1

Application of Data Mining Techniques to Storage Management and Online Distribution of Satellite Images

Denise N. Rotondi Azevedo[1] and José M. Parente de Oliveira[2]

[1] National Institute for Space Research
denise@dss.inpe.br
[2] Technological Institute of Aeronautics
parente@ita.br

The evolution of computational systems and data capacity storage brought a spread of information to many society niches. This information, normally heterogeneous and dispersed in the organizations, can bring in itself a set of indirect knowledge with a great potential of use. The knowledge extracted from this information can be valuable for profile evaluations, detection of problems and opportunities, in the decision making, etc.

In this context, the concepts of Knowledge Discovery in Databases (KDD) came up, having, as part of the process, the Data Mining techniques which have as their main objective patterns discovering in large data sets. The Data Mining algorithms and techniques deal with two dimensions of problems: the great amount of data that makes any discovering process of known patterns very complex and, on the other hand, the search of a priori unknown patterns that can be very useful in organizations.

This chapter describes the use of knowledge discovery techniques by means of the use of traditional Data Mining algorithms in a real context of a Satellite Images Processing and Distribution System by Internet.

The system described stores about 450,000 images and distributes an average of 10,000 images a month, usually delivering images to the users in up to 4 minutes. However, in some cases, this delivery time can take up to 20 minutes or more. This discrepancy in the delivery time occurs whenever a user requests an image that has not been processed or has not been requested previously. In this case, the user has to wait for the image processing. Today, the choice criteria of the images that will be processed a priori are not automated and depend on the operators who take this decision based on empirical criteria.

The use of Data Mining techniques has helped in automating the choice of which images would be previously processed and stored. It represented an improvement in the customer service and leaded to a better use of the storage space and the processing resources.

The chapter is organized as follows. Section 1.1 presents the concepts used in the chapter. Section 1.2 presents a characterization of the research problem

N. Nedjah et al. (Eds.): Innovative Applications in Data Mining, SCI 169, pp. 1–15.
springerlink.com

considered. Section 1.3 describes how data mining techniques have been applied. Section 1.4 presents some concluding remarks.

1.1 Introduction

Data Mining can be seen as a stage of a larger process known as Knowledge Discovery in Databases (KDD). A known definition for the knowledge discovery is [2]: "KDD is a process, in several stages, not trivial, interactive and iterative, for identification of new valid understandable and potentially useful patterns from large data sets".

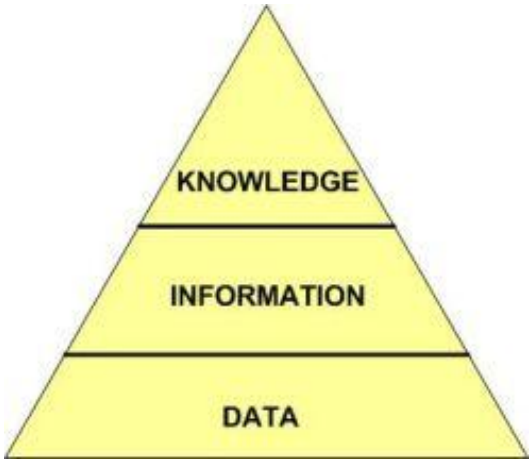

Fig. 1.1. Data, information and Knowledge pyramid

The data, information and knowledge concepts can be seen in the following Figure 1.1 [3]. Thus, the use of Data Mining is intended to support the discovery of patterns in databases in order to transform information in knowledge [3], to assist the decision making process or to explain and justify it. According to Witten and Frank [4], Data Mining can be defined as an automatic or semiautomatic patterns discovery in great amounts of data, where these patterns can be perceived as useful.

Some authors divide the complete process of data mining in the following stages: problem definition, application of Data Mining techniques (prototyping), implementation and evaluation of investment return [3][1].

The stage of data mining application can still be divided in three great iterative stages: pre-processing, pattern extraction and post-processing [3][5].

1.1.1 Pre-processing

The data mining is usually carried out on great and complex masses of data that are centralized in one database, dispersed through several databases, or

distributed in many formats and/or medias. In any case, a pre-processing stage is necessary.

