УДК 378

CLUSTERING AND MAKING DECISIONS METHODS FOR INTELLIGENT SYSTEMS

Miguel A. Garay Garcell[1], Marcello Sandi Pinheiro[2], Vanessa Teixeira de Oliveira Sandi[3]

Research Center for Systems Engineering Havana Institute of Technology, Ciudad de La Habana, Cuba (1), Catholic University of Brasilia Brasilia, Brasil (2), CTIS Software House of Informatics, Catholic University of Brasilia, Brasilia, Brasil (3)

Represented by the member of the Editorial Board Professor S.V. Mishchenko

Key words and phrases: clustering; intelligent systems; hypermedia; multimedia; systems engineering.

Abstract: The processes involved in the classification of objects and word problems continue being a complex and ill-structured problem. In this paper, a software tool for Hypermedia Intelligent Tutoring Systems is proposed. This tool helps us to define the complexity levels for a set of proposed problems to students. In this way, the intelligent systems designers and professors can use it to organize the problem base. On the other hand, this tool constitutes an important element during training and learning process because it permits to organize the transit from simple to complex in problem solving. It uses different methods of classification, like as statistical methods, experts' proofs and others. The evaluations of parameters can involve objective and subjective elements. In this relation, a particular version of Delphi Method is developed. The system was implemented on Delphi 4.0 for Windows.

Introduction

Since ancient times, the mankind history has been full of situations in which the man had necessity of classifying objects with similar or analogous characteristics (natural phenomenon, concepts, things and others). Classification is an essential tool in science. The concept Classification appears frequently in all branches of science. By Classification, Clustering, Cluster Analysis or Taxonomy we understand a process of grouping, organizing or ordering different objects into certain classes. Etymologically, Class (Classe in French and Classis, from Latin) means a group, set or kind sharing common attributes, a division or rating based on grade or quality. Classify – to arrange in classes, to assign some thing to a category. In this relation, the concept Classification is the act or process of classifying, systematic arrangement in-groups or categories according to established criteria. The science of Classification in broad sense is usually restricted to biological classification and specifically to the classification of plants and animals. The term is derived from the Greek: Taxis («Arrangement») and Nomos («Law»). Taxonomy is, therefore, the methodology and principles of systematic botany and zoology, and sets up arrangements of the kinds of plants and animals in hierarchies of superior and subordinate groups. In the present paper we state by objective to design a software tool for classification of optimization problems. Hypermedia Intelligent

Tutoring Systems (HITS) design is a complex and ill-structured problem. In it, must be solved the following problems:

- User identification and evaluation;
- A problem bank or base must be constructed;
- The problems selected by HITS must increase the scientific and cognitive interest of students;

This tool helps us to define the complexity levels for a set of proposed problems to students. In this way, the intelligent systems designers and professors can use it to organize the problem base. On the other hand, this tool constitutes an important element during training and learning process because it permits to organize the transit from simple to complex in problem solving. It uses different methods of classification, like as statistical methods, experts' proofs and others. The evaluations of parameters can involve objective and subjective elements. The software was implemented in Delphi 4.0.

Information measure

The problem addressed is the partition of a population of S things into classes, where each thing is characterized by measurement values $x\,[d,\ s]$ on D different attributes ($s = 1, 2, ..., S$; $d = 1, 2, ..., D$). The aim is to produce classes such that things in the same class are sufficiently similar to be treated as equivalent for some proposes. This is the classic taxonomic problem, and has been attacked by many authors (e.g. Sokal and Sneath, 1963, Lance and Williams, 1967a, 1967b, Wallace and Boulton, 1968, 1970, 1973). In an earlier paper Wallace and Boulton (1968) suggested that progress on this problem could be accelerated if an attempt was made to define a criterion or figure of merit designed to measure the degree to which a particular classification of a population achieved its aims. As a possible criterion, they defined the «**information measure**» of a classification. In 1970 they defined the information measure only for non-hierarchic classifications, that is, classifications in which each class is defined and its properties stated independently of other classes. Later, they derived the information measures for two kinds of hierarchic classification, and discuss the computer strategies used to generate optimum hierarchic classifications.

