Recent Advances in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics
Volume II: Applications

Editors

Krassimir T. Atanassov
Władysław Homenda
Olgierd Hryniewicz
Janusz Kacprzyk
Maciej Krawczak
Zbigniew Nahorski
Eulalia Szmidt
Sławomir Zadrożny

SRI PAS IBS PAN

Recent Advances in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics Volume II: Applications

Systems Research Institute
Polish Academy of Sciences

Recent Advances in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics
Volume II: Applications

Editors

Krassimir T. Atanassov
Władysław Homenda
Olgierd Hryniewicz
Janusz Kacprzyk
Maciej Krawczak
Zbigniew Nahorski
Eulalia Szmidt
Sławomir Zadrożny

IBS PAN **SRI PAS**

Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
www.ibspan.waw.pl

ISBN 9788389475367

The hierarchical agglomerative approach to cluster data series

Maciej Krawczak[1, 2] **and Grażyna Szkatuła**[1]
[1]Systems Research Institute, Polish Academy of Sciences
Newelska 6, Warsaw, Poland
[2]Warsaw School of Information Technology
Newelska 6, Warsaw, Poland
{krawczak, szkatulg}@ibspan.waw.pl

Abstract

In this paper we developed a methodology for grouping data series. First, a technique for reduction of dimensionality was applied. The technique is based on the previously developed concept of upper and lower envelopes, aggregation of the envelopes and extracting essential attributes. Next, the values of essential attributes were nominalized and symbols were assigned. For reduced symbolic representation of data series we introduced a definition of conditions domination within each pair of cluster. The developed hierarchical and agglomerative method is characterized both by high speed of computation as well as extremely good accuracy of clustering.

Keywords: data series, cluster analysis, essential attributes.

1 Introduction

Nowadays there are collected many data series characterized by huge number of objects (also called examples) and each object is characterized by a large number of attributes. Analyzing such data very often we have to overcome the "curse of dimensionality" of the problem. The attributes in a data set can be numerical or categorical; the categorical attributes can be either ordinal or nominal. In a case of the ordinal attributes some order relationship between elements of the set of its values have to be distinguished, otherwise we can say about the nominal nature of the attributes. Often nominal attributes are considered in a symbolic way.

Recent Advances in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics. Volume II: Applications *(K.T. Atanassow, W.Homenda, O.Hryniewicz, J.Kacprzyk, M.Krawczak, Z.Nahorski, E. Szmidt, S. Zadrożny, Eds.), IBS PAN - SRI PAS, Warsaw, 2010.*

Within data series analysis there are two main problems, namely clustering and classification of objects.

In this paper we will consider a clustering problem of data series of high dimensionality. In order to reduce the high dimensionality of the data series some effective technique developed by Krawczak and Szkatuła (2009a, 2009b, 2010a, 2010b) was applied. Next, the values of the new reduced attributes were changed into intervals, and then the primary symbolic values were assigned to the ordinal intervals (Krawczak and Szkatuła, 2010c). In order to obtain the symbolic representation of the considered data series we named each difference of two primary symbols by different letters of the alphabet.

In this way we could reduce the high dimension of each object to representation characterized by a short sequence of symbols.

Our aim is to group the data series using the new symbolic representation. To do that we developed a new technique based on introduced relation of dominance between the clusters.

However one can find some algorithms specialized to analysis of long chains of symbols. The algorithms found applications in text analysis or in bioinformatics (Apostolico *et al.*, 2002), Gionis and Mannila, 2003), (Lin *et al.,* 2007). Most of algorithms dealing with symbolic data are based on introduction of some distance between objects, e.g. Wang (2010), Domingo-Ferrer and Solanas (2008).

Our approach to cluster analysis with symbolic data differs from algorithms known in the literature and the efficiency of it seems to be higher than those known in literature. The developed algorithm has several features common with standard ones, namely our algorithm is hierarchical and agglomerative ("bottom-up"). Hierarchical clustering (defined by Johnson in 1967) is starting with N clusters (each containing one object). This kind of algorithms can find the closest (most similar) pair of clusters and merge them into a single cluster. This kind of hierarchical clustering is called *agglomerative* because it merges clusters iteratively. The main weaknesses of agglomerative clustering methods are that they can never undo what was done previously. In our algorithm instead of measure of distance between objects, in the paper, we introduced a definition of the condition's dominance which allowed merging smaller clusters in order to get larger ones.

