2011 11th IEEE International Conference on Data Mining Workshops

Interpretability of Sudden Concept Drift in Medical Informatics Domain

Gregor Stiglic, Peter Kokol
Faculty of Health Sciences
University of Maribor
Maribor, Slovenia
{gregor.stiglic, kokol}@uni-mb.si

Abstract— **Concept drift is usually met in rapidly changing environments, especially in sequential data classification, where different types of concept drift occur on regular basis. This paper presents an approach to dynamic visualization of sequential data characteristics aiming to improve the comprehensibility of concept drifts that result in significant change of classification performance. The proposed approach is applied to sequential multi-label hospital discharge dataset containing diagnosis information for more than two million patients. Our experimental results demonstrate visualization of the anomalies in diagnosis coding through time that can explain the differences in sudden changes of class distribution or classification performance.**

Keywords- concept drift; multi-label classification; interpretability of classifiers

I. INTRODUCTION

We live in a rapidly changing environment where more and more data is stored in huge data repositories on daily basis. Consequently it is necessary to update data mining systems with a new knowledge that is available through the incoming data. The changes in data distribution or in the concept of the predicted class are also known as the concept drift [1]. As described by Tsymbal there are two types of concept drift: sudden or abrupt and gradual concept drift. Some authors also include recurring contexts in a group of concept drifts [2, 3]. There are numerous papers on detection and adaptation of classifiers to different types of concept drifts in data. However, apart from the work by Pratt and Tschapek [4], we are not aware of any research work in the field of concept drift visualization or visual interpretation. Some research has been done in the field of multivariate streams visualization, presenting techniques that could be applied to concept drift detection, but were initially created for visual observation and exploration of data stream characteristics [5, 6, 7]. Most of them are based on clustering techniques. There are also approaches like the one by Wei et al. [8] aiming to detect anomalies in time series. Such approaches could also be useful for concept drift detection and interpretation, but they are mostly applied and specialized to univariate time series signals.

Our work demonstrates an approach to concept drift visualization and interpretation using dynamic charts also known as motion charts [9]. Motion chart allows an efficient and interactive visualization of multivariate longitudinal data. Therefore it is well suited for visualization of data characteristics that change over time. This paper presents a real-world example from medical informatics to demonstrate the interpretation of concept drift. Two measures of co-morbidity along with a simple statistical test are used to visualize the relation between descriptive and target attributes. Existence of a concept drift is demonstrated by a static and dynamic ensemble of Naïve Bayes classifiers on the problem of diagnosis presence classification. In Healthcare Information Systems (HIS) it is of high importance to allow the most efficient way of entering the diagnosis and procedure codes for the end user. Therefore the developers of HIS aim to minimize the time spent documenting patient information and the number of human errors [10]. This can be done by prediction on which diagnosis and procedure codes should be offered or automatically entered in HIS based on the already present diagnosis codes for each hospitalization event.

The paper is organized as follows. In section II we present the data that is used in the study. Section III presents two human disease networks related measures and is followed by section IV, where we describe experimental settings. The results in section V demonstrate the effectiveness of the proposed visualization technique. Conclusions and further work plans are presented in section VI.

II. HOSPITAL DISCHARGE DATA

This study uses medical domain dataset to demonstrate the appropriateness of motion chart visualization technique for interpretation of concept drift as a consequence of significant changes in data distribution. This section introduces one of the largest freely available datasets in medical informatics - i.e. National Hospital Discharge Survey (NHDS) data [11]. The dataset contains hospital discharge records for approximately 1% of US hospitals. Only short-stay hospitals (hospitals with an average length of stay for all hospitalization events of less than 30 days) with at least six beds or those whose specialty is general (medical or surgical) or children's general were included in the survey. In our study we used data from 10 consecutive years from 2000 to 2009 that were joined in a single dataset containing 3,106,176 hospitalization events. Removing records for patients aged less than 18 years old reduced the number of records to 2,509,113.