This stage should be based on a deep analysis of the problem, context and available data. Many tasks must be carried out and some decisions must be taken.

The main tasks in pre-processing are:

a) Extraction and integration: dispersed data in different sources and medias must be integrated and extracted, tables in databases may be joined and "desnormalized";
b) Transformations: data can be codified, normalized, enriched with external information, etc. aiming at a better adjustment to the problem goal, data mining algorithms and techniques or to the processing requirements.
c) Cleaning: the data can contain incorrectness and it should be completed or cleaned to prevent absent data or to inconsistent data (outliers)[3].
d) Data Selection and reduction: most of the time it is not needed to use all the available data items. Sometimes the use of the complete amount of data, or the use of all existing attributes can affect the processing and even the results in a bad way. So, selections and reductions can be applied in the attributes, in the attributes space and in the sample number.

The problem analysis and the pre-processing stages are very important and sensitive in the Data Mining Process. A wrong decision in data choice, or the use of incomplete attributes, or any deformation in the data can invalidate the patterns extraction or can even take to inconsistent patterns.

1.1.2 Pattern Extraction

This stage is the mining process itself. Normally, the characteristic of the problem, its domain and expected results define the model to be used for the pattern extraction. The models can be predictive or descriptive. "A predictive Model calculates some value that represents a level in the future activity, a descriptive model discovers rules that are used to group items in categories"[1].

Therefore, the problem classification, the expected results and the characteristics of the data must guide the choice of the mining techniques and the most appropriate algorithms.

The main mining techniques are:

a) Classification: prediction of a category or discrete class. From classes already known the classification algorithm can be trained to characterize new records for predicting its category. Several algorithms are used: C.45 (decision tree), Network Bayes, Neural Networks, etc;
b) Regression: it is also associated to prediction, and has similar objectives to the classification, but it is applied to continuous values. This technique aims to seek a function that maps values of a set of attributes in a target attribute value, so that it can estimate the value of this target attribute in future samples. The regressions can be linear or nonlinear. Statistical algorithms and neural networks are used in this task;

c) Association: related to descriptive models, it looks for the identification of patterns that occur simultaneously in collections of data;
d) Clustering: technique of samples aggregation that has as main objective to maximize the intra-groups similarity and to minimize the inter-groups similarity [3].

1.1.3 Post-processing

Once realized the pattern extraction, in the last stage, the Post-processing, the results are analyzed. Many times, different patterns are found in data and a deep analysis of the results is needed in order to filter the less important or very abnormal results. Usually, a problem domain expert conducts this analysis.

The data mining is intended to validate hypotheses or to look for new patterns inside the data. Thus, for this purpose, it must be verified:

- If patterns had been found according to the domain specialist's expectations;
- Which are the differences between the specialist's knowledge and the found patterns;
- If the new patterns found are relevant, logic and useful.

It is important to remind that data mining is in fact an experimental process and, therefore, the stages should not be completely tight, they should be iterative.

1.2 Problem Definition

The Data Center of the National Institute for Space Research in Brazil has a satellite image data base (CBERS and Landsat Satellites) of approximately 120 Tbytes with a growing rate of about 30 Gbytes a day.

These data - raw data and products - are stored in a Library and are addressed by an image Catalogue which maintains for each satellite image a metadata that describes it and a sub-sampled image (browse image) for its visualization. Through this catalogue the user can search images, request the desired images and make download of the chosen ones - a product request.

Figure 1.2 shows CBERS sub-samples images that are visualized by users in the catalogue.

Using a Catalogue System, users may put requests for a product (processed scene). When the product is ready for download, an e-mail is sent to the user advising that he/she can download the processed scene.

Currently, a user receives an ordered product in a short period of time, elapsing, most of the time, up to 4 minutes between the moment of the request and the reception (download) of the image (0-4 minutes: 46% of the cases). Figure 1.3 summarize the delivery times.

