Master's Thesis
Computer Science
September 2012

Concept Drift in Surgery Prediction

Beyene, Ayne and Welemariam, Tewelle

School of Computing
Blekinge Institute of Technology
SE – 371 79 Karlskrona
Sweden

This thesis is submitted to the School of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:
Authors:
Ayne Assegahegne Beyene
Address: Lindblomsvagen 98, 37233 Ronneby, Sweden
E-mail: me.ayne@gmail.com
Tewelle Welemariam
Address: Lindblomsvagen 96, 37233 Ronneby, Sweden
E-mail: tewemit@gmail.com

University advisors:
Marie Persson, Ph.D
E-mail: marie.persson@bth.se
School of Computing
Niklas Lavesson, Ph.D
E-mail: niklas.lavesson@bth.se
School of Computing

School of Computing
Blekinge Institute of Technology
SE – 371 79 Karlskrona
Sweden

Internet	: www.bth.se/com
Phone	: +46 455 38 50 00
Fax	: +46 455 38 50 57

Abstract

Context: In healthcare, the decision of patient referral evolves through time because of changes in scientific developments, and clinical practices. Existing decision support systems of patient referral are based on the expert systems approach. This usually requires manual updates when changes in clinical practices occur. Automatically updating the decision support system by identifying and handling so-called concept drift improves the efficiency of healthcare systems. In the state-of-the-art, there are only specific ways of handling concept drift; developing a more generic technique which works regardless of restrictions on how slow, fast, sudden, gradual, local, global, cyclical, noisy or otherwise changes in internal distribution, is still a challenge.

Objectives: An algorithm that handles concept drift in surgery prediction is investigated. Concept drift detection techniques are evaluated to find out a suitable detection technique in the context of surgery prediction. Moreover, a plausible combination of detection and handling algorithms including the proposed algorithm, Trigger Based Ensemble (TBE)+, are evaluated on hospital data.

Method: Experiments are conducted to investigates the impact of concept drift on prediction performance and to reduce concept drift impact. The experiments compare three existing methods (AWE, Active Classifier, Learn++) and the proposed algorithm, Trigger Based Ensemble(TBE). Real-world dataset from orthopedics department of Belkinge hospital and other domain dataset are used in the experiment.

Results: The negative impact of concept drift in surgery prediction is investigated. The relationship between temporal changes in data distribution and surgery prediction concept drift is identified. Furthermore, the proposed algorithm is evaluated and compared with existing handling approaches.

Conclusion: The proposed algorithm, Trigger Based Ensemble (TBE), is capable of detecting the occurrences of concept drifts and to adapt quickly to various changes. The Trigger Based Ensemble algorithm performed comparatively better or sometimes similar to the existing concept drift handling algorithms in the absence of noise. Moreover, the performance of Trigger Based Ensemble is consistent for small and large dataset. The research is of twofold contributions, in that it is improving surgery prediction performance as well as contributing one competitive concept drift handling algorithm to the area of computer science.

Keywords: Concept drift, Concept Drift in Surgery Prediction, Concept Drift Handling Algorithm, Trigger Based Ensemble

Summary of Contribution

In healthcare systems, the decision of patient referral evolves through time because of changes in scientific developments and clinical practices. In the case of elective patient referrals, which are non-emergency cases, general practitioners at the primary cares refers a patient to a hospital. However, since the general practitioners are not experts on all specialties unnecessary patient referrals occurs. Thus, hospitals need patients to be extensively examined at the primary care before the referrals. Similarly, the general practitioners need support on what examinations to conduct before referring a patient. An automated decision support system would improve the efficiency of patient referrals. Existing decision support systems of patient referrals are based on the expert systems approach. This usually requires manual updates when changes in clinical practices occur. Automatically updating the decision support system by identifying and handling so-called concept drift improves the efficiency of healthcare systems. Among other hospital treatments (drug, therapies, etc), surgery is one of the most expensive treatments. It needs improvements in referring a patient for surgery, i.e. surgery prediction. In the state-of-the-art, concept drift in surgery prediction has not been investigated yet. On the other hand, the existing concept drift handling approaches are studied for specific domains and specific drift types. Developing a more generic technique which works regardless of restrictions on how slow, fast, sudden, gradual, local, global, cyclical, noisy or otherwise changes in internal distribution, is still a challenge.

In this research, concept drift in surgery prediction is investigated. A plausible combination of the existing detection, handling and the proposed algorithm are evaluated to find a suitable surgery prediction concept drift handling approach. A number of experiments are conducted to investigate the existence of surgery prediction concept drift, its impact on the performance of surgery prediction and handling the impact. The experiments evaluate and compare four algorithms including the proposed algorithm, Trigger Based Ensemble. Real-world dataset from the orthopedics department of the Belkinge hospital is used in the experiment.

The result of the experiment showed that concept drift exists in surgery prediction, concept drift has a negative impact on the performance of surgery prediction and handling concept drift improves surgery prediction performance. Moreover, the relationship between temporal changes in data distribution and surgery prediction concept drift is identified.

The authors of this thesis conclude that with the help of automated decision support systems, which are capable of handling concept drift in surgery prediction, general practitioners will be able to make patient referrals with better performance. This will have a great contribution in the improvement of healthcare productivity.

Contents

List of Figures

List of Tables

Acknowledgement

We would like to express our deepest and sincere gratitude to our university advisors Dr. Marie Persson and Dr. Niklas Lavesson for their continuous guidance, support, motivation, invaluable comments and immense knowledge. Their guidance was paramount in providing a well rounded experience aligned with our long-term career goals.

We would like to thank Blekinge Hospital for providing us data which played a key role for the thesis success. We would also like to thank the UCI machine learning repository for providing a real-world dataset that helped us achieve a better result.

1 Introduction

Improvements in information management, data warehousing technologies and storage costs have provided medical centers, telecommunication industries, banks, and other service providers with the advantage of collecting and storing large volumes of data. The application of data mining and machine learning techniques on such data has helped the service providers to improve their decision making process. Healthcare is one of these domains that use machine learning and data mining techniques in the analysis of clinical parameters for diagnosis, prediction of the effectiveness of surgical procedures, and discovery of the relationships among clinical and diagnosis data (Magoulas and Prentza, 2001, Khaing, 2011). In addition to the data growth, the healthcare service demand is also increasing.

The healthcare service demands and demographic trends are related variables. Unproportional growth of population, related to birth and death rates, causes increase in healthcare service demand. This effect is visible in the medical and technological improvements to treat patients at increasingly higher ages. For example, statistics on hip-replacement surgery in Sweden, from 1992 to 2003, shows the population aged 85 years and older increased by 32% and the number of total hip-replacements in this age group also increased by 200%[1]. Such increase in life expectancy and service demand is expected to occur in developing countries where 80% of the world's population lives (McKenzie et al., 2011). Due to the unproportional growth of population, the healthcare domain has less professional expertise than demanded. The increase in the service demand has made patient referral an important process. Among other treatments (drug, therapies, etc), surgery is one of the most expensive treatments in secondary care units (Persson and Lavesson, 2009). Hence, with the help of supervised learning an intelligent decision support systems can improve the performance of patient referral at primary care units.

Supervised learning is used to learn the concept of surgery to improve the process of patient referral. One example of the application of supervised learning is the identification of surgical indicators by mining existing hospital data for surgery prediction (Persson and Lavesson, 2009). However, the decision of patient referral evolves through time because of changes in scientific developments, and clinical practices. Consequently, the prediction model will not work properly because the concept it learned before is not valid any more (Alippi et al., 2011). The existing decision support systems for patient referral require manual updates when changes

[1] Molin and Johansson (2005)

in clinical practice occur. Thus, automatically updating the decision support system by handling the so-called concept drift improves the efficiency of healthcare systems. In the state-of-the-art, there are only specific ways of handling concept drift; developing a more generic technique which works regardless of restrictions on how slow, fast, sudden, gradual, local, global, cyclical, noisy or otherwise changes in internal distribution, is a challenge (Elwell and Polikar, 2011). Thus, a learning algorithm that automatically handles concept drift in surgery prediction is investigated.

This research investigates a learning algorithm that handles concept drifts in surgery prediction. The result shows concept drift affects prediction performance negatively. Therefore, the selected existing concept drift handling algorithms and the proposed algorithm, Trigger Based Ensemble (TBE) are evaluated on a real hospital and other domain data. Consequently, TBE performed better or sometimes similar to the existing concept drift handling algorithms in the absence of noise.

1.1 Aim and Scope

The research aimed at investigating a concept drift handling approach to predict surgery without significantly loosing the prediction performance over time. The existing concept drift handling algorithms are identified and categorized based on their handling techniques. Moreover, the available concept drift detection methods are identified and compared to get a better concept drift detection method in surgery prediction.

The scope is limited to investigating a concept drift handling algorithm for surgery prediction. The investigation uses real-world datasets of different domains with different types of concept drift and two synthetic datasets. The real-world datasets are the hip-replacement dataset from the orthopedics department of Blekinge hospital [2] and the poker-hand dataset from the UCI machine learning repository. The synthetic datasets are purposely generated, based on the hip-replacement dataset, to introduce different types of concept drift.

1.2 Outline

The outline of the research is as follows. Chapter 2 presents a thorough discussion of the background including elective surgery, supervised learning, and concept drift. Chapter 3 gives a deeper insight into the related work and current research gap. Chapter 4 presents the research methodology, and its design, used to address the research questions and hypotheses. The existing concept drift detecting and handling algorithms, including the proposed algorithm, are discussed in chapter 5. Chapter 6 describes the experimental design, environment and results. In

[2] Table 6.1 shows the hip-replacement dataset.

chapter 7, a thorough analysis and discussion is made based on the results of the experiment. Finally, Chapter 8 concludes the research with a discussion on the completed work and possible directions for future work.

