Contextual Knowledge Granularity

Janina A. Jakubczyc and Mieczysław L. Owoc
University of Economics, Wrocław, Poland

janina.jakubczyc@ue.wroc.pl mieczysław.owoc@ue.wroc.pl

Abstract

Knowledge granularity is often regarded as one of the essential factors of knowledge repositories basically in terms of ways of knowledge gathering and storing as well as its usability. The aim of this paper is to discuss the importance of this phenomena in the case of contextual classification. This kind of directed granulation by context gives possibility to generate new and intelligent knowledge structure, to see a problem through many context simultaneously perspectives or through one context perspective. A part of the investigation of usability of context-based approaches in creation knowledge structures interrelationships between knowledge granularity and effectiveness of classification tasks is discussed.

Keywords: knowledge granularity, context, classification, contextual classification, knowledge representation, knowledge application

Introduction

Nowadays domain knowledge plays the crucial role in decision-making processes at any level of management. Sometimes, a decision-maker needs knowledge covering many aspects of intelligent structures and using many resources while in particular situation his expectations are reduced to very simple and precisely expressed knowledge items. Therefore, in particular situations we apply knowledge more or less compound and - what is important – different in terms of its heterogeneity and consistency. Phenomena of knowledge granularity (widely identified with the ability to represent and operate on different levels of detail on data structures) appears and in our opinion very often is contextual dependent.

The main goal of this paper is investigation on importance of context-based approaches in formulation of knowledge structures and from the other hand discovering relationships between different perspectives of knowledge granularity and its impact on classification accuracy. Therefore one can say contextual granulation is basically oriented on defined goal(s) and its usability is validated through classification accuracy (models, decision trees and eventually knowledge granules).

The origin of the problem is discussed in the first section stressing flexibility of knowledge granularity interpretations. Very useful in this area is a proposal presented as theory of granularity. The second part is devoted to presentation of approaches representing contextual situations classification. The core section of the paper includes examples of contextual classification where knowledge granularity plays essential role. In the last part finding research is presented apart of future topics.

Knowledge Granularity Concepts

Granularity as a intriguing phenomena comes from photography to describe accuracy of pictorial presentation on film (the higher level means more details). Intuitively we use this term to express some level of aggregation components in different domains; therefore we may observe granular models in physics, computing and business etc. Applying a concept of granularity in information technologies one can say that granularity of information resources refers to size, decomposability and the extent to which a resource is intended to be used as part of a larger re-source (Wagner, 2002).

In global meaning, granularity is widely identified as the ability to represent and operate on different levels of detail in data structures (including information and knowledge) – compare the work of C.M. Keet (Keet, 2008). An essence of typical approaches to the discussed problem is expressed in theory of granularity (TOG) proposed by Mani (Mani, 1998) and extended by Keet (Keet, 2008). Core categories applied in this theory embraces: domains, entities, relations, levels, perspectives, types of granularity, contents of a granular level and granular reasoning. A general idea of chosen static components introduced into TOG is presented in Figure 1.

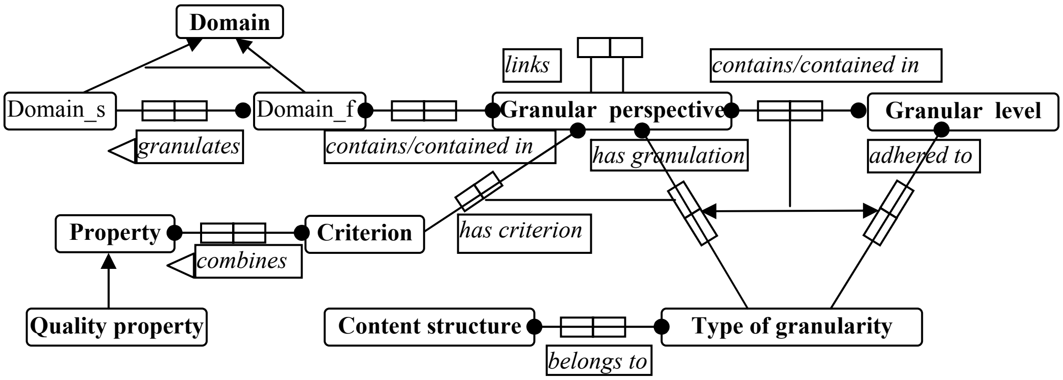

Figure 1: Perspectives and levels in theory of granularity
Source: own elaboration based on Keet, 2008, Mach and Owoc, 2009

The starting point is a domain which consists of many entities described via properties and interrelated themselves; some concepts of the domain_f can be used in more universal way as parts of ideas expressed in the domain_s . Concepts of granularity levels and granularity perspectives allow for creating and maintaining more complex combinations of entities as a base of formulating different types of granularity (including perspectives and levels of granularity. Considering different perspectives of granularity we take into account criteria of its differentiation as well as its properties. The presented approach can be extended by theories which are oriented on more particular applications or a way of formalization, for example data warehouse technology in order to support analytical processing (Kamble, 2004) or fuzzy sets and fuzzy logic to serve the problem of similar entities and clustering (Zadeh, 1997; Pawlak 1998).