Terminology

Let suppose that we have certain set of data in table $N \times n$ object – characteristic, where N – number of objects, and n – number of characteristics. According to this point of view, to classify objects, it is necessary to define an adequate measure of similarity between the objects and then implement it creatively to determine the corresponding classes. Suppose we have certain set of objects, then we mean by classification the partition or division of certain set of objects into subsets of analogous objects. These subsets are known as classes. Two classes C_A and C_B are equal, written $C_A = C_B$, if they consist of the same elements, i.e. if each member of $C_A \subset C_B$. The intersection of two classes C_A and C_B, denoted by $C_A \cap C_B$ is the set of elements which belong to both C_A and C_B, in example, $C_A \cap C_B = x : x \in C_A$ and $x \in C_B$. If $C_A \cap C_B = 0$, then C_A and C_B are said to be disjoint or non-intersecting. In relation with these formulations appears the concept of equivalence relations. A relation R in a set A, i.e. a subset R of $A \times A$, is termed an equivalence relation if it satisfies the following axioms:

Reflexive: For every $a \in A$, $[a, a] \in R$

Symmetric: If [a, b] ∈ R, then [b, a] ∈ R

Transitive: If [a, b] ∈ R and [b, c] ∈ R, then [a, c] ∈ R

Accordingly, a relation R is an equivalence relation if it is reflexive, symmetric and transitive. In this relation, A class $\mathbf{C_A}$ of non-empty subsets of a set **A** is called a partition of **A** if:

1. each $a \in A$ belongs to some member of C_A and
2. the members of C_A are pair-wise disjoint.

In this relation, we can represent the relations between elements of any set in form of boolean matrix. Let r_{ij} the relation between node i and node j. If exists a certain link between i and j, then $r_{ij} = 1$ and $r_{ij} = 0$ in another way. Another important concept is the Euclidean Metric or Distance Function. Let R^m denotes the product set of m copies of the set R of real numbers, i.e. consists of all m-tuples $a_1, a_2, ..., a_m$ of real numbers. The function d defined by $d(p, q) = \sqrt{\sum (ai - bi)^2}$, where: $i = 1, 2, ..., m$; $p = a_1, a_2, ..., a_m$ and $q = b_1, b_2, ..., b_m$; is a metric, called the Euclidean Metric or Distance on R^m.

Classification can be defined according to one of following principles:

1. At first, the classes of objects are partitioned, then the main characteristics for these objects are described or defined or determined with precision;
2. At first, the characteristics of objects and its values are given, then the classes of objects are building or designed or defined;

Let X be a non-empty set. A real-valued function d defined on $X \times X$, i.e. ordered pairs of elements in X, is called a metric or distance function on X if it satisfies, for every $a, b, c \in X$, the following axioms:

1) $d(a, b) \geq 0$ and $d(a, a) = 0$;

2) (symmetric) $d(a, b) = d(b, a)$

3) (triangle inequality) $d(a, c) \leq d(a, b) + d(b, c)$

4) If $a \neq b$, then $d(a, b) > 0$

The real number $d(a, b)$ is called the distance from a to b.

Classes are defined as collections of objects whose intraclass similarity is high and interclass similarity is low. Because the notation of similarity between objects is fundamental to this view, clustering methods based on it can be called similarity-based methods. Many such methods have been developed in numerical taxonomy, a field developed by social and natural scientists, and in cluster analysis, a subfield of pattern recognition. Various similarity measures and clustering algorithms have been developed in the last years. Another view recently developed in Artificial Intelligence postulates that objects should be grouped together not just because they are similar according to a given measure, but because as a group they represent a certain conceptual class. This view, called Conceptual Clustering, states that clustering depends on the goals of classification and the concepts available to the clustering system for characterizing collections of entities. Clustering is the basis for building hierarchical classification models.

Problem solving

Problem solving process continues being a complex and ill-structured problem. From etymological point of view, the problem concept is related to:

- Proposition or difficulty of getting certain solution with uncertainty;
- Set of facts and circumstances that make difficult to get certain objective;
- A question raised for inquiry, consideration, or solution; a source of perplexity, distress, or vexation;
- Difficulty in understanding or accepting;
- Proposition directed to inquire the way of getting certain result when some data are known.

