# Knowledge Uncertainty and Composed Classifier

Dana. Klimešová, Eva Ocelíková

***Abstract*** — The paper deals with the relations between knowledge management, uncertainty and the context evaluation on the background of computer science, artificial intelligence and the new possibilities of information technologies that can help us to carry out the knowledge management strategies. The paper discuss the problem of wide context (temporal, spatial, local, objective, attribute oriented, relation oriented) as a tool to compensate and to decrease the uncertainty of data, classification and analytical process at all process to increase the information value of decision support. The contribution deals with a problem of creating the composed classifier with boosting architecture, whose components are composed of classifiers working with k - NN algorithm (k - th nearest neighbour).

***Keywords*** — Boosting architecture., contextual modelling, composed classifier, knowledge, knowledge management, uncertainty.

## I. Introduction

The ways of managing and distributing data and particularly data sources has rapidly changed. Data are collected and processed and during the last couple of years the data flows in and between organizations have extremely increased. In the connection with these facts also the data management tools and techniques are continually changed [10]. It includes the automated knowledge acquiring, proper handling of large volumes of data, new accesses to data interpretation and effective exchange of information between and among various institutions.

The knowledge is an issue changing the society and economy today. Knowledge has become the most important resource. So, knowledge management is an essential process improving competitive advantage.

Knowledge management tries to compensate the loss of stable procedural knowledge, the loss of customer-related or project-related experiences and know-how eventually the loss of the middle management information analysis and routing services.

Together with the globalisation of business, an enormous market pressure enforces ever-shorter product life cycles. On the other hand, modern information technology allows worldwide geographically dispersed development teams, virtual enterprises [3] and close cooperation with suppliers, customer companies, and outsourced service providers. The area in Computer Science that is most influenced by the concept of knowledge is Artificial Intelligence (AI) and Knowledge Based Systems (KBS) [6].

In the early 1980s the development of a KBS was seen as a process of transferring human knowledge to an implemented knowledge base. This transfer was based on the assumption that the knowledge, which is required by the KBS, already exists and only has to be collected and implemented [5].

Since the knowledge is specified independently from the application domain, reuse of the knowledge is enabled for different domains and applications. Besides knowledge modelling and knowledge representation is also an important field of research in computer science and AI.

Some observations can be made about modelling view of the building process of a KBS.

- ❑ The model is only an approximation of reality.
- ❑ The modelling process is a cyclic process. New observations may lead to a refinement and modification. On the other hand, the model may guide further acquisition of knowledge – contextual understanding.
- ❑ The modelling process is dependent on the subjective interpretation of the knowledge engineer. Therefore this process is faulty and an evaluation of the model with respect to reality is indispensable for the creation of an adequate model.

## II. Uncertainty

The dimensionality of data and the complexity of objects structure hierarchy are rapidly growing and consequently with these aspects increase the uncertainty entering into the processing.

Data uncertainty plays a special role in the environment of Internet and Web Services [3]. It is quite another situation than in case of the closed system, where the user has full control over all steps of processing from data input to presentation of results. In frame of open interoperable system with access to web sources with a great number of existing databases the user control gets completely lost.

A great number of existing databases offer a variety of data sets covering different thematic aspects like topographic information, cadastral data, statistical data,

Manuscript received in March 14, 2007
Revised Manuscript received in July 1, 2007.

This work was supported by the project Information and knowledge support of strategic control - MSM 6046070904.

D. Klimešová is with the Czech University of Life Sciences, Prague, Faculty of Economics and Management, Kamýcká 129, 165 21 Prague, CZECH REPUBLIC, (e-mail: klimesova@pef.czu.cz ) and with the Institute of Information Theory and Automation, and Czech Academy of Sciences, Pod vodárenskou věží 4, 182 00 Prague 8, ( e-mail:klimes@utia.cas.cz ).

