21st IEEE International Symposium on Computer-Based Medical Systems

Effectiveness of Local Feature Selection in Ensemble Learning for Prediction of Antimicrobial Resistance

Seppo Puuronen
Dept. CS & ISs,
University of Jyväskylä,
Finland
sepi@cs.jyu.fi

Mykola Pechenizkiy
Dept. of CS, Eindhoven
University of Technology,
The Netherlands
m.pechenizkiy@tue.nl

Alexey Tsymbal
Corporate Technology,
Siemens AG, Erlangen,
Germany
alexey.tsymbal@siemens.com

Abstract

In the real world concepts are often not stable but change over time. A typical example of this in the biomedical context is antibiotic resistance, where pathogen sensitivity may change over time as pathogen strains develop resistance to antibiotics that were previously effective. This problem, known as concept drift (CD), complicates the task of learning a robust model. Different Ensemble Learning (EL) approaches (that instead of learning a single classifier try to learn and maintain a set of classifiers over time) have been shown to perform reasonably well in the presence of concept drift. In this paper we study how much local feature selection (FS) can improve ensemble performance for data with concept drift. Our results show that FS may improve the performance of different EL strategies, yet being more important for EL with static integration of classifiers like (weighted) voting. Further, the improvement of EL due to FS can be explained by its effect on the accuracy and diversity of base classifiers. The results also provide some additional evidence that diversity can be better utilized with the dynamic integration of classifiers.

1. Introduction

Nosocomial infections and antimicrobial resistance (AR) are highly important problems that impact the morbidity and mortality of hospitalized patients as well as the cost of care. It is known that 3 to 40 percent of patients admitted to hospital acquire an infection during their stay, and that the risk for hospital-acquired infection, or *nosocomial infection,* has risen steadily in the recent decades [3].

To fight nosocomial infections, at first a microbiological investigation is normally conducted. In this investigation pathogens are first isolated, and, for each isolated bacterium, an antibiogram is then built (which represents bacterium's resistance to a series of antibiotics). The user of the test system can define the set of antibiotics used to test bacterial resistance. The result of the test is presented as an antibiogram, which is a vector of couples "antibiotic/resistance". The information included in the antibiogram is used to prescribe an antibiotic with a desired level of resistance for the isolated pathogen.

Today, virtually all important bacterial infections throughout the world are becoming resistant. Infectious microorganisms are developing resistance faster than scientists can create new drugs. This problem is called *antibiotic resistance*, also known as antimicrobial resistance or drug resistance [17].

The problem of understanding antibiotic resistance is in direct connection with the problem known in machine learning and data mining as *concept drift (CD)* [18, 21] where changes in the hidden context (those "not visible" to the learning process) can induce more or less radical changes in the target concept.

Ensemble learning (EL) is among the most popular and effective approaches to handle CD. In EL a set of concept descriptions (base classifiers) built over different time intervals is maintained, and the predictions of those base classifiers are combined using a form of voting, or the most supported classification is selected [11, 15, 16, 20]. However, there is a problem with majority of current ensemble approaches; they are not able to deal with local CD, which is a common phenomenon in real-world data. For example, only a particular bacterium may develop resistance to certain antibiotics, while the resistance of the others can remain the same; or a particular bacterium can change resistance depending on the season.

EL can be enchanced e.g. replacing the combination method. One approach is to use an instance-level combination of base classifiers' classifications using dynamic integration. In dynamic integration, each base classifier receives a weight for its vote and this weight is proportional to base classifier's local accuracy in the neighbourhood of the current instance, instead of using global classification accuracy as in static weighted voting.

The idea to use dynamic integration to handle (local) CD was introduced in [19] with experiments

1063-7125/08 $25.00 © 2008 IEEE
DOI 10.1109/CBMS.2008.22

focusing on the same problem of antibiotic resistance in nosocomial infections. The encouraging results motivated further development of the idea for better handling of CD in predicting AR.