2 Background

An elective patient referral is a non-emergency case that is commonly submitted by a general practitioner at the primary care. In Sweden, the general practitioner (GP) refers a patient when surgery seems necessary. The referral is assessed by a surgeon specialist usually at the hospital. The patient is then added to a surgery waiting list, depending on medical priority, to meet an appropriate surgeon specialist. Subsequently, the patient meets the surgeon specialist to decide together about the surgery need. If they decide to have surgery then the patient will be added to another surgery waiting list in order to be scheduled for surgery. However, if the surgeon specialist determines that the referral is unnecessary, the patient is sent back to the primary care. The unnecessary referrals result in having many patients on the surgery waiting list when they could be treated at the primary care. Therefore, hospitals need the referred patients to be extensively examined by the GPs. On the other hand, the GPs are not experts on all specialties. Thus, the GPs need support on what examinations to conduct before referring a patient and whether to refer the patient or not. The examinations can give valuable information to make decisions about the patient referral.

The elective patient referral contains information that indicates patients surgery need. The indicators can be used as an input to develop an intelligent decision support system at the primary healthcare care. The intelligent decision support system predicts patients surgery need and assists the GPs in making an efficient decision. Thus, unnecessary patient referrals can be reduced through the usage of intelligent decision support system. Machine learning and data mining techniques are used in supporting decisions with the analysis of clinical parameters for diagnosis, prediction of the effectiveness of surgical procedures, medical tests and surgery prediction (Magoulas and Prentza, 2001, Khaing, 2011, Persson et al., 2010, Persson and Lavesson, 2009).

Data mining is the process of discovering knowledge or structural patterns in data. Some of the techniques used to find structural patterns or knowledge discovery in such data are within the field of machine learning. These techniques take examples as an input. Examples are a set of instances characterized by a predefined set of features called attributes (Witten et al., 2011). Supervised learning is one of the prominent techniques of machine learning used to find structural pattern from a set of given examples (Goodacre et al., 2004, Witten et al., 2011).

2.1 Supervised Learning

A supervised learning is a technique where learning algorithms are provided with a set of inputs along with their corresponding outcomes (Goodacre et al., 2004, Witten et al., 2011). In the classification context, the outcomes are called target classes. The algorithms learn by finding relationships between the features and the target classes. Classification is a supervised learning where learning algorithms are expected to learn a way of classifying unobserved examples based on previous observation. The structural pattern or knowledge learned by the learning scheme is called concept (Witten et al., 2011).

Concept can be a target class or the distribution of an example at a given time with its target classes (Ouyang, 2011, Witten et al., 2011). The learning algorithms train a set of examples with their respective target class to learn concept. The training dataset is called training set. The target class of a training set is compared to the predicted target class to identify the algorithm's accuracy.

The accuracy of classifiers depends on the appropriateness of the algorithms selected and the amount and quality of the training set. However, regardless of the learning algorithms used, a prediction model does not always work in dynamic environments because the concept acquired during the training phase might be changed (Alippi et al., 2011). The medical domain is one of such domains that is dynamic, complex and changes along with the scientific development and clinical practice. Thus, the usage of machine learning techniques to predict surgery need is susceptible to concept drift.

2.2 Concept Drift

Concept drift is a problem in machine learning, where prediction models lose their performance over a period of time (Wang et al., 2011). The models lose their performance as the target class and/or the data distribution of a dataset is changed.(Tsymbal, 2004, Wang et al., 2011, Elwell and Polikar, 2011).

Figure 2.1 depicts a supervised learning model that is built from a certain training set at time, t. The learning model predicts the target class correctly at t and $t + 1$. A test data drawn at time $t + 2$ is a different population and at time $t + 3$ the concept itself is changed. Therefore, the prediction model at time $t + 2$ and $t + 3$ becomes susceptible to make wrong surgery prediction.

There are two types of concept drifts, real and virtual. A real concept drift is a change in the concept of the target class. A virtual concept drift is a change in the data distribution. (Stiglic and Kokol, 2011). In real concept drift, models need replacement as the old concepts become irrelevant, whereas in virtual concept drift models need additional learning as the error of models may no longer be

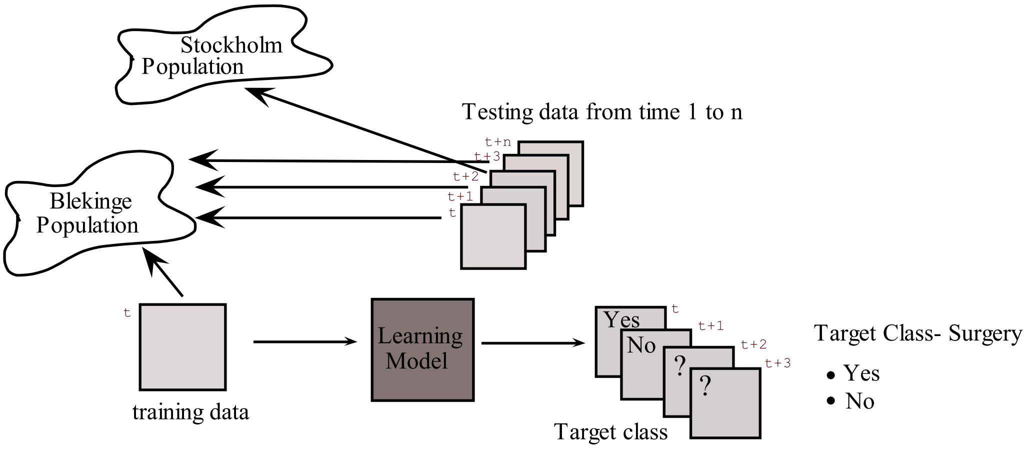


Figure 2.1: Supervised Learning with Concept Drift

acceptable (Tsymbal, 2004, Masud et al., 2010).

Real concept drift can occur due to changes in a hidden context (Tsymbal, 2004, Masud et al., 2010). A Hidden context is insufficient, unknown or unobservable features in a dataset (Widmer and Kubat, 1996). Virtual concept drift can occur while the target concept remains the same (Tsymbal, 2004, Wang et al., 2011). The need to change the current model due to virtual concept drift shows that the learner is provided with additional data (Elwell and Polikar, 2011). In surgery prediction, concept drift occurs due to scientific development, change in clinical practice, and change in data distribution/ pattern (Perner, 2009).

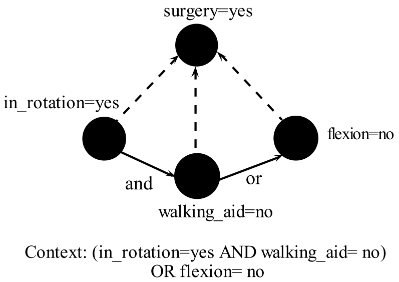
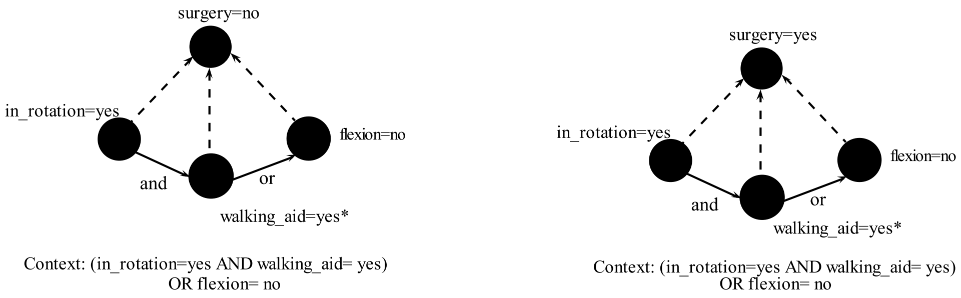


Figure 2.2: Orginal Model Data Observation

Figure 2.2 shows the concept learned by the model from previous data. The probability of the target class depends on the previously observed data. The current value of walking-aid is changed to yes in both Figures 2.3-(a) and 2.3-(b). Figure 2.3-(a) shows the occurrence of both real and virtual concept drifts, i.e., the target class is changed from No to Yes. On the contrary, in Figure 2.2:b, the change does not affect the target class i.e. virtual concept drift.



(a) Real and Virtual Concept Drift

(b) Virtual Concept Drift

Figure 2.3: Concept Drift

There are different types of change in concept drift. The common types are sudden, gradual and recurring. The sudden changes are abrupt when affecting the classification model. For instance, when a specific surgery technique is stopped on a legal basis because the treatment is discovered to be hazardous to the surgery outcome and to the patient's health, sudden change. The gradual changes evolve slowly through time, such as when a specific surgery technique is tried out on a specific health situation and proven to be successful by accident, and hence gradually learned by other surgeons. The recurring changes are hidden contexts that reoccur, either cyclically or in an unordered manner, for instance: when surgery and drugs are competing treatments for a specific problem and drugs along with surgery technique gradually is changed but cyclically replace each other. Thus, a learning algorithm should be able to handle these changes in surgery (Widmer and Kubat, 1996).

Currently, there are different approaches to handle concept drift. The approaches can be generalized as:

1. learners that adapt to changes at regular intervals and

2. learners that detect the occurrence of concept drift before adapting (Gama et al., 2004a, Ouyang, 2011). There are different approaches to detect concept drift. Among those approaches are probability distribution, features of classification models, characteristics relevance, classification accuracy, error rate, precision, recall and time stamp (Zhenzheng et al., 2011, Gama et al., 2004a).

An ideal concept drift handling technique should be able to adapt quickly to changes, robust to noise and distinguish it from concept drift, and recognize and treat recurring contexts (Tsymbal, 2004).

3 Related Work

The related research works are discussed chronologically by commenting on their (a) main assumptions made, (b) their advantages and disadvantages and (c) their coverage of properties that a concept drift handling technique should fulfill. In this section, the existing concept drift handling approaches are generally grouped into: single classifier approaches and ensemble classifiers approaches. A detailed discussion on the possible categorizations of concept drift handling approaches is presented in section 5.

I. Single classifier approaches

In this section, the popular as well as the pioneer single classifier approaches used to classify streaming data are discussed.

A. STAGGER

STAGGER is the first technique designed to cope with concept drift (Ouyang, 2011, Schlimmer and Granger, 1986a). It uses a concept description consisting of class nodes connected to attribute-value nodes by probabilistic arcs. These probabilities are updated when new training examples arrive. It also adds nodes corresponding to new classes and new features. To cope with concept drift, STAGGER decays its probabilities over time.

