper, as an example, a real study case that provided successful results thus proving right the usefulness of the algorithm and their results.

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

0 INTRODUCTION

This paper accounts some of the deeds and latest achievements cropped by the joined efforts of the API La Rioja and API León members; this time teaming up with researchers from the Canadian McGill University and French INRIA (**) (read details in [3]); in order to tackle the nowadays increasingly vital point of revealing the unexpected features of an allegedly under control manufacturing process.

The mainspring of our research is not other but to discover the latent models that support the industrial processes we deal with, in order to raise their quality standards. We know about the processes from their performance, written in numerous records stored in huge databases, enormous not only in dimension but also in number. The use of computer databases allows us to numerically retrieve a picture of the real on going process during long enough periods of manufacturing operation. These records convey all the subtleties that no other approach might imagine to feature; thus providing a unique opportunity to get to know the process in minute detail. Knowing a process from its recorded data allows us either to confirm previous assumptions or to reject old prejudices not supported by the evidence of the observed cases.

Nevertheless, a glimpse at the contents of these databases seems neither to be very interpretive, nor quite enlightening. We should not falter in dismay at such an obscure outset; we have just got the raw material that must be distilled into useful information by suffering a number of suitable set of analysis actions, although no standard operating procedure can be easily established. Eventually our efforts should bear fruit and the boiled down data might gather the high quality information that quality control engineers should take advantage of so as to properly adjust the levels of the commanding actions involved in the formerly unknown key factors; thus gaining overall quality, dwindling rejections and saving both energy and money.

As for the tools at hand, sundry methods have been proposed in the bulk of the literature to aid in this matter, though they fail at large when it comes to dealing with industrial databases: such is their size. We feel compelled to follow a methodology [1] in order to face these datasets. Following this methodical approach, at an intermediate position in the list of analysis actions we should apply to the dataset, we find the cluster analysis techniques, about which we will focus in this paper.

Cluster analysis aims at finding groups in the dataset, that is, to identify the different kinds of behavior observed, grouping the cases with their next of kin, so as to zoom out from the grainy close-up of a mass of discrete identities, to the glossy picture of the superstructures that nestle and hold the body of specific cases. In themselves, cluster analysis methods are already very useful as their results are immediately liable to interpretation. Moreover, they can be further used in subsequent analyses for a deeper dig into the structures within the data.

We may classify the mainstay of available methods according to the nature of their fundaments:

- Neural Classifiers: Self Organizing Maps, Multilayer Perceptrons…
- Hierarchical Methods: Agglomerative, Divisive and Regression Trees…
- Partitioning Methods: K-means, K-NN…
- Model Based Algorithms: M-Clust …
- …

Amongst the above-mentioned methods, Hierarchical Agglomerative Trees provide a particular kind of result of the utmost interest: the complete set of nested relationships that the observed cases support. That is to say, not just a transversal cut at a specified level, as Partitioning Methods and some other techniques would supply, but the whole set of classifications ranging from a monolithic single root to as great many leafs as observed cases are considered. We may quite easily agree about the usefulness of such an analysis, but we should understand as well that the method grows inapplicable as the number of considered cases increases: we must, inter alia, measure the similarities between cases; irrespective of the computer used, the procedure becomes sooner or later inappropriate due to the heavy CPU usage requirement. Of course, we cannot even try to think about facing a massive data set armed with this method alone: the size of the former exceedingly surpasses the capabilities of the latter.

An alternative approach can be taken to avert this barrier. Instead of sewing up the completely unraveled tree, just building the top few branches may suffice, for that is where the analyst may find most frequently the more meaningful and useful picture. Thus, instead of building the tree from individual cases we may start off by arranging the mass of cases into more complex intermediate structures; as many as we could handle later on (the interested reader may find out the details in [3] and [4]).

1 AN EXPERIENCE FROM A GALVANIZING LINE.

1.1 The study case.

In order to manufacture galvanized steel products efficiently and economically, it is imperative to control the process status with the utmost care. Nonetheless, 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 its measure requires laboratory analysis 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 the commanding action levels; here is where our methods fit in.

