SUPPORT AND CONFIDENCE PARAMETERS TO INDUCT DECISION RULES TO CLASSIFY TEHRAN'S SEISMIC VULNERABILITY

Samadi Alinia, H.[a *], M. R. Delavar [b], Y.Y. Yao [c]

[a] MSc. Student, Dept. of Surveying and Geomatics Eng., College of Eng., University of Tehran, Tehran, Iran – alinia@ut.ac.ir

[b] Center of Excellence in Geomatics Eng. and Disaster Management, Dept. of Surveying and Geomatics Eng., College of Eng., University of Tehran, Tehran, Iran - mdelavar@ut.ac.ir

[c] Department of Computer Science, University of Regina, Canada - yyao@cs.uregina.ca

Commission VII, WG VII/5

KEY WORDS: Granule network, decision rules, classification, entropy measurement, seismic vulnerability, geospatial information system

ABSTRACT:

Rule induction is an area of machine learning in which formal rules are extracted from a set of observations or decision trees. Inducted rules can be expressed as relationships between concepts in terms of their intensions and extensions, such as sub-concepts and super-concepts, disjoint and overlap concepts, and partial sub-concepts. Existing algorithms for construction of decision trees cannot effectively deal with missing values. To overcome the limitation, this paper presents granule network as an improved decision trees to induct more informative classification rules in which all objects in the universe are classified correctly with minimum uncertainty. To achieve this, some quantitative measures such as coverage and support are used to estimate the quality of the granules and their relations at each step. This paper is concentrated on seismic vulnerability classification using granule tree by applying quantity measures. It is implicated on 177 urban areas of north of Tehran, capital of Iran and the results are used to classify Tehran's seismic vulnerability. The achieved results of the classification are successfully verified the proposed methodology.

1. INTRODUCTION

Tehran, capital of Iran, has several known and unknown active faults hence huge earthquakes will permeates human settlement there. Production of seismic vulnerability map could help local and national disaster management organizations to create and implement a plan to promote awareness of earthquake vulnerability and implementation of seismic vulnerability reduction measures in Tehran (Aghataher et al., 2005; Alinia and Delavar, 2009).

Production of seismic vulnerability map generally depends on various criteria. So, for knowledge discovery of seismic vulnerability of urban areas, from a dataset with minimum entropy, granular computing model is proposed.

This algorithm selects more confidence pair attribute-value of a granule at each step depending on its quantity on coverage and support measurements. The algorithm will continue until all objects are correctly classified in which, each object is associated with a unique class label.

The concept of seismic vulnerability can be exemplified at two parts, extension, i.e., a set of objects as instances of pre-species category of seismic vulnerability and intension, consists of all properties or attributes with more effective impacts in seismic vulnerability, that are valid for all those urban areas where the concept applies. In this paper, objects are urban areas and five grades of vulnerability for these objects are considered. An object must be described in terms of a fixed set of attribute, each with its own set of possible values. For example in this paper, 'Slop' might be attributes with sets of possible values {low, moderate, high, very high}. The knowledge mined from a sample dataset is represented in the form of rules. More certainty and strongest rules are extracted automatically from granule decision trees then they are implied to test the dataset to evaluate method and at last all urban areas are classified.

The main objective of this paper is to exhibit how granule network approach can be used for the induction of seismic classification rules and howmuch accurate of the method is. GIS is used to develop spatial data layers where decision rules are used for classification of urban areas.

The scheme of this work starts with the design of the granule decision tree and identifies a subset of inputs, in the form of an information table, to induct classification rules. Induction the more usefulness rules and more applicability needs some measure such as coverage, confidence and entropy. Extracted Rules with higher coverage and confidence are applied to a test dataset to evaluate the precision of classification with respect to an expert opinion. At the final stage total urban areas are classified in terms of seismic vulnerability. In this paper seismic vulnerability classification of Tehran Metropolitan Area based on granular computing model is undertaken.

The basic ideas of granular computing, i.e., problem solving with different granularities, have been explored in many fields, such as artificial intelligence, interval analysis, quantization, rough set theory, Dempster-Shafer theory of evidence, divide and conquer, cluster analysis, machine learning, databases, and many others (Zadeh, 1997). The term "granular computing" was first suggested by T.Y. Lin (Zadeh, 1998). A number of researchers have examined granular computing. Yao and Yao presented a granular computing view to classification problems and proposed a granular computing approach for classification. (Yao and Yao, 2002).

