An Expanding Context against Weighted Voting of Classifiers

Vagan Terziyan (*), Boris Omelayenko (*), Seppo Puuronen (**)

(*) Department of Artificial Intelligence, State Technical University of Radioelectronics
14 Lenin Avenue, 310166, Kharkov, Ukraine
phone: +380 302 214; fax: +380 572 479 113;
e-mail: vagan@kture.cit-ua.net *;* vagan@jytko.jyu.fi

(**) Department of Computer Science and Information Systems,
University of Jyväskylä, P.O.Box 35, SF-40351, Jyväskylä, Finland
phone: +358 14 603 028; fax +358 14 603 011; e-mail: sepi@jytko.jyu.fi

Abstract

The paper describes the use of a context to improve the classification accuracy. The two-dimensional context is considered based on contextual features and contextual examples. The quality function is defined which evaluates the context in the interval [0,1] from the point of view of its effect to classification accuracy. The key idea presented in the paper is the idea of *expanding context*: if one puts the predictions of the classifiers in order according to the quality of their contexts then it is assumed the existence of a trend among these predictions. This trend can be successfully used to *extrapolate* the opinions and find the integrated one outside all the opinions in the point of optimal quality. The approach fits well for a continuous outcome classification (regression). One can create the ensemble of classifiers using the same learning algorithm several times in the contexts of different quality. For some families of classifiers, for example Nearest Neighbor, or linear regression, classifier is constructed from its context and is equal to the context. When classifying a new instance one should build the sample set of examples (quality – prediction) based on predictions of the classifiers. By extrapolating this set of points one can obtain the value of prediction for maximal (equal to 1) quality. It is experimentally shown that for some datasets the extrapolation of expanding contexts performs better than voting.

1. Introduction

Data Mining or knowledge discovery is the process of finding previously unknown and potentially interesting patterns and relations in large databases [Fayyad et al., 1997]. Classification or pattern recognition is the most important task in Data Mining. The databases targeted by Data Mining tools usually store information about many features of thousands of objects, or multiple dimensions. The number of dimensions provided by features is very high as well as the number of objects. Processing all this information by learner, or classifier, is usually computationally impossible. That's why the traditional Data Mining scheme includes special preprocessing stage to select features and objects to be processed by the classifiers. This stage consists of two parts: feature selection and example selection. Relevant features are selected on the feature selection stage, and relevant, or representative, examples are selected on the example selection stage.

Each classifier consists of some rules and algorithms to classify, distance and information measures, etc. There are a number of combined and adaptive classification schemes that adapt themselves to match trends in the input data. In the case of homogeneous adaptive classifies we

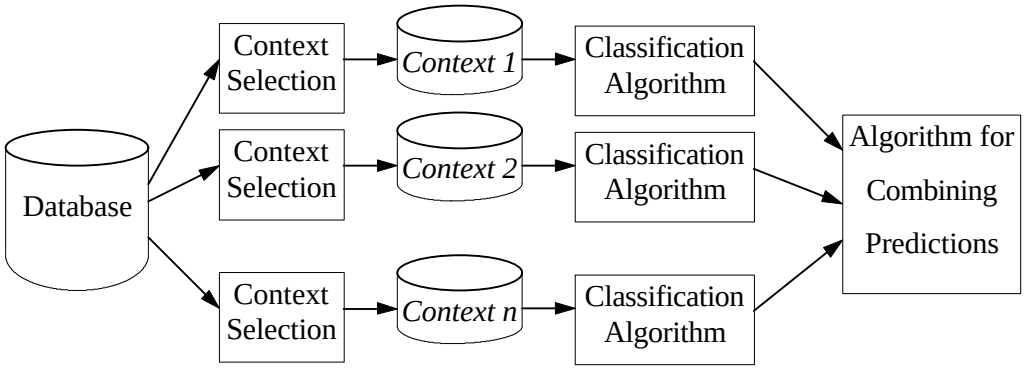

Figure 1. Data Mining Scheme

will have a set of the same classifiers, which will perform and reconfigure themselves differently on different sets of input features and examples. For, example, the dynamic selection algorithm [Terziyan, Tsymbal & Puuronen, 1998] will select different basic classification algorithms and distance measures on slightly different input data. That's why we say that the output of data preprocessing (feature selection and example selection) has the two-dimensional *context* for each classifier.