It was shown in [5] how data mining techniques can be a useful set of tools to deal with this kind of processes. The aim of that work was to improve the quality of galvanized steel with an optimized predictive model of the steel mechanical features by means of applying data mining techniques to the process databases. The size of the data set was 311,456 cases, a relatively small set but big enough to start causing trouble to traditional techniques. 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.

1.2 Preclustering the data set.

In order to perform the hierarchical agglomerative tree 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 considered, and the results confirmed, that a Self Organizing Map would have the job done. We show in Fig. 1 the resulting Kohonen Map of twenty cells.

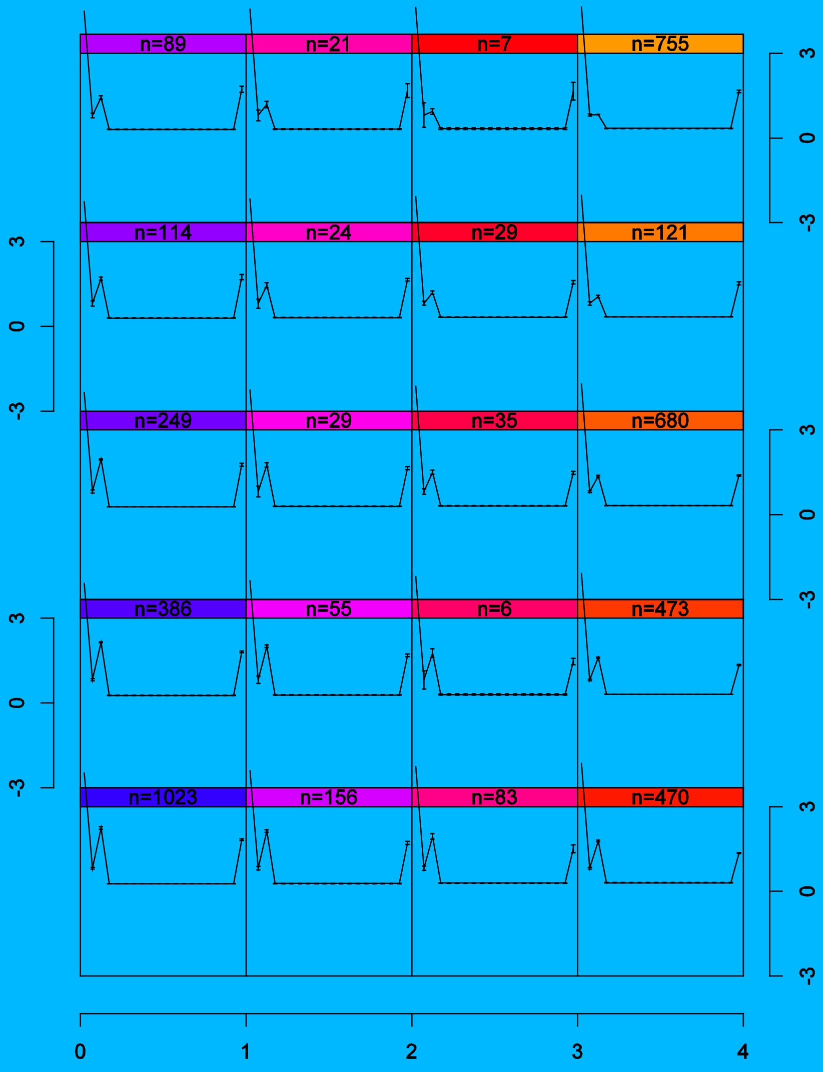

Fig. 1 – Self Organizing Map obtained from the Galvanizing Line.

1.3 Building the tree from the Preclustering objects.

The cells of the Self Organizing Map constitute bins that contain the samples generated by the subpopulations present in the dataset. In order to sew up the top few branches of our hierarchical tree, we must use the complex objects that represent these subpopulations. This job is done by our "*CiTree*" algorithm [4]. We show in Fig. 2 the resulting agglomerative tree.

