View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Servicio de Difusión de la Creación Intelectual

Thesis Overview

KNOWLEDGE EXTRACTION IN LARGE DATABASES USING ADAPTIVE STRATEGIES

Waldo Hasperué
Advisor: Laura Lanzarini
Facultad de Informática, Universidad Nacional de La Plata
PhD Thesis, March 2012
whasperue@lidi.info.unlp.edu.ar

Motivation

The general objective of this thesis is the development of an adaptive technique for extracting knowledge in large databases.

Nowadays, technology allows storing huge volumes of information. For this reason, the availability of techniques that allow, as a first stage, analyzing that information and obtaining knowledge that can be expressed as classification rules, is of interest. However, the information available is expected to change and/or increase with time, and therefore, as a second stage, it would be relevant to adapt the knowledge acquired to the changes or variations affecting the original data set.

The contribution of this thesis is focused on the definition of an adaptive technique that allows extracting knowledge from large databases using a dynamic model that can adapt to information changes, thus obtaining a data mining technique that can generate useful knowledge and produce results that the end user can exploit.

The results of this research work can be applied to areas such as soil analysis, genetic analysis, biology, robotics, economy, medicine, plant failure detection, and mobile systems communications. In these cases, obtaining an optimal result is important, since this helps improve the quality of the decisions made after the process.

CLUHR

When the data domain to be treated is continuous, hyper-rectangles are a convenient way of representing the data and have been used in several data mining works. Hyper-rectangles are a powerful way of representing data, since they can describe in an almost natural manner the data subsets that they represent. This is because the boundaries of each hyper-rectangle formed in the data model can be used as clauses in the IF-THEN rules resulting from the knowledge extraction process. At the same time, and due to their features, they can be easily handled in the input data domain, which enables contracting, merging and dividing with simple operations.

While working on this doctoral thesis, a technique called CLUHR was defined and implemented. This technique extracts knowledge from large volumes of information as classifications rules that help the end user understand the data being used and make relevant decisions. These are strict rules, formed by the hyper-rectangles resulting from a data model that is created, and they represent the boundary between two hyper-rectangles in a fully rigid way.

The system generates hyper-rectangles from the initial data and, based on an iterative process that removes overlaps, it makes decisions regarding which overlaps should be removed based on the computation of a set of overlap indexes that are proposed and developed in this thesis. These decisions cause hyper-rectangles to change their size or be divided. This optimization process continues until the intersection volume between the different classes of hyper-rectangles is minimized (or removed). Finally, the rules resulting from the hyper-rectangles obtained are generated.

Building the model with expert monitoring

To build the data model, the classification algorithm has to make numerous decisions. Any decision regarding which overlap to solve is made by computing the overlap indexes. This index can be calculated in many different ways, and the elements used to calculate it can be weighed differently also. On the other hand, once the overlap that is to be removed has been selected, a decision has to be made regarding the need - or not - to divide the hyper-rectangles, reduce their volume, which hyper-rectangle to divide, or how to carry out this division. In this sense, the technique proposed in this thesis does not make any decision in particular, since the correct option will to a great extent depend on the problem.

In any case, all these decisions can be made before building the model, configuring the algorithm to build and running the model in a fully automated manner to achieve the end result. Even though the technique proposed in this thesis is entirely flexible in the sense that an expert, before running the process, can determine which indexes to use and how to carry out the divisions, customizing the automated process, in real problems it is never possible to build an automatic process that is able to improve the "on-line" decisions that would make a human being, even more so when this human component is an expert on the domain of the problem.

The strategy presented in this thesis can be used either as a fully automated process or an interactive process with the participation of an expert in the domain of the problem, with the following features:

- Flexibility in the use of indexes.
- Possibility of assigning different weights to each class.
- Freedom to decide how to carry out a division or volume reduction.
- Possibility of running the process in a fully automated mode, fully expert-monitored mode, or partially expert-monitored mode.