The illustration for the above is presented in Figure 1. Different feature and example selection algorithms marked as Context Selection use the same database and produce different contexts. These contexts are marked as Context 1, ... , Context *n* in the Figure 1. Classification algorithms do not use the whole data, but they perform strictly in the correspondent context. The classifiers opinions are combined on the last stage of the scheme.

The goal of this paper is to investigate some applications of the theory of context in managing data mining tasks.

2. Assumption of Extrapolated Context

Let us suppose that each classification algorithm A_i works in some context. Thus *n* classification algorithms will use *n* different contexts. Let us denote the prediction made by the *i*-th algorithm A_i with P_i. We also define monotone quality evaluation function *Q* for evaluating the context of the classifier. The main assumption of the paper is:

If one puts the predictions of the classifiers in order according to the quality of their contexts then it is assumed the existence of a trend among these predictions.

This assumption leads to the main idea of the paper:

If the classifiers in their contexts form a trend of their (possibly incorrect) opinions then the classifier in ideal (best possible) context (virtual classifier) will give the correct (best possible) opinion.

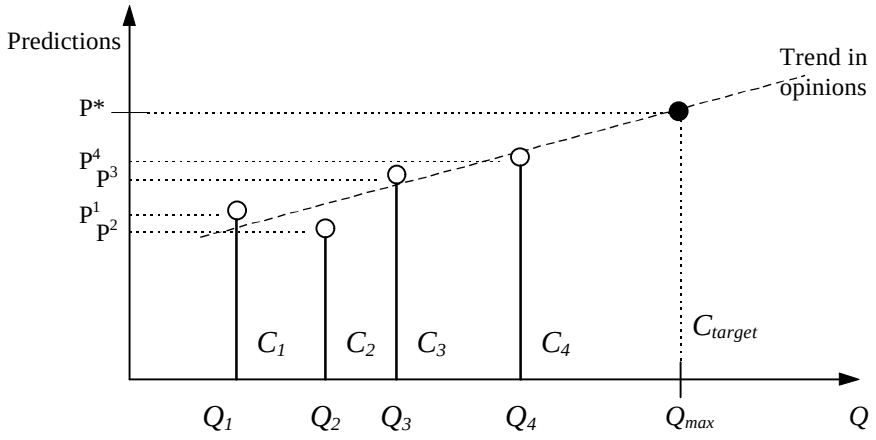

Figure 2. The notion of extrapolated context

We name this assumption as the notion of *extrapolated context*. This notion is illustrated in Figure 2. We assume that there are few homogeneous classifier, each of which works in its own context *C*. For some families of classifiers, for example Nearest Neighbor, or linear regression, classifier is constructed from its context and is equal to the context. These contexts are expanding and the context C_2 is larger than C_1. Moreover, C_2 includes C_1: $C_1 \subset C_2$. The quality evaluation Q_i of these contexts if presented with the function *Q* on the horizontal axis. Opinions P_i of the classifiers are marked on the vertical axis in the Figure. The main assumption – the trend in opinions is presented with the dash line; the virtual classifier in the context of the best quality $Q*$ will have the opinion $P*$, very close to the real class and marked in the Figure 2 with the black circle. Under the basic idea the context of the best quality will consists of only the target object. Nearest Neighbor classifier of this context will consist of only one example with the same features as the target example and the class, defined using the trend in opinions.

This is a novel way of combining multiple opinions in Data Mining, and it stands opposite to voting methods, which used to interpolate the opinions.

This idea makes quite responsible the selection of the quality function *Q*. Two possible interpretations of this function in terms of Data Mining tasks and the corresponding two ways of understanding the notion of context in classification tasks is discussed in the next chapter.