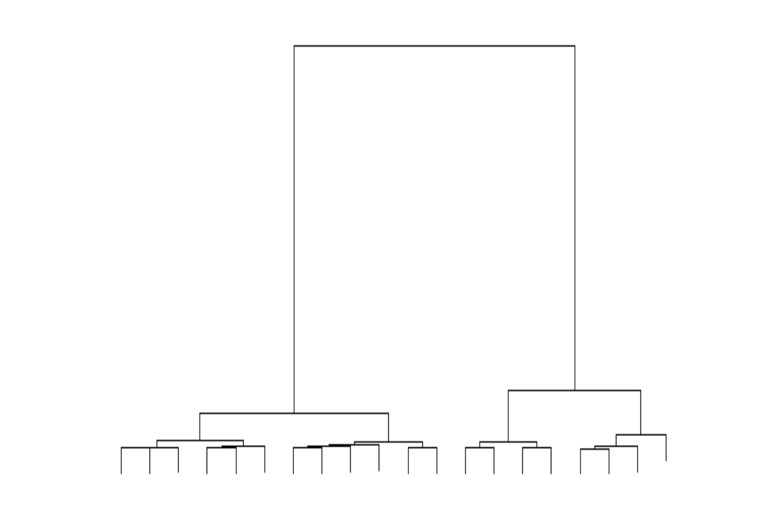

Fig. 2 - Top few branches of the galvanizing line CiTree dendrogram.

A mere visual inspection may suggest the presence of may be two or 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 at the data set projected onto a Linear Discriminant plane at Fig. 3. The discrete resulting shapes suggest that the hierarchical clustering algorithm did a good enough job.

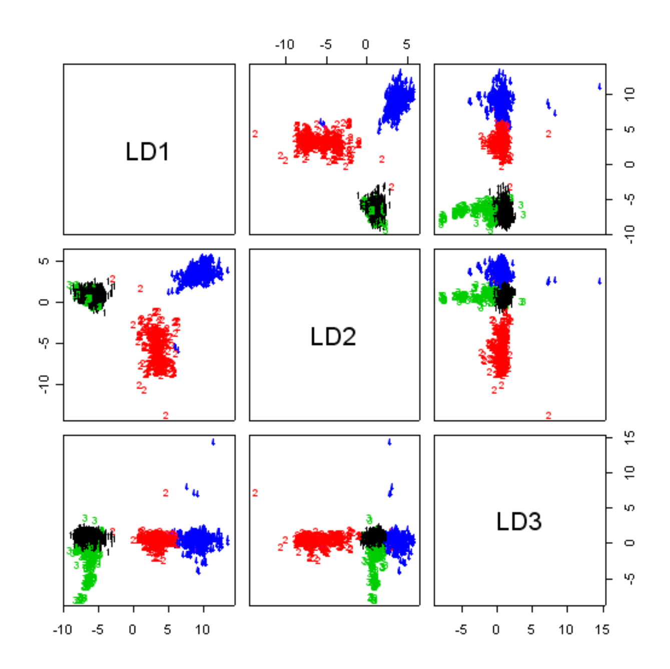

Fig. 3 - LDA projection of the galvanizing line according to the CiTree results.

1.4 Further processing of the clustering results.

As we mentioned in the introduction, the clustering results may suffice for themselves as far as they describe the different kinds of behavior present in the state space of our process. Nevertheless, they also represent a starting point to revise again our assumptions and analyses, taking now into account what in the tree is displayed. As an example, we may consider the outlier tests. Were all the samples we used in our line of analysis actions correct or had any – or many – of them suffered any sort of perturbation? We may now answer to this question. We feed our PAELLA [2] algorithm -- for outlier identification in non-normal samples -- these fresh results recently obtained by the CiTree algorithm. What we do get can be seen in Fig. 4.; a Linear Discriminant Analysis projection of out data set according to the PAELLA results. It is clear in the picture that some of our samples, those in red, are considered nominees to belong to the outlier category. That means that we should have taken the needed precautions in the rest of analysis actions about the effect of these outliers on the corresponding results.

In addition to what has already been said, the PAELLA results may also determine which were the most perturbed variables, or the ones that had a stronger influence in their outlying character. A detailed analysis of the components of the discriminant functions obtained in the linear discriminant analysis, shows that the Carbon and Manganese concentration levels claim responsibility, with a 94.64 percent of influence, for the appearance of perturbations. This new information should make the engineers pay special attention on these variables and exquisitely care their measure procedures.