On the other hand, a solution is an action or act to dissolve a doubt. In this relation, we can say that a problem exists when a learning person is in from of a situation that he can‘t solve immediately. To have a problem implies having certain information on its characteristics and structure. It is necessary to define the objective of problem solving process, to specify in what conditions the problem must be solved and by means of what mechanisms or instruments and with what resources the problem will be solved. Mathematically, we can say that a problem can be represented by means of the following expression: <given *V*, it's necessary to get *W* > or <*V*; *W*>, where: *V* – are the given conditions; and *W* – are the objectives to obtain. The objective *W* defines a desired situation or state. The concept of problem structure also results an important concept, from the theoretical and practical point of view. Simon (1973) gives a set of criteria for defining a problem structure. They are:

- Existence of a problem space with one start state defined and all possible intermediate states;
- All transformations of state can be represented in a space of problems;
- All relevant knowledge can be represented on a space problem;
- The problems include the actions of real world in accord to the state transformations and its effects.

In relation with the elements before analyzed, the principle of maximum standardization of software express, formulate the idea of obtaining design solutions that can be adapted to many objects. Hypermedia Intelligent Tutoring System (HITS) design and development is a multistage process that has been developed during several years. In this relations, it's necessary the research groups don't limit its analyze to concrete conditions of one specific system, otherwise they must achieve a required level of generalization. On this base, it's convenient to establish during system development process a clear difference between elements of general and particular character. This research will permit, a posteriori, a rational adaptation of system to others analogous objects. During project design, it's necessary to analyze carefully the main parameters for every object and word problem in order to classify them. This element is very important during the process of structuring and identification. Our experience shows that the expert criteria and evaluations tend to converge during the research process.

Computer classification problem

The classification of objects and word problems consists in the determination of several levels of complexity. These levels are defined in accord to the characteristics of every specific problem results an interesting computer problem. It is possible to define a problem in function of a set of parameters. In this relation, it's necessary to define that set in one domain or knowledge area of analyzed object or word problem. This requires consulting groups of experts in this knowledge area with the objective of establishing the more representative or significant parameters. Here, a method to divide in n levels of complexity the different objects and word problems starting from values of parameters used for characterizing them is proposed. Specifically, a software tool on Delphi 4.0 of Borland Corporation was developed. The proposed method consists in the following: Suppose we have m objects or word problems that must be classified in different classes in accord to the values reached for the more relevant parameters of each object or word problem. These parameters can be of subjective and objective character on their activity. In this relation, it's possible to design or construct a matrix that relates the problems and its parameters. Let $A = //Aij//$, where $i = 1, 2, ..., n$; In this matrix, *Aij* is a value correspondent to object or word problem «*i*». How was observed before, on the matrix there exists objective and subjective evaluations. The subjective evaluations require a special treatment. In this work, it is proposed to use different methods for making a scientific evaluation of values for each parameter. The methods analyzed are: 1) Delphi Method; 2) Kendall Tau Method; and others.

Procedure

1. Select for each column «*j*» «of matrix $A = //Aij//$ the value: max {*Aij*} and min {*Aij*}.
2. Determine the difference between max {*Aij*} and min {*Aij*}.
3. Calculate the middle value $Aij = [\max \{Aij\} - \min \{Aij\}] / 2$.
4. Classify the set of objects or word problems in function of its location. In this relation, «*n*» intervals of variation or complexity levels for a set of problem are defined.
5. When all parameters have been analyzed it is necessary to define in what class must be located each problem. This step can be made in different ways: a) locate a problem in the complexity level or class that have been repeated more times during the evaluation process. b) use the decision theory methods for deciding in what complexity level locate each problem. (See Table 1).

The development system was designed in a way such that it permits to use different decision-making methods for different problem situations. Thus, for example, the system permits to use decision-making methods and criteria (see [14], [15], [16]). The hole analyze of these criteria permit to select the alternative more convenient in from of different decision situations. In 1994 was developed experimentally a Problem Classifier at Havana Institute of Technology. This classifier has been used in different academic courses, such as Mathematics, Physics, Chemistry, Biology, and others. On the other hand, Classifier can classify objects. For example, it can be used in the classification of sugar factories, transport objects, etc. This software tool can be used if it is possible to define objects or word problems as a function of certain set of parameters and weights. These can change according to a relative importance of each parameter. A parameter or attribute is used in order to measure the efficiency in relation with a determined objective. The attributes initially can be vague and later can be defined with mayor precision depending of knowledge level acquired by experts.