2 Symbolic representation of data series

We considered a set of data series $\left[x_1(n), x_2(n), \ldots, x_M(n)\right]^T$ for $n = 1, 2, \ldots, N$, M denotes the number of the primary attributes. Our aim is to group N objects into prescribed number of clusters.

In order to reduce the dimensionality of the problem the following procedures were developed:

1) The m-step upper envelopes generation and/or m-step lower envelopes generation, where m denotes a number of sequent data values, $m << M$, are generated. We used the approach developed by Krawczak and Szkatuła (2009a, 2009b, 2010a, 2010b).
2) The envelopes (upper and/or lower) were aggregated. In results we obtained the reduced form of the envelopes, now the dimension of each envelop (representing the object) is equal $\left\lfloor \frac{M}{m} \right\rfloor$.
3) In order to get the further reduction of the objects representations the essential attributes were extracted. The heteroassociative neural network was applied to obtain K essentials attributes, $K << \left\lfloor \frac{M}{m} \right\rfloor$.
4) After application of the nominalization of the essential attributes and next after some rearranging of nominal values of the essential attributes we obtained the symbolic representation of the data objects in the following form:

- We have a finite set of data objects $U = \{e^n\},\ n = 1, 2, \ldots, N$.
- The objects are described in the form of conditions associated with the finite set of attributes $A = \{a_1, ..., a_K\}$.
- The set $V_{a_j} = \{v_{j,1}, v_{j,2}, ..., v_{j,L_j}\}$ is the domain of the attribute $a_j \in A$, $j = 1, ..., K$, where L_j - denotes number of values of the j-th attribute.
- Each object $e^n \in U$ can be described in the form of conjunction of K elementary conditions in the following manner

$$e^n = (a_1 = v_{1,t(1,n)}) \wedge ... \wedge (a_K = v_{K,t(K,n)}) \tag{1}$$

where $v_{j,t(j,n)} \in V_{a_j}$ and $j = 1, ..., K$. The index $t(j,n)$ for $j \in \{1, 2, ..., K\}$ and $n \in \{1, 2, ..., N\}$ denotes that the attribute a_j takes value $v_{j,t(j,n)}$ in the object e^n.

For example, for the *j*-th attribute the set $V_{a_j} = \{v_{j,1}, v_{j,2}, ..., v_{j,L_j}\}$ using letters of the alphabet, can have the following symbolic form for $L_j = 9$

$$V_{a_j} = \{a, b, c, d, e, f, g, h, i\}.$$

An exemplary data object for a given $n \in [1, N]$ can be written as follows:

$$e^n = [(a_1 = b) \wedge (a_2 = d) \wedge (a_3 = f) \wedge (a_4 = c) \wedge (a_5 = e) \wedge \\ (a_6 = f) \wedge (a_7 = c) \wedge (a_8 = k) \wedge (a_9 = a) \wedge (a_{10} = g)]$$

or shortly

$$e^n = [\mathrm{b,d,f,c,\ e,f,c,k,a,g}].$$

3 Basic elements of the hierarchical agglomerative approach applied

The task of clustering can be formulated as follows: we want to splits the set of objects *U* into non-empty, disjoint subsets $\{C_1, C_2, ..., C_C\}$, $\bigcup_{g=1}^{C} C_g = U$, (called *clusters*) so that objects in the same cluster are similar in some sense. The set of clusters on *U* is denoted by *C(U)*.

If a certain object belongs to a definite cluster then it could not be included in another cluster, by assumption. Basic elements of proposed method were introduced below.

Consider a attribute a_j, $j = 1, ..., K$ and no empty sets $A_{j,t(j,k)}$ and $A_{j,t(j,n)}$, where $A_{j,t(j,k)} \subseteq V_{a_j}$, $A_{j,t(j,n)} \subseteq V_{a_j}$.