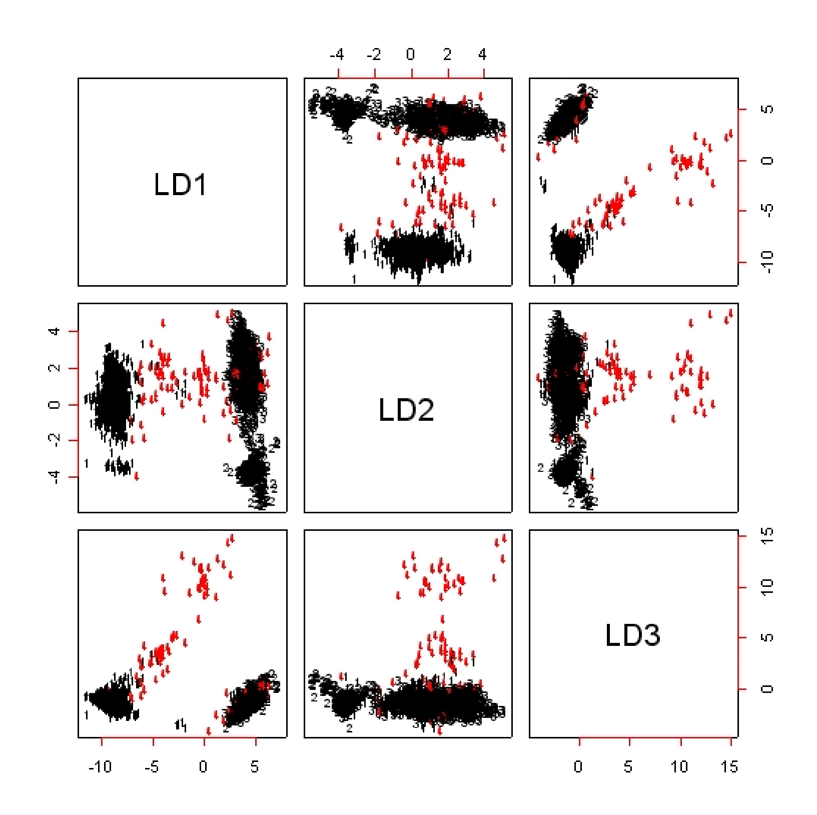

Fig. 4 – Outlier identification based on the CiTree results.

What is really good about the combined use of the Citree and PAELLA algorithms is that Citree dramatically speeds up the previous clustering phase that used to strangle the PAELLA algorithm, thus killing two birds with one stone.

2 CONCLUSIONS

We were able to build a hierarchical agglomerative clustering dendrogram of a large data set by means of preclustering the raw cases into a set of more complex structures that eventually were used as leafs by the CiTree algorithm. This approach represents an advance in the fight for clustering massive data sets, where the number of cases to be considered surpasses the capabilities of traditional algorithms.

The CiTree approach makes use of basic and fast clustering algorithms in order to neglect the calculation of the bottom branches of the hierarchical tree, that most frequently do not provide much sense or information to the analyst, so as to avoid the time consuming, and in many cases inadvisable, task of sewing the whole tree up from the individual cases.

We showed as well a real case experience from a galvanizing line, where the use of the CiTree algorithm remarked on the presence of four different clusters, clearly separated at the LDA graphical output. The suggestions born from visual interpretation of the shape of the tree were ratified by the numerical analysis based on the Fowlkes-Mallows index, thus showing the usefulness of the visual character of the dendrogram.

These results could have been useful in themselves but we also showed how they could be further used to help 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.

3 REFERENCIAS

3.1 Referencias Bibliográficas

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*". Tesis Doctoral dirigida por Prof. Dr. Ordieres Meré, J.B. 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.

NOTA:

(*) Joaquín B. Ordieres Meré, Eliseo P. Vergara González and Francisco J. Martinez de Pisón Ascacibar are professors at the Universidad de La Rioja, Departamento de Ingeniería Mecánica, Área de Proyectos de Ingeniería.

(**) We refer here to our colleagues Antonio Ciampi – McGill University – and Yves Lechevallier – INRIA – that in merry fellowship with their spanish companions developed the CiTree algorithm about which we refer to in [4] and that supports this paper.