B. Concept Versioning

Concept versioning (Klenner and Hahn, 1994) is a tracking method for gradual or evolutionary concept drifts. Using a frame representation, the method copes with such drifts by either adapting its existing concept class, generalizing the violated conditions, or by creating a new version of concept class. The method creates a new version when attribute values and ranges, present in the current concept descriptions, are dissimilar or changed. This is done based on a measure that accounts for both quantitative (e.g., value differences) and qualitative (e.g., increasing trend) information.

C. FLORA

The FLORA framework (Widmer and Kubat, 1996) tracks concept drift by keeping a sequence of instances over a dynamically adjusted time-window. It uses instances to construct three sets of rule-based concept descriptions: rules covering the positive instances, rules covering the negative instances, and potential rules (i.e) fuzzy. Based on the instances entering and leaving the window, FLORA updates the rules, to either move them from one set to another or to remove them. FLORA works with only one instance at a time; so it has a limitation on the speed of arriving data (Ouyang, 2011). Moreover, it has a built-in forgetting technique with the implicit assumption that those instances falling outside their training window are no longer relevant, and the information carried by them can be forgotten (Elwell and Polikar, 2011).

D. MetaL (B) and MetaL (IB)

Although not deliberately designed to deal concept drift, the MetaL (B) and MetaL (IB) systems (Widmer and Kubat, 1996) use meta-learning with naive Bayes and instance-based learning, respectively, to cope with recurring contexts. Meta-learning techniques identify contextual attributes by maintaining co-occurrence and frequency counts entire history of the learner over a fixed window of time. For instance, the concept of warm is not same in summer and winter. In this case, the season or average temperature is a contextual attribute that identifies the relevant concept of warm. In moving from one season to the next, the contextual variable is used to better focus on those features relevant for learning and prediction.

II. Ensemble approaches

The use of classifier ensembles is a common way of boosting classification accuracy. Due to their modularity, ensemble approaches also provide a natural way of adapting to change by modifying the ensemble members. The following sections describe five specific adaptive ensemble algorithms.

A. Streaming Ensemble Algorithm(SEA)

Streaming Ensemble of Algorithms (SEA) (Street and Kim, 2001) is based on a fixed number of ensemble classifiers each constructed from relatively small subsets of data, read sequentially in blocks. Once the ensemble is full, new classifiers are added only if they satisfy some quality criterion, based on their estimated ability to improve the ensemble's performance. Since the ensemble size is fixed, one of the existing classifiers is replaced by the new one. However, because of the replacement of the ensembles recurrent concepts may not be addressed.

B. Accuracy Weighted Ensemble (AWE)

According to Wang et al. (2003), a simple ensemble might be easier to use in changing environments than a single adaptive classifier. Wang et al. (2003) propose an ensemble of classifiers called Accuracy Weighted Ensemble (Wang et al., 2003). Wang et al. (2003) maintains a fixed-size collection of classifiers built from batches of training samples, but it weights each classifier based on their performance on the most recent batch. One of the drawbacks of evolving ensemble of classifiers, in general, is that it builds a new base classifier for each batch of new data. This is not a good idea in terms of performance.

C. Paired Learners

Bach and Maloof (2008) deal concept drift using paired learners, a stable on-line learner and a reactive learner. They argue that a learner that is too reactive may have difficulty acquiring any target concept; while a learner that is too stable may not learn a new concept (Bach and Maloof, 2008). Another research very close to this, named paired evaluators method, deals with sudden changes of concept drift (Furuhata et al., 2010). Two evaluators are used, stable evaluator and reactive evaluator, which are used to evaluate the performances of underlying classifiers. The difference between these two studies is on handling windows and on dealing with specific types of concept drift.

D. Active Classifier

Active classifier focuses on learning an accurate model with as few labels as possible. It studies how to label selectively instead of asking for all true labels. It contains four active learning strategies to explicitly handle concept drift. They are based on random labeling, fixed uncertainty strategy, variable allocation of labeling efforts over time and randomization of the search space (Zliobaite et al., 2011). It also contains the Selective Sampling strategy, which is adapted from (Cesa-Bianchi et al., 2006). An active classifier encapsulates all the active learning strategies and allows to have benchmark streaming data experiments through stored, shared, and repeatable settings for synthetic and real data (Zliobaite et al., 2011). Active classifier uses EDDM as the drift detection technique and when a change is detected, the old classifier is replaced by the new one. In such an approach, recurrent concept drifts may not be handled.

E. Learn++

Elwell and Polikar (2011) introduce a relatively more generic ensemble of classifiers. It incrementally learns in the presence of concept drift, called Learn++ or Learn++.NSE. The classifiers are incrementally trained (with no access to previous data) on incoming batches of data, and combined with weighted majority voting. Classifiers capable of identifying previously unknown instances get more credits while classifiers that misclassify previously known data are penalized. In other words, the penalty of misclassifying previously known data has more weight than misclassifying an unknown data. Finally, the voting weights are determined as log-normalized reciprocals of the weighted errors (Elwell and Polikar, 2011). The core task in Learn++.NSE is determining the voting weights based on the time-adjusted accuracy competencies of the classifiers. However, since the weight adjustment mechanism used is based on the history of their classification accuracies, classifiers may get penalized or rewarded wrongly because of noisy input. Adding a noise detector before updating the classifiers would make this framework more effective. On top of that, a new set of classifiers is created for each new data chunk. So, the ensemble size can become extremely large considering lifelong learning.

In general, the current approaches are either tested on synthetic data or are studied for specific drift types in specific environments (Elwell and Polikar, 2011). For example, techniques used effectively in spam filtering may not be good enough in surgery prediction or weather forecasting. Developing an adaptive learning model that handles concept drift in dynamic environments with the treatment of concept drift and noise is a research area which demands improvements in the state-of-the-art (Wang et al., 2011, Ouyang, 2011).

4 Research Methodology

A research methodology is a way of solving research problems systematically along with research methods (Kothari, 2008). Research methodology studies how a research is done scientifically by answering different questions, such as why the research is conducted, how the research problem is defined, how hypotheses are formulated, what kind of data is collected and method is used to collect the data, and technique(s) used data analysis (Kothari, 2008). Researchers design different methodologies depending on the nature of the research problems. Thus, the design of this research methodology includes literature review, problem definition, research questions and hypotheses formulation, research design, and experimental result analysis and interpretation. Figure 4.1 shows the research methodology of the research.

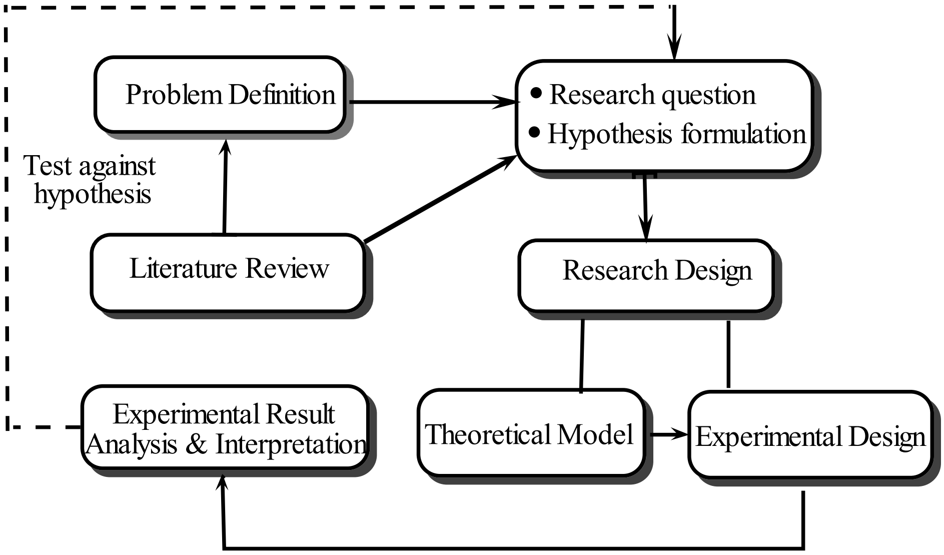


Figure 4.1: Research Methodology

4.1 Literature Review

An extensive literature review is conducted from different perspective, such as the healthcare domain, data mining, machine learning and research methodologies. The state-of-the-art healthcare systems are deeply reviewed with more emphasis on the automation and performance improvement of patient diagnosis and referrals. Then the literature review is deepened and narrowed down with respect to the evolving decisions on referrals for surgery. Surgery is one of the most expensive treatments among others (drug, therapies, etc) in secondary care, hospitals (Persson and Lavesson, 2009). The literature review included studying the effect of such evolution on the application of decision support systems in surgery prediction or patient referral. A further extensive literature review is done to address the state-of-the-art of data mining and machine learning with respect to handling concept drifts, particularly surgery prediction concept drift. As part of the literature review, the research gap is identified, the existing concept drift detection and handling approaches are investigated and categorized into possible categories. This continued with evaluating, comparing and contrasting these categories with respect to criteria like prediction accuracy, addressing different concept drift types (sudden, gradual, recurrent, etc), memory usage, and treatment of noise. Following the theoretical modeling and hypotheses formulation of the problem of surgery prediction concept drift, the literature review is continued in finding suitable statistical tests to test and analyze the experimental results with respect to the hypotheses and research questions.

4.2 Problem Definition

The decision of patient referral is evolving over time. A prediction model developed based on historical hospital data will not work properly on new data when the concept it learned is changed. The prediction model needs to adapt itself to handle such concept drifts. However, there is no universal way of handling concept drifts that can be directly applied in all such evolving domains (Elwell and Polikar, 2011). Developing a more generic technique which works regardless of restrictions on how slow, fast, sudden, gradual, local, global, cyclical, noisy or otherwise changes in internal distribution, is still a challenge (Elwell and Polikar, 2011). Thus, there is a need for investigation on handling surgery prediction concept drift.

4.3 Research Questions

The purpose of this research is reflected through research questions. The research questions are framed to provide a guideline for conducting the research. An empirical approach is used to answer the research questions. The outcome of the empirical approach is verified through an experiment. Generally, the research questions are descriptive analysis of data for inferential study and tested through hypotheses. The research questions raised in this research are:

- *RQ1:* What is the impact of concept drift on classification performance? Elaboration: To identify the relationship between concept drift and classification performance in the context of surgery prediction.