Example

Suppose that we have certain set of problems in determined knowledge area and desire to model several activities in sugar factories. In this relation, we can establish a hierarchical multilevel representation (see Fig. 1); in other words, we can create a problem situation to analyze the whole operation of sugar factories. If we are developing a Hypermedia Intelligent Tutoring System (HITS), we can use this approach to construct a hierarchical problem bank or problem base of system. The problems will different according to their complexity level. In this relation, the problems proposed to students can transit from general to specific and from simple to complex. The problems will propose to students step by step in accord to their grade of efficiency in different stages of training process. In this relation, to classify the problems defined in the problem bank is an important scientific and technical problem. To classify the problems it is necessary to use the methods analyzed before in this work In these conditions, an implementation of a software tool to classify a problem set constitutes a more efficient way. We can classify the problems in «n» complexity levels or classes in function of different parameters and weights. In the present example, to classify different sugar factories' models we can take: 1) Number of equations; 2) Number of variables; 3) Capacities of sugar factories; 4) Demand and others, in quality of parameters. These parameters and its values can be defined by experts. In this relation, we can use the Delphi Method in order to minimize the data processing and improve the quality in decision-making process.

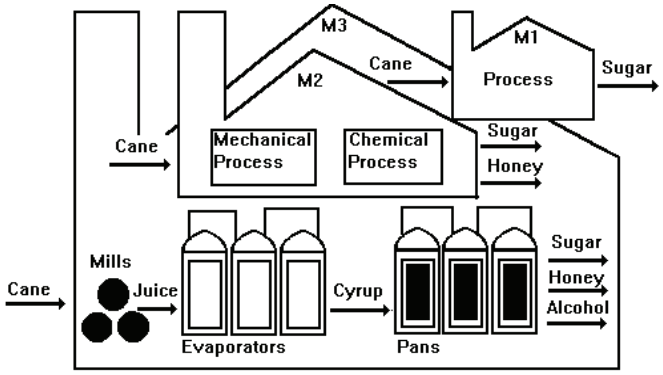

Fig. 1 Hierarchical multilevel representation of a sugar cane factory

Table 1

Complexity Levels of parameters

Classes	Inferior limit	Superior limit
1	$\min_i \{Aij\}$	$1/3\ [\max_i\{Aij\} + 2\ \min_i\{Aij\}]$
2	$1/3[\max_i \{Aij\}+ 2\min_i\{Aij\}]$	$1/3[2\ \max\ \{Aij\} +\min_i \{Aij\}]$
3	$1/3[2\ \max_i \{Aij\} + \min_i \{Aij\}]$	$\max_i\{Aij\}$

Computer example

In this session, the windows used by the Problem Classifier will be presented. Fig. 2 shows the software interface of Problem Classifier. In this case, a classification of algebra word problems is developed. Fig. 2 shows the selected parameters in order to classify the problem set. Here, the selected parameters were a) number of equations; b) number of variables; c) complexity of problem. It can be introduce to analyze other factors, such as: a) characteristics of right hand coefficient; b) interpretation of problems; c) time of problem solution, and other factors can be considered. System out consists of a definition of membership of each problem to one of defined complexity levels or classes. Each problem is located in one class or complexity level in function of values assigned to parameters and certain weight associated to specific parameter.

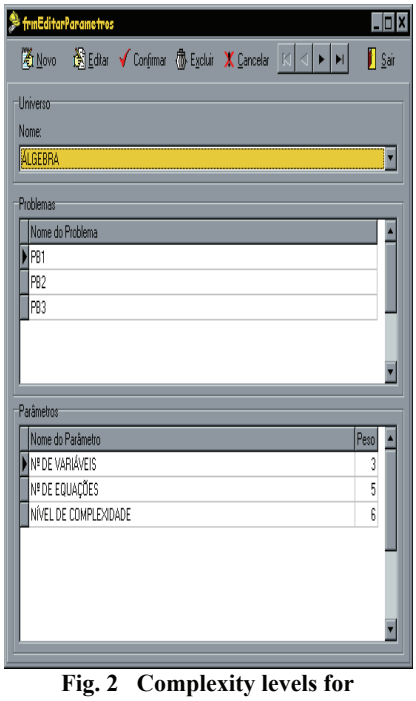

Fig. 2 Complexity levels for algebra problems

Conclusion

In the present work the concept of problem and problem structure was analyzed. On this conception, the idea of defining a structure of problems in function of certain set of parameters and decision-making criteria was formulated. In this relation, the