Granular computing splits the feature space into a set of subspaces (or information granules) such as classes, subsets, clusters and intervals (Lin, 1997). This technique can be applied to rule mining. In granular computing, information granules are first constructed and computations are subsequently carried out with the granules (Yao, 2000). Several methods are proposed to granulation of the universe like clustering (Zhong et al., 2007), and fuzzy sets (Zadeh, 1997).

2. DATA PREPARATION

2.1 Study area

Tehran as a mega city is located in an earthquake prone region based on the historical records available. Therefore, it is predicted that huge earthquake will be happen in Tehran soon. (Aghataher et al., 2005; Alinia and Delavar, 2009).

Among several known and unknown active faults, the north Tehran fault situated towards the northern side of Tehran has the potential to generate MW = 7.2 respectively, which according to the earthquake scenarios developed under the JICA-CEST project "Study on the Seismic Microzoning of the Greater Tehran Area", 1999-2000, could produce many victim. Seismologists believe a strong earthquake will strike Tehran in the near future because the city has not experienced a disastrous earthquake since 1830 (JICA, 2000).

In this study the result of north Tehran fault hazard analysis is applied to the vulnerability assessment process and activation of other faults have been ignored. It is assumed that the northern fault of Tehran is activated and then a seismic physical vulnerability gained.

A pilot area of Tehran Metropolitan Area located in the north of Iran was selected for the purpose of this study. This area contains 177 urban areas. The study area is located between 51˚ 23' N, 51˚ 33' N Latitude and 34˚ 46' E, 35˚ 49' E Longitude.

2.2 Data

Five effective factors in assessment of seismic vulnerability of the areas including Slope, percentages of weak buildings less than 4 floors, percentage of more than 4 floor buildings, percentage of buildings built before 1966, percentage of buildings built between 1966 and 1988 were considered. The data is obtained from Statistical Center of Iran.

A dataset of 20 urban areas as input to granule decision tree is considered. Objects are selected based on a stratified random sampling. To estimate precision of the approach, the inducted classification rules are extracted from 50 test urban areas in Tehran's District 1.

All spatial and non-spatial data on the urban area were converted to ArcGIS database format. Figure 1 shows the map of the study area in which the location of the north Tehran Fault is shown.

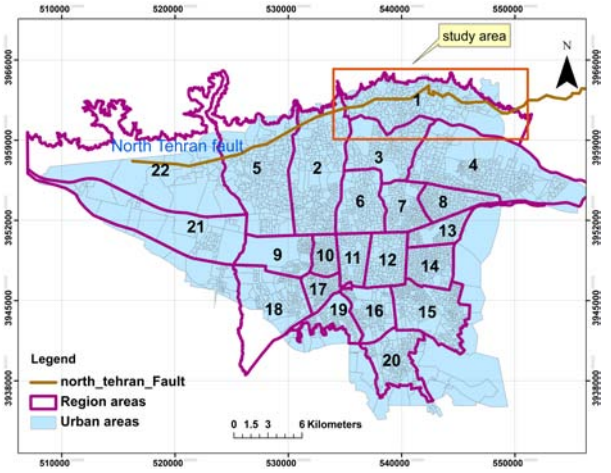

Figure 1. Study area: Tehran's District 1

2.3 Information table:

Information tables are base knowledge in granular computing models. An information table provides a convenient way to describe a finite set of objects called a universe by a finite set of attributes (Pawlak, 1991; Yao and Zhong 1999).

An information table has the following tuple:

$$S = (U, At, L, \{V_a \mid a \in At\}, \{I_a \mid a \in At\})$$

where U is a finite non-empty set of objects,
At is a finite non-empty set of attributes,
L is a language defined by using attributes in At,
Va is a non-empty set of values of $a \in$ At,
$I_a : U \rightarrow V_a$ is an information function that maps an object of U to exactly one possible value of attribute a in Va.