The idea of expanding context is close to the idea of metalevel context described in [Terziyan & Puuronen,1997]. In [Terziyan & Puuronen,1997] a multilevel semantic network is used to represent knowledge within several levels of contexts. The zero level of representation is semantic network that includes knowledge about basic domain objects and their relations. The first level of presentation uses semantic network to represent contexts and their relationships. The second level presents relationships of metacontexts i.e. contexts of contexts, and so on at the higher levels. The topmost level includes knowledge which is considered to be “truth” in all the contexts. Thus a

semantic metanetwork is the hierarchical set of semantic networks above each other so that relations of each previous level are context objects of the next level. This paper presents our understanding of the notion of expanding context in terms of semantic networks.

The context is also used in Dynamic Selection classification approach [Terziyan, Tsymbal & Puuronen, 1998]. The nearest neighbors of the target example (which will be classified) is used there to select the appropriate classification technique. In this approach context not just affects the classification algorithm but causes the selection of different algorithms. The notion of expanded context is a natural improvement of this approach.

3. Two Dimensions of Context in Data Mining

As it was mentioned above, data preprocessing stage in Data Mining selects relevant features and relevant examples for each classifier. They both form the two-dimensional context in which the classifier works: relevant features are the first dimension and relevant examples are the second. In this chapter we discuss the way used by standard preprocessing algorithms to create these contexts and a method for decontextualization, based on the notion of expanding context.

3.1. Context of Relevant Features

The formal definition of context (actually context of features) in terms of Data Mining is presented in [Turney, 1996a]. The main idea of the definition is: A primary feature is informative about the class when considered all by itself, without the remaining features. A contextual feature is only relevant when considered in some non-empty context; a contextual feature is irrelevant when considered in isolation. The definition from [Turney, 1996a] shows that straightforward usage of the definition for feature selection is computationally prohibitive.

Five basic strategies for using contextual information from [Turney, 1996b] are so-directed with the notion of expanding context. For example the strategy of contextual expansion is similar to a two-stage expanding context: the first, narrow context of target example, is presented with primary features and wider context – with contextual. But the expansion has a different form from the proposed in present paper. One of the open questions from [Turney, 1996b] is: Is the presented list of strategies complete? Present paper adds another strategy to this list.

The excellent survey of feature selection methods [Dash & Liu, 1997] covers 31 feature selection methods, classified by five evaluation measures for feature selection quality and two strategies for finding optimal feature subset, called generalization measures. Most of the methods use various heuristics for search, and that's why do not provide comprehensive selection for relevant features. These heuristics are widely used in practical feature selection methods and lead to subjective selection of the sets of relevant features by each algorithm. Different feature selection algorithms,

or even the same algorithm in some cases, can give different sets of features, considered as relevant. And these different sets will force correspondent classification algorithms to work in different context, defined by relevant features.

Feature selection algorithm usually iteratively generates sets of relevant features and evaluate the 'quality' of these sets with a heuristic quality evaluation function. Four basic steps in a typical feature selection method are defined in [Dash & Liu, 1997]:

(1) a generation procedure to generate the next candidate feature subset;

(2) an evaluation function to evaluate the subset under examination;

(3) a stopping criterion to decide whether to stop; and

(4) a validation procedure to check whether the subset is valid.

These steps allow making the following conclusions:

(a) generation procedure is usually heuristic, because exhaustive search of feature subsets is computationally very difficult and prohibited: different feature selection algorithm, and even the same algorithm will generate different subsets;

(b) evaluation function and validation procedure are always heuristic and different evaluation function will select different subsets;

(c) stopping criterion can interrupt selection procedure any time;

Thus, different feature selection algorithms will generate different subsets of 'relevant' features. We can say that context of features is subjective to the feature selection algorithm which refined it.

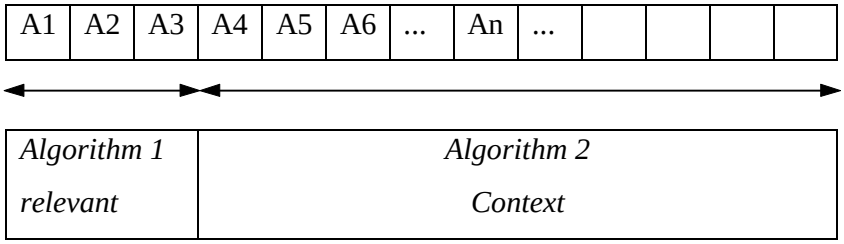

Figure 3. Relevant and Contextual Features

Figure 3 presents an illustration for the process of refining relevant and contextual features. Relevant features in the figure are features A1-A3, and they are processed by classification algorithm 'Algorithm 1'. Features, which feature selection algorithm marked as not 'very relevant' are considered as context and are processed by classification algorithm 'Algorithm 2'. This illustration allows making the following view to the philosophy of context:

The context of an object is such subset of object's features that is reasonable to process by another classification algorithm than one used for relevant features.