E. Ocelíková is with Technical University Košice, Faculty of Electrical Engineering and Informatics, Letná 9, 042 00 Košice, SLOVAK REPUBLIC, ( e-mail: ocelike@ccsun.tuke.sk)

digital maps, aerial and satellite images including temporal data. Data collection is changing from digitising own data to retrieving and transferring from existing databases coming from task processing and result presentation.

To deal with such data sets, the user requires an uncertainty description that has to be added by the producer. User needs an appropriate uncertainty model for this purpose, integrated in GIS [4]. From the philosophical point of view the uncertainty is quite natural part of our life and the surrounding world. Usually we meet uncertainty in the sense of valuation.

Uncertainty is a real and universal phenomenon in valuation and the sources of uncertainty are rational and can be identified. Valuation is the process of estimating the value and estimation will be affected by uncertainties. The input uncertainties will translate into an uncertainty of the valuation.

Actually – the uncertainty arises from imperfect understanding of the events and processes in the world around. From another point of view the fact of uncertainty is very stimulating for the research on the field of *defining*, *measuring*, *modelling* and *visualizing* uncertainty and data quality analysis. The uncertainty opens the space for further questions like: *where, why and when* and the answers to this question can help us to do better decisions [7].

To gain the relevant answer it is necessary incorporate the various contexts into the analysis of objects, phenomena, events and processes and connect up uncertainty into the knowledge-construction and decision-making process through context cognition.

Open systems are using frequently various models but the model is only an approximation of reality and the modelling process is dependent on the subjective interpretation of the knowledge. It means that new observations may lead to a further refinement, modification, or completion of the already constructed model. And the model may guide further acquisition of knowledge and the knowledge is the base for decision support. Moreover, besides knowledge modelling also knowledge representation is very important field of research.

### A. Uncertainty Management

**Data are not perfect from many reasons:**

- ❑ Incomplete data
- ❑ Precision of measurements
- ❑ Discreet description of connective phenomena
- ❑ Inherent part reflecting our understanding of things [11]

On the other hand the current top level of GIS usage, it is control GIS, where the large ability is aided to implement knowledge models from different branches of scientific investigation, wide context implementation including less evident connections, models of trends, objects and expected or predicted relations.

To reduce uncertainty of data it is mainly the question of the proof of recognized quality assurance. Some users often take the pragmatic approach to the cost versus accuracy. Sometimes, without the relevance testing, the resolution of data is used for the whole set of different task. Then the problem of over-defined and under-defined objects brings the difficulties [8].

Especially uncertainty of a geographic object can be modelled through uncertainty of its geospatial, temporal, thematic and others attributes. Uncertainty of relations takes into consideration spatial, temporal and spatial- temporal relations.

To add suitable attribute or to spread the net of relations reduce the uncertainty of the object. The special case is to model objects uncertainty using spatial-temporal approach to the objects and incorporate spatial-temporal relationships. The dynamics of object is very powerful tool to obtain exact results about the object and phenomenon behaviour to support further decision [16].

The decision making process is always associated with some level of uncertainty which can rise from:

- ❑ Definition of the problem
- ❑ Data used
- ❑ Sequence of operations used to obtain result
- ❑ Understanding of result

GIS is shifting very fast from desktop GIS to network GIS. Great advantage of network GIS is ability to provide GIS services in a networked environment, typically through the Internet.

With this technology, all GIS components, data components and functional objects, can be distributed across the network. In this component-oriented framework the user has no problem with the increasing complexity of information structures and quality demands and is able understand objects and phenomena and theirs expressions in various context and provide richer analysis with different aspects of modelling.

### B. Context Understanding

The contextual modelling deals with different types of context information. It is possible consider context as follows [13]:

Context as the reflection of object or phenomena using different interpretation through the system of cognition: **perception, conception, and interpretation.**

Context as the reflection of selected facts is concerned with validity of statements and the system of argumentation: **identification, analysis- coordination, and synthesis – decision.**

Context as the reflection when hypothesis stays instead of experience in the system of abduction – instinct based context: **recognition of patterns, coordination by intuition, and judgement due to synthetic inference.**

Context as the reflection concerning validity of statement using knowledge generating system – knowledge based context.