In this work our focus is on studying how much (local) FS can improve ensemble performance in the presence of concept drift. The rest of the paper is organized as follows. In Section 2 we consider the problem of handling CD, describe related work and the focus of this study. In Section 3 we describe the dataset used and the experimental setup. In Section 4 the main results of our experiments are presented. Finally, in Section 5 we briefly conclude with a summary and present some directions for further research.

2. Handling concept drift: related work and our approach

2.1. The problem of handling concept drift

A difficult problem with learning in many real-world domains is that the concept of interest may depend on some hidden context, not given explicitly in the form of predictive features. Changes in the hidden context can induce more or less radical changes in the target concept, which is generally known as CD [21]. A typical example of this in the context of antibiotic resistance is the pathogen sensitivity change over time as new pathogen strains develop resistance to antibiotics that were earlier effective. An effective learner should be able to track such changes and quickly adapt to them. Kukar [13] states that even in most strictly controlled environments some unexpected changes may happen e.g. due to failure and replacement of some medical equipment, or due to changes in personnel, causing the necessity to change the model.

There are a few strategies to detect and handle CD, see e.g. [9, 18] for an overview. Our focus is on EL approaches to address this problem. EL maintains a set of concept descriptions as an ensemble of base classifiers and their classifications are usually combined using a form of voting, or the most supported classification is selected. Street and Kim [16] and Wang *et al.* [20] suggest that simply dividing the learning data into sequential blocks of fixed size and building an ensemble on them may be effective for handling CD. Stanley [15] and Kolter and Maloof [12] build ensembles of incremental learners in an online setting, starting to learn new base classifiers after fixed intervals, while continuing to update the existing ones. All incremental ensemble approaches use current learning data and some criteria to dynamically delete, reactivate, or create new base classifiers of the ensemble.

At present, the most common combining approach for handling CD is weighted voting, where each base classifier receives a weight proportional to its relevance to the current concept [11, 15, 16]. In weighted voting, lower weights can be assigned to base classifiers simply because their global accuracy decreases, even if their local accuracy in the stable parts of data remains high.

In the real world, CD may be local; for example only a particular bacterium may develop resistance to certain antibiotics, while resistance of the others could remain the same. In the case of local CD, base classifiers should not be discarded (or lower weights assigned to them) simply because their global accuracy decreases. In [19] we proposed to apply dynamic integration of classifiers at an instance level, according to the local accuracies, in order to handle local CD for predicting AR in nosocomial infections. Also two synthetic data sets, representing simulated gradual and abrupt CDs respectively, were considered in the experimental study.

2.2. Local FS for handling CD

In this section we briefly overview some related work on local FS for handling CD and discuss the focus of our study.

The task of identifying relevant features has mostly been tried to be solved globally by covering the whole domain space with a single subset of features. However, it was shown that the feature space is often heterogeneous, where the features that are important for learning are different in different contexts (regions of the feature space) [1].

There are two main approaches to solve the above problem [2]. First, a more complex global model can be generated to include the necessary information, for example by adding extra artificial features defining the context. Second, the data mining problem can be divided into subproblems and the solution of the whole classification task can be guided by the heterogeneity of the feature space. This decomposition approach is potentially highly beneficial, because most widely used data mining techniques rely on measures that are computed over all the features and all the examples at hand, and are inevitably diffused by an averaging effect over the entire problem [2]. Our focus in this paper is on the second approach trying to fit a simple model to local regions instead of trying to build a single global model for the whole problem space.

In the context of a changing environment, it was also shown that features may have completely different importance over time, thus motivating context-specific feature selection, too. For example, Fawcett [7] demonstrates it for the problem of spam detection, and in [5] a periodic feature reselection strategy to improve the performance of a spam filter is presented.

In this paper we study how ensemble learning in changing environments can be improved with periodic context-specific feature selection. We extend the experimental setting used in [19] with a few common feature selection techniques and study their effect on the learning performance, for the same antibiotic resistance data.

3. The AR dataset and experimental setup

The *data* for our analysis were collected in the N.N. Burdenko Institute of Neurosurgery, Moscow, over the years 2002-2004, using a bacterial analyzer Vitek-60 (developed by bioMérieux, *www.biomerieux.com*).