In this case Algorithm 1 must be a precise algorithm to lead the classification. It has harder time complexity and handles noise and irrelevant features poorly, because it obtains only high-relevant and not noisy ones. For example it can be an ensemble of k-NN classifiers, which do not handle noisy and irrelevant features well, or it can be a decision tree. Algorithm 2 has to work with more noisy data, with irrelevant features presence. This can be Bayesian inducer, Winnow, Boosting or other algorithm with not very high time complexity.

Actually, some classification algorithms try to reduce the number of considered features themselves. For example, C4.5 algorithm tries to remove from the decision tree the attributes that provide low information gain about the target concept, and do not consider them in the future classification.

Winnow classifier [Blum, 1997] forms a large set of very simple classifiers, each of which is based on two different features. Then it updates weights of classifiers according to their accuracy and prunes classifiers with low weights. This will remove classifiers based on irrelevant attributes and exclude irrelevant features from consideration. Such an idea to start with a wide context (full possible context) and then narrow it to the optimal size lays closely to the idea of expanding context.

In the Nearest Neighbor family of classifiers the context of features is presented with the distance measure, the function which calculated the distance between examples in the space of their features. Classical Euclidean metric gives equal weight to every dimension of the space. But the distance along more important dimension (which presents more relevant attribute) must influence on the distance more than the distance along less relevant dimension (or feature). This problem is extensively discussed in [Wilson & Martinez, 1997], where new heterogeneous distance functions are introduced: Heterogeneous, Interpolated and Windowed. Attribute values are divided into ranges and then interpolated in these functions. Authors state the problem of selecting these ranges as an open question. Present research is strongly connected with [Wilson & Martinez, 1997] and proposes a look on this problem from the area of contexts. During our experiments we met the similar problems of understanding the bounds of context. And, possibly, the proposed notion of extrapolated context will improve the area of distance functions, too.

Each feature selection algorithm uses its own evaluation function to evaluate quality of generated subset. It applies the function to generated subsets and selects the ‘best’ subset according to this evaluation function. We can use similar evaluation functions, described in [Dash & Liu, 1997] to evaluate quality of feature subsets generated by different (or even the same) feature selection methods. For this we apply the evaluation function to feature subsets, generated by different

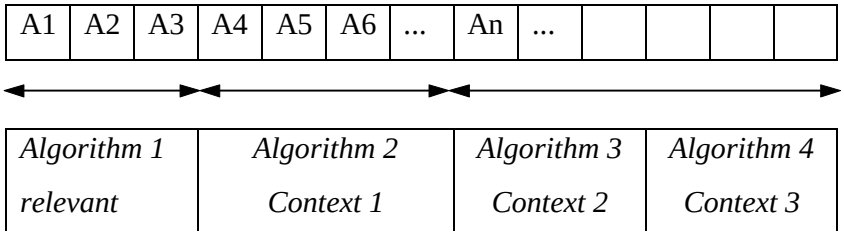

Figure 4. Multiple contexts

feature selection methods. Possible evaluation functions from [Dash & Liu, 1997] used to evaluate the selected feature subset, are:

(1) Distance measure between conditional class probabilities;

(2) Information measure for amount of information about the class provided by the feature subset;

(3) Dependency, or correlation measure between features of the subset;

(4) Consistency measure of generated on the feature subset hypotheses.

The quality of the context, however, can be higher than the quality of relevant features. For example, three relevant features can contain less information about the class than ten context features. Thus in such case the evaluation function '(2) Information measure' will say that context of relevant features is more important than the relevant features themselves.