In such a context, we should focus on knowledge granularity. Granularity phenomena can be defined including many approaches (Mach&Owoc, 2008). One of the them is applied in the e-learning area (Duncan, 2003). Learning courses (in fact knowledge existing in courses) are divided into ”knowledge pieces” according to audience familiarization with presented topics or aims of the course. Therefore from logical point of view we may separate knowledge presenting definition of some phenomena, put some procedures how to classify some objects or give examples of the procedural knowledge. Very close is the concept of learning sets used in machine learning where “knowledge items” express certain regularities of the investigated process and are defined at the some level and perspective. In this paper we try to consider exactly this point of view.

Taking into account technological aspects we should consider granularity computation orientation. Granularity computation, as one of the promising streams in the research of artificial intelligence, is very important in knowledge discovery, image compression, semantic Web service and the like. Let us stress different roots of granular computing; some concepts are derived from: rough set theory, cluster analysis, machine learning, the database and information retrieval. That was a reason to develop and implement several models rooted in the mentioned disciplines.

Knowledge granularity seems to be crucial point of the whole process broadly termed as knowledge management (Mach&Owoc, 2010). Classical stages defined in this process are: knowledge acquisition, knowledge storing, knowledge dissemination and knowledge applications. The phenomenon of knowledge granularity appears in the first stage (we may gather knowledge from many resources with different levels of details) and then is re-formulated during storing stage (usually several models of stored knowledge can co-exist). Also in the next two stages: knowledge dissemination and knowledge applications granules of knowledge can be formulated from different perspectives and levels including many purposes and users.

An Essence of Contextual Classification

The contextual granulation is a partition of observation (mentioned earlier learning set) what finally leads to knowledge granulation (Barg et al., 2000). Granulation criterion is the context in which the phenomenon (problem, concept) can occur. Differently than is commonly assumed in the area of granulation (Zadeh 1997, Pawlak 1998), this type of granularity do not have to have a hierarchical structure but more often overlapping structures are the results.

This means that single observation can be seen through the prism of multiple contexts simultaneously or through single context. The context can be known or unknown, may be included in the learning set or elsewhere. This article is limited to the context contained in the data set.

The way the context can be identifed is two folded. The first one concerns exploiting expert or domain knowledge. The context discovered that way most often relates to the whole analyzed phenomena. The second concerns an automatic way by using definition of possible types of attributes presented by P. Turney (Turney, 1993). In this case the subject of context is single decision feature on the contrary to the whole analyzed problem. The roles assigned to the features in the light of context may be ‘contextual’ or ‘context-dependent’. In short it can be said, that contextual attributes are not relevant to the described concept. They are features that are relevant only to the context-sensitive features what means that context-sensitive feature can be described by contextual properties. The primary feature does not depend of any other features what means that they are not context-dependent and are only decision features.

The identified context - according the roles assigned to attributes - relates to single context-depended attribute. Thus the contextual granulation is the partition of learning set according to the identified contextual features (for example few simple contexts: localization, seasons, age which can be considered as granularity criteria) for the former and for the letter the partition is conducted according contextual feature via context-depended feature. Every pair of contextual feature-value splits the learning file into subsets directly or by context-depended-feature-value. The number of subsets is determined by finite number of contextual feature values and context-depended feature-value or by the first one only.

The way of creating granules that are dependent on the identified primary, contextual and context-sensitive features is as follows:

1. Separation of the features, which directly influence the classification task (decision features) and the features which do not (the non-decision features); this step is conducted by creation of a decision tree on the basis of the entire learning set
2. Identification of context-sensitive features from the set of decision features discovered above; this step stand for searching for contextual features from non-decision features for each decision attribute by applying decision tree algorithm
3. Selection of contextual and context-sensitive features that can be used to partition the learning file, according to the assumed level of classification accuracy.
4. Granulation of the learning set according to the structure of values of the contextual feature and its corresponding context-sensitive feature.

The maximum number of contexts is determined by the cardinality of set of decision features. The context stand for single decision feature that is explained by contextual attributes.