classification problem of objects and word problem was treated. The developed software tool permitted to find a computer solution to process of subjective evaluations by means of different expert methods, such as Delphi Method, Kendall Tau Method and others. The software tool was implemented in Delphi version 4.0 of Borland Corporation. Problem Classifier has been used for classifying different objects and word problems and it can be used also to classify the academic state of each student during a training session. It is an interesting question because the academic state of students can change dynamically during the training sessions using a Hypermedia Intelligent Tutoring System. Problem Classifier has been used to classify word problems in different subject matters, such as Physics, Mathematics, Biology, Chemistry and others. Problem Classifier has been introduced successfully in several Latin American universities and schools in Cuba, Mexico, Brazil and Ecuador. This software tool can help teachers and professors to organize the problems banks in general and specifically to work with a Hypermedia Intelligent Tutoring System.

References

1. Boulton, D. M. and Wallace, C. S. (1970), «A Program for Numerical Classification,» The Computer Journal, Vol. 13, No. 1, p. 63.

2. Lance, G. N. And Williams, W. T. (1967a), «A General Theory of Classificatory Sorting Strategies», I, Hierarchical Systems, The Computer Journal, Vol. 9, No.4, p. 373.

3. Wallace, C. .S. and Boulton, D. M. (1968), «An information measure for classification», Computer Journal, Vol. 11, No. 2, p. 185.

4. Dunn, J. C. (1974)«A fuzzy relative of the ISODATA process and its uses in detecting compact well-separated clusters», Journal of Cybernetics, No. 3, pp. 32-57.

5. Späth, H. (1975) «Cluster analysis algorithms». Wiley, New York, 226 pp.

6. Maxrov, N.V., A.A Modin, and E.G., Yakovenko, (1974) «Design parameters on modern MIS in enterprises», Chapter 3, pp. 39 – 71 Ed. Nauka. Moscow.

7. Garay M.A, C. Sotolongo (1982), «Método para la clasificación de Empresas por computadoras.» Revista Ingenierна Industrial. Junio/1982. ISPJAE, MES. La Habana. Cuba.

8. Chiang, L.A.(1994), «Clasificador de Problemas.» Congreso Internacional Informбtica 94. Memorias del Congreso. Palacio de las Convenciones. La Habana. Cuba.

9. Savage L.J. (1951), «The Theory of Statistical Decision,» Journal of the American Statistical Association, Vol. 46, pp. 56 – 67.

10. Wald A., (1950), «Statistical Decisions Functions,» John Wiley & Sons, New York.

11. Hurwicz L. (1951), «Optimality Criteria for Decision Making Under Ignorance,» Cowles Commission Discussion Paper, Statistics, No. 370.

12. Keeney, R. L. & Raiffa H., (1976), «Decisions with multiple objectives: Preferences and Value Tradeoffs,» John Wiley & Sons, Inc. New York.

13. Garay Garcell, Miguel (1991), «La inteligencia artificial en la enseñanza de la Modelación Matemática». II Congreso Mundial de Educación y Entrenamiento en Ingeniería y Arquitectura. Palacio de Convenciones. Ciudad de La Habana. Cuba. Septiembre/1991.

14. Lozano Reyes, F., Sandi Pinheiro, M., Garay Garcell, M.A., Garcia de la Vega, D, Chiang, L.A. (1999), «A software tool for classification of objects and word problems in hypermedia intelligent tutoring systems», ICECE 99, IEEE, Rio Palace Hotel, Rio de Janeiro, Brazil.

15. Sandi Pinheiro, M., Reyes, Garay Garcell, M.A. (1999), «A software tool for classification of objects and word problems in hypermedia intelligent tutoring systems»,

Journal of Computer Applications on Engineering Education, Vol. 8, No. 3 / 4 , 2000 (CAE 20-264) pages 235 – 239. Ed. Magdy F. Iskander. John Wiley & Sons. Oct - December, 2000. USA. Online. ISSN: 1099 - 0542 and Print ISSN: 1061-3773.

16. Garay Garcell, M.A., Sandi Pinheiro, M., «Clustering methods in Hypermedia intelligent tutoring systems. Monte Carlo Resort hotel, Las Vegas, Nevada, USA. The 2001 International Conference on Internet Computing. June 25-29, 2001. Proceedings of Conference.