- *RQ2:* When does change in data distribution cause surgery prediction concept drift?
  Elaboration: To investigate the relationship between temporal changes in data distribution and surgery prediction concept drift.
- *RQ3:* What kind of concept drift handling algorithm performs better in surgery prediction concept drift, trigger based or evolving algorithms?

4.4 Hypotheses

Hypotheses are derived from theory and used as means to validate or invalidate the theories through subsequent experiments in the research. The hypothesis delimit the research by providing a focal point to the research. The investigation of concept drift handling algorithm and impact of concept drift on prediction performance is limited to orthopedics department surgery prediction.

Basic Hypothesis: Concept Drift occurs in surgery prediction.
H1: Concept drift impacts classification performance negatively in surgery prediction.
Null hypothesis H_{01}: Classification performance is not significantly affected by the occurrence of concept drift.
Alternate Hypothesis H_{11}: Classification performance decreases significantly in the occurrence of concept drift.
Dependent Variable: Prediction error rate - a measure of classification performance
Independent Variable(s): Patients' record, classification algorithms, and experiment environment

H2: Concept drift can be detected in surgery prediction.
Null hypothesis H_{02}: The variance of classifier's performance between batches is the same.
Alternate Hypothesis H_{12}: The variance of classifier's performance between batches is different.
Dependent Variable: Prediction error rates variance in batches.
Independent Variable(s): Patients' record, classification algorithms, and experiment environment

H3: Handling Concept Drift can reduce the impact of Concept Drift on surgery prediction.
Null hypothesis H_{03}: There is no significant improvement in prediction performance by handling concept drift.
Alternate Hypothesis H_{13}: Handling concept drift significantly improves predic-

tion performance
Dependent Variable: Prediction error rate - the probability of errors made over set of instances.
Independent Variable(s): Patients' record, classification algorithms, and experiment environment

4.5 Research Design

The research design provides a mechanism for defining a conceptual framework for the research. The research design depends on the purpose, problem and contribution of the research (Kothari, 2008). The research design is composed of a theoretical model and experimental design.

4.5.1 Theoretical Model

The problem of handling concept drift in surgery prediction is framed theoretically. The theoretical model builds a conceptual framework of the real world problem. An algorithm that handles concept drift is modeled based on the investigation of existing concept drift detection and handling algorithms. The theoretical model is followed by an experiment to validate the theories.

4.5.2 Experimental and Sample Design

An experiment is found as an appropriate method to address the research questions by testing the hypotheses in the context of handling concept drift in surgery prediction. The experiment tests the theoretical model against the reality by manipulating the variables in a certain research environment to observe their effect on other variables. The experiment includes exploration of new findings by trying the plausible combinations of the existing detection and handling algorithms as well as the theoretical modeled algorithm. The experimental design of the research is discussed in detail in Section 6.1.

Moreover, to execute the research design an appropriate target population, sampling and data collection technique are designed.

A. Target population
The research focuses on real world hospital data, that is patients who undergo surgery and patients who did not undergo surgery. The inclusion criteria is patients who are seen at hospital outpatient clinic as a non-emergency case, elective referrals.

B. Sampling Technique

A real world hospital data is scarce and difficult to find because of many reasons. One of the reason is the confidentiality of the data. The sampling technique used is a non-probability sampling technique called convenience sampling. The convenience sampling technique is selected for two main reasons. First, the availability of data from Blekinge Hospital. Second, to increase the diversity of the sample data by including an existing hospital data that is collected by Persson et al. (2010). Persson et al. (2010) collected the data for the research of automatic identification of surgical indicators. Thus, sampling technique targets the orthopedics department, i.e. to have similar attributes with the existing hospital data.

C. Data Collection

The sample data are drawn from the orthopedics department of Blekinge Hospital. A delegated person with the medical domain knowledge collected the data. The data are collected through direct observation by reviewing patients who are seen at the outpatient clinic for the first and second time. The direct observation method has the advantages of obtaining hospital data that are independent of respondents, free from subjective bias and relatively less demanding for active cooperation from others. The collected data are discussed in-depth in the dataset preparation section of experimental design section, Section 6.1.1.

4.6 Experimental Result Analysis and Interpretation

The collected data are analyzed and interpreted quantitatively to draw a valid conclusion from the experimental result. The analysis includes hypotheses testing. The hypotheses are tested statistically based on the experiment result. The statistical tests used for each sub-experiment are explained below.

4.6.1 Statistical Test

The statistical tests are used to validate the experimental results in reference to the hypotheses. The research focuses on non-parametric tests because the parametric tests assume data are distributed normally. The non-parametric tests are Wilcoxon rank sum test(Mann-Whitney U test), Levene test and Friedman test with Nemenyi post-hoc test. The statistical tests are conducted in the R environment.

The Wilcoxon rank sum test (Mann-Whitney U test) is a non-parametric test used to compare two independent samples. The test ranks the performance difference of two algorithms. The differences are ranked according to their absolute values; average ranks are assigned in case of ties. (Wohlin et al., 2000, Demšar, 2006)

The Levene test is a non-parametric type of test used to test the equality of variances across independent samples, i.e, whether samples have equal variances or not. The equal variances across samples is called homogeneity of variance. The test of Levene is the absolute deviations of the data points from the estimated group center (Carroll and Schneider, 1985).

The Friedman test is another non-parametric type of test used for 'one factor with more than two-treatments'. The Friedman test ranks the algorithms for each dataset separately, the best performing algorithm gets a rank of 1, the second best get a rank of 2 and so forth. In case of ties, average ranks are assigned.

Let r^m be the rank of the j^{th} algorithm of k algorithms on the i^{th} of N data sets. The Friedman test compares the average ranks of the algorithms, $R_j = \frac{1}{N}\sum_i r_i^j$. If the performance of the algorithms is equivalent, then their ranks R_j will be equal. The Friedman test is distributed according to x^2 with $k-1$ degrees of freedom, as shown in equation (4.6.1.1). However, if the ranks are not equal, the null hypothesis will be reject and a post-hoc test, Nemenyi test, will be conducted.

$$\chi_F^2 = \frac{12N}{k(k+1)}[\sum \frac{R_j^j - k(k+1)^2}{4}] \tag{4.6.1.1}$$

The Nemenyi test does a pair-wise comparison of the algorithms to determine which algorithm differ significantly. In the Nemenyi test, all algorithms, k, are compared to each other over N data sets. The performance of two algorithms is significantly different when the corresponding average ranks differ by at least the critical difference, as equation (4.6.1.2)

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{4.6.1.2}$$

where the critical value q_α depends on the significance level α and k (Demšar, 2006).

4.6.2 Experimental Result and Statistical Analysis

The experiment has three sub-experiments: concept drift detection, concept drift handling and surgery prediction concept drift handling algorithms comparison and selection.

A. Sub-experiment 1 - Concept drift detection
The first sub-experiment is a kind of one-shot study without control groups. The experiment evaluates the degree of prediction performance changes over a sequence of batches or a group of examples in order to test the null hypothesis, H_{01}. If the null hypothesis is rejected, the effect of concept drift on the prediction performance will be measured by comparing the prediction performance of a classifier(s) on the batches. Hence *RQ1* and *RQ2* are addressed in this sub-experiment. The result of the experiment is statistically tested by using the Wilcoxon rank sum

test and Levene test for *RQ1* and *RQ2* respectively. Both the Wilcoxon rank sum test and Levene test are evaluated with with 0.05 significance level.

B. Sub-experiment 2 - Handling concept drift
The second sub-experiment is a controlled experiment with *one-factor with two treatments* type of design. The factor is surgery prediction error rate while prediction without handling concept drift, and prediction with handling concept drift are the treatments. The experiment has a balanced design; it has surgery patient examples for each treatment. By statistically testing the measurements of the factor on the treatments, *RQ3* is answered. The hypothesis, H_{03}, is evaluated by comparing the mean error rates (the dependent variable) of the treatments. The result of the experiment is statistically tested using Wilcoxon rank sum test with 0.05 significance level.

C. Sub-experiment 3 - Evaluation of TBE
The third sub-experiment is done to evaluate the concept drift handling algorithms. The results of the experiment are statistically tested using the Friedman test with the post-hoc test Nemenyi. The overall performance of the algorithms is evaluated using the Friedman test with 0.05 significance level. When the performance of the algorithms is not equal, the Nemenyi test post-hoc test is used. The Nemenyi test is used to do a pair-wise comparison of the performances of the algorithms to check if they differ significantly.

5 Concept Drift Detection and Handling Approach

Today, the mining and classification of sequence of data having concept drift is a challenging research area; the challenge about how to design a reliable and sensitive concept drift detection method is not solved yet (Ouyang, 2011). The ideal approach is to identify the time when concept drift occurs, then update the classification model by re-training it timely. Nonetheless, the issue is far more difficult than it seems. The current methods dealing with concept drift are investigated in the next section.

5.1 Existing Approach

Effective learning in environments with hidden contexts and concept drift requires a learning algorithm that can detect context changes without being explicitly informed about them, can quickly recover from a context change and adjust its hypotheses to a new context, and can make use of previous experience in situations where old contexts and corresponding concepts reappear.

5.1.1 CD Detection Technique

The main point of concept drift detection method is about directly comparing the similarities and differences between concepts of different times. This is often a very difficult task. Currently, most of the concept drift detection methods continuously track concept drift from two aspects: reasons that may cause concept drift or possible effects after concept drift. Such detection techniques include the following.

A. Probability distribution
Uniformly distributed or stationary data streams assume all data processed is generated by the same probability distribution function, but in the case of evolutionary data streams, with concept drift, the probability distribution of data may be continuously changing over time. Hence, the occurrence of concept drift can be detected by monitoring the change in the probability distribution of the data.

B. Characteristics relevance
The occurrence of concept drift may cause changes in the relevance of sample characteristics (attributions); previously relevant characteristics may no longer be relevant (Ouyang, 2011). By tracking the relevance of various characteristics, one

can decide if concept drift has occurred.

C. Features of classification models
Some classification models are very sensitive to changes in the sample probability distribution. For example, rule-based classification systems and support vector machine (SVM) based classification systems are sensitive sudden change in the number of classification rules and the number of support vectors respectively. Such changes may imply concept drift.