This efficiency is due to some factors: use of distributed processing in different stages of the process; use of an automatic Library to store the data, pre-generation of products following certain established criteria or by demand;

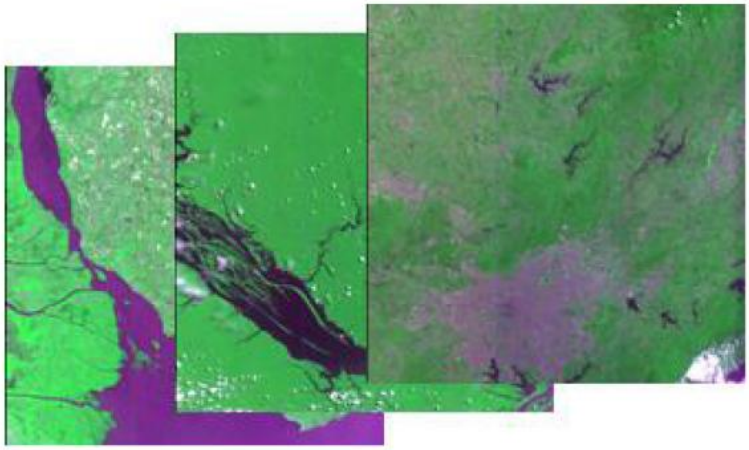

Fig. 1.2. CBERS - Browse Image

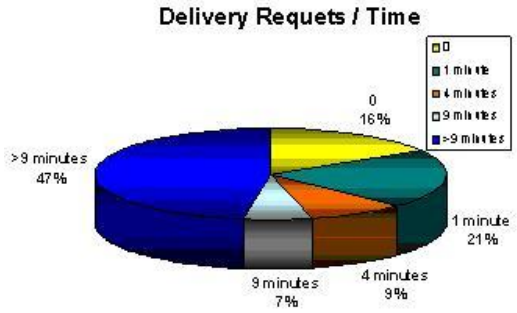

Fig. 1.3. Delivery Request Time

and the storage of part of processed images in disk (online access) - the other part of processed images are usually stored in Library tapes (near line access).

The entire process can be described in the following steps, as can be seen in Figure 1.4.

1. At the Reception Station:
   a) The raw satellite data is received using an antenna;
   b) The raw data are sent to the Processing Station.
2. At the processing station:
   a) Using the Catalogue Generation Sub-system, the received raw satellite data are stored in a library and processed to generate the catalogue database (metadata and browse images);
   b) Using the Pre-processing sub-system, the raw data are preprocessed and stored in disks. As there is a limit in the disk space, it is kept just the