Adaptability

As for adaptive behavior, the new data that are entered to the data model cause the hyper-rectangles to change, which in turn will eventually cause new splits or joins. This means that there is no need to rebuild the model using the entire data set, but only adapting it by modifying its internal structure upon the arrival of new data. Once the internal structure of the model is updated, the set of existing rules is also updated.

The adaptive strategy proposed is characterized for allowing the user to act as a monitor and take part in the decision made by the algorithm when creating or adjusting the hyper-rectangles. Thus, an expert in the domain of the problem can contribute extremely valuable information during the creation of the data model.

At the same time, as well as offering the various alternatives and possible solutions with varying validity levels, the model itself can make suggestions to the user regarding specific actions that can be done based on the knowledge acquired and data dynamic trends, if they are being updated.

Thus, an adaptive strategy has been established that is capable of building a first model from a database and then gradually adapt as new information is acquired. Even though this strategy can operate in a fully automated manner, it also offers the possibility of having and expert taking part of the process, who can participate with various levels of involvement in building the model.

Results

Firstly, various problems are analyzed with two-dimensional databases. These problems allow a detailed analysis of the operation of the strategy proposed for different spatial distributions of the data used for building the model. These problems used as examples are artificially built databases that allow representing in a figure the spatial distribution of their data.

A series of problems with two, three, and up to four data classes, problems with various spatial distributions, problems that can be easily solved, and problems that require several hyper-rectangle split operations to build the data model, are also analyzed. The resources used by the strategy to build a model are also analyzed. The resource measured is the number of times the database is examined while building the model.

CLUHR was tested with 13 databases from UCI. This testing process was carried out to compare the results obtained with those from methods C4.5 (Quinlan, 1993), EHS-CHC (García et al., 2009) and PSO/ACO2 (Holden et al., 2008). Using the 10-fold cross-validation test, the accuracy achieved by the data model built, the number of rules extracted, the average number of clauses per rule, and the number of times the database is examined to build the data model, were measured.

As regards the accuracy obtained by the data model (Table 1), the number of rules extracted (Table 2), and the average number of clauses per rule (Table 3), it cannot be concluded that CLUHR is either better or worse than the other strategies used for comparison. It may be slightly worse than PSO/ACO2, since this method allows the presence of "ambiguity" in the set of rules. These rules, on the other hand, have a certain order of execution. Also, since this is an optimization strategy, the expectation is that a particle representing an optimal result will be found.

Even though CLUHR does not seem to be better than the other techniques that were studied, yielding similar results regarding accuracy, number of rules extracted and average number of clauses per rule, it does present two great advantages:

- The same as C4.5, CLUHR is a deterministic strategy, meaning that the same input always produces the same output, which is something that EHS-CHC and PSO/ACO2, or any other optimization strategy for that matter, can ensure.
- The number of times the database is examined with CLUHR to build the data model is much lower than that of the optimization strategies and slightly lower than in the case of C4.5, with an average that is twice lower than that of C4.5 (Table 4).

The incremental aspect of CLUHR was compared against the ITI technique (Utgoff et al., 1996). The main problem with the ITI technique, as with any decision-tree-based technique, is that data build-up and node decision function re-assessment make, sooner or later, sub-tree restructuring necessary. When the sub-tree to be redone has as its root a node from one of the first levels, the restructuring work to be done is significant, since a large portion of the database has to be re-examined. The worst-case scenario would be when the node to be restructured is the root node of the tree, resulting in a full restructuring process with the corresponding full examination of the database.

In CLUHR, the addition of new data causes changes in only one hyper-rectangle. If the affected hyper-rectangle overlaps with other hyper-rectangles, the latter are also modified.

The results obtained show that CLUHR requires much less computational effort than that required by decision trees based on the ITI algorithm (Table 5).

Future Work

As regards the overlap indexes used to decide which overlap was to be removed, the six indexes that are proposed in this thesis proved to be capable of successfully solving numerous problems.

In this realm, the development of new indexes that measure other aspects of an overlap can be studied, for instance, indexes that simultaneously measure the characteristics of more than two classes, or indexes that simultaneously measure aspects with more than one dimension.