Each NHDS record contains the personal characteristics of the patient, including birth date or age, gender, race, and

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.104

marital status; administrative information, including length of stay, and discharge status; medical information, including diagnoses, surgical and nonsurgical procedures. Each hospital discharge record can have up to 7 diagnoses with an additional admitting diagnosis, where available. Age, gender and all diagnosis codes were the only attributes that were used in our study. The output of the predictive model was one of the diagnosis codes selected by user, while age, gender and the remaining diagnosis codes were used as the input variables. The International Classification of Diseases, 9th Revision, Clinical Modification, or ICD-9-CM was used for coding diagnoses. ICD-9-CM coding uses taxonomy of five-digit codes, where the first three digits represent the general diagnosis and are followed by two additional digits describing a more detailed subgroup of the general diagnosis. Using very detailed five-digit codes would result in high number of diagnoses with extremely low support even with over two million hospitalization events in the dataset. Therefore all five digit codes were collapsed to more general three digit codes that still represent a wide variety of medical conditions (1188 binary attributes).

From the concept drift perspective it is important to note that NHDS data contains information on month of discharge for each record. For classical data stream analysis we would need more accurate information in form of exact date of discharge. However, due to privacy concerns, only the month of discharge is available. Therefore, we grouped records by months and obtained 120 sequential subsets. On average there are 20,909 hospital discharge events available in each month. The number of records is approximately 50% lower in 2008 and 2009 due to reduced funding of NHDS project.

III. Human Disease Networks

To improve the interpretability of concept drift in hospital discharge domain, we use two basic concepts that are usually used when constructing human disease networks: morbidity, representing the support for a single diagnosis in the given population; and co-morbidity, the support for co-occurrence of two diseases. In our experiments we use two co-morbidity measures (relative risk and phi-correlation) from a human disease network study by Hidalgo et al. [12].

The relative risk measure connecting diseases i and j, can be calculated as follows:

$$RR_{ij} = \frac{C_{ij}N}{M_i M_j} \quad (1)$$

where C_{ij} is co-morbidity of two compared diseases, M is morbidity of a single disease and N represents the number of all patients.

Relative risk measure is intrinsically biased towards overestimation of relationships between rare diseases and underestimates the co-morbidity of more frequent diseases. This bias can be reduced by introduction of a ϕ-correlation measure, defined as:

$$\phi_{ij} = \frac{C_{ij}N - M_i M_j}{\sqrt{M_i M_j (N - M_i)(N - M_j)}} \quad (2)$$

Both measures are used as attributes for visualization using motion chart and are calculated for each visualized disease code w.r.t. the class attribute disease code. The same approach to measure the relatedness between different attributes can be used for any multi-label dataset and is not restricted to medical informatics domain.

IV. Experimental Settings

To demonstrate the existence of the concept drift in hospital discharge data, we build two classifiers using Weka [13] machine learning environment. Based on their characteristics we call them Static and Dynamic Ensemble of Naïve Bayes Classifiers. The static classifier is trained only on the first 12 months of data and does not change after that. The dynamic classifier is updated with each new batch of incoming patients – i.e. once a month. We use 25 Naïve Bayes classifiers built using spread subsample filter from Weka that takes care for balanced instance sampling from the majority class. For each minority class sample a random sample is chosen from the majority class. Updateable Naïve Bayes classifier is used to allow adaptations for each new batch of data. Majority voting is used in the evaluation phase.

A binary classification for different target classes (i.e. ICD9-CM diagnosis codes) is performed to observe the performance metrics over 119 months. We use two performance metrics that are frequently used in multi-label classification. Area Under ROC Curve (AUC) is based on true positive rate and false positive rate of a classifier and it is known that AUC is a very effective measure for highly imbalanced class attributes [14]. In our case we are especially interested in a single class – i.e. presence of a disease. Therefore we should observe precision and recall of the classifier that are integrated in the F-measure metric [15]. There are also other metrics that focus on highly unbalanced classes or data with extremely small number of predicted class samples, but since this paper focuses on visualization, we only used two of them. Our own implementation of a special case in prequential performance evaluation was used by adapting the evaluation classes in Weka environment. In contrast to typical prequential performance estimation [16] where each sample is first evaluated and then used in training set, we had to adapt to batches of consequential samples. Therefore each individual batch of records was first used for testing before it was used for training, and performance metrics were incrementally updated.