D. Classification accuracy (classification error)
Most integrated classification algorithms use classification accuracy as an indicator to judge whether a concept drift has occurred or not. This includes: window based algorithms (Widmer and Kubat, 1996), random decision tree based methods (Fan and Streamminer, 2004), accuracy-weighted ensembles (Wang et al., 2003), methods using high performance basic classifiers in place of relatively low performance basic classifiers (Wang et al., 2003, Street and Kim, 2001). Many single-classifier algorithms adaptively adjust the size of a sliding window according to changes in the classification accuracy (classification error) (Jose et al., 2006, Klinkenberg and Joachims, 2000) or update the current classification model (Widmer and Kubat, 1996). Moreover, indicators derived from the classification accuracy, like recall, precision, etc. or a combination of them can be used to detect the occurrence of concept drift under this category (Klinkenberg and Renz, 1998).

E. Time stamp
Taking the sample's time stamp as additional input attribute, it can be used to determine whether concept drift occurs or not, according to a rule with the time stamp attribute. This technique can be used in the learning of time-changing concepts. The series of algorithms: CD3, CD4 and CD5 proposed in Hulten et al. (2001) use time stamp as an additional input attribute when constructing their decision trees. In the case of constantly distributed data streams, the time stamp attribute is not relevant to other properties, and it does not appear in any path of the decision tree. However, when a concept drift occurs, the time stamp attribute will appear in the tree. When a classification path contains the time stamp attribute value, it shows the path, i.e. classification rules, is relevant to time stamp; when the path representing previous time appears in the tree it shows that the classification rule is outdated and thus can not be used for classifying the data any more. Depending on different criteria of comparison, the current concept drift handling approaches can be categorized in different ways.

5.1.2 Single Classifier Approach and Ensemble Approach

This division is based on the number of classifiers used in the classification system.

5.1.3 Sample Manipulation Based and Classifier Manipulation Based Approach.

The dimension of grouping concept drift handling approaches is based on how the learners adapt. The adaptation mechanisms are either related to training set formation or a design and parametrization of the base learners.

5.1.4 Evolving Approach

Some of the techniques discussed above employ change detection mechanisms, still these are not the triggers of adaptation ('detect and cut'), but rather a tool to reduce computational complexity. First we discuss ensemble techniques, which make the largest group, and then other evolving techniques.

A. Adaptive ensembles
An evolving technique is most commonly used for handling concept drift is classifier ensemble. The final decision is received by combining or selecting the classification results of many models based on combination rules. The rules determine how the individual model's results or weights are treated. There are a number of ensembles for concept drift with ideas not specific to particular type of base learners (Kolter and Maloof, 2003, Karnick et al., 2008, Street and Kim, 2001, Wang et al., 2003, 2006). On the other hand, there are other base learner specific ensembles where the classifier combination rules usually depend on the base learner specific parameters of the learned models (Wu et al., 2005, Law and Zaniolo, 2005, Klinkenberg, 2004). In both cases adaptivity is achieved by the combination rules on how the weights are assigned to the individual model outputs at each point in time.

In adaptive ensemble learners, more attention is given to model the evaluation and fusion rules, while little attention is given to building the model. While there are many options on how to build diverse base classifiers, the implicit target is to have at least one classifier in the ensemble trained for every distinct concept. This can be done using many training set selection strategies.

The straightforward strategy is to break historical data into blocks, which has instances sequential in time. Such approaches are suitable for sudden and to some degree for incremental drifts. Another strategy can be using different-sized training windows (Kolter and Maloof, 2003, Scholz and Klinkenberg, 2007).

Another strategy to construct diverse base classifiers is to use similar training data, but different types of base learners like: Naive Bayes, decision tree, or SVM (Zhang et al., 2008, Rodríguez and Kuncheva, 2008). Such techniques build individual classifiers from what has already been seen in the past.

B. Instance weighting

Instance weighting forms another category of evolving adaptation techniques. Instance weighting based algorithms can contain a single learner (Zhang et al., 2008, Bifet et al., 2009). However, here, the adaptivity is achieved not by combination rules, but by a systematic training set formation. Often, the main idea is boosting, giving more attention to the misclassified instances.

C. Feature space

There are some models manipulating feature space to achieve adaptivity. For example, Forman (2006) use ideas from transfer learning to achieve adaptivity. New features are added to the training instances, which contain information from the past model performances. Wenerstrom and Giraud-Carrier (2006) use dynamic feature space over time.

D. Base model specific

There are also models where adaptivity is performed by controlling model specific parameters or model design. Nunez et al. (2007) maintain variable training windows by adjusting the internal structure of decision trees. Kelly et al. (1999) adjust regression parameters to achieve adaptivity. Syed et al. (1999) transfer and combined past support vectors with the recent training data.

E. Trigger based approaches

The other category of approaches uses triggers to determine how the models or sampling must be updated at a given time.

F. Change detectors

In the literature, the most common trigger technique is change detection, which is often implicitly associated to a sudden concept drift. The detection can be based on monitoring: the raw data (Patist, 2007), the parameters of the learners (Su et al., 2008) or the outputs (error) of the learners (Gama et al., 2004b, Jose et al., 2006).

G. Training window

Heuristics can be used for determining training window sizes (Widmer and Kubat, 1996, Bach and Maloof, 2008). The heuristics is related to error monitoring. To determine the size of the training window, an action look-up table is used. An action is associated with each possible value of a trigger. Base learner specific methods can also be used to determine the training windows (Hulten et al., 2001, Zhang et al., 2008).

H. Adaptive sampling
Another category of trigger based methods uses instance selection, where arriving unlabeled testing instances are inspected. Relying on the relation between the predefined prototypes and the testing instance (Katakis et al., 2009) a training set for a given instance is selected.

5.2 The New Approach-Trigger Based Ensemble (TBE)

The problem of concept drift in surgery prediction is conceptualized theoretically to handle changes in concept. The theoretical model combines trigger and ensemble-based approach to handle the concept drifts. Street and Kim (2001) and Wang et al. (2003) studies that an ensemble built by dividing the data into sequential blocks of fixed size examples is effective to handle concept drift. The ensemble based algorithm handles recurrent concepts by retaining the old concepts in the ensemble (Tsymbal et al., 2006). There are several empirical evaluations that suggest ensembles perform better than single classifier (Kolter and Maloof, 2003, Xiang et al., 2009). However, such ensembles are creating new classifiers for each block of examples without detecting the occurrence concept drift. This will result in unnecessary growth of the number of classifiers and increase of memory usage. Managing the ensemble size by creating new classifiers only when a concept drift is detected can improve the size, memory, and performance.

Thus, the problem of concept drift in surgery prediction is framed by dividing the dataset into sequential blocks of fixed size. Each block of patient example, batch, is an m-dimensional vector of attributes in some predefined vector space $x = R^m$ and a class label $y \in \{yes, no\}$. Each batch, b, containing n examples, we get a sequence of batches $[(x_i, y_i), (x_{i+1}, y_{i+1}), ...(x_n, y_n)]$, where the i^{th} example will be represented by (x_i, y_i). Each incoming patient example is represented as (x, y) in TBE.

5.2.1 Concept Change Detection

Parametric approaches assume data distribution is normal in nature and can be modeled statistically based on means and covariance. The performance of classification algorithms are commonly evaluated through an error and a reject rate. Errors are unavoidable due to the existence of uncertainty and noise in classification tasks. Subsequently, a reject rate is introduced as threshold to avoid excessive misclassification (Markou and Singh, 2003). Recently many studies suggest classification errors as a detection method for concept changes (Jose et al., 2006, Nishida and Yamauchi, 2007a, Gama et al., 2004a, Nishida and Yamauchi, 2007b). The classification error based detection methods have the ability to detect concept drift from a small number of examples and have low computational costs (Nishida and Yamauchi, 2007b). Thus, the research detects concept changes in surgery prediction by monitoring the classification error rates.

To monitor the surgery prediction classification error rates, early drift detection method proposed by (Jose et al., 2006) is adapted. According to PAC learning, if the distribution of an example is similar to another example, the classification error rate decreases as the number of examples increases (Mitchell, 1997). With a large number of examples, $(n > 30)$, the Binomial Distribution is closely approximated by a Normal Distribution with the same mean and variance (Jose et al., 2006). Thus, each batch is chosen to have more than 30 examples [1] in order to make the detection method valid, and to monitor the classification error rates statistically based on means and variance.

A binomial distribution gives the probability of observing an error for each instance i in the batching sequence of n examples. The error rate is the probability of misclassifying instance (p_i), with standard deviation $s_i = \sqrt{\frac{p_i(1-p_i)}{i}}$ (Jose et al., 2006). The average distance between two errors $(p\prime_i)$ and its standard deviation $(s\prime_i)$ is calculated. The distance between each incoming of batches is calculated to indicate the warning and drift level of concept drift in surgery prediction.

Jose et al. (2006) define two thresholds for warning and drift levels (Equation 5.2.1.1 and Equation 5.2.1.2 respectively).

$$\frac{(p\prime_i) + 2 \times (s\prime_i)}{(p\prime_{max}) + 2 \times (s\prime_{max})} < \alpha \tag{5.2.1.1}$$

$$\frac{(p\prime_i) + 2 \times (s\prime_i)}{(p\prime_{max}) + 2 \times (s\prime_{max})} < \beta \tag{5.2.1.2}$$

These thresholds are used to track the warning and drift levels. Each time a warning level, α are stored in advance to a possible change of context. If the error rate falls below the warning threshold, the warning is treated as a false drift. But if the drift level is reached, then it is assumed a true concept drift is detected. The EDDM detects both gradual and sudden concept drifts. Thus, the concept drift in surgery prediction is monitored based on the warning and drift levels proposed by (Jose et al., 2006).