In certain cases, smaller hyper-rectangles can be merged into a single, larger one, which would result in fewer rules in the data model.

Also of interest is researching the possibility of merging hyper-rectangles from a same class while the model is being built, or if this merging process should be considered as an additional activity that should be carried out as the final stage of the simplified-rule extraction process. Improving this aspect would benefit the procedure for extracting simplified rules, since a smaller number of rules would have to be used.

In this regard, on of CLUHR's weaknesses is that the end process for extracting a set of simplified rules uses a greedy, order O(n2) procedure. Even though different approximations have been proposed to perform this work in a more efficient manner, the possibility of somehow "marking" those hyper-rectangle faces that represent a boundary within the data space is of interest. Since from each of the faces of a hyper-rectangle a clause is extracted for the rule corresponding to that hyper-rectangle, if the faces representing data space boundaries were identified, the corresponding rules would be directly excluded and no simplification method would be required.

References

Quinlan John Ross C4.5: Programs for Machine Learning. - Morgan Kaufmann Publishers, Inc., 1993. - ISBN 1-55860-238-0.

Holden Nicholas & Freitas Alex A. A hybrid PSO/ACO algorithm for discovering classification rules in data mining. Journal of Artificial Evolution and Applications. - 2008. – 2008:1-11.

García Salvador. A First Approach to Nearest Hyperrectangle Selection by Evolutionary Algorithms. Proceedings of the 2009 Ninth International Conference on Intelligent Systems

Design and Applications, 2009. - pp. 517-522.

Utgoff Paul E. ID5: An Incremental ID3. In Proceedings of ML. - 1988. - pp. 107-120.

Tables

Table 1. Model accuracy achieved by each of the studied strategies and for each of the tested databases.

	C4.5	EHS-CHC	PSO/ACO2	CLUHR
E. coli	0.7964 (0.0141)	0.7948	-	0.7891 (0.0160)
Glass	0.6576 (0.0302)	0.6287	0.7095 (0.075)	0.6215 (0.0360)
Haberman	0.7103 (0.0202)	0.7122	-	0.7356 (0.0064)
Image	0.8586 (0.0155)	-	0.9667 (0.0117)	0.8538 (0.0135)
Ionosphere	0.9054 (0.0151)	-	0.8806 (0.0491)	0.8777 (0.0169)
Iris	0.9420 (0.0077)	0.9267	0.9467 (0.0526)	0.9300 (0.0079)
Liver	0.6418 (0.0300)	0.6167	-	0.5918 (0.0211)
Pima	0.7434 (0.0093)	0.7384	-	0.5595 (0.0191)
Sonar	0.7235 (0.0247)	0.7650	0.7505 (0.0911)	0.6666 (0.0283)
Vehicle	0.7111 (0.0099)	-	0.7305 (0.0445)	0.6819 (0.0171)
Vowel	0.6008 (0.0158)	-	0.8616 (0.0347)	0.7120 (0.0132)
Wine	0.9141 (0.0145)	0.9490	-	0.9530 (0.0113)
Wisconsin	0.9446 (0.0047)	0.9599	0.9487 (0.0253)	0.9251 (0.0102)
Forest covertype	0.7063 (0.0187)	-	-	0.6928 (0.0149)

Table 2. Number of rules extracted by each of the studied strategies and for each of the tested databases.

	C4.5	EHS-CHC	PSO/ACO2	CLUHR
E. coli	12.1 (1.45)	11.1	-	12.62 (1.44)
Glass	14.8 (0.79)	12.2	20.4 (1.35)	15.17 (1.30)
Haberman	10.7 (3.62)	4.4	-	4.29 (0.33)
Image	10.6 (0.70)	-	21.9 (0.99)	10.93 (0.47)
Ionosphere	10.2 (2.04)	-	3.6 (0.97)	3.98 (0.37)
Iris	4.0 (0.47)	3.4	3.0 (0.00)	3.21 (0.12)
Liver	23.9 (4.46)	9.8	-	17.79 (2.21)
Pima	13.2 (1.40)	11	-	10.45 (0.91)
Sonar	10.9 (1.60)	10.3	4.4 (1.58)	4.14 (0.20)
Vehicle	31.0 (2.31)	-	37.8 (1.2)	32.35 (2.03)
Vowel	32.8 (2.20)	-	29.0 (0.82)	31.74 (0.78)
Wine	5.1 (0.57)	3.6	-	3.18 (0.11)
Wisconsin	11.9 (1.79)	3.8	10.2 (1.87)	9.63 (1.39)
Forest covertype	39.7 (2.35)	-	-	41.25 (2.05)