Each instance of the data set represents one sensitivity test and contains the following features: pathogen that is isolated during the bacterial identification analysis, antibiotic that is used in the sensitivity test and the result of the sensitivity test itself (sensitive *S*, resistant *R*, or intermediate *I*), obtained according to the guidelines of National Committee for Clinical Laboratory Standards (NCCLS) [8]. The information about sensitivity analysis is related to a patient, his/her demographical data (sex, age) and hospitalization in the institute (main department, days spent in ICU, days spent in the hospital before the test, etc.).

Each instance of microbiological test in the data set corresponds to a single specimen that may be blood, cerebrospinal fluid (liquor), urine, etc. In this study we focus on the analysis of meningitis cases only, and the specimen is liquor. For the purposes of our analysis we picked up all 4430 instances of sensitivity tests of meningitis cases during the years 2002-2004.

After a discussion with medical experts, we have formed new binary features for antibiotics and pathogens corresponding to their tree-like categorization. For antibiotics the 35 original features were grouped into 4 major categories, and for pathogens the 16 original features were categorized into 7 major groups.

Each instance included 34 features that contained information corresponding to a single sensitivity test augmented with data concerning the antibiotic used, the isolated pathogen, clinical characteristics of the patient and his/her demographics.

This data set was considered earlier in [14] where it was experimented with using different techniques including various inductive learning and dimensionality reduction algorithms and the principle of natural clustering. Besides, it was shown that there is a significant level of CD pertinent to this domain. A few interesting findings were discovered, including seasonal context recurring with winter and spring models, which corresponds to yearly spring and winter infection outbreaks.

The *experimental setup* was as follows. We used an implementation based on the machine learning library WEKA 3.4.2 [22]. Default settings were always used in the WEKA learning algorithms and feature selection techniques.

In our analysis we consider the classification problem aimed to predict the sensitivity of a pathogen to an antibiotic based on data about the antibiotic, the isolated pathogen, and the demographic and clinical features of the patient.

To build ensembles, we divide the data into blocks ("chunks") corresponding to a certain time interval. We use a sliding window approach, and thus, when the window shift is less than the size of the window, the data blocks are not mutually exclusive. With all the ensembles we used the simple "replace the loser" ensemble pruning strategy. With this strategy, if the ensemble size is greater than or equal to 25, the worst classifier is replaced with a new one trained on the most recent data.

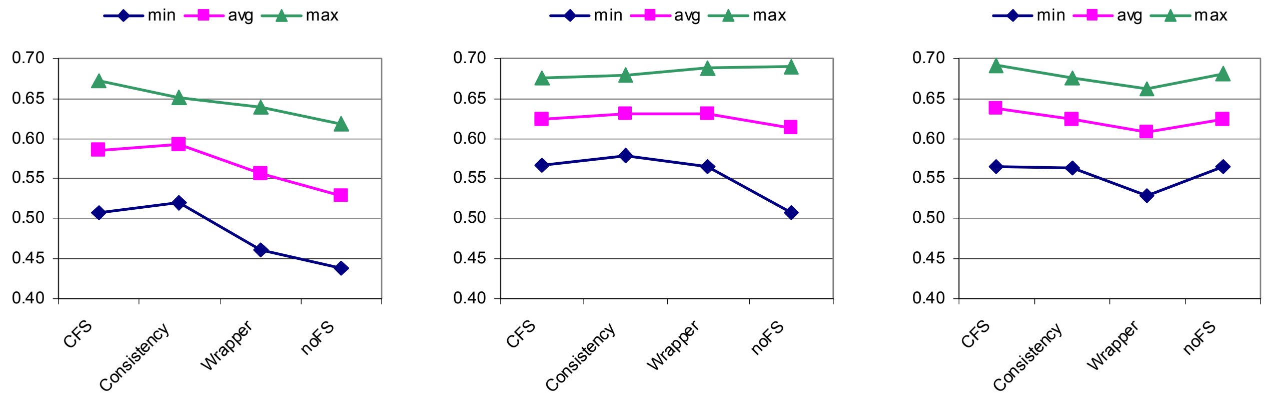
So-called progressive evaluation was used in order to evaluate different integration techniques. The most recent window before the current test block was used to calculate the validation estimates for dynamic integration and weighted voting. In order to avoid overly optimistic validation estimates for the last classifier in the ensemble, which is built on data from the current validation block, 10-fold cross validation was used.