5.2.2 Concept Drift Handling Algorithm

The patient examples arrive in batch, b, over time. A knowledge base is initialized by creating a base classifier from the available batch b of data. For each new training dataset, the existence of concept drift is tracked using EDDM detection

[1] The batch size used depends on the total number of instances included in the experiment, i.e. 50 and 500. See Section 6.1.3.

method. If a drift is detected, the algorithm adds a new classifier. Otherwise, the classifiers will be trained by updating their weights based on their performance on the current dataset, a classifier that performs less gets less and that performs better gets better weight. When the buffer size is full, less relevant classifiers will be pruned based on their error rate and generation time. Pruning helps the model to maintain the ensemble's overall competency and preserve memory and computation time in a long-term data-mining applications (Bach and Maloof, 2008, Wang et al., 2003). The final decision of the ensemble is obtained based on majority voting of the current classifiers. This method can be applied with any learning algorithm. It can be directly implemented inside online and incremental algorithms, and can be implemented as a wrapper to batch learners. The goal of the proposed method is to detect concept drift from a sequences of examples with a uniform distribution. Those sequences of examples are denoted as context. From the practical point of view, the method chooses the training set which is more appropriate to the actual class-distribution of the examples.

```
Input            : For each dataset $D^t$ t = 1,2, ...
Training Data: $x^t(i)\epsilon X; y^t(i)\epsilon Y; i = 1, ..., no\_inst$
Supervised learning algorithm: BaseClassifier
for element of $b$ in $D^t$ do
    if $t > 1$ then
        Detect the occurrence of change
        if Detected then
            if buffer_full then
            Remove the poorest and oldest classifier, and add a new classifier
            else add new classifier
            else
                Compute error of the existing ensemble on new data
                for $i \leftarrow 0$ to $no\_inst - 1$ do
                    $error = \sum \frac{1}{no\_instance}$
                end

            end
        end
        Update and normalize instance weights
        for $i \leftarrow 0$ to $no\_inst - 1$ do
            if correctlyClassified then $inst\_weight(i) = \frac{1}{no\_instances}$
            $total_weight+ = \frac{1}{no\_instances})$
            else $inst\_weight(i) = (\frac{1}{error}) \times (\frac{1}{no\_instances})$
            $total_weight+ = (\frac{1}{error}) \times (\frac{1}{no\_instances})\}$
            for $i \leftarrow 0$ to $no\_inst - 1$ do
                $inst\_weight(i) = \frac{inst\_weight(i)}{total\_weight}$
            end

        end
    else
        Initialize $inst\_weight(i) = \frac{1}{no\_inst}$
    end
    Call Base Classifier with $D^t$, obtain $h^t : X \rightarrow Y$
    Evaluate all existing classifiers on new data $D^t$
    Compute the weight of each classifier based on its current accuracy on the
    new data
    Normalize and update the weight of each classifier k
    Obtain the final hypothesis based on majority vote
end
```

Algorithm 1: TBE Pseudocode

6 Experiment

6.1 Experimental Design

The experimental design of the research is composed of datasets preparation, algorithm selection criteria, algorithm evaluation criteria, algorithm performance comparison criteria, experimental environment and experimental result.

6.1.1 Datasets Preparation

There is a shortage of suitable and publicly available real-world benchmark data sets intended for the research in data stream classification. Most of the available benchmark datasets are not suitable for evaluating data stream classification algorithms. The datasets contain too few examples and do not show sufficient concept drift. For this reason, it has become a common practice, for researchers, to publish results based on both real-world and synthetic data sets. Similarly, for this experiment a real-world hospital data, hip-replacement data, from the orthopedics department of Blekinge Hospital is used. Synthetic data are also generated to obtain valid experimental results by including more examples in the research. The synthetic data are generated by using STAGGER and SEA concept generators. The synthetic data generators, STAGGER and SEA, are discussed below in Section B and C respectively. Additionally, a poker-hand dataset, from different domain is used to increase the validity and generality of the experimental results.

A. Hip-replacement Dataset

The hip-replacement dataset includes a two years patient referrals seen at the outpatient clinic of the Orthopedics department at Blekinge Hospital. The data are drawn from year 2008 and 2011. A delegated person reviewed all patient referrals seen at the outpatient clinic in year 2011 through direct observation. Accordingly, a total of 151 complete patient records are identified. Moreover, a total of 80 patient referrals are included in the hip-replacement dataset from year 2008. These data are secondary data drawn from two consecutive months. The patient records collected from year 2011 are further preprocessed. However, the patient records from year 2008 are already preprocessed by the researchers of automatic identification of surgical indicators, Persson et al. (2010).

The identified patient records from year 2011 are preprocessed by removing attributes that are related to the social affair and less relevant for the classification. A noisy data, incorrectly value, are also removed from the dataset. Some of the numerical attributes are discretizated to improve classification accuracy and reduce the learning complexity. Discretization divides the numerical values of the continuous attributes into intervals (Witten et al., 2011, Yang et al., 2010). The patient records are also preprocessed in a similar way as the secondary dataset and validated by the expert of the field. Finally, a total of 222 patient records(instances) are included in the research after the preprocessing. These instances have 11 attributes, one of which is the class value. The list of attributes included in the experiment are shown in Table 6.1.

Table 6.1: Orthopedics Dataset Statistics

Attribute	Range[a]	Mean/Ratio[b]	Missing[c]
Age	38...92	71(10.14)	0
Pain duration	long, short	0.52, 0.17	0.31
Walking-aid	0,1,2,3,4	0.20, 0.04, 0.07, 0.04	0.33
Walking distance	short, long, medium,limited	0.18, 0.09 0.23, 0.05	0.45
in-rotation	no,some,yes	0.36, 0.15, 0.21	0.28
out-rotation	no,some,yes	0.17, 0.19, 0.33	0.30
x-ray gradation	1,2,3,4	0.01, 0.08, 0.07, 0.05	0.79
diagnosis	M160, M161, M167, M706 T840F, T841F	0.55, 0.07, 0.01, 0.01, 0.01, 0.01	0.36
abduction	narrow,normal wide	0.05, 0.08, 0.01	0.87
had surgery	0,1	0.18, 0.04	0.78
surgery(target)	yes, no	0.75, 0.25	0

[a] The range value is the minimum and maximum value of the numerical variables and the possible values of the nominal variables.
[b] The mean and standard deviation of the numeric variables and the ratio of instances belonging to each class of the nominal variables.
[c] The ratio of missing value of each variables.

The walking aid variable indicates the level of aid needed for walking. A patient may be able to walk with; no aid (0), a cane (1), crutches (2), a walking frame (3), or he/she needs a wheel chair (4). Similarly, the walking distance variable indi-

cates the distance a patient can walk. The variable walking distance is discretized in intervals; < 100 (Limited) , < 300 meters (short), < 1 kilometers (medium), ≥ 1 kilometer (long)(Persson et al., 2010).
The variables in-rotation and out-rotation represent the different types of hip mobility. The patient record is presented with the degree of mobility for each of these variables. As a result, the variables are discretized in intervals of 0 degrees (no), $\leq$ 10 degrees (some),and $>$ 10 degrees (yes).
The pain duration variable shows estimated time that the patient has been experiencing pain from the affected joint. The variable pain duration is discretized in interval, < 1 year (short), ≥ 1 year (long). Patient's previous surgery history, contralateral hip joint replaced with a joint prosthesis, is described by using the variable had surgery. The values are yes and no are discretized as 1 and 0 respectively.
The x-ray variable (x-ray) indicates the grade of osteoarthritis as seen on AP pelvic radio-graph and has the following grades: 0 (none), 1 (doubtful), 2 (minimal), 3 (moderate), and 4 (severe) (Persson et al., 2010). The variable abduction indicates the ability of the hip muscles to move the leg away from the central line of the body. The variable is discretized as $> 10(narrow)$, $15 - 30(normal)$ and $> 30(wide)$. The variable diagnosis represents the cause of illness. The variable values are presented based on the 10^{th} revision of the International Statistical Classification of Diseases and Related Health Problems, (ICD-10). The identified diagnoses are: M160(Bilateral primary osteoarthritis of hip), M161(Unilateral primary osteoarthritis of hip),M167 (Other unilateral secondary osteoarthritis of hip), M706 (Trochanteric bursitis), T840f (Mechanical complication of internal joint prosthesis) and T841f (Mechanical complication of internal fixation device of bones of limb or Mechanical complication of int fix of bones of limb). Finally, surgery variable is a the class value that indicates whether surgery was carried out or not with values yes or no.

The hip-replacement dataset is used as a baseline to generate synthetic data. The synthetic data are generated by using concept generators, STAGGER and SEA, to introduce different types of concept change. The concept drift generators are selected based on literature review and types of concept drift they introduce.

A. STAGGER

STAGGER is concept generator introduced by Schlimmer and Granger (1986b). STAGGER creates sequence of data with gradual, abrupt concept drift and noise free examples(Minku et al., 2010, Gama et al., 2004b). The STAGGER is used to generate 10,000 hip-replacement instances from the 222 real-world hip-replacement examples.

B. SEA

SEA is used to generate larger hip-replacement dataset from the 222 real-world examples. SEA concept generator simulates recurring and abrupt concept drifting to the hip-replacement dataset(Kolter and Maloof, 2007, Minku et al., 2010). SEA also introduces noise to the dataset SEA generates. The type of noise it generates is a class noise, incorrect class values. A total of 10,000 instance is generated from the 222 real-world hip-replacement examples. The SEA concept generator is configured to add 10% class noise in the dataset.

C. Poker-Hand dataset

The poker-hand dataset consists of 1,000,000 instances and 11 attributes. Each instance of the poker-hand dataset is an example of a hand consisting of five playing cards drawn from a standard deck of 52. Each card is described using two attributes, suit and rank, for a total of 10 predictive attributes. There is one class attribute that describes the "Poker Hand". The order of cards is important, which is why there are 480 possible Royal Flush hands instead of 4 (Bifet et al., 2009). In the poker-hand dataset, the cards are not ordered, that a hand can be represented by any permutation, which makes it very hard for propositional learners, especially for linear one (Bifet et al., 2010).

The poker-hand dataset is used increase the generality of the research. The number of instances included in the experiment are limited to 10,000 because of the limitation in computational resource. Thus, the first 10,000 instances are selected as the poker-hand dataset has already random cards.

6.1.2 Algorithm Selection Criteria

The focus of the research includes studying the possible concept drift handling algorithms. The investigation includes the properties which an ideal concept drift handling algorithm should have, discussed in the background section. Thereof, algorithms from different learning paradigms and with better ability to handle a diversity of drifts are selected. From single classifier, active classifier with early drift detection method is chosen. From ensemble methods, the Accuracy Weighted Ensemble (AWE), Learn++, and the proposed algorithm, TBE, are chosen.