Using context it is possible to derive new quality of information that can be used to support decision. To apply the context the composed classifier provides good frame for this purpose.

## III. COMPOSED CLASSIFIER

Nowadays many classification technologies and algorithms are developed and increased requirements are taken on these technologies in regard to increased precision, to achieve shorter classification time and so on. One of the possible solutions how to increase fruitfulness of classification is utilization of composed classifiers.

Composed classifier is a composition of component classifiers, which predictions are connecting by combining classifier, unlike contrast to simple classifiers. There are several architectures for possible combination of classifiers. Main architectures for combination of classifiers are:

- ❑ Stacked Generalization
- ❑ Boosting
- ❑ Recursive Partitioning

**Stacked Generalization (SG)** [21] is a level architecture for combination of classifiers, in which classifiers on a higher level combine prediction of classifiers immediate on the lower level, figure 1, where C1 – Cn are component classifiers, C0 is combining classifier.

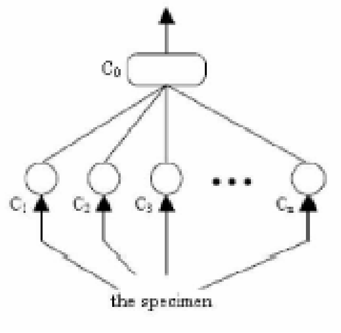


Fig. 1. Architecture SG

**Boostin***g* tries to increase the precision of given classifier by creating a complementary component classifier [17] by filtration of a training set. On recovery resultant prediction is used to vote between existing classifier and new created component, fig. 2, where C1 is given basic classifier, C2 and C3 are supplementary classifiers, C0 is combining classifier.

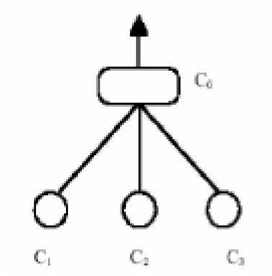


Fig. 2. Boosting architecture

**Recursive Partitioning (RP***)* is method for connecting of classifiers, in which domain space is divided recursively into many areas [20]. One classifier for prediction is applied in every of these areas, figure 3, where C0 – C8 are component classifiers.

Common feature of these three architectures is the fact that they trying to reduce error of classification by combined prediction of component classifiers.

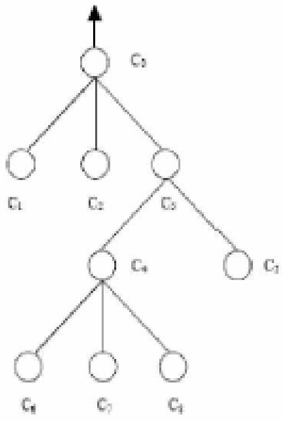


Fig. 3. Architecture RP

### *A. Composite Classifier Design Criteria*

There are three primary criteria that have been applied to creating composite classifiers:

**Accuracy of the component classifiers** - the accuracy criterion arises from a desire to make the component classifiers independently accurate. By “independently,” we mean when applied alone as a single classifier.

**Diversity of the component classifiers** - the diversity criterion arises from the simple observation that combining the predictions of a set of classifiers that all make the same errors cannot lead to any improvement in the accuracy of the composite prediction.

**Efficiency of the entire composite classifier** - the efficiency criterion is less often considered in composite classifier work, but arises from the general requirement that a classifier should use only reasonable amounts of time and memory for training and application.

## IV. EXPERIMENTS

In this work was used composed classifier based on architecture Boosting, which forms more-accurately composed classifier by combine existing basic k-NN classifier with the others supplementary classifiers. Supplementary classifiers were created with methods CR (Coarse Reclassification), DM (Deliberate Misclassification), CF (Composite Fitness) and SF-F (Composite Fitness-Feature Selection) [19]. These methods verified various opinions to create accurate supplementary classifier.