The base classifiers were formed using three commonly used learning algorithms available in WEKA; Naïve Bayes ("NB"), C4.5 decision tree ("J48"), and k-NN ("IBk", *k*=11). For feature selection in the ensembles we consider three feature subset evaluators; the correlation-based feature selection ("CFS") [10], consistency-based feature selection ("Consistency") [4], and the more orthodox accuracy-based feature selection ("Wrapper"). To guide the search over the candidate feature subsets we use the WEKA's Genetic Search, with the default parameters.

4. Main results

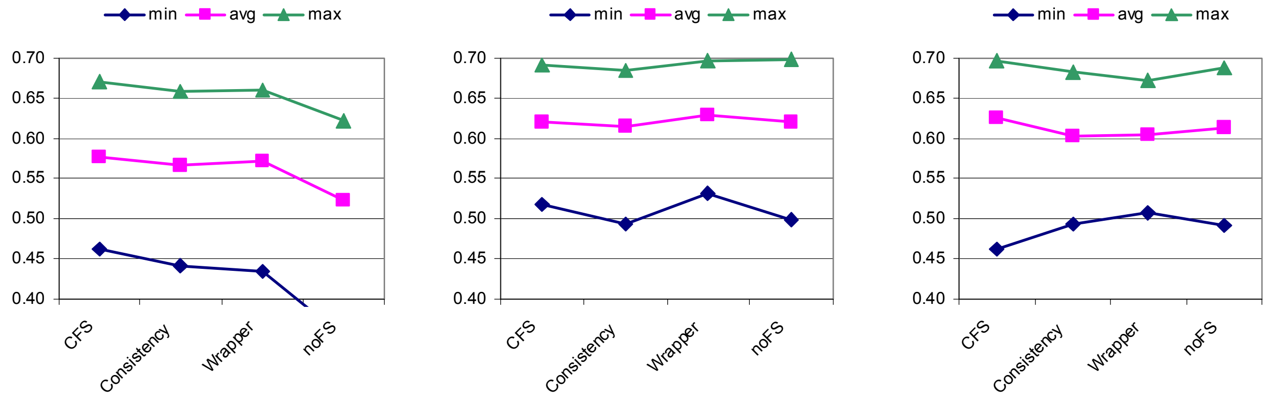
In this section we present a concise summary of the most important findings related to the impact of local FS for EL. These include the average accuracy results for base classifiers with different FS strategies (Figures 1 and 2), the corresponding diversity results (Figure 3), and the corresponding accuracy results for ensembles with different integration strategies; static and dynamic (Figures 4 and 5).

The results are presented for two scenarios; with non-overlapping chunks (window size is equal to shift size and equals to 3 months) and with overlapping chunks (with window size of 3 months and shift size 1 month).