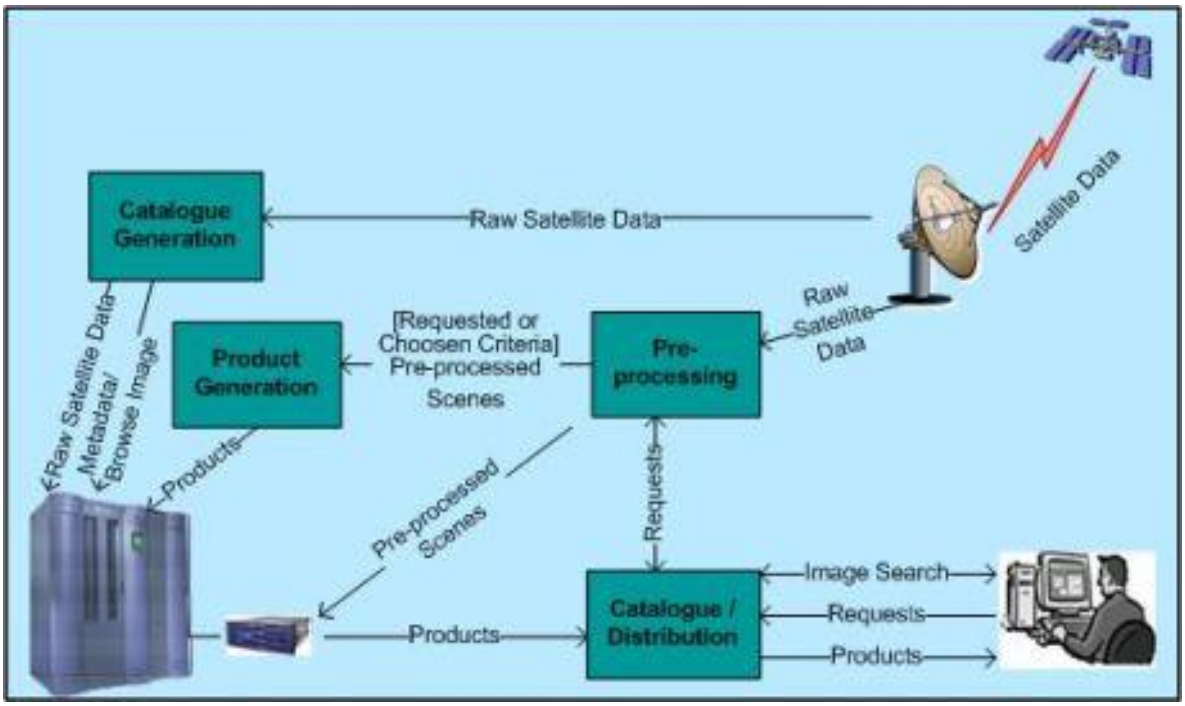

Fig. 1.4. Satellite Data Processing System

images for the last month of reception. The reprocessing step may cut the product generation time in a half;

c) And, as the last step, using defined criteria - normally better images -, part of these images is entire processed and the generated products stored. It makes, for those images, the delivery time immediate.

The products are also stored by demand, in other words, only the first user that requests a specific product (image) has to wait for the processing time. Using this process, although in most of the time the delivery of a request is almost immediate, for some requests this delivery can take up to 4 hours. Additionally, a lot of images that would never be requested or that would be requested just one time are being pre-processed, entirely processed or stored in the disk causing processing and storage waste. The main objective of the work developed has been to maximize the speed in the attendance of a user request, through the pre-processing of images with greater potential of use, aiming at the optimization of the employed resources - processing and storage. To accomplish such an objective, it has been required the study of usage behaviors of satellite images through Data Mining techniques and the correspondent classification of these images in relation to the identified usage behaviors.

1.3 Application of Data Mining Techniques

After the definition of the problem, among the several approaches found in the literature, a prototype was defined where the following iterative phases for the work have been identified:

- Definition of the environment and tools to be used;
- Definition of the data mining techniques and algorithms to be used;
- Data pre-processing;
- Patterns extraction using the defined tool;
- Post-processing and analysis of the results.

Afterwards, it is described in details the results of each data mining application stage.

1.3.1 Definition of the Used Environment and Tools

For the development of the prototype, the computational learning environment Weka (Waikato Environment Knowledge Analysis), which implements a relatively complete set of Data Mining algorithms, has been used. This tool has interfaces where different algorithms can be exercised interactively and a graphical support that aids in the visualization of results and data. It has also an interface to Java language that has been used in the phase of prototype implementation.

Mondrian tool has been also used as a support for the graphical analysis of data.

1.3.2 Description of Used Data

The prototype was made using real data from the Image Satellite Database (Catalogue) of National Institute for Space Research. This work used data from 2004 and 2005, as explained in the sampling step.

As represented by Figure 1.5, the catalogue data is currently composed by some distinct databases: Users database, Scene database (metadata of the stored scenes) and Request database (requests and their items).

Also, mainly the descriptive data of the images (metadata) and the information on its use (number of requests) have been used. From the original databases the following tables have been used: Scene (Scenes DB - 38 attributes), Grid