To prepare input dataset of granule decision tree, a table of objects and its available information should be considered. The information table of this paper is constructed from 20 urban areas in 20 rows and 6 columns of 5 attributes which describe objects and unique attribute class taking class labels as its values. Columns of decision classes are the grade of seismic vulnerability filled by some seismic experts. Each record in a column is a value of attribute in human-language form. Table1 illustrates the information table of this research. For simplification, the titles of all criteria are given as follows:

Slop.	Slope (Degree)
build_less4.	Percentage of Weak buildings less than 4 floors
build_more4.	Percentage of more than 4 floor buildings
Bef-66.	Percentage of buildings built before 1966
Bet-66-88.	Percentage of buildings built between 1966 and 1988

The objective of the classification is to determine the degree of seismic vulnerability in a randomly selected object. Five classes for discerning levels of seismic vulnerability between the groups of urban blocks are considered. Including very high vulnerable, high vulnerable, moderate vulnerable, low vulnerable and very low vulnerable. In order to comfort the process it is necessary to brief the label of the classes. Including Very high vulnerable=1, high vulnerable=2, moderate vulnerable=3, low vulnerable=4 and very low vulnerable=5.

object	slope	build_less4	Bef-66	Bet-66-88	build_more4	class
u1	Very high	high	low	high	low	5
u2	Very high	moderate	low	high	low	5
u3	low	moderate	moderate	moderate	low	3
u4	Very high	low	low	Very high	low	4
u5	low	low	low	high	low	2
u6	low	low	low	moderate	low	1
u7	Very high	high	low	Very high	low	5
u8	moderate	moderate	low	Very high	low	4
u9	moderate	low	low	low	low	1
u10	low	moderate	low	moderate	low	2
u11	low	low	low	high	low	2
u12	low	low	low	moderate	very high	1
u13	moderate	moderate	moderate	high	low	4
u14	moderate	low	low	moderate	low	2
u15	high	low	low	moderate	low	2
u16	high	low	low	low	low	1
u17	low	low	low	Very high	high	3
u18	Very high	low	low	moderate	high	2
U19	high	low	low	Very high	moderate	3
u20	high	low	low	moderate	high	2

Table 1. Seismic Vulnerability Information table

3. GRANULAR COMPUTING APPROACH

A granular computing approach is proposed as a granule network to extend the existing classification algorithms. In the approach, Based on a measure of connection between two partitions such as H (Ψ | Φ), one selects an attribute to divide the universe into a partition (Yao and Yao, 2002; Quinlan, 1983).In a granule network, each node is labelled by a subset of objects. The arc leading from a larger granule to a smaller granule is labelled by an atomic formula. In the Decision Language, an atomic formula is given by a = v, where $a \in At$ and $v \in Va$. In addition, a smaller granule is obtained by selecting those objects of the larger granule that satisfy the atomic formula. The family of the smallest granules thus forms a conjunctively definable covering of the universe.

3.1 Induction classification rules

To construct a granule network it is required that first dividing the universe into grouping or partitions of the same class with atomic formula of attribute-values. A rule can be expressed in the form, Φ ⇒ Ψ, where Φ and Ψ are intensions of two concepts. In many studies of machine learning and data mining, a rule is usually paraphrased by an if-then statement, "if an object satisfies Φ then the object satisfies Ψ." The interpretation suggests a kind of cause and effect relationship between Φ and Ψ (Yao, 2001). For this, some measures for single granule, relationship between two granules and relationship between a granule and a family of granules is applied automatically by the granule network algorithm. In the algorithm, less entropy and high coverage is two important quantitative measures to select granules as decision tree nodes. Then, the active node and non-active node at a level should be characterized by considering two conditions. A granule is non-active if it has two conditions: including (i) the granule be a subset of unique class and (ii) union of all granules at low level and non-redundant cover the solution of the root granule. An active granule will be further divided through efficient measures. After union of all inactive granules construct a covering solution of universe set, construction of decision granule tree would be stopped. So the tree provides information in the form of "IF - THEN" statements. At the last stage of induction classification rules, to estimate the relative classification accuracy, the decision rules are applied on the test dataset.

3.1.1 Measurements on granules

To estimate granule and therefore a rule in various aspects, some measures are used:

3.1.1.1 Measures of a single granule: indicates the relative size of the granule. A granule defined by the formula is more general if it covers more instances of the universe. The quantity may be viewed as the probability of a randomly selected object satisfying formula (Yao and Yao, 2002).