These context features can then be divided into more relevant and contextual. This is illustrated in Figure 4. Algorithm 1 from the Figure divides all the features into relevant and contextual (see Figure 3). If the context is more informative than the set of relevant features then another algorithm (Algorithm 2 in the Figure) refines more relevant part of the context, marked as Context 1, and the context of the second level, marked as Context 2 and Context 3. Algorithm 4 in turn will refine third-level context and so on. There will be a good idea to use different classification algorithms on each context, as it was discussed above.

3.2. Context of Relevant Examples

The second dimension of context in Data Mining is the context of relevant examples. Preprocessing stage of must filter irrelevant and noisy examples. Examples with many missing attribute values are also good candidates for filtering.

Some classification algorithms themselves try to narrow their context of examples. For example, AdaBoost iteratively re-weight training examples to concentrate on examples which it learned badly. For learning AdaBoost uses a set of so-called 'weak' learners which predict slightly better than random guessing. AdaBoost starts with a distribution with equal weights of examples, or with the wide context for weak learners; and continue with transformed and narrowed set of examples or narrowed context. This method is also the use of the idea of expanding context.

And, finally, Nearest Neighbor family of classifiers is completely based on the context of the target example (example with unknown class). In the *k*-Nearest Neighbor classifier the class of the target example is defined by voting of classes of *k* its nearest neighbors from the training set. The algorithm consists of training examples, distance measure between examples in the space of their attributes and voting mechanism. The last two aspects are usually fixed and we can say that it is fully defined by the examples, which represent the context of examples:

The Nearest Neighbor classifier coincides with its context.

The problem of selecting and treating the context of examples is very important for this family of classifiers because they have no mechanism to refine the context, opposite to more sophisticated classifiers.

In [Skalak, 1997] the accuracy of ensembles of k-Nearest Neighbor classifiers was significantly improved by special selection of relevant training examples, presented to the classifier. The selection algorithm selects very small set of training examples (prototypes), typically one per class. The main idea of this approach is: 'The accuracy of a standard nearest neighbor classifier can be improved or exceeded through composite-classifier architectures that incorporate a small number of diverse component nearest neighbor classifiers, each of which stores only a small number of prototypes' [Skalak, 1997, p. 11].

Early research in this area assumed that one wide context, presented with great number *k* of considered neighbors, will improve the classification. But [Skalak, 1997] investigates ensemble of classifiers, each of which stands in a very narrow context, as narrow as possible. These ensembles work faster than classifiers in wide context and outperform the latter. In contextual terms this result sounds as follows:

It is better to use few Nearest Neighbor classifiers in narrow context selected for each classifier than to use single Nearest Neighbor classifier in a wide context.

Let us consider an example of expanding context of neighbors presented in Figure 5. The nearest context of an object is presented with its one nearest neighbor. This context is the narrowest. The second context, is wider that the first one and includes it; it is presented with two additional neighbors. The third, wider context is presented with three additional neighbors. It also includes first and second contexts. If we denote these contexts as *Context-1*, *Context-2* and *Context-3* then *Context-0* will be the 'zero' context of the object to be classified. This context includes nothing but the target object. In Figure 2 these contexts are denoted with C_1, C_2, ..., and *Context-0* is denotes with C_{target}. Nearest neighbor classifier has natural measure for quality of the context. Evaluation function for quality is the distance function between examples in the space of their features, the same as used to find nearest neighbors.