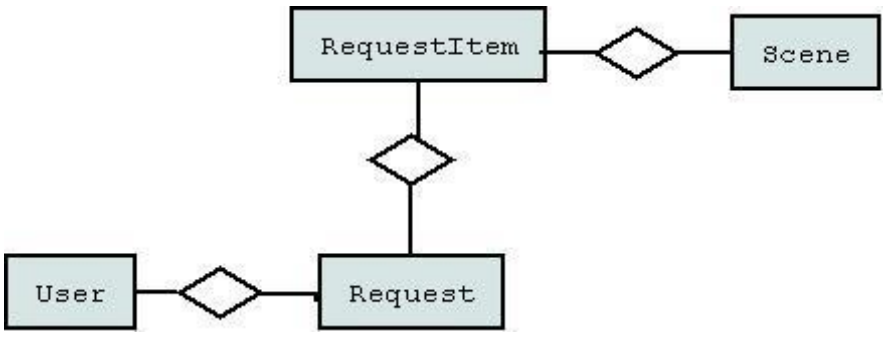

Fig. 1.5. Catalogue Database

(Scenes DB - 6 attributes), Request (Request DB - 14 attributes) and RequestItem (Request DB - 22 attributes).

1.3.3 Description of Applied Pre-processing

In the development, some databases tables, using a specific program, have been pre-processed and converted generating Arrf files (format used by the Weka software). In this first conversion stage, several tables of the different databases have been joined and related in a process of "denormalization" and some attributes have been discarded.

After the analysis of the presented data, it was decided for another set of modifications, reductions and enrichment of the data aiming at better adjusting them to the algorithms that would be used and to the defined software structures. The following processing tasks have been carried out:

- The attribute Season was added because of the interest in the seasonality and to facilitate the application of the classification algorithms that could supply information on the data characteristics. The experience showed that some phenomena are seasonal as, for example, agriculture of some craps that happens in specific months of the year, the cloud months in the forest, and so on;
- To avoid that the cloud coverage indexes for a quadrant (quarter of scene) could have an incorrect effect in the classification results, increasing the decision tree complexity, it has been adopted the attribution of just one cloud cover index for the whole scene (an average). This index was reduced to discrete values from 0 to 10;
- In relation to the Cloud Coverage, it was kept also the lesser quadrant index. This information was kept trying to minimize the effect caused by the use of

Table 1.1. Scenes

Attribute	Description	Type
Sensor	Satellite Instrument	Nominal
Path	Position in a grid	Numeric
Row	Position in a grid	Numeric
Region	Geographic region	Nominal
Month	Acquisition month (image date)	Numeric
Season	Acquisition season	Nominal
Indicative of Earth or Water	An indicative if the scene was over sea or earth	Nominal
Cloud Coverage Index (average)	Cloud cover index of the scene (0-10)	Numeric
Cloud Coverage Quadrant	Cloud cover index of the lesser cloudy quarter of scene	Numeric
Scenes Class (Used, Not Used)	Classes of use	Nominal
Scenes Class (Not used, Used, Very Used)	Classes of use	Nominal

the average. An user may request a cloudy scene because the area of interest is in a part of the image that is clear, in this case it is possible that the average index is bad but this scene is still interesting for the user;
- For the final classification of a scene, two approaches were analyzed: the use of classes Not Used and Used to indicate the use or not of an image and the use of classes Used, Not Used and Very Used, quantifying the use and trying to determine images with greater potential of interest.

Table 1.1 shows the final version for the Image Table attributes (Scenes).

1.3.4 Sampling

Analyzing the data base used, at the moment of processing, there were 100,352 valid images where 77.69% were Not Used images, 16.87% were images Used (up to 10 requests) and 5.44% were images Very Used (more than 10 requests).

From these data two types of sampling has been carried out:

- Random data sampling by class: 50% of images Not Used (38,983 scenes), 60% of images Used (10,156 scenes) and 80% of images Very Used (4,365 scenes), a total of 53,504 scenes. From these scenes two files were generated, one for training the algorithms and another for validation, in the ratio of 2 registers for training and 1 register for validation as Rezende advises [5];
- Sampling per year: all images of the year 2004 (44,647 images) have been used for the training set and a random sampling of scenes of the year 2005 (18,478 scenes) has been used for the validation set.

1.3.5 Application of Data Mining Algorithms and Results

The application of data mining algorithms and their results is a macro step that can be divided into sub-steps:

- Definition of Data Mining techniques and algorithms: in this sub-step, after the data analysis, the problem is identified and the algorithms to be applied are selected;
- Using the defined patterns extraction tool: in this sub-step the algorithms chosen to be tested are effectively applied to the data through the chosen tool and patterns are obtained, compared and filtered;
- Results: in this latter sub-step, the obtained patterns are consolidated and analyzed.