Visualization using motion charts in this study includes the following information for attributes: ϕ-correlation and relative risk measures w.r.t. the attribute class, one out of the 19 possible groups for ICD-9 disease groups, chi-square value and significance (p-value), morbidity of selected disease. Chi-square test was chosen as a simple and robust identifier of the attribute discriminative power. Including the time dimension, we can visualize up to five values. One can interactively change the information to be displayed on x and y-axes, as color and size of the bubbles. To avoid too many objects in the motion chart, we apply filtering of disease codes based on their maximal frequency. All disease codes

with a frequency of over 100 in any time point were selected for visualization.

V. Case Study Results

In our experiments we demonstrate the efficiency of motion charts for visualization of attribute characteristics during the concept drift. Experimental observation of classification performance on different target classes shows that there are different types of concept drifts present in NHDS data – i.e. gradual, sudden and recurring. In this paper we demonstrate the visualization of a sudden concept drift that occurs in classification of diagnosis code D585 (Chronic kidney disease). Figure 1 demonstrates the classification performance for the static and dynamic ensemble over 119 months. The moment of concept drift can be clearly seen at the end of the year 2005 where the classification performance of the dynamic ensemble suddenly improves.

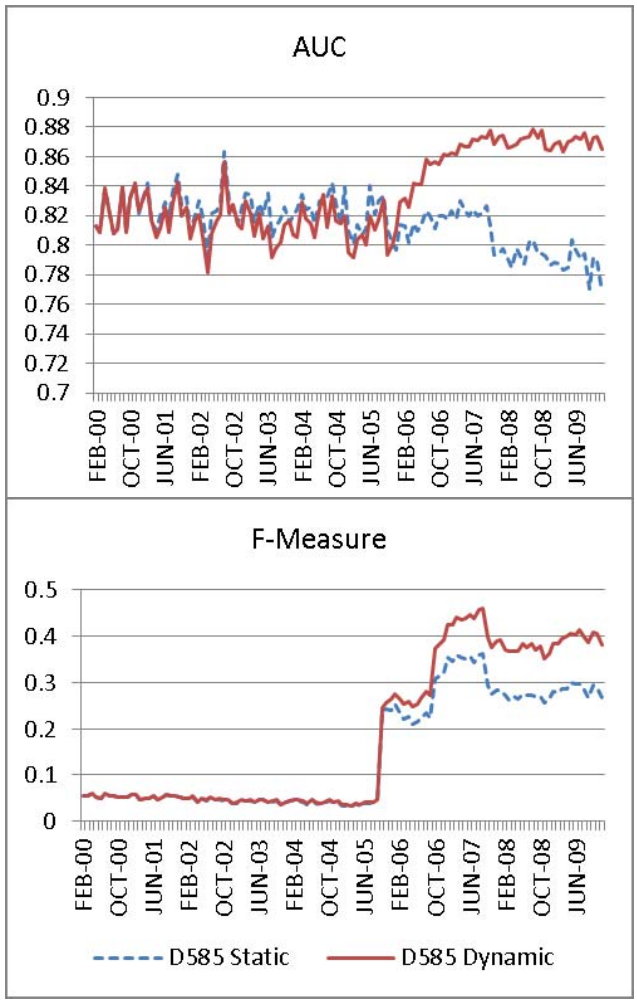

Figure 1. Classification performance for static and dynamic ensemble of classifiers.

The same class attribute is used for visualization using motion charts. Due to simplicity and ease of integration we choose the motion chart implementation by Google. Our visualization of attribute characteristics for classification of D585 diagnosis is available at http://ri.fzv.uni-mb.si/icdm11. Two screenshots of the motion chart visualization can be found in Figure 2 and Figure 3. The first figure displays the characteristics of selected attributes approximately one month prior to the concept drift. The differences the occurred after the concept drift can be observed by comparing the first figure to the second one that was taken approximately one month after the concept drift occurred.