The AWE algorithm works well on data with reoccurring concepts as well as different types of drifts. AWE improves its performance gradually over time and is best suited for large data streams (Wang et al., 2003). Active classifier is a single classifier with concept drift detectors, DDM and EDDM. Moreover,the algorithm is designed to detect concept drifts by considering changes in data distribution. The Learn ++ is an incremental ensemble learning algorithm that learns from consecutive batches of data without making any assumptions on the nature or rate of drift (Elwell and Polikar, 2011).

6.1.3 Algorithm Evaluation Criteria

An interleaved Test-Then-Train evaluation method is chosen for the experiment. The evaluation criteria uses each example to test the model before it is used for training, based on which the accuracy is incrementally updated. The model is first tested on examples it has not seen previously. Thus, the evaluation method has the advantage that no holdout set is needed for testing, making maximum use of the available data (Bifet et al., 2009).

The number of instances used for surgery prediction and algorithms performance evaluation is 444 and 10,000 respectively. The batch size used is 50 and 500 for the 444 and 10,000 instances respectively. The batch size is selected based on literature review by maintaining the data normality and avoiding issue related to batch size. The issue related to the batch size is: a large batch size is a stable learner that is suitable for gradual drifts while a small batch size adapts quickly to concept changes, appropriate to abrupt drifts (Gama et al., 2004b, Zhu et al., 2010). Therefore, the batch size is adjusted approximately to the proposition of the number of instances.

To make the comparison more significant, similar parameter values are set for all algorithms. For the ensemble and active classifier algorithms, the default settings are used to evaluate the classification performance. The algorithms default settings are used by assuming the algorithms have the best default setting. Moreover, tunning each algorithm to different settings of its own subjects the experiment to have imbalanced results. For instance, in Learn++ the number of classifiers can not be restricted whereas AWE provides the option. Thus, restricting only the ensemble size of AWE might lead to biased results, positively or negatively.

6.1.4 Algorithm Performance Comparison Criteria

Many measures of performance are possible depending on the nature of the task in question. In the case of supervised concept learning and classification tasks, where each instance has an associated class name, the most obvious metric is the percentage of correctly classified instances (Quinlan, 1986).

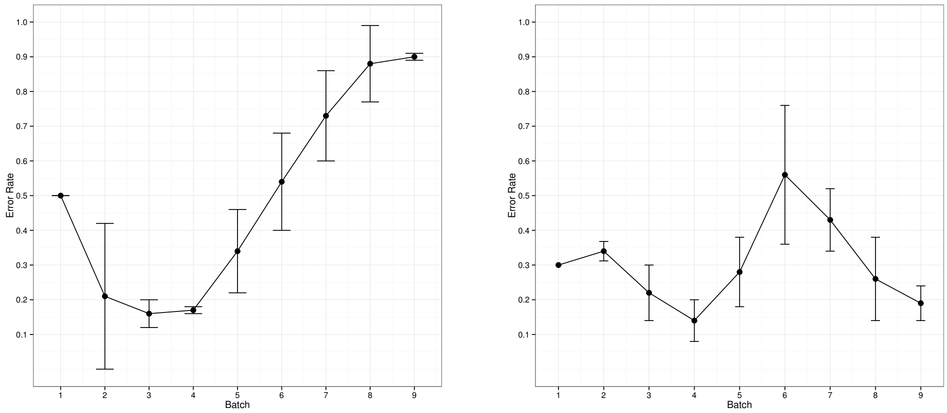
The key independent variables whose variation affects the dependent variable, i.e. prediction performance, are the dataset, the concept drift detection methods and the concept drift handling algorithm.

6.1.5 Experimental Environment

All the evaluated algorithms are implemented in Java using the Massive Online Analysis (MOA) framework. The proposed algorithm, TBE, and the data block evaluation procedures are implemented by the author of this thesis while the rest of the algorithms were already included in the environment. The experiments took place on a machine equipped with an Intel Pentium Core 2 Duo P9300 @ 2.26 GHz processor and 3.00 GB of RAM. Each algorithm was tested on three data sets, described in the previous section, using the Data block evaluation procedure.

6.2 Experimental Results

According to the research questions and hypotheses described in Section 4.3 and Section 4.4 respectively, the experimental results are organized into tables and plots. Table 6.2 presents classification error-rate before and after handling the impact of concept drift. The impact of concept drift in surgery prediction and the relationship between concept drift and temporal changes in data distribution is shown in Figure 6.1. The relationship between concept drift and temporal changes is presented based on the standard deviation between two consecutive batches classification error rate. The improvement of surgery prediction performance after handling concept drift is shown in Figure 6.1b. Finally, the four concept drift handling algorithms(Active Classifier, AWE, Learn ++ and TBE) performance and their rank is shown from Table A.1 to A.3. Moreover, the four concept drift handling algorithms performance is presented from Figure 6.2 through 6.4. These plots will be analyzed in the ensuing sections with respect to the research questions and hypotheses.



(a) Before handling concept drift

(b) After handling concept drift

Figure 6.1: Hip-replacement dataset prediction performance Before and After handling concept drift

Figure 6.1a, depicts the error rate of surgery prediction over sequences of batch before handling concept. The figure on the right side, Figure 6.1b, depicts the classification error-rate of surgery prediction after handling concept drift. The error bars represents the deviation between two consecutive batches, current batch (b), and $b - 1$, performance.

Table 6.2: Performance of Hip-replacement Data set(on manually modified data set)

Data batch	Instances	Error rate without concept drift handling(SD)[a] Before Handling	Error rate with concept drift handling(SD)[b] After Handling
1	50	0.50(-)	0.30(-)
2	100	0.21(0.21)	0.34(0.23)
3	150	0.16(0.04)	0.22(0.08)
4	200	0.17(0.01)	0.14(0.06)
5[c]	250	0.34(0.12)	0.28(0.10)
6	300	0.54(0.14)	0.56(0.20)
7	350	0.73(0.13)	0.43(0.09)
8	400	0.88(0.11)	0.26(0.12)
9	450	0.90(0.01)	0.19(0.04)
	Mean error-rate	**0.49**	**0.30**

[a] The error rate of each batch before handling concept drift. SD is the standard deviation is between two consecutive batches (For instance, between batch 1 and batch 2, between batch 2 and batch 3 ans so on.)

[b] The concept drift handling algorithm is TBE.

[c] A sudden increase in error-rate

Table 6.3: Wilcoxon Rank Sum Test (sub-experiment 1)

data	before and after handling CD
p-value	0.007

Table 6.4: Classical Levene's Test (sub-experiment 1)

data	Error
p-value	0.035

Table 6.5: Wilcoxon Rank Sum Test (sub-experiment 2)

data	Concept Drift Handler and No Concept Drift Handler
p-value	0.046

From Figure 6.2 to Figure 6.4 shows the performance of AWE, Active Classifier, LearnNSE and TBE on different datasets.

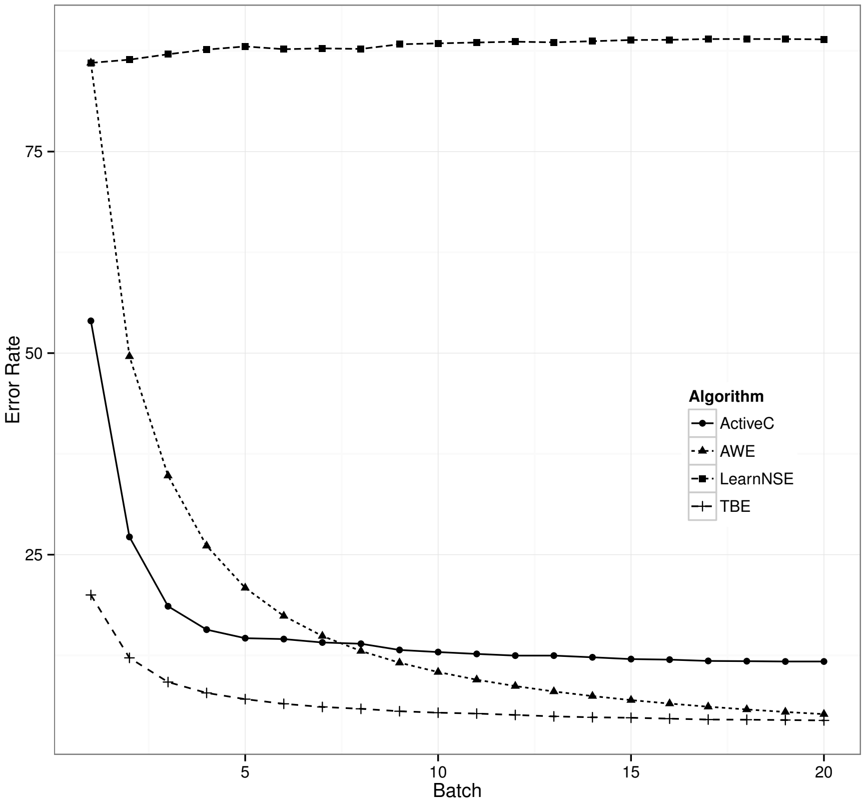


Figure 6.2: Performance Comparison on Handling CD (STAGGER Hip-replacement Dataset)