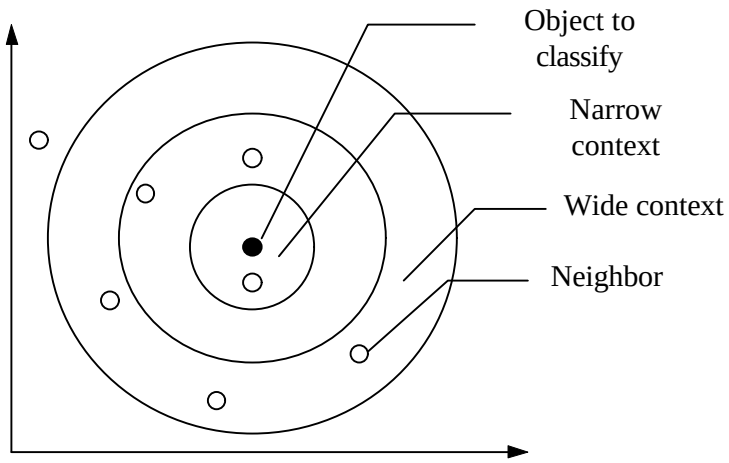

Figure 5. Expanding Context of Neighbors

The quality of *Context-0* is the best, because it is the smallest context and the closest to the target example. The correspondent virtual classifier in *Context-0* consists of the only example – the target example to be classified. If we refer to the basic assumption about the trend in classifier opinions (see Figure 2), then the class this virtual closest neighbor in *Context-0* calculated using the trend will give the correct class of target object.

4. Discussion

There are few successful efforts of elicitation and usage of context in machine learning. A number of batch learning tasks was treated well. For example, [Harries, Horn, 1996] achieved significant accuracy improvement (up to 30%) due to extracting and processing stable domain objects in drifting context. Many different approaches for usage of context in machine learning is described in [Harries, Sammut, 1998]. In papers [Turney, 1993a] and [Domingos, 1997] good results are also reported.

In the present we will consider an experimental example of expanding context in the Nearest Neighbor classifier. We used the Vowel data set from the UCI Machine Learning Repository of Databases. This data set presents coded information about similar vowel sounds pronounced by different people. Seven speakers train the classifier and the other seven are used to test the classifier. Experimental results are presented in Table 1, where seven test speakers are listed in the first column. Sounds are presented with 10 continuos features, and three features (with one irrelevant) describes the speaker. Test data set contains 463 examples, 67 per speaker, hence the accuracy is measured with about 2% error.

For each test example we marked its five expanding contexts of growing width. The width, considered as Euclidean distance grows linearly. The narrowest context, Context-1, consists of

the neighbors, for which the distance between neighbor and target example is less than half of the maximal normalized attribute value. The next Context-2 includes the examples with the distance less than the maximal normalized attribute value, and so on. The second distance is greatly less than maximal possible distance between examples because of ten-dimensional space. We met the problem with definition of the coefficient of distance growth (one half), similar to mentioned above problems from [Wilson & Martinez, 1997]. So, we used heuristic coefficient. The accuracy of the contexts are presented in Table 1 in the columns marked C1, C2, C3, C4, and C5. C1 corresponds to the narrowest context, C5 is the widest. The number of examples from each context determines context's size, average size for each context is presented in the bottom row in the Table. We see, that lineal grow of context width in 10-dimensional space causes approximately logarithmic growth of context size for Vowel dataset.

The classification is made by the Nearest Neighbor classifier, which uses all examples from the context. This is not usual way of building the Nearest Neighbor classifier, but is the only appropriate way. Really, all nearest neighbors of the target example belong to Context-1; if we will use outer examples from each context then the method will be not the Nearest Neighbor at all and we will loose the examples from inside of the context.

It is clear from the Table that sometimes context C1 has the best accuracy, but sometimes C2 do, and even C5. Lower accuracy of C1 shows how different is the test speaker from the nearest training speaker. In real tasks we can not understand which context perform better and have to unite the opinions from all the contexts. But the contexts have not a diverse opinions, so usually robust simple voting performs badly, as presented in the Simple vote column of the Table.

Let us assume that there is a trend in opinions of classifiers build on the contexts. We used a heuristic quality evaluation function *Q*, based on the average size of the context. Quality of the context C1 is 0.05 (ratio to the size is 150), quality of C2 is 0.2 (ratio is 300), quality evaluations for C3, C4 and C5 are 2.5, 4 and 5 respectively (ratio is 50). The estimated quality is equal to zero.

The trend estimated from all five contexts is presented in the Table in the Trend column. Generally, it performs better than simple voting with average accuracy 35.6% versus 28.4%. This result is statistically significant.

We also tried another, more sophisticated extrapolation scheme. We extrapolated opinions of the contexts C1-C3, C1-C4 and C1-C5. Then these opinions for the target example were integrated by voting. The results are presented in the Trends column of the Table and show that this scheme performs slightly worse than simple trend (33.4% versus 35.6%). But this result is not statistically significant.