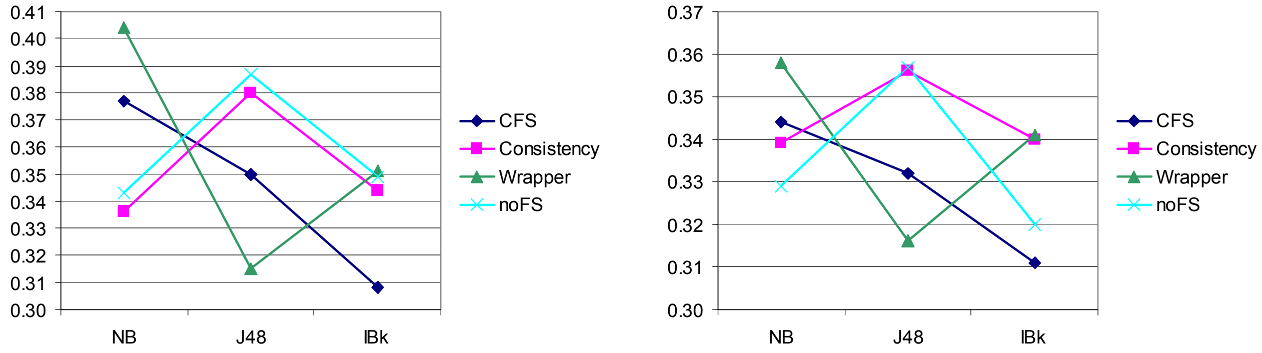
NB J48 IBk

Figure 1. Minimal, average, and maximal accuracies of base classifiers (non-overlapping chunks)



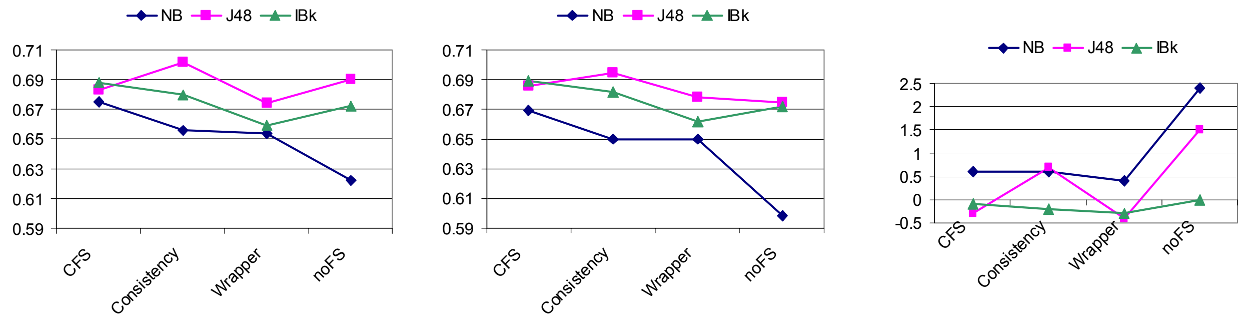
NB J48 IBk

Figure 2. Minimal, average, and maximal accuracies of base classifiers (overlapping chunks)



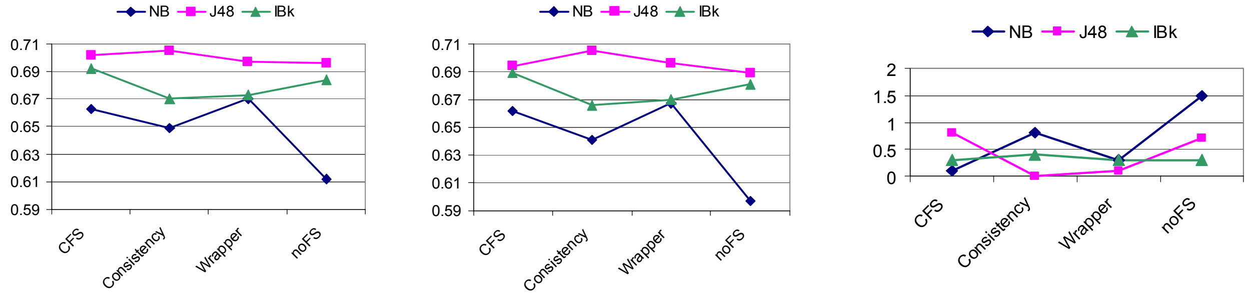
Non-overlapping chunks Overlapping chunks

Figure 3. Diversity of base classifiers in ensembles



DV WV DV – WV

Figure 4. Ensemble accuracies with dynamic (DV) and static (WV) integration of classifiers, and difference in their performance (DV – WV) (non-overlapping chunks)



DV WV DV – WV

Figure 5. Ensemble accuracies with dynamic (DV) and static (WV) integration of classifiers, and difference in their performance (DV – WV) (overlapping chunks)