### *A. Algorithm ID3*

Algorithm ID3 [20] is the best-known algorithm generating decision tree by method from the top to down. Finishing criterion of this algorithm is that every subspace contains only examples of the one class.

If a set of attributes is sufficient, the decision trees constructed by the mentioned progress, correctly classify the training examples. For classification of the new example is needed by monitoring the way from radical element of decision tree until the end node. At every interior node follows branch corresponding to the value of testing attribute. A class near the terminal node introduces prediction of the class for existing examples.

### B. *Classifier k-NN*

The classification rule of the k-nearest neighbour [19] is non-parametric statistical criterion. This algorithm designates class of unknown quantities sample according to base the classes to the nearest neighbour. Algorithm operates with constant numbers of attributes and it doesn't need to know statistical distribution of training set. At classification by choosing distance metric calculate distance of testing sample to all placing training samples. Then the sample is assigned from training and set into the class of the nearest neighbour.

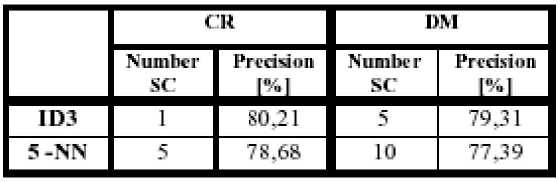
Partial methods of boosting and verification of improvement precision were tested on data from remote sensing of the Earth by LANDSAT TM (7 spectral channels). The data set consists of 368 152 pixels (specimens) of the Earth surface, where one of them represents area of 30 x 30 meters, representing a total of 332 sq km of land. A 7-dimensional vector characterizes every pixel. These partial components are describing the brightness of the seven spectral bands.

### C. *Evaluation*

The best method for creating supplementary classifiers independently on applications combining classifiers by general results on testing database was the already mentioned method Composite Fitness-Feature Selection. In the three cases for methods Coarse Reclassification, Deliberate Misclassification and Composite Fitness is better combining classifier ID3. Additionally ID3 needs smaller numbers of supplementary classifiers (1-3) on achievement of this precision. In the aggregate the best average precision has composed classifier with 10 supplementary classifiers combination with 5-NN classifier.

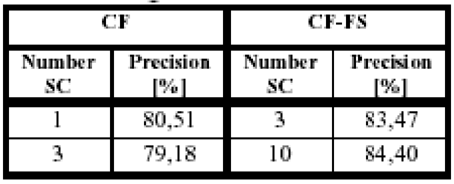
The best average accuracies for existing combining classifiers are mentioned in percentages in Table 1, where **number SC** is the number of Supplementary Classifiers.

TABLE I.
THE BEST RESULTANT ACCURACIES FOR SINGLE METHODS CR AND DM



| | CR | | DM | |
|---|---|---|---|---|
| | Number SC | Precision [%] | Number SC | Precision [%] |
| **ID3** | 1 | 80,21 | 5 | 79,31 |
| **5 -NN** | 5 | 78,68 | 10 | 77,39 |

TABLE II.
THE BEST RESULTANT ACCURACIES FOR SINGLE METHODS CF AND CF-FS



| CF | | CF-FS | |
|---|---|---|---|
| Number SC | Precision [%] | Number SC | Precision [%] |
| 1 | 80,51 | 3 | 83,47 |
| 3 | 79,18 | 10 | 84,40 |

Methods for generation suitable supplementary classifiers were mentioned for existing k-NN classifier. These classifiers were combined by duo combining classifiers - ID3 and k-NN. By comparing the combined classifiers ID3 and k-NN with majority voting, on the average algorithm ID3 reaches for all methods and has better results with smaller number of supplementary classifiers.