Definition of Data Mining Techniques and Algorithms

For solving the identified and delimited problem, it was necessary to discover the potential use of the images which would help to get to an automatic criteria for identification of the scenes that should be pre-processed or not, and that should be stored at the disk or not (online), aiming at the customer service improvement without processing overload and with storage optimization. This

problem was identified as being a classification problem. Thus, the scenes have been classified in function of its use and two approaches already described were adopted: Scenes were divided into two groups, where the first one contained Used and Not Used scenes and second one contained Not Used, Used and Very Used scenes. The objective was, at the first moment, to understand which attributes were important in the categorization of the use of a scene and, after that, to verify if it was possible to classify scenes in agreement with these attributes for the chosen category.

An algorithm that implements a decision tree has been chosen, since this kind of tree was capable of explaining which attributes were important, how it classified the data and how it influenced in the determination of the usability of a scene. According to Witten [4], the data explanation is so important or even more important than its classification.

Weka software and the J4.8 algorithm that is an implementation of C4.5 algorithm have been used. The C4.5 algorithm in turn is the ID3 - a classifier that generates a decision tree to classify the test instances - with some modifications.

After that, it was used also a neural network algorithm to classify the data and a comparison of the results was made.

Patterns Extraction Using the Defined Tool

As the first step, before the application of the Classification algorithms, aiming at validating the chosen classes, the clustering algorithm SimpleKMeans's implemented by the Weka tool was applied. The obtained results are shown in Table 1.2.

Table 1.2. Clustering classes

Class Attribute (Clusters)	Samples	Incorrectly Classified	Error (%)
(0) NotUsed (1) Used	35,673	14,809	41.51
(0) NotUsed (1) Used (2) VeryUsed	35,673	20,428	57.26

Then, the J48 algorithm was executed with pruning (a cut in the number of instances in the leaves - minimum of 10 and 50 instances). This choice didn't bring great loss in the classification capability or in the obtained information since the number of instances was relatively large (44,647 training instances).

For the decision tree training, two experiments with attributes that were equivalent in semantic but different in type and granularity have been carried out. The first one prioritized the original numeric attributes: the location in a geographical grid (numerical path and row) and the month of the year. The second experiment used nominal attributes, where the spatial references were the geographical areas that included a set of paths and rows, and the temporal

Table 1.3. J48 - Numeric attributes

Run Information					
Algorithm	J48 (Confidence 0.25 Pruning 10)				
Tree Size	369/723 (Number of Leaves/size of tree)				
Training Set - Results					
Class (Samples)	Correct Classified		Incorrect Classified		
	Samples	(%)	As Class	Samples	(%)
NotUsed (34,174)	32,847	96.12	Used	1,269	3.71
			VeryUsed	58	0.17
Used (7,962)	5,004	62.85	NotUsed	2,554	32.08
			VeryUsed	404	5.07
VeryUsed (2,511)	1,892	75.35	NotUsed	48	1.91
			Used	571	22.74
Total (44,647)	39,743	89.02		4,904	10.98
Test Set - Results					
Class (Samples)	Correct Classified		Incorrect Classified		
	Samples	(%)	As Class	Samples	(%)
NotUsed (14,230)	13,608	95.63	Used	606	4.26
			VeryUsed	16	0.11
Used (2,896)	1,179	40.71	NotUsed	1512	52.21
			VeryUsed	205	7.08
VeryUsed (1,352)	969	71.67	NotUsed	38	2.81
			Used	345	25.52
Total (18,478)	15,756	85.27		2,722	14.73

references were the seasons representing about three months each. It was expected with this second experiment made with the transformed attributes, the obtainment of a data use characteristics description.

A summary of the obtained results is showed in Table 1.3 and Table 1.4.