We say that the condition $(a_j \in A_{j,t(j,k)})$ *dominates* the condition $(a_j \in A_{j,t(j,n)})$ if the clause $A_{j,t(j,k)} \supseteq A_{j,t(j,n)}$ is satisfied, denoted by $(a_j \in A_{j,t(j,k)}) \succ (a_j \in A_{j,t(j,n)})$. Let us notice that condition $(a_j \in \{a, b, f\})$ dominates the condition $(a_j \in \{a, f\})$, i.e. $(a_j \in \{a, b, f\}) \succ (a_j \in \{a, f\})$.

We assume that there is *lack of mutual dominance* two conditions $(a_j \in A_{j,t(j,k)})$ and $(a_j \in A_{j,t(j,n)})$ if the first condition does not dominates the second and the second condition does not dominates the first, denoted by $(a_j \in A_{j,t(j,k)}) \prec\succ (a_j \in A_{j,t(j,n)})$. Let us notice that there is lack of mutual dominance of two conditions $(a_j \in \{a,b,f\})$ and $(a_j \in \{a,c\})$, i.e. $(a_j \in \{a,b,f\}) \prec\succ (a_j \in \{a,c\})$.

The *cluster* C_g can be expressed as follows:

$$(a_1 \in A_{1,t(1,g)}) \wedge ... \wedge (a_K \in A_{K,t(K,g)}) \tag{2}$$

where $A_{j,t(j,g)} \subseteq V_{a_j}$, for $j = 1,...,K$.

We say that the object $e^n = (a_1 = v_{1,t(1,n)}) \wedge ... \wedge (a_K = v_{K,t(K,n)})$ *belongs to the cluster* C_g if conditions:

$$\begin{gathered}(a_1 \in A_{1,t(1,g)}) \succ (a_1 = v_{1,t(1,n)}) \\ \cdots \\ (a_K \in A_{K,t(K,g)}) \succ (a_K = v_{K,t(K,n)})\end{gathered} \tag{3}$$

are satisfied.

Let's consider two clusters: $C_{g_1} : (a_1 \in A_{1,t(1,g_1)}) \wedge ... \wedge (a_K \in A_{K,t(K,g_1)})$ and $C_{g_2} : (a_1 \in A_{1,t(1,g_2)}) \wedge ... \wedge (a_K \in A_{K,t(K,g_2)})$, for $A_{j,t(j,g_1)} \subseteq V_{a_j}$, $A_{j,t(j,g_2)} \in V_{a_j}$, $j = 1,...,K$.

We say that the cluster C_{g_1} and the cluster C_{g_2} are ω*-distinguishable* for the set of attributes $\{a_j : j \in I_k\}$, $card(I_k) = \omega$, if two conditions are satisfied:

1) $(a_j \in A_{j,t(j,g_1)}) \prec\succ (a_j \in A_{j,t(j,g_2)})$, $\forall j \in I_k$

2) $(a_j \in A_{j,t(j,g_1)}) \succ (a_j \in A_{j,t(j,g_2)})$ or $(a_j \in A_{j,t(j,g_2)}) \succ (a_j \in A_{j,t(j,g_1)})$, $\forall j \in \{1,...,K\} \setminus I_k$. (4)

We assume that the cluster $C_{g_1} : \{e^n :\ e^n \in U,\ n \in J_{g_1} \subseteq \{1,2,...,N\}\}$ and the cluster $C_{g_2} : \{e^n :\ e^n \in U,\ n \in J_{g_2} \subseteq \{1,2,...,N\}\}$ are ω-distinguishable for the set of attributes $\{a_j : j \in I_k\}$, $card(I_k) = \omega$.

The ω - *conditional action rule* is defined in the following manner:

$$\underset{j \in I_k}{\wedge} (A_{j,t(j,g_3)} := A_{j,t(j,g_1)} \cup A_{j,t(j,g_2)}) \text{ and}$$

$$\underset{j \in \{1,2,...K\} \setminus I_k}{\wedge} (A_{j,t(j,g_3)} := \ dom \ \{A_{j,t(j,g_1)}, A_{j,t(j,g_2)}\}) \qquad (5)$$

$$\Rightarrow ((C_{g_1}, C_{g_2}) \to (C_{g_3}))$$

where *dom* - dominant condition.