Each context subject-independent may be represented by simple or complex structures (Fig.2). In the case of one-attribute context, the learning set is partitioned into number of granules according to attribute values (Fig 2, Context 1). If context sensitive feature is described by many contextual attributes the granulation is settled by each path in the decision tree that model the context (Fig.2, Context 2). Each path is called contextual situation that represent single granule.

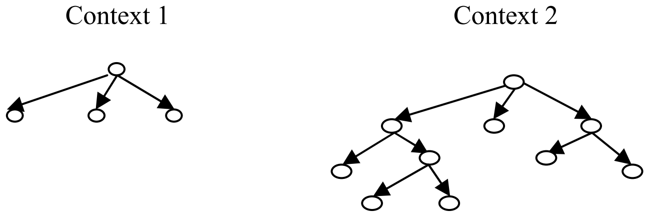

Figure 2: The representation of context and contextual situations
Source: own elaboration

Instead of building one description (model) for entire learning set we create model for each contextual situation for given context (Figure 3). Each context is represented by the set of classifiers. If the number of identified context is higher than one, single observation can be seen in many contexts simultaneously. The contexts are independent, so each divides the learning set separately.

As we can see each context is the set of contextual situations where each of them is represented by separate decision tree. The received structure divides the knowledge about analyzed phenomena into contextual knowledge and context-depended knowledge. The contextual knowledge may be used to control and choose appropriate model or group of models.

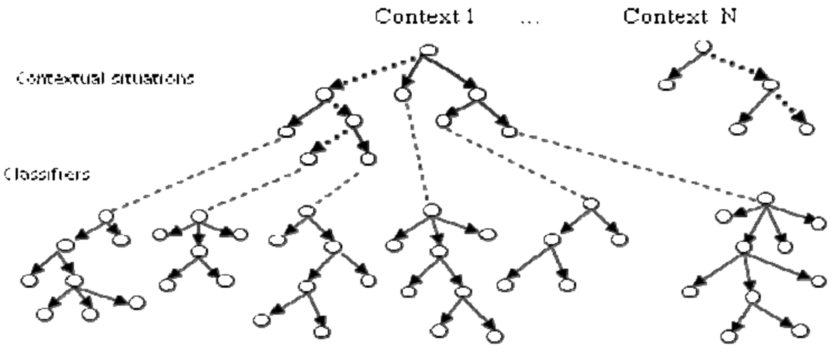

Figure 3: The structure of contextual knowledge and context-dependent knowledge
Source: own elaboration

This approach has a number of advantages. The one of them is an usage of additional knowledge included in non-decision attributes, that wouldn't be used by models. So we can say about more complete utilization of information included in the data set. The advantage and disadvantage alike can be the partition of learning set into context depended subsets. On the one hand it can be treated as some solution for computational difficulties and difficult classification problems. From the other hand it can result overmuch fragmentation that can obscure description of the problem (for example: the level of generality may be too low, the interpretation may be difficult, and so on).

The possibility of more complete data analysis is another good point, what means that we can make comparisons of different contexts and classifiers according to the described attributes, classification accuracy and complexity. We can look for the best classification model. We can determine which attribute are more frequent than the other, choose the best matching context and classifier for the analyzed case.

Implementation of Knowledge Granularity in the Contextual Tasks

Two examples are prepared to illustrate importance and relationships between knowledge granularity and classification tasks where context was included. The former is the case where contexts are automatically identified and the granulations for chosen contexts are acceptable as a whole. This means that all granules have high level of quality according to classification task. The latter example refers to the contexts that were recognized by domain experts. The conducted granulation for each context can't be seen as a whole because the quality of some granules for some context are unacceptable.

Example 1

This example deals with the prediction of active and non active bank client ("2nd competition," 2009). The classification accuracy obtained on the base of decision tree was not acceptable for the level of non active client that is 67% while for active client the accuracy is acceptable (81%). Then we provided the test if decision attributes are context-depended features by building models that explain each decision feature by non decision features. It turned out that all decision features are contextually dependent, according to the level of classification accuracy. The complete results are contained in Table 1. For each context-dependent feature there are specified attributes of contextual situation. To simplify the perception of these examples we assume that identified contexts adopt the names of context-dependent features. The best classifier according to classification accuracy is model A but not many worse is model D but is more simple.

The number of contextual situations is in the range of 6-15 and it determines the number of granules. Assuming simplicity as the one of criterion, we should focus at the lowest ones. But maybe the more interesting criterion should present the user when looking at the given structures that represent all contexts. The user has the possibility to explain the rationale of these contexts and given structures in the light of his domain knowledge. He can compare contexts. For example in the Table 1 we can notice that three pairs: B and C, F and G, H and I, have the first attribute the same. Thus it is possible to join each pair into one contextual structure or limit them to shared features. The frequencies of attributes occurrence may be also of some importance. This type of analysis would be conducted much more profoundly by the user.