It can be seen from Figure 1 that average (also *min* and *max*) accuracies of NB base classifiers are higher with all applied FS (*CFS*, *Consistency* and *Wrapper*) than without FS (*noFS*). With *J48* all applied FS results in at least in as high average and minimal accuracies but a bit lower maximal accuracy than without FS. With *IBk* only *CFS* increases a bit average and maximal accuracies while the other two FS techniques (*Consistency* and *Wrapper*) decrease accuracy a bit or keep it same. Comparing Figures 1 and 2 we can see that the patterns of base classifiers accuracies are similar for *NB* and *IBk* except for *min* accuracies that are at significantly lower level. With *J48* it seems that FS has almost no effect except decreasing the minimal accuracies level.

It can be seen from Figure 3 that with non-overlapping chunks only with *NB* base classifiers diversity increases when *Wrapper* or *CFS* FS is applied. As discussed above, in these situations also the accuracies are higher. On the other hand it is interesting that *Consistency* FS that results in the same diversity as *noFS*, still increases the average accuracy of *NB* (6.5%) and *J48* (2.5%) classifiers. Further with *J48* base classifiers *CFS* and *Wrapper* produce less diverse ensembles but still the average accuracies are at least at the same level.

The diversities of overlapping chunks follow similar patterns of behavior as with the non-overlapping chunks, the main difference being that the diversities are of a lower level. Further analysis of how FS impacts accuracy and diversity of base classifiers suggests the explanation of ensemble performance with different parameters.

In Figures 4 and 5 we can see that while without FS dynamic integration (DIC, DV in figures) outperforms the static integration (SIC, WV in figures), with any FS this difference becomes insignificant. Please note that FS may improve both DIC and SIC, but has a much bigger effect on SIC.

With both DIC and SIC, FS strategies have the same ranking, i.e. wrapper-based FS was beneficial (yet, not better than *Consistency* and *CFS*) only with *NB*, whilst with *J48* and *IBk* it even deteriorated the performance of ensembles in 3 out of 4 combinations. Consistency FS was beneficial with every combination of base classifier and integration method. However, *CFS* was better than *Consistency* for *J48* and *IBk*. Comparing Figures 4 and 5 we can see that the trends are similar for *noFS* vs. FS patterns, although the effects are smaller.

Likely because DIC utilizes the diversity of classifiers better than SIC, FS is more important for

SIC (i.e. average accuracy of base classifiers is more important than diversity for SIC than for DIC).

6. Conclusions and future directions

Predicting AR in nosocomial infections is a recognized important and challenging problem. It has a direct connection to the CD problem that has been under active study in DM/ML areas in the recent decades.

In our previous studies we demonstrated that EL and especially the dynamic integration of classifiers approach can handle CD in AR prediction reasonably well. The focus of this paper was on evaluating the impact of local FS on EL for predicting AR.

Our results demonstrated that FS may improve the performance of different EL strategies, with a more clear effect for EL with static integration of classifiers like (weighted) voting, and that the improvement of EL due to FS can be explained by its effect on accuracy and diversity of base classifiers.

The directions of our future work include the development of new FS techniques for handling (local) CD and application of developed approaches to other biomedical data that is known to contain concept drift.

Acknowledgments: This research is partially supported by the Academy of Finland.

7. References

[1] Apte C., Hong S.J., Hosking J.R.M., Lepre J., Pednault E.P.D., Rosen B.K., Decomposition of heterogeneous classification problems. In: Proc. Advances in Intelligent Data Analysis, IDA-97, Springer, 1997, 17-28.

[2] Atkeson C., A. Moore, S. Schaal, Locally weighted learning. AI Review, 11, 1997, 11-73.

[3] Brossette SE, Sprague AP, Jones WT, Moser SA. A data mining system for infection control surveillance, Methods of Information in Medicine 39 (4-5), 2000, pp. 303-310

[4] Dash M., Liu H., Motoda H., Consistency based feature selection, In: 4th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, PAKDD'00, Springer, 2000, 98 - 109.