From the dynamics of the changes in ϕ-correlation and relative risk there are two attributes that stand out towards the end of 2005 – D403 and D404. Both diagnoses represent similar conditions related to hypertension and chronic kidney disease (Table I). It takes some further examination to find the reasons behind the sudden changes, especially in correlation between D404 and our class attribute D585. Examining the documentation supporting the NHDS data it is possible to track which codes changed from year to year. It has to be noted that changes in ICD9-CM codes become effective on October 1 of the calendar year. The reason of sudden change in classification performance can therefore be explained through changes of the title and description of D403 and D404. For example, D403 was titled "Hypertensive renal disease" prior to 2005 and "Hypertensive chronic kidney disease" from 2005 on. It is interesting to note that a detailed description mentioned the following "D403 includes any condition classifiable to 585 with any condition classifiable to 401" even before the new code name was introduced. It can be inferred that the new name brought some more attention in 2005 and users of HIS became more aware of this code in hypertension related chronic kidney disease cases. The same explanation can be found for code D404. It should be noted that some information is lost by using only three digit instead of five digit codes. But on the other hand, using the full ICD-9 codes would mean that there would be an extremely small number of samples in most groups that would cause problems in significance testing. The problem of changes in the coding system and issues that this problem can cause is not new to healthcare informatics community. More similar use-cases can be found in a study by Cruz-Correia et al. [17].

TABLE I. Disease codes used in visualizations

Diagnosis code	Description
D401	Essential hypertension
D403	Hypertensive chronic kidney disease
D404	Hypertensive heart and chronic kidney disease
D428	Heart failure
D583	Nephritis and nephropathy, not specified as acute or chronic
D584	Acute renal failure
D585	Chronic kidney disease (CKD)
D588	Disorders resulting from impaired renal function

VI. Conclusions

In many concept drift detection algorithms it is problematic to set a pre-defined thresholds for upper and lower classification performance boundaries to detect if concept drift has actually occurred. In such cases