Observing the tables, it can be noticed that the first experiment showed a tree with better index of classification in the training stage, however larger and more complex; while the second experiment resulted in a tree with a little increase in the classification error, however smaller and more explanatory. It could still be observed that, although the percentage of scenes classified correctly was larger in the case of the numeric attributes, the second experiment classified the test sample better. It is worth pointing out that the second experiment used more generic attributes and a larger pruning in the tree. This might have made possible a better categorization of the new scenes.

The tables also show, through the classification index per classes, that the correctness classification index of Not Used scenes was above the average of the general classification correctness, and that there was a great dispersal in the results of class Used (less than 10 downloads). They still show that the classification error in the scenes Very Used occurs mainly in the Used scenes column. As the problem focused mainly on the scenes Not Used and Very Used

Table 1.4. J48 - Nominal attributes

Run Information					
Algorithm	J48 (Confidence 0.25 Pruning 10)				
Tree Size	68/104 (Number of Leaves/size of tree)				
Training Set - Results					
Class (Samples)	Correct Classified		Incorrect Classified		
	Samples	(%)	As Class	Samples	(%)
NotUsed (34,174)	32,492	95.08	Used	1,627	4.76
			VeryUsed	55	0.16
Used (7,962)	3,628	45.57	NotUsed	3,856	48.43
			VeryUsed	478	6.00
VeryUsed (2,511)	1,712	68.18	NotUsed	137	5.46
			Used	662	26.36
Total (44,647)	37,832	84.74		6,815	15.26
Test Set - Results					
Class (Samples)	Correct Classified		Incorrect Classified		
	Samples	(%)	As Class	Samples	(%)
NotUsed (14,230)	13,869	97.46	Used	358	2.52
			VeryUsed	3	0.02
Used (2,896)	945	32.63	NotUsed	1,743	60.19
			VeryUsed	208	7.18
VeryUsed (1,352)	994	73.52	NotUsed	84	6.21
			Used	274	20.27
Total (18,478)	15,808	85.55		2,670	14.45

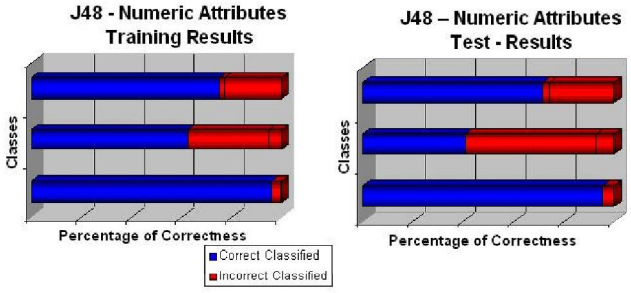

Fig. 1.6. Numeric Attributes - Results

these data characteristics presented by the decision tree did not disturbed the expected results.

Figure 1.6 and 1.7 summarize and consolidate the results. For the choice of numeric attributes (geographical grid and the month of the year), Figure 1.6 shows graphically the percentage of samples that have been corrected (blue) an incorrect (red) classified in both phases: training phase and testing phase. Figure 1.7 shows the same information for nominal attributes (geographical areas and seasons).

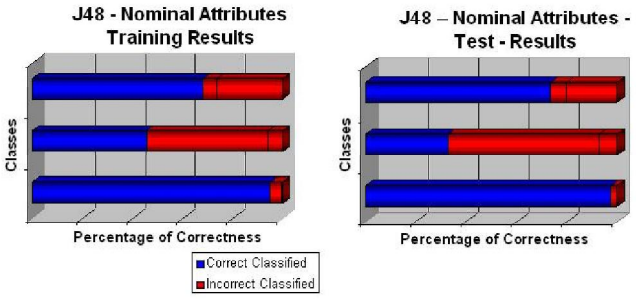

Fig. 1.7. Nominal Attributes - Results

A Weka implementation of Neural Network algorithms were also applied, using the following configurations for the Network: 5 neurons in one hidden layer, 25 neurons in one hidden layer and 25 and 5 neurons in two hidden layers.

The best results obtained with these algorithms are described next.

Results

Some considerations to the application of the algorithms are the following:

Neural Network Algorithm. The results of the application of Neural Network algorithms, for many different designs of the network, were very similar to the J48 results. As the decision tree explains better the data and had better performance, the decision tree has been chosen for the prototype.