3.1.1.2 Confidence: Confidence or absolute support is defined as the fraction of instances that are correctly classified by the rule among the instances for which it makes any prediction. Thus, it is a measure of the correctness or the precision of the inference. The quantities can be computed by fraction of number of samples that satisfies the THEN part of the rule, to the number of samples that satisfy only the IF part of association rule (Yao and Yao, 2002).

3.1.1.3 Coverage: coverage is a measure of the applicability or recall of the inference. It indicates fraction of data in a class correctly classified by the rule (Yao and Yao, 2002).
The quantities can be computed by fraction of number of samples that satisfy the THEN part of the rule, to the size of training data with the same class label as the rule consequent.

3.1.1.4 Conditional entropy: it provides a measure that is inversely related to the strength of the inference. This measurement which depends on the confidence, if an object satisfies the formula of attribute-value, one can identify one equivalent class in which the object belongs with no measure of uncertainty. In this case, confidence of the formula for at least one equivalent class is 1.

4. VULNERABILITY OF THE DECISION TREE

To produce a minimum uncertainty of the seismic vulnerability classification, it is required to find a subset of attribute-value with high coverage, confidence and minimum entropy.
In this paper, construction of the tree are continued until all granules reach to zero entropy in which, union of all inactive granules be equal to the universe set.
From 20 training data, 15 rules are inducted at 4 levels. Each non-redundant active node should be divided until all objects are classified correctly. Union of non-active granules at 4 levels covers the universe set. Each node of the granule decision tree is labelled by a value of attribute and each branch is labelled by a value of the parent attribute.
Seismic vulnerability decision tree of this research is illustrated in Figure 2.

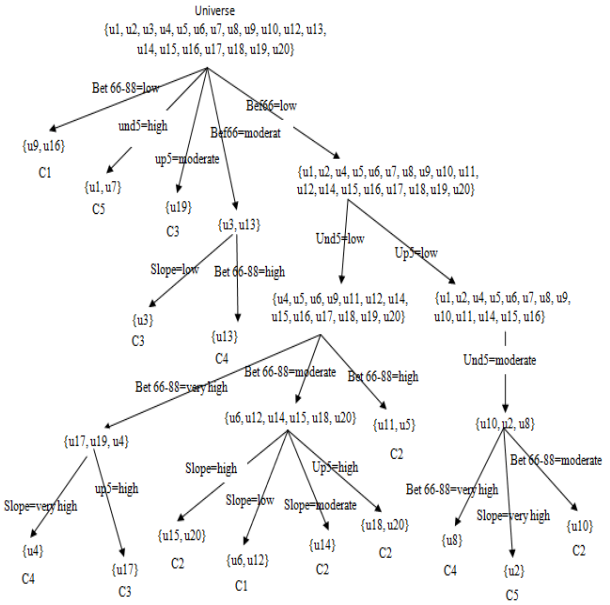

Figure 2. The developed granule decision tree

In the tree (Figure 2), each non-active granule is labelled by its vulnerability class. Union of all inactive granules form a non-redundant covering solution of the consistent classification problem. It means, the collection of the subsets derived by deleting any one of the granules is not a covering.

5. SIESMIC VULNERABILITY CLASSIFICATION

Producing seismic vulnerability map using the inducted rules, follows five sequential steps including: (1) preparing data grids of value of attributes (2) applying rules on the test data (3) integrating rules to produce seismic vulnerability (4) analysis of the difference between model outputs and expert opinion and (5) producing seismic vulnerability map for all of the urban areas.

In Step 1, processing of spatial data and generation of the data layers were established within a GIS exhibiting different attributes. All data about urban areas in vector format converted to raster format to facilitate applying the rules. The projection of all layers were defined in Universal Transverse Mercator (UTM) WGS 1984 zone 39N. In step 2, Inducted rules are applied on the spatial data which were prepared in step 2. In step 3, decision rules in step 2 are arranged by multiplying the rules to their risk weight then degree of seismic vulnerability of test area are produced by overlaying the rules. In step 4, the difference between the results and the expert classification to evaluate the uncertainty of the expert opinion. The cells are compared one by one. If the quantity of the difference being acceptable, in step 5 the inducted rules are applied to whole study areas.

Figure 3 illustrate the expert opinion classification on 50 urban areas. The figure demonstrates a 5 level classification.