Table 3. Average number of clauses per rule extracted by each of the studied strategies and for each of the tested databases.

	C4.5	PSO/ACO2	CLUHR
E. coli	4.32 (0.30)	-	4.65 (0.15)
Glass	5.68 (0.75)	3.11 (0.18)	5.37 (0.18)
Haberman	4.54 (1.27)	-	2.54 (0.06)
Image	4.31 (0.58)	2.8 (0.27)	3.74 (0.10)
Ionosphere	5.36 (0.89)	3.33 (0.79)	5.17 (0.18)
Iris	2.25 (0.27)	0.93 (0.14)	2.08 (0.05)
Liver	6.80 (1.30)	-	5.01 (0.06)
Pima	4.55 (0.27)	-	5.27 (0.12)
Sonar	3.99 (0.43)	2.6 (0.63)	16.27 (0.72)
Vehicle	7.10 (0.34)	3.85 (0.18)	7.38 (0.33)
Vowel	5.69 (0.18)	4.2 (0.25)	8.13 (0.27)
Wine	2.46 (0.17)	-	4.08 (0.09)
Wisconsin	4.31 (0.39)	1.21 (0.07)	3.59 (0.11)
Forest covertype	6.67 (0.82)	-	6.49 (0.48)

*Table **¡Error! No hay texto con el estilo especificado en el documento.**. Resources used: number of times the database is examined. Strategies EHS-CHC and PSO/ACO2 require over 2000 times.*

	C4.5	CLUHR	Significance
E. coli	4.19 (0.39)	3.53 (0.33)	+
Glass	5.64 (1.10)	3.97 (0.37)	+
Haberman	3.61 (1.26)	5.28 (0.32)	-
Image	3.84 (0.35)	1.67 (0.06)	+
Ionosphere	5.78 (0.73)	2.47 (0.14)	+
Iris	2.02 (0.13)	1.5 (0.06)	+
Liver	6.59 (1.48)	5.20 (0.50)	+
Pima	3.74 (0.24)	4.97 (0.39)	-
Sonar	4.03 (0.49)	2.41 (0.17)	+
Vehicle	5.98 (0.24)	5.06 (2.56)	=
Vowel	5.54 (0.13)	2.99 (0.11)	+
Wine	2.34 (0.10)	1.20 (0.01)	+
Wisconsin	3.19 (0.35)	3.02 (0.32)	=
Forest covertype	5.71 (0.72)	5.24 (0.45)	=
Total			+7

Table 5. Resources used: number of times the database is examined.

	ITI	CLUHR	Significance
E. coli	5.19 (0.95)	0.59 (0.45)	+
Glass	14.50 (2.56)	0.44 (0.11)	+
Haberman	12.84 (2.24)	0.99 (0.25)	+
Image	2.37 (0.44)	0.19 (0.15)	+
Ionosphere	1.71 (0.30)	1.59 (0.43)	=
Iris	0.25 (0.05)	0.90 (0.50)	-
Liver	14.58 (2.70)	0.61 (0.16)	+
Pima	21.53 (3.69)	1.31 (0.37)	+
Sonar	5.11 (0.90)	0.06 (0.05)	+
Vehicle	14.42 (2.48)	1.07 (0.32)	+
Vowel	31.20 (5.30)	0.11 (0.05)	+
Wine	1.32 (0.26)	0.26 (0.41)	+
Wisconsin	1.65 (0.30)	8.31 (2.11)	-
Total			+8