In result a new cluster $C_{g_3} : (a_1 \in A_{1,t(1,g_3)}) \wedge ... \wedge (a_K \in A_{K,t(K,g_3)})$ contains the following objects $\{e^n : \ e^n \in U, \ n \in J_{g_1} \cup J_{g_2}\}$.

We proposed a hierarchical agglomerative approach to cluster nominal data. The bottom level of the structure has singular clusters while the top level contains one cluster with all objects. During iteration two clusters are heuristically selected. These selected clusters are then merged to form a new cluster, see Example 1.

Example 1. Let's consider two clusters C_{g_1} and C_{g_2} shown in Table 1. The cluster C_{g_1} contains objects *e1* and *e2*. The cluster C_{g_2} contains objects *e3*, *e4* and *e5*. Let us notice that the cluster C_{g_1} and C_{g_2} are *4*-distinguishable for the set of attributes $\{a_j\}$, $j \in I_k = \{1,4,5,10\}$, $card(I_k) = 4$.

Table 1

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
C_{g_1} : *e1, e2*	e	e	g∨h	e∨d	d	g	g	f∨g	f	f
C_{g_2} : *e3, e4, e5*	f	e	g	f∨e	e	g	g	g	g∨f	g

The way of construction of a new cluster C_{g_3} is shown in Table 2.

Table 2

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
C_{g_3} :*e1,e2, e3, e4, e5*	e∨f	e	g∨h	e∨d∨f	d∨e	g	g	f∨g	f∨g	f∨g

New cluster $C_{g_3} : (a_1 \in \{e, f\}) \wedge (a_2 \in \{e\}) \wedge ... \wedge (a_K \in \{f, g\})$ contains objects *e1*, *e2*, *e3*, *e4* and *e5*.

□

Suppose we have a finite set of objects $U = \{e^n\}$, $n = 1, 2, \ldots, N$. The objects are described in the form of conditions associated with the finite set of attributes. We want to splits the set of objects U into non-empty, disjoint subsets $\{C_1, C_2, ..., C_C\}$, $\bigcup_{g=1}^{C} C_g = U$.

Basic elements of proposed algorithm were introduced below.

Step 1. U – set of objects, K - number of attributes, C - waited number of clusters. Each object creates one-element cluster in the initial set of clusters *C(U)*, $card(C(U)) = N$, $\omega = 0$.

Step 2. From pair of ω-distinguishable clusters we create new cluster. If $card(C(U)) = C$, go to Step 4; otherwise, if it exists pair of ω-distinguishable clusters, repeat Step 2; otherwise, go to Step 3.

Step 3. $\omega := \omega + 1$; if $\omega \leq K$, repeat Step 2; otherwise, go to Step 4.

Step 4. STOP.

Some basic elements of the method used to generated clusters are presented in Example 2.

Example 2. We considered the data showed in Table 3. The objects *e1,...,e6* are described in the form of conditions associated with the set of attributes $\{a_1, \ldots a_5\}$.

Table 3

Object \ Attribute	a_1	a_2	a_3	a_4	a_5
e1	c	b	a	a	b
e2	b	a	b	a	c
e3	d	b	c	a	b
e4	d	a	a	b	a
e5	b	a	b	b	a
e6	d	b	c	a	b

Each object creates one-element cluster in the initial set of clusters *C(U)*, $card(C(U)) = 6$, $K = 5$, $C = 2$, $\omega = 0$.

From a pair of *0*-distinguishable clusters is created new cluster (Table 4). The newly formed clusters are shaded in the tables 4, 5, 6 and 7.

Table 4

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5
e1	c	b	a	a	b
e2	b	a	b	a	c
e3, e6	d	b	c	a	b
e4	d	a	a	b	a
e5	b	a	b	b	a

Pair of *1*-distinguishable clusters does not exist. From each pair of *2*-distinguishable clusters is created new cluster.