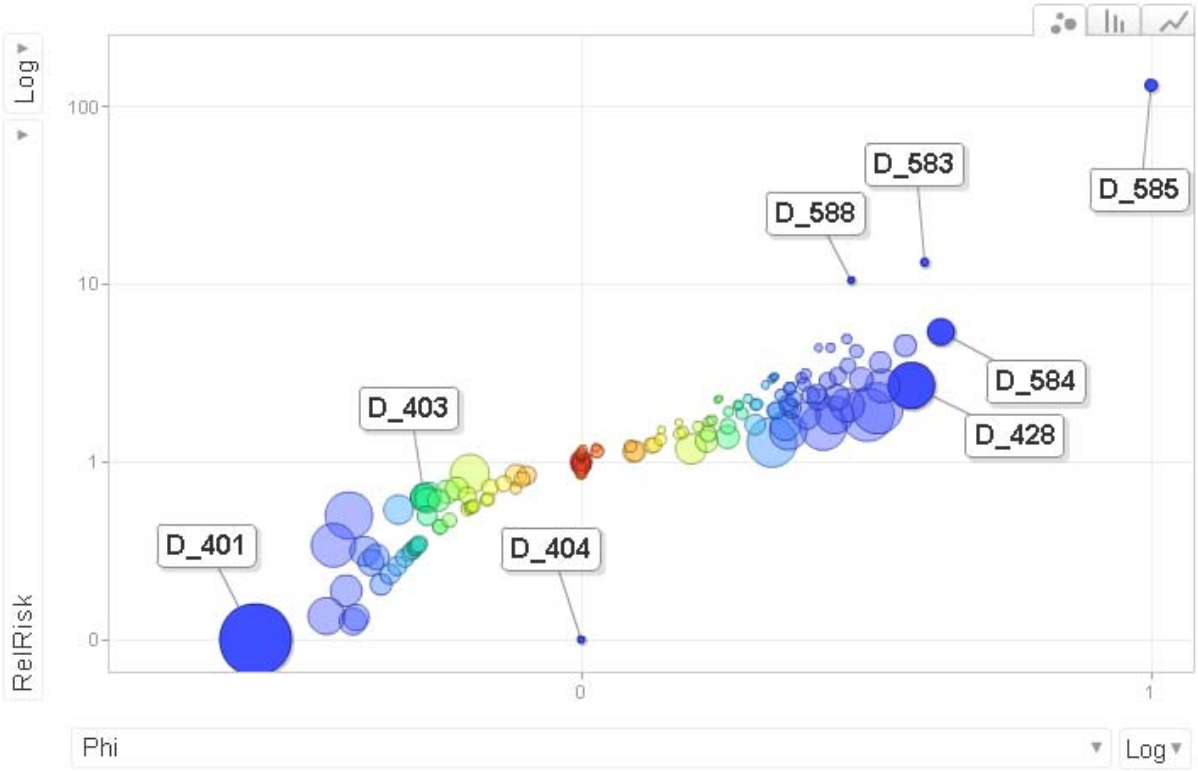

Figure 2. Visualization of calculated attribute characteristics for target class D585 in August 2005. Color ranging from blue (low) to red (high) represents p-value of the chi-square test. Size of the bubble represents the prevalence of the disease in the population and ranges from 0 to 10,000.

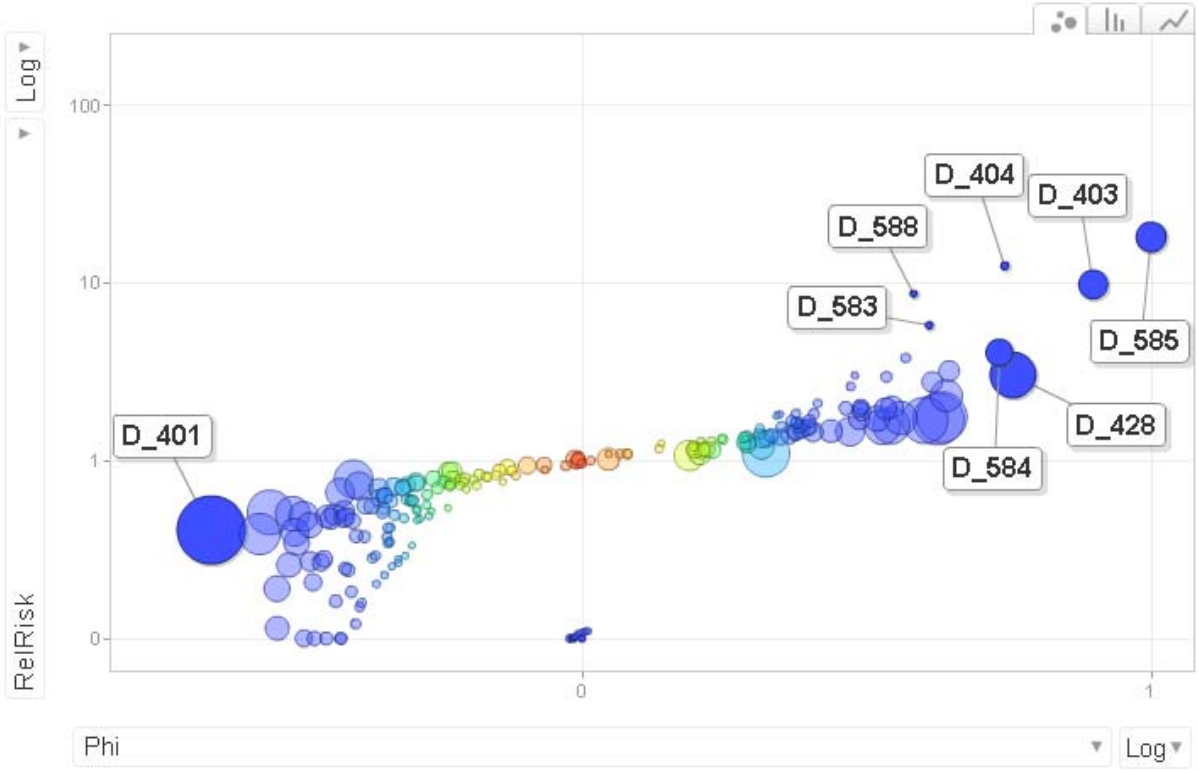

Figure 3. Visualization of attribute characteristics for target class D585 in October 2005

visualization techniques can be used to help experts in setting of the optimal thresholds for effective concept drift detection. Another way of using concept drift visualization is to improve the ability to detect abnormal changes of attribute value distributions or correlations between the observed attribute and the class attribute. This paper proposes an approach using motion charts that are able to dynamically display multiple attribute values. In our case specific characteristics for selected attributes are calculated mainly focusing on correlations and discriminative power w.r.t. the class attribute.