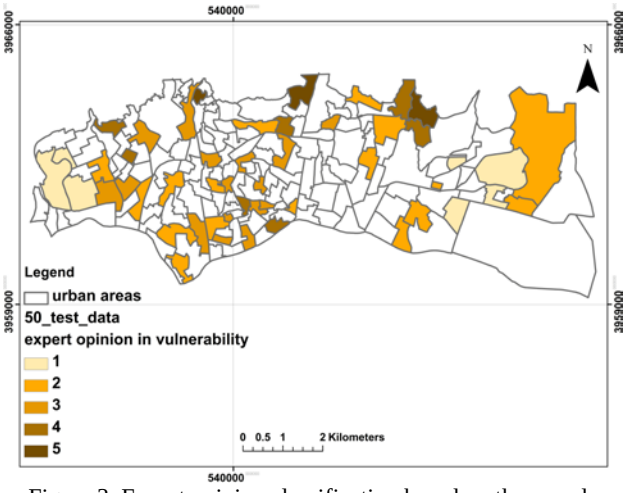

Figure 3. Expert opinion classification based on the sample

Physical vulnerability of the 50 urban areas predicted by granule decision rules is illustrated in Figure 4.

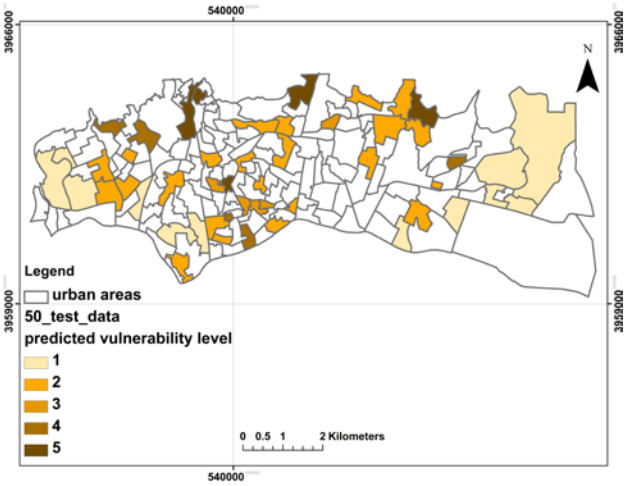

Figure 4. Seismic vulnerability classification of test areas based on included rules

Comparing Figures 3 and 4, it is clear that most of blocks have similar values of the seismic vulnerability. The relative accuracy of decision making in each class of seismic vulnerability is as follow: 48% with zero difference, 34% with one levels difference, 16% with two levels differences and 2% of cells have three levels of differences.

In next step, seismic vulnerability of the study area by applying decision rules is computed. Figure 5, illustrates the urban seismic vulnerability classification. The figure shows that 10 urban areas have not been classified using the inducted rules. It is obvious that they are dependent on distribution of sampling and training data, in this situation these data do not correspond to none of the inducted rules. Other reason could be due to binary definition of the value attributes and categorizing the data.

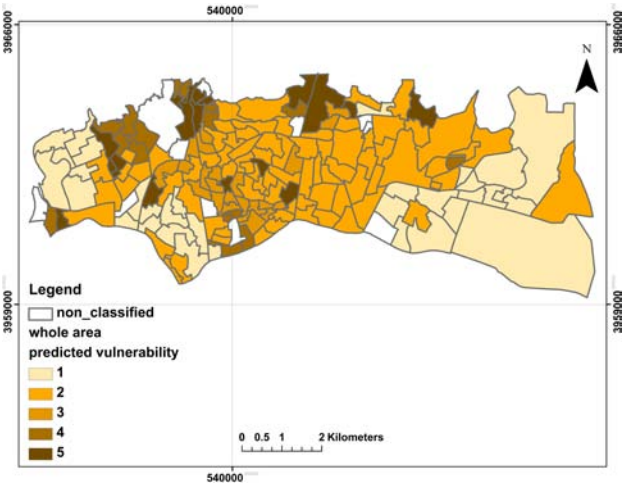

Figure 5. Seismic vulnerability map of all urban areas

6. CONCLUSIONS

This paper has proposed a new approach to analyze classification of physical vulnerability against earthquake considering northern fault of Tehran is activated. In this research for seismic vulnerability assessment, some information about effective parameters to vulnerability in each urban area including slope, weak buildings less than 4 floors, percentage of buildings with more than 4 floors, percentage of buildings built before 1966, and percentage of buildings built between 1966 and 1988 are considered.