Table 5

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5
e2	b	a	b	a	c
e1, e3, e6	c ∨ d	b	a ∨ c	a	b
e4	d	a	a	b	a
e5	b	a	b	b	a

Table 6

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5
e2	b	a	b	a	c
e1, e3, e6	c ∨ d	b	a ∨ c	a	b
e4, e5	b ∨ d	a	a ∨ b	b	a

Table 7

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5
e1, e3, e6	c ∨ d	b	a∨ c	a	b
e2, e4, e5	b ∨ d	a	a ∨ b	a ∨ b	a ∨c

$card(C(U)) = 2$, STOP.

In this way, two clusters C_1 and C_2 were formed. The cluster C_1 contains objects *e1*, *e3* and *e6*. The cluster C_2 contains objects *e2*, *e4* and *e5*.

The hierarchical clustering dendrogram representing the entire process of going from individual objects to one big cluster is shown below.

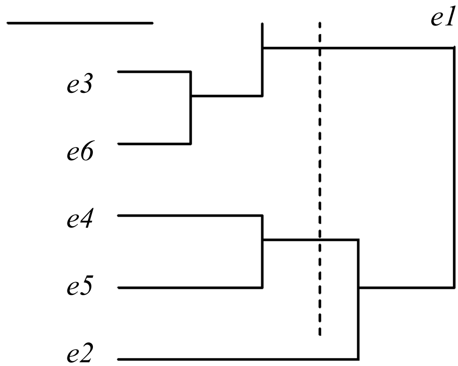

Figure 1: Dendrogram

In the dendrogram shown above, every object creates one-element cluster in the initial set the clusters *C(U)*, $card(C(U)) = 6$. Objects *e3* and *e6* are the most similar (*0*-distinguishable clusters) and join to form the new cluster, similarly objects *e4* and *e5* (*2*-distinguishable clusters). The last two clusters to form are C_1: {*e1, e3, e6*} and C_2 : {*e4, e5, e2*}.

Next, it is possible to create descriptions of created clusters (on basis of Table 7) in the following form

Table 8

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5
C_1	c	*	*	*	*
	*	b	*	*	*
	*	*	c	*	*
	*	*	*	*	b
C_2	b	*	*	*	*
	*	a	*	*	*
	*	*	b	*	*
	*	*	*	b	*
	*	*	*	*	a ∨ c

These descriptions can be represented in the form of rules of "IF *certain conditions are fulfilled* THEN *membership in a definite cluster takes place*" type. In our case, the conditional part of the rules will contain the disjunction of conditions related to the subset of attributes selected for the description of the objects. The rules obtained are shown below.

IF $(a_1 \in \{c\}) \vee (a_2 \in \{b\}) \vee (a_3 \in \{c\}) \vee (a_5 \in \{b\})$ THEN C_1
IF $(a_1 \in \{b\}) \vee (a_2 \in \{a\}) \vee (a_3 \in \{b\}) \vee (a_4 \in \{b\}) \vee (a_5 \in \{a,c\})$ THEN C_2

4 Conclusions

In this paper we considered streaming data characterized by a huge dimensionality. A procedure for reduction of data dimensionality, developed by the authors in earlier papers, was successfully applied. In order of simplification the objects data the description was changed into categorical form, namely into symbolic form using alphabet letters.

For such objects description we introduced and developed the algorithm based on the idea of dominations of conditions within each pair of cluster. Each cluster is described by a conjunction of conditions associated with attributes describing objects.

The primary solved examples shows that the efficiency of the proposed method seems to be effective even though the approach is not so complicated.

Acknowledgements

The research was partially supported by the Ministry of Science and Higher Education (Poland) under Grant Nr N N519 384936.

References

[1] Apostolico R., Bock M. E., Lonardi S. (2002). Monotony of surprise in large-scale quest for unusual words. In: Proceedings of the 6th International conference on research in computational molecular biology, Washington, DC, April 18-21, 22-31.