17. Garay Garcell, Miguel A. and Sandi Pinheiro, Marcello «On Classification Problems in Hypermedia Intelligent Tutoring Systems», Pre-prints SIT' 2001, Keynote Speaker. Symposia in Informatics and Telecom- munication. University of La Coruña, Campus Elvica, Faculty of Informatics, A Coruña. Sept. 12-14th, Spain. ISBN: 84 – 931 933 – 8 – 0.

Кластеризация и методы принятия решений для интеллектуальных систем

Мигель А. Гарай Гарсел[1], Марсело Санди Пинейро[2], Ванесса Тексейра де Оливьера Санди[3]

Исследовательский центр системотехники,
Гаванский технологический институт, Циудад де ла Габана, Куба (1),
Бразильский Католический университет, Бразилиа, Бразилия (2),
Кафедра программирования и информатики,
Бразильский Католический университет Бразилии (3)

Ключевые слова и фразы: кластеризация; интеллектуальные системы; гипермедиа; мультимедиа; системный инжиниринг.

Аннотация: Процессы, связанные с задачами классификации объектов и слов, остаются сложными, плохо-структурируемыми проблемами. В данной статье предлагается инструмент создания гипермедийных интеллектуальных обучающих систем. Этот инструмент помогает определить уровни сложности для ряда задач, предлагаемых студентам. Таким образом, разработчики систем интеллекта и преподаватели могут использовать его для создания проблемной базы. С другой стороны, этот инструмент составляет важный элемент в процессе обучения и тренировки, т.к. он позволяет осуществлять переход от простого к сложному при решении задачи. В нем используются различные методы классификации, такие как статистические методы, экспертные оценки и другие. Оценка параметров включает объективные и субъективные элементы. В связи с этим разработана соответствующая версия метода Делфи. Система была выполнена в Delphi 4.0 для Windows.

Klasterisierung und Methoden von den Beschlußfassungen für die Intellektualsysteme

Zusammenfassung: Die Prozesse, die mit den Aufgaben der Klassifizierung von den Objekten und den Schichten verbunden sind, bleiben als komplizierte, schlechtstrukturierte Probleme. Es wird in diesem Artikel das Instrument der Schaffung von den hypermedien intellektuellen Bildungssystemen vorgeschlagen. Dieses Instrument hilft bei der Bestimmung der Kompliziertheitstufen für die den Studenten

vorschlagenden Aufgaben. Auf diese Weise können die Ausarbeiter der Intellektualsysteme und die Lehrer dieses Instrument für die Schaffung der Problemgrundlage anwenden. Von anderer Seite bildet dieses Instrument ein wichtiges Element im Prozeß der Ausbildung und des Trainierens, denn es erlaubt der Übergang vom Einfachen zum Komplizierten bei der Aufgabelösung zu verwirklichen. Es werden dabei die verschiedene Klassifikationsmethoden benutzt: statistische Methoden, Experteneinschätzungen u.s.w. Die Parametereinschätzung enthält objektive und subjektive Elemente. In diesem Zusammenhang ist die entsprechende Version der Delphi-Methode ausgearbeitet. Das System wurde in Delphi 4.0 für Windows erfüllt.

Clastérisation et méhodes de l'adoption des résolutions pour les systèmes intellectuels

Résumé: Les processus liés aux problèmes de la classification des objets et des mots sont complexes, mal structurés pour la création des systèmes intellectuels hypermédias de l'enseignement. Cet article aide à définir les niveaux de la complexité pour une série de problèmes proposés aux étudiants. Ainsi, les élaborateurs des systèmes intellectuels et les professeurs peuvent l'utiliser pour créer une base de problèmes. De l'autre côté, cet outil présente un élément important dans le processus de l'enseignement et de l'apprentissage, puisqu'il permet de réaliser le passage à partir des problèmes simples vers les problèmes complexes. On y emploie de différentes méthodes de la classification comme, par exemple, les méthodes statistiques, de la valeur d'expert et d'autres. L'appréciation des paramètres comporte les éléments objectifs et subjectifs. Donc, on a élaboré la version correspondante de la méthode Delphi. Le système a été exécuté dans Delphi 4.0 pour Windows.