[5] Delaney S.J., Cunningham P., Tsymbal A., Coyle L. A case-based technique for tracking concept drift in spam filtering, Knowledge-Based Systems 18 (2-3), Elsevier, 2005, 187-195.

[6] Fan W. Systematic data selection to mine concept-drifting data streams. In: Proc. 10th Int. Conf. on Knowledge Discovery and Data Mining KDD'04, 2004.

[7] Fawcett T., "In vivo" spam filtering: A challenge problem for data mining, In: CoRR 2004.

[8] Ferraro M.J., et al. Methods for dilution antimicrobial susceptibility tests for bacteria that grow aerobically: approved standard; Performance standards for antimicrobial susceptibility testing. Wayne, PA: National Committee for Clinical Laboratory Standarts, NCCLS, 2005. (Documents M7-A6 and M100-S14, *www.nccls.org*).

[9] Gama J. Learning with local drift detection. In: Proc. 2nd Int. Conf. on Advanced Data Mining and Applications, ADMA 2006, LNCS 4093, Springer, 2006, 42-55.

[10] Hall M.A., Correlation-based feature selection for discrete and numeric class machine learning, In: Proc. Int. Conf. on Machine Learning, ICML 2000, Morgan Kaufmann, 2000, 359 - 366.

[11] Klinkenberg R. Learning drifting concepts: example selection vs. example weighting, Intelligent Data Analysis, Special Issue on Incremental Learning Systems Capable of Dealing with Concept Drift, 8 (3), 2004.

[12] Kolter J.Z., Maloof M.A. Dynamic weighted majority: a new ensemble method for tracking concept drift. In: 3rd Int. Conf. on Data Mining ICDM'03, IEEE CS Press, 2003, 123-130.

[13] Kukar M. Drifting concepts as hidden factors in clinical studies. In: Proc. 9th Conf. on Artificial Intelligence in Medicine in Europe, AIME'03, Springer, 2003, 355-364.

[14] Pechenizkiy M., Tsymbal A., Puuronen S., Shifrin M., Alexandrova I. Knowledge discovery from microbiology data: many-sided analysis of antibiotic resistance in nosocomial infections. In: 3rd Int. Conf. on Professional Knowledge Management: Experience and Visions (WM05), Springer, LNAI 3782, 2005, 360-372.

[15] Stanley K.O. Learning concept drift with a committee of decision trees, Tech. Report UT-AI-TR-03-302, Dept of Computer Science, Univ. of Texas at Austin, USA, 2003.

[16] Street W., Kim Y. A streaming ensemble algorithm (SEA) for large-scale classification. In: 7th Int. Conf. on Knowledge Discovery KDD'01, 2001, 377-382.

[17] The Problem of Antibiotic Resistance, NIAID Fact Sheet. National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, U.S. Dept of Health and Human Services, USA, April 2006 (available at *www.niaid.nih.gov/factsheets/antimicro.htm*).

[18] Tsymbal A. The problem of concept drift: definitions and related work, Tech. Report TCD-CS-2004-15, Department of Computer Science, TCD, Ireland, 2004.

[19] Tsymbal A., M. Pechenizkiy, P. Cunningham, S. Puuronen, Handling local concept drift with dynamic integration of classifiers: domain of antibiotic resistance in nosocomial infections. In: Proc. 19th IEEE Int. Symp. on Computer-Based Medical Systems CBMS'06, IEEE CS Press, 2006.

[20] Wang H., Fan W., Yu P.S., Han J. Mining concept-drifting data streams using ensemble classifiers. In: Proc. 9th Int. Conf. on Knowledge Discovery and Data Mining KDD'03, 2003, 226-235.

[21] Widmer G., Kubat M. Learning in the presence of concept drift and hidden contexts, Machine Learning, 23 (1), 1996, 69-101.

[22] Witten I., Frank E. Data Mining: Practical Machine Learning Tools With Java Implementations, San Francisco: Morgan Kaufmann, 2000.