About the final results the user decides, who can choose for each classification case the most appropriate model of the contextual situation. He may also create the ranking list according to the more appropriate contexts in given situation. There is also the possibility to create an ensemble of classifiers that can make the classification decision collectively (using some kind of voting schema, for example).

Table 1: The results of possible structures of analyzed phenomena

Context-dependent attributes	**Attributes of contextual situations**	**Number of contextual situations**	**Classification accuracy**		
			A	**N**	**Average**
A	aa, pp	13	**79**	**75**	77
B	ii, hh, ll	12	**81**	**72**	**77**
C	ii, bb,	15	83	71	77
D	gg, mm	6	**78**	**74**	**76**
F	cc, kk	10	80	72	76
G	cc, dd	9	**78**	**73**	75
H	bb, ff	8	**78**	**74**	76
I	bb, ll	13	80	73	77
All			81	67	74

Source: own elaboration

It should be noted that because of acceptable level of classification accuracy we did not analyze granules (models) for single contextual situations. As we can see it can be important when some of granules for contextual situations are not acceptable towards classification accuracy.

Example 2

The second example concerns the problem of returns for mail order companies ("DCM Competition," 2010) that become more and more troublesome and costly. The task is difficult because received classification model gives not acceptable level of classification accuracy (value is 65% for client with high return risk). So we should conduct the structural analysis according the possible context included in the learning set. In this case the context is proposed by the expert including: client activity, client localization and age of a client. The each context divides learning set into number of granules according to the pair context-value. The results for context 'client activity' are presented in the table 2.

Table 2: The results for context 'client activity'

Context: 'client activity'	Classification accuracy	
	High risk	Low risk
high	44	84
medial	63	81
low	64	87
All	65	81

Source: own elaboration

As we can see, the classifiers (granules) for received structure do not have acceptable classification accuracy for high return risk clients that is even lower than for single model. So we can say that ‘client activity’ fail to better describing prediction for such clients.

The next possible structure is indicated by the ‘client localization’. The results for this context are presented in Table 3.

Table 3: The results for context ‘client localization’

Localization Code	Classification accuracy	
	High risk	Low risk
10 - 40; 70 - 90	74 - 80	73-84
50 - 60	65 - 67	79 - 83
All	65	81

Source: own elaboration

In the Table 3 we have the results grouped by the level of classification accuracy. The localization code for the range 50-60 doesn’t improve the results of the general classifier. The significant improvement brings the localization values 10-40 and 70-90. In this situation we can employ only the best granules for cases with specified values of localization. Overflow cases should be supported by general classifier or contextual classifiers convertibly because they are alike.

The next under examination is context ‘age of client’. The results are presented in Table 4. As we can see for the clients marked ‘older’ and ‘pensioner’ there is no improvement. For the rest ones the improvement is significant.

Table 4: The results for context ‘client age’

Age	Classification accuracy	
	High risk	Low risk
very young	86	78
young	71	74
middle	70	79
older	55	99
pensioner	26	97
All	65	81

Source: own elaboration

In this situation we can employ the first three granules for client classification. The rest two are of not use.

When we consider few context that indicate different structures for examined phenomena, there emerge a problem which structure should be used or how to join granules from different contexts. The solutions for this problem are not obvious. The additional knowledge may suggest that one or another solution is the preferred one. The ranking using the real results can be in use also.

Conclusions

The main advantage of the presented idea is the increase the informativeness of the description enriched by the contexts. In addition, there are more complete exploiting knowledge included in data and better use of knowledge generated from contextual granules of observation, through the possibility of choosing the most adequate knowledge granule (models, decision trees) or a combination of more contextual granules in order to make classification decision. This approach stress user (expert) participation in the process of data mining.

There is worth to mention that knowledge granulation might denote more intelligent approach to the problem solving. Another issue is that this approach reveals the open problem of representativeness of learning sets. Even though we realize that the goal of collecting data and learning task are different - the applied approaches ignore it.

Our experiences presented by two chosen typical examples from implementation of contextual granulation approach can be expressed as follows. Implementation of this approach should be taken, when there is no way to improve quality of descriptive model of analyzed problem or there is a need to more profound problem insight. The goal for this granulation is better description of classification problem. Each contextual granule is assessed by quality of classification model. Identified or known contexts may give satisfying description (it means of acceptable accuracy) of all contextual situations or only for some of them.