The main focus of this research was inducting classification rules with maximum certainty with expert's opinion to classification. The uncertainty which has been discussed focuses on vagueness of determining the influencing rules with less entropy to expert's opinion in seismic vulnerability mapping in Tehran.

In order to model and pay attention to more precisely evaluate the physical seismic vulnerability, granular computing model was used. In this method, contrary to the classical approach which focuses on the selection of a suitable partition, i.e., a family of granules defined by values of an attribute, at each step, and the selection of a single granule. To vulnerability assessment of the urban area using expert opinion, constructing granule decision tree to induct rules with high confidence and coverage quantities was performed. Applying the extracted rules to the test data, led to classification of the study area with nearly 50% with high certainty. The results in some blocks are different and this difference could be due to uncertainties in experts' opinions in filling the column of decision attribute of information table. In classification of all of the urban area some urban area did not belong to any particular class. It could be due to uncertainty in the training data sampled. One of the main advantages of using this approach is removing data with similar attribute-value which is labeled differently by experts due to uncertain expert's knowledge.

It is recommended that, in order to make a good decision without losing some data unclassified, training dataset should cover all possible categorize of the data.

The main contribution of the paper is the formal development of the granule cantered strategy for mapping seismic vulnerability based on minimum entropy with expert opinion.

References:

Aghataher, R., Delavar, M.R. and Kamalian, N. (2005), "Weighing of Contributing Factors in Vulnerability of Cities against Earthquakes." *Proc. Map Asia Conference*, Jakarta, Indonesia.

Alinia, S.H. and Delavar, M.R., (2009), "Granular computing model for solving uncertain classification problems of seismic vulnerability" spatial Data quality from process to Decision, Edited by Rodolphe Devillers and Helen Goodchild., pp.132-134

JICA, 2000. The Study on Seismic Microzoning of the Greater Tehran Area in the Islamic Repoblic of Iran, Final Report. Japan International Cooperation Agency (JICA).

Lin, T.Y., 1997. Granular computing, Announcement of the BISC Special Interest Group on Granular Computing

Pawlak, Z, 1991. Rough Sets: Theoretical Aspects of Reasoning about Data, *Kluwer Academic Publishers*, Dordrecht.

Quinlan, J.R. "Learning efficient classification procedures and their application to chess end-games", in: *Machine Learning: An Artificial Intelligence Approach*, Vol. 1, Michalski, J.S., Carbonell, J.G., and Mirchell, T.M. (Eds.), Morgan Kaufmann, Palo Alto, CA, pp.463-482, 1983.

Zadeh, L.A., 1997. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*, Elsevier, pp. 111-127

Zadeh, L.A., 1998. Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. *Soft Computing*, pp. 23-25

Zhong, W., He, J., Harrison, R., Tai, P.C., Pan, Y. 2007 Clustering support vector machines for protein local structure prediction. *Expert Systems with Applications* pp. 518–526

Yao, Y.Y., 2000. Granular computing: basic issues and possible solutions. In: Wang, P.P. (ed.) Proceedings of the 5th Joint Conference on Information Sciences, Atlantic City, New Jersey, USA. *Association for Intelligent Machinery*, vol. I, pp. 186–189

Yao, Y.Y. and Zhong, N., 1999. Potential applications of granular computing in knowledge discovery and data mining, Proceedings of World Multiconference on Systemics, Cybernetics and Informatics, 573-580.

Yao, J.T. and Yao, Y.Y., 2002. Induction of Classification Rules by Granular Computing, Proceedings of the Third *International Conference on Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence* pp. 331-338.

Yao, J.T. and Yao, Y.Y., 2002. A granular computing approach to machine learning, *Proceedings of the First International Conference on Fuzzy Systems and Knowledge Discovery* (FSKD'02), Singapore, pp. 732-736.

Yao, Y.Y, 2001. On Modelling data mining with granular computing, *Proceedings of COMPSAC, 25th Annual International Computer Software and Applications Conference (COMPSAC'01),* pp.638-643