Table 1. Experimental results

Accuracy (%) of three contexts and their combinations on Vowel dataset								
	Separate contexts					Combinations		
Part of the dataset	C1	C2	C3	C4	C5	Simple vote	Trend	Trends
Speaker 1	47.0	**53.0**	51.5	22.7	10.6	42.4	**50.0**	42.4
Speaker 2	37.9	**42.4**	31.8	12.1	09.1	27.3	**51.5**	47.0
Speaker 3	25.8	36.4	**53.0**	13.6	16.7	**34.8**	**34.8**	25.8
Speaker 4	34.8	18.2	16.7	25.8	**36.4**	**34.8**	22.7	28.8
Speaker 5	**33.3**	25.8	21.2	01.5	10.6	25.8	25.8	**27.3**
Speaker 6	**44.8**	19.4	10.4	0.00	01.5	10.4	**17.9**	**17.9**
Speaker 7	**63.1**	35.4	23.1	15.4	09.2	23.1	**46.2**	44.6
Average	40.9	32.9	29.7	13.0	13.4	28.4	**35.6**	33.4
Average size	7	63	134	189	223			

5. Conclusions

In the paper a notion of context for Data Mining was given. This is the notion of expanding context. The idea of this notion is: if the classifiers in their contexts form a trend of their (possibly incorrect) opinions then the classifier in ideal (best possible) context (virtual classifier) will give the correct (best possible) opinion.

The extrapolation of opinions made in different expanding contexts stands opposite to usual ways of opinion integration by interpolation. We analyzed the two-dimensional context of Data Mining algorithms: context of relevant features and context of relevant examples, and some attempts to treat this context from the literature.

For some datasets extrapolation of expanding contexts works better than widely used voting.

6. References

Blum, A. (1997), Empirical Support for WINNOW and Weighted-Majority Algorithms: Results on a Calendar Scheduling Domain, *Machine Learning* **26**(1), pp. 5-24.

Dash, M, Liu, H. (1997) Feature Selection for Classification. Intelligent Data Analysis, Vol. 1 (3), Elsevier Science, http://www.elsevier.com/locate/ida

Domingos, P. (1997) Context-sensitive feature selection for lazy learners. Artificial Intelligence Review 11, p. 227-253. Special Issue on lazy learning, edited by D. Aha.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1997): Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press

Harries, M., Horn, K. (1996) Learning stable concepts in domains with hidden changes in context. In M. Kubat, G. Widmer, eds., Learning in context-sensitive domains (Workshop Notes). Thirteenth International Conference on Machine Learning, Bari, Italy

Harries, M., Sammut, C. (1998) Extracting Hidden Context. Machine Learning 32(2), Special Issue on Context-Sensitive Learning

Skalak, D. (1997), Prototype Selection for Composite Nearest Neighbor Classifiers, Ph. D. Thesis, Department of Computer Science, University of Massachusetts Amherst, 259 pp.

Terziyan V., Puuronen S., (1997) Multilevel Context Representation Using Semantic Metanetwork, In: Context-97 - Proceedings of International and Interdisciplinary Conference on Modeling and Using Context, Rio de Janeiro, Brazil, Febr. 4-6, 1997, pp. 21-32 (to appear also in the Special Volume of Book in the Kluwer Series).

Terziyan V., Tsymbal A., Puuronen S. (1998) The Decision Support System for Telemedicine Based on Multiple Expertise, *International Journal of Medical Informatics*, Elsevier, V. 49, No.2, 1998, pp. 217-229.

Turney, P. (1996a) The identification of context-sensitive features: a formal definition of context for concept learning. In Proceedings of the Workshop on Learning in Context-Sensitive Domains at the Thirteenth International Conference on Machine Learning ICML-96, Bari, Italy, July 3-6, 1996

Turney, P. (1996a) The management of context-sensitive features: a review of strategies. In Proceedings of the Workshop on Learning in Context-Sensitive Domains at the Thirteenth International Conference on Machine Learning ICML-96, Bari, Italy, July 3-6, 1996

Wilson, D. R., Martinez, T. R. (1997) Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research 6, p. 1-34.