There can be applied three levels of granulation; the bottom granule represents knowledge separated by contextual situation, the second level represents knowledge about the entire context with the all contextual situation for given context; top granule may be created by a composition of the granules from the bottom or from the second level of contextual knowledge structure thus may stay for joined contexts or contextual situations.

The presented idea of contextual granulation gives the possibility to more profound analysis of examined classification problem. The extended examination may concern classification models (significance of attributes in different models, their frequency for example). We can also consider a verification of known and identified contexts. We are working now on some principles that can determine how to join different contexts and contextual situation. Such profound analysis may be of particularly importance when the problem concerns service for human users, for example internet store customers, on-line students, and web users of other sorts.

References

Barg, M., Fekete, A., Greening, T., Hollands, O., Kay, J., Kingston, J. H., et al. (2000). Problem-based learning for foundation computer science courses. *Computer Science Education, 10*(2), 109-128.

Duncan, C. (2003). Granularisation. In A. Littlejohn (Ed.), *Reusing online resources: A sustainable approach to eLearning*. Kogan Page, London

Kamble, A. S. (2004). *A data warehouse conceptual data model for multidimensional information*. PhD thesis, University of Manchester

Keet, C. M. (2008). *A formal theory of granularity*. PhD Thesis, KRDB Research Centre, Faculty of Computer Science, Free University of Bozen-Bolzano, Italy

Mach, M. A., & Owoc, M. L. (2008). Granularity of knowledge from different sources. *Intelligent Information Processing IV. IFIP Advances in Information and Communication Technology, 288*, 50-57, DOI: 10.1007/978-0-387-87685-6_8

Mach, M. A., & Owoc, M. L. (2009). About dimensions and measures of knowledge granularity. In R. Tadeusiewicz & A. Ligeza (Eds.), *Computer and methods systems*. Kraków: AGH.

Mach, M. A., & Owoc, M. L.(2010). Knowledge granularity and representation of knowledge. Towards knowledge grid. *Intelligent Information Processing V, IFIP - The International Federation for Information Processing, Volume 340*

Mani, I. (1998). A theory of granularity and its application to problems of polysemy and underspecification of meaning. In A. G. Cohn, L. K. Schubert, & S. C. Shapiro (Eds.), *Principles of knowledge representation and reasoning, Proceedings of the Sixth International Conference (KR98)*, San Mateo: Morgan Kaufmann

Pawlak, Z. (1998). Granularity of knowledge, indiscernibility and rough sets. *Proceedings of 1998 IEEE International Conference on Fuzzy Systems*, 106-110

Roussev, B. (2003a). Empirical evidence justifying the adoption of a model-based approach in the course web applications development. *Journal of Information Technology Education, 3*, 73-90. Retrieved from http://www.jite.org/documents/Vol2/v2p073-090-79.pdf

Roussev, B. (2003b). Teaching introduction to programming as part of the IS component of the business curriculum. *Journal of Information Technology Education, 2*, 349-356. Retrieved December 15, 2010 from http://www.jite.org/documents/Vol2/v2p349-356-43.pdf

Turney, P. D., (1993). Robust classification with context-sensitive features. *6th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, Edinburgh, UK, 1 - 4 Jun 1993

Wagner, E. (2002): Steps to creating a content strategy for your organization. *eLearning Developers' Journal.* eLearning Guild. October 29, 2002. Retrieved December 15, 2010 from http://www.elearningguild.com/pdf/2/102902MGT-H.pdf

Zadeh, L.A. (1997): Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and System, 19*(1).

DMC 2004 Competition. (2010, June 24) Retrieved from http://www.data-mining-cup.de/en/review/dmc-2004/

2nd competition. (2009, May 1) Retrieved from http://www.eunite.org/knowledge/Competitions/2nd_competition/Submissions_and_Results/Submissions_and_Results.htm

Biographies

Janina A. Jakubczyc, PhD in Economics, associated professor of Computer Science at Wroclaw University of Economics, Department of Artificial Intelligence Systems has over 30 year of teaching experience in artificial intelligent systems, data discovery and computer science.

Dr. Janina A.Jabubczyc’s interests include knowledge management, data mining with particularly area of classifier ensembles and contexts in concept learning, education and academic-industrial relations. Her activities contain management of postgraduate studies for human capital development.

Mieczyslaw L. Owoc, PhD habilitatus in Economics, associated professor of Computer Science at Wroclaw University of Economics, the Head of Department of Artificial Intelligence Systems, has over 30 years of teaching experience and research in databases and intelligent systems.

He authored over 120 publications mostly oriented on artificial intelligence methods and knowledge management topics. His current research is in modern information technologies including cloud and grid computing with focus on knowledge validation and knowledge grid.