## IV. CONCLUSION

The contribution deals with more abstract level for reflection and understanding of the various modelling processes. In this paper, the problem of wide spatial and temporal context is discussed. Our decisions are becoming increasingly dependent on understanding of complex relations and phenomena in the world around and context modelling is able to incorporate new requirements and produce more valuable results. The main goal has been to show selected aspects of this process and compare the increasing possibilities of the sources with the difficulties of data contextual structuring, implementation and evaluation.

The paper shows the architecture of composed classifier gives the space where it is possible to incorporate the additional aspects and refine our decision rules.

## ACKNOWLEDGMENT

The Project Information and knowledge support of strategic control - MSM 6046070904 supports this work.

## REFERENCES

[1] Aamodt, A. and Nygard, M., 1995. Different roles and mutual dependencies of data, information and knowledge. *Data & Knowledge Engineering,* 16, 191-222.

[2] Benedikt J., Reinberg S., Riedl L., 2002. A GIS application to enhance cell-based information modeling. *Information Sciences* 142 (2002): 151-160.

[3] Bernbom, G., 2001. Information Alchemy: *The Art and Science of Knowledge Management,* EDUCAUSE Leadership Series #3. San Francisco: Jossey-Bass. Graham, Ricci.

[4] Bolloju N., 1996. Formalization of qualitative models using fuzzy logic. *Decision support systems 17*(1996): 275-289.

[5] Cornelis, B., and Brunet, S., 2000. A policy-maker point of view on uncertainties in spatial decisions. *Spatial data quality,* Chapter 12, pp. 168-185.

[6] Fensel, D., Decker, S., Erdmann, M., and Studer, R., 1998. Ontobroker: Transforming the WWW into a Knowledge Base. In *Proceedings of the 11th Workshop on Knowledge Acquisition Modeling and Management,* Banff, Canada, April 18-23.

[7] Fuller R., 2000. In: Introduction to Neuro-Fuzzy systems. *Advances in soft computing,* Physica-Verlag Heidelberg. 289 pages.

[8] Klimešová D., 2006. Study on Geo-information Modelling, 5 (2006), *WSEAS Transaction on Systems,* pp. 1108-1114.

[9] Klimešová D., 2004. Geo-information management. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* 35 (2004), 1, pp. 101-106.

[10] O'Leary, D. 1998. Knowledge Management Systems: Converting and Connecting. *IEEE Intelligent Systems*, May/June 1998, pp. 30-33.

[11] Parent C., Spaccapietra S., Zimanyi E.,2000. MurMur: database management of multiple representations. *Proceedings of AAAI-2000 Workshop on Spatial and Temporal Granularity*, Austin, Texas.

[12] Peuqeut, D.J., 2002. *Representations of Space and Time.* The Guilford Press.

[13] Power, D., J., 2002. *Decision Support Systems:* Concepts and Resources for Managers, Quorum Books Published 2002.

[14] Yao T., Journel A. G., 1998. Automatic modeling of (cross) covariance tables using fast Fourier transform. *Mathematical Geology,* 30(6): 589-615.

[15] Zerger A. 2003. Examining GIS decision utility for natural hazard risk modelling. *Environmental modelling & software*, 17 287-294.

[16] Zhang J., Goodchild M. 2002. *Uncertainty in geographical information.* Taylor & Francis, London, pp. 127-130.

[17] FreundY.,SchapireR.,1996. Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning.* Morgan Kaufmann, San Francisco, 1996

[18] Ocelíková,E.,Kristof,J.2001.Classification of Multispectral Data. *Journal of Information and Organizational Sciences,* Vol. 25, Number 1, Varazdín, 2001, pp. 35-41.

[19] Skalak, D.B., 1997. Prototype Selection for Composite Nearest Neighbour Classifiers. *CMPSCI* Technical Report pp.96-89.

[20] Utgoff, P.E. :, 1989. Perceptron Trees: A Case Study in Hybrid Concept *Representations. Connection Science* 1: 377-391.

[21] Wolpert D., 1992. Stacked Gene-ralization. *Neural Networks* 5:241-259.