Decision Trees Algorithm. Although the application of the clustering algorithm has presented a better result when it considered two classes (Used and Not Used), and, in consonance, the decision tree has classified the images with larger precision, nevertheless, it was preferred to use three classes (Used, Not Used and Very Used), as explained above. Table 1.5 shows the incorrect classification index:

Table 1.5. Incorrect classification index

Class	**Incorrect Classified**		
	As Very Used	As Used	As Not Used
Very Used (7%)		22.74%	1.91%
Used (18%)	5.07%		32.08%
Not Used (75%)	0.17%	3.71%	

Because of the data characteristics, it was obtained a better result of the decision tree for the training set when all scenes of whole year (2004) were used. The generated tree was tested with a random scenes set of the year 2005.

From the decision tree generated through the nominal attributes, some inferences concerning the data use characteristics could be made:

- The main cut in the tree was made on cloud coverage index in two main points: less than 30% with larger percentage of Used and Very Used scenes and above 80% with most of the scenes Not Used. These values - 30% and 80% of coverage - brought quite important information enriching the intuitive choice criteria;
- Another important cut in the tree was the indicative of Water or Land or if scenes are inside the country or not (Region = Other);
- 10% or less for the cloud coverage index normally classifies the scenes as Very Used;
- When season was winter, for the same values in the other attributes, the images were always classified as higher level than the other season's images.

Many inferences made using the results were in agreement and close to the known patterns in the data behavior and, sometimes it is interesting to prove patterns that are known by the experience, or logic, or even intuitively through an algorithm information, but in this case, the values and limits found are more precise and can be useful in decision making over the data. It was obtained an index of about 97% of correctness for the scenes that certainly will never be requested and an index of about 73% for the scenes that probably will be used many times and has to be in an online archive.

Comparing the results with the criteria previously being used, the work realized can:

- Provide an economy of:
  1. About 50% in processing time;
  2. About 10% in the storage space (12 Tbytes).
- Improve the delivery time:
  1. The number of requests delivered in 0 minutes has been increased to 41% of the total (today is 17%);
  2. The number of requests delivered in 9 minutes has been increased to 43% of the total (today is 36%);
  3. The number of requests delivered in more than 9 minutes has been decreased to 16% of the total (today is 47%).

1.4 Conclusion

This chapter presented the use of Data Mining algorithms in a real context of a Satellite Images Processing and Distribution System by Internet.

A prototype using Data Mining techniques, mainly classification algorithms, was developed in order to determine the use potentiality of each image received through the Satellite, automating the choice criteria of each image to be preprocessed or entirely processed a priori.

The main objective with the use of data mining techniques is the improvement of the distribution services rendered, as well as the improvement of the processing resources and storage administration.

The results obtained from the pre-processed data presented to the algorithms showed patterns that are close to the patterns already known, and the analysis of the decision tree emphasized some characteristics in the classified data, bringing some new and important knowledge of values for some known patterns.

As already said, for the considered problem, the use of a classifier algorithm, especially if centered in the Not Used and Very Used images, seemed to be quite interesting in the election of scenes to be discarded or processed/pre-processed and in the management of the data storage (online and near line images). The application of Data Mining techniques allows an intelligent automation of the process, bringing economy in the processing effort, in the storage space, and decreasing the request delivery time of images.

References

1. Braga, L.P.V.: Introdução á Mineração de Dados, 2. Edição revista e ampliada, Rio de Janeiro: E-Papers Serviços Editoriais, p. 212 (2005)
2. Fayyad, B.M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview. In: Knowledge Discovery and Data Mining. AAAI Press, Menlo Park (1996)
3. Goldschmidt, R., Passos, E.: Data Mining - Um Guia Prático. Elsevier, Rio de Janeiro (2005)
4. Witten, H.I., Frank, E.: Data Mining - Practical Machine Learning Tools Techniques and storage diseases, 2nd edn. Elsevier, MA (2005)
5. Rezende, S.O.: Sistemas Inteligentes: Fundamentos e Aplicações. Manole, São Paulo (2003)