This study demonstrates the effectiveness of visualization technique in selection of attributes with abnormal dynamics that can be further examined for a detailed interpretation of concept drift. Further improvements are possible by including the classification related metrics like precision-recall or sensitivity-specificity visualization that can include additional information visualized in form of different colors and size of bubbles. Additionally, it would also be possible to use the proposed visualization approach for educational purposes when an insight into classification performance of single classifiers in an ensemble is needed.

REFERENCES

[1] A. Tsymbal, "The problem of concept drift: Definitions and related work," Technical Report TCD-CS-2004-15, Trinity College Dublin, 2004.

[2] G. Widmer, and M. Kubat, "Learning in the presense of concept drift and hidden contexts," Machine Learning, vol. 23, no. 1, pp. 69-101, 1996.

[3] M.B. Harries, C. Sammut, K. Horn, "Extracting hidden context," Machine Learning, vol. 32, no. 2, pp. 101-126, 1998.

[4] K.B. Pratt, and G. Tschapek, "Visualizing concept drift," Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 735-740, 2003.

[5] T. Munzner and L. Berry, "Dynamic Exploration of Time Series Datasets Across Aggregation Levels," Proc. IEEE Symposium on Information Visualization, pp. 215-216, 2004.

[6] H. Hochheiser and B. Shneiderman, "Dynamic query tools for time series data sets: timebox widgets for interactive exploration," Information Visualization, vol. 3, no.1, pp.1-18, 2004.

[7] P. Kranen, H. Kremer, T. Jansen, T. Seidl, A. Bifet, G. Holmes, B. Pfahringer, "Clustering Performance on Evolving Data Streams: Assessing Algorithms and Evaluation Measures within MOA," icdmw, pp.1400-1403, 2010 IEEE International Conference on Data Mining Workshops, 2010.

[8] L. Wei, N. Kumar, V. Lolla, E. Keogh, S. Lonardi, and C. Ratanamahatana, "Assumption-free anomaly detection in time series," In J. Frew, editor, Proceedings of the 17th International Conference on Scientific and Statistical Database Management (SSDBM'05), pp. 237-242, 2005.

[9] J. Al-Aziz, N. Christou, I. Dinov, "SOCR motion charts: An efficient, open-source, interactive and dynamic applet for visualizing longitudinal multivariate data," J Stat Educ, vol. 18, pp. 1–29, 2010.

[10] K.J. O'Malley, K.F. Cook, M.D. Price, K.R. Wildes, J.F. Hurdle, C.M. Ashton, "Measuring diagnoses: ICD code accuracy," Health Serv Res, vol. 40, no.5 Pt 2, pp.1620-39, 2005.

[11] National Center for Health Statistics, National Hospital Discharge Survey (NHDS) data, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland, available at: http://www.cdc.gov/nchs/nhds.htm (2008)

[12] C.A. Hidalgo, N. Blumm, A. Barabási, and N.A. Christakis, "A dynamic network approach for the study of human phenotypes," PLoS Comput Biol, vol. 5, 2009.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, vol. 11, no. 1, 2009.

[14] A.P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145–1159, 1997.

[15] C.J. Van Rijsbergen, "Information Retrieval," 2nd edition. London: Butterworths, 1979.

[16] J. Gama, R. Sebastiao, and P.P. Rodriguez, "Issues in evaluation of stream learning algorithms," In KDD '09, pp. 329-338, 2009.

[17] R. Cruz-Correia, P. Rodriguez, A. Freitas, F. Almeida, R. Chen, "Data quality and integration issues in electronic health records," In V. Hristidis, Information discovery on electronic health records, London: CRC Press, pp. 55-95, 2010.