[2] Dunn J. C. (1973). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics* 3: 32-57.

[3] Bezdek J. C. (1981). "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York.

[4] Dempster A. P., Laird N. M., and Rubin D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM algorithm", *Journal of the Royal Statistical Society*, Series B, vol. 39, 1:1-38.

[5] Gionis A., Mannila H. (2003). Finding recurrent sources in sequences. In: Proceedings of the 7th International conference on research in principles of database systems, Tucson, AZ, May 12-14, 249-256.

[6] Johnson S. C. (1967). Hierarchical Clustering Schemes, *Psychometrika*, 2:241-254.

[7] Krawczak M., Szkatuła G. (2010a). On time series envelopes for classification problem. Developments of fuzzy sets, intuitionistic fuzzy sets, generalized nets, vol. II, 2010.

[8] Krawczak M., Szkatuła G. (2010b). Time series envelopes for classification. In: Proceedings of the conference: 2010 IEEE International Conference on Intelligent Systems, London, UK, July 7-9 2010, 156-161.

[9] Krawczak M., Szkatuła G. (2010c). Redukcja wymiarowości szeregów czasowych. Studia i materiały Polskiego Stowarzyszenia Wiedzą, No. 31, 32-45.

[10] Kumar N., Lolla N., Keogh E., Lonardi S., Ratanamahatana C., Wei L. (2005). Time-Series Bitmaps: A Practical Visualization Tool for Working with Large Time Series Databases. In proceedings of SIAM International Conference on Data Mining (SDM '05), Newport Beach, CA, April 21-23, 2005.

[11] Lin J., Keogh E., Wei L., Lonardi S. (2007). Experiencing SAX: a Novel Symbolic Representation of Time Series. Data Min Knowledge Disc, 2, 15, 107–144.

[12] MacQueen J. B. (1967). “Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability”*, Berkeley, University of California Press, 1:281-297.

[13] Nanopoulos A., Alcock R., & Manolopoulos Y. (2001). Feature-based Classification of Time-series Data. International Journal of Computer Research, 49-61.

[14] Oja E. (1992). Principal components, minor components and linear neural networks. Neural Networks, vol.5, 927-935.

[15] Wei L., Keogh E. (2006). Semi-Supervised Time Series Classification. In: *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006),* 748 - 753, Philadelphia, PA, U.S.A., August 20-23, 2006.

The papers presented in this Volume 2 constitute a collection of contributions, both of a foundational and applied type, by both well-known experts and young researchers in various fields of broadly perceived intelligent systems.
It may be viewed as a result of fruitful discussions held during the Ninth International Workshop on Intuitionistic Fuzzy Sets and Generalized Nets (IWIFSGN-2010) organized in Warsaw on October 8, 2010 by the Systems Research Institute, Polish Academy of Sciences, in Warsaw, Poland, Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences in Sofia, Bulgaria, and WIT - Warsaw School of Information Technology in Warsaw, Poland, and co-organized by: the Matej Bel University, Banska Bystrica, Slovakia, Universidad Publica de Navarra, Pamplona, Spain, Universidade de Tras-Os-Montes e Alto Douro, Vila Real, Portugal, and the University of Westminster, Harrow, UK:

Http://www.ibspan.waw.pl/ifs2010

The consecutive International Workshops on Intuitionistic Fuzzy Sets and Generalized Nets (IWIFSGNs) have been meant to provide a forum for the presentation of new results and for scientific discussion on new developments in foundations and applications of intuitionistic fuzzy sets and generalized nets pioneered by Professor Krassimir T. Atanassov. Other topics related to broadly perceived representation and processing of uncertain and imprecise information and intelligent systems have also been included. The Ninth International Workshop on Intuitionistic Fuzzy Sets and Generalized Nets (IWIFSGN-2010) is a continuation of this undertaking, and provides many new ideas and results in the areas concerned.

We hope that a collection of main contributions presented at the Workshop, completed with many papers by leading experts who have not been able to participate, will provide a source of much needed information on recent trends in the topics considered.

ISBN-13 9788389475367
ISBN 838947536-7
9 788389 475367