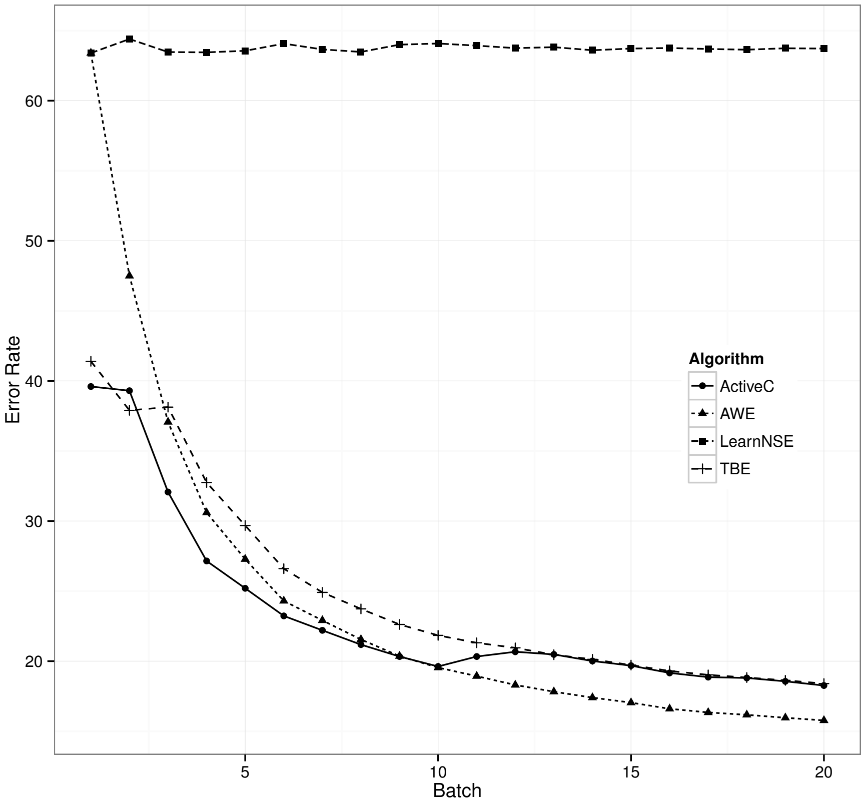


Figure 6.3: Performance Comparison on Handling CD (SEA Hip-replacement Dataset)

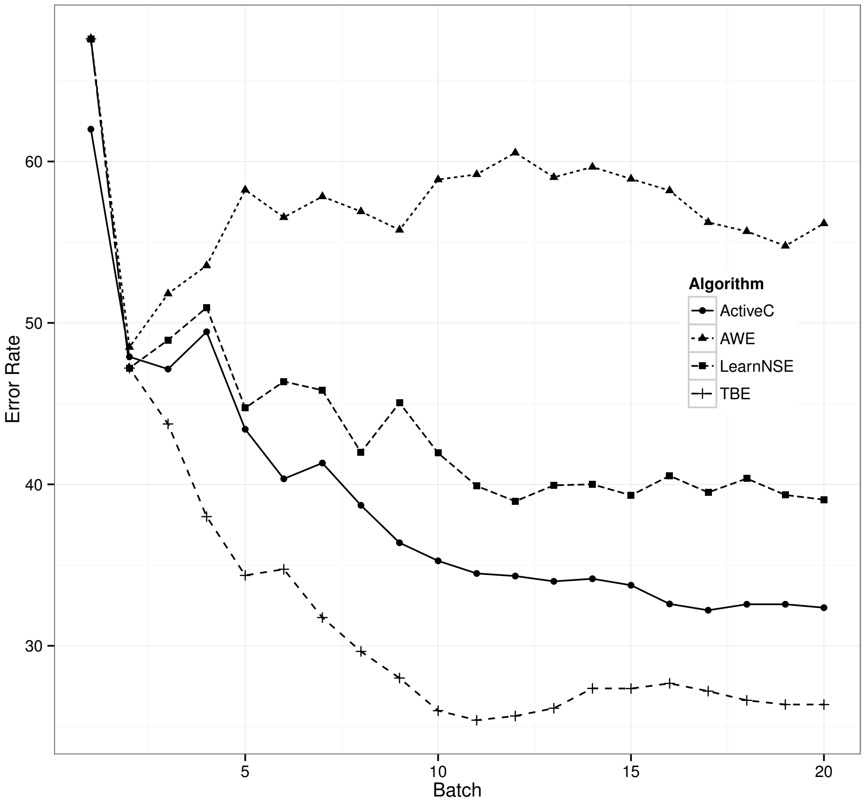


Figure 6.4: Performance Comparison on Handling CD (Poker-Hand dataset)

7 Analysis and Discussion

In the research methodology section, it was hypothesized that: a concept drift exists in surgery prediction, the existence of concept drift has a negative impact on surgery prediction, and the impact can be reduced by using concept drift handling methods.

To start with analyzing the existence of concept drift in surgery prediction, Figure 6.1a illustrate the performance of classification accuracy over the sequence of batches without handling concept drift. The experimental result in Column 3 of Table 6.2 is used to evaluate the degree of prediction performance changes over the sequence of batches in the occurrence of concept drift. According to PAC, if the distribution of examples is similar across batches, the classification error rate decreases as the number of examples increases (Mitchell, 1997). Similarly, in Figure 6.1a, the classification error rate decreases in the first few batches as the number of instances increase. However, the error rate increases suddenly at the 5-th batch due to changes in the concept.

The first sub-experiment tests the null hypothesis (H_{01}), classification performance is not significantly different in the occurrence of concept drift. The hypothesis testing table for classification performance is shown on Table 6.3. The test result indicates that the occurrence of concept drift affected classification performance significantly with the p-value of 0.007. Thus, the null hypothesis (H_{01}), classification performance is not significantly affected by the occurrence of concept drift, is rejected and the alternate hypothesis (H_{11}) is accepted. The impact is also visible in Figure 6.1-(a). As a result, *RQ1* is answered, i.e. concept drift has a negative impact on classification performance.

The relationship between temporal changes in data distribution and surgery prediction is shown based on the standard deviation of the classification error rate of two consecutive batches. As discussed in section 5.2.2, a sequence of instances are treated in batches that arrive at different points in time. The standard deviation between the classification error rates of two consecutive batches is depicted in Figure 6.1a and Figure 6.1b with error bars. A variance in prediction performance indicates concept change. The Levene test used in the second sub-experiment to test the null hypothesis (H_{02}), the variance of classifiers performance between batches is the same. The hypothesis testing table for classification performance variation is shown on Table 6.4. Accordingly, the test result indicates that the variance of classification performance differs between batches with the p-value of 0.035.

Hence, since the p-value is < 0.05, the null hypothesis (H_{02}) is rejected and the alternate hypothesis (H_{12}) is accepted. This answers *RQ2*, significant temporal changes in the data distribution indicate concept drift in surgery.

The performance of classifier is ensured whenever the training and test data are drawn independently from an identical distribution (McAllester, 1998). Therefore, if the distribution of examples across batches is similar then classifiers performance should not decrease. However, Figure 6.1a depicts the error rate of the algorithm increases as the number of examples increases. Moreover, the hypothesis test shows a significant variation in classification performance. The variance in the surgery prediction performance indicates change in concept. Similarly, the error bars in Figure 6.1a also illustrate how the classification performance deviates between temporals as concepts changes. Thus, concept drift can be detected by tracking the performance classifiers across the batches. The Early Drift Detection Method (EDDM) is one of the concept drift detections method that compute the standard deviation of misclassified instance to track change in concept. The EDDM uses the probability of misclassifying each instance and their standard deviations to monitor changes, as discussed in section 5.2.1, equations 5.2.1.1 and 5.2.1.2.

The remaining experiment tests *H3*, handling concept drift can reduce the negative impact of concept drift in surgery prediction. Figure 6.1a and Figure 6.1b illustrate the prediction performances before and after handling concept drift, respectively. In Figure 6.1a, the curve depicts that the error rate is increasing after the occurrence of the concept drift, at the 5-th batch, whereas Figure 6.1b depicts how the accuracy recovers from decreasing. Thus, The negative impact of concept drift in surgery prediction is reduced through a concept drift handling algorithm, as can be view in Figure 6.1b. Furthermore, the experimental results in Column 4 of Table 6.2 is statistically analyzed in the third sub-experiment to test $H3$, i.e. if handling concept drift significantly improves prediction performance or not. If the handling improves prediction performance significantly then the performance of the plausible combinations of the existing detection and handling algorithms including TBE is evaluated. The evaluation and comparison of the algorithms contributes in addressing *RQ3*. The Wilcoxon rank sum test is used to test the null hypothesis (H_{03}), handling concept drift improves prediction performance significantly. Table 6.5 shows the p-value is 0.046. Since the p-value < 0.05, handling concept drift improves prediction performance significantly. Therefore, null hypothesis (H_{03}) is rejected and the alternate hypothesis, H_{13} is accepted. This shows that concept drift handling improves classification performance in general. However, the amount of improvement in prediction performance depends on the type of concept drift handling method used. This is asked in *RQ3*. To answer this question, another sub-experiment is ran with the algorithms and datasets explained in the experimental design section, Section 6.1. As a result, four handling algorithms including TBE, and two different hip-replacement datasets with different characteristics are used. Moreover, to strengthen the validity of the experiment, another data from different domain, namely poker-hand dataset is used.

The results summarized in Figure B.1 and Figure B.2 of Appendix B, show surgery prediction performances of the four algorithms, TBE, AWE, Learn++ and Active classifier on the synthetic hip-replacement dataset that is generated by using SEA and STAGGER respectively. The performance of the algorithms are measured based on the algorithms error rate, the probability of misclassifying instance.

Table A.1 illustrate the performance of the four algorithms, AWE, Learn++, Active Classifier and TBE. The performance of each algorithm is ranked. The average ranks provide a fair comparison of the algorithms. The Friedman test checks whether the average ranks are significantly different from the mean rank, 2.5. The average rank is computed based on the algorithms error rate rank. The test result is used to evaluate the significance difference between the performances of the algorithms. Similarly, Figure B.1 illustrate the results of a Friedman test with the Nemenyi post-hoc test for hip-replacement dataset generated using STAGGER. The result of the tests is used to evaluate the significance difference between the algorithms. The Nemenyi test makes pair-wise tests of algorithm performance. The result shows:

- The error rate of AWE and Active Classifier is not significantly different with p-value of 0.9064.
- The error rate of Active Classifier is significantly lower than Learn++ with p-value of 0.0062.
- The error rate of TBE is significantly lower than Active Classifier with p-value of 0.0003.
- The error rate of AWE is significantly lower than Learn++ with p-value of 0.0005.
- The error rate of TBE is significantly lower than than AWE with p-value of 0.0040.
- The error rate of TBE is significantly lower than Learn++ with p-value of 0.0000.

From the above results, the error rate of TBE is significantly lower than the other algorithms with 99.9% confidence interval. AWE and Active Classifier have similar performance.

In addition to the statistically test, Figure 6.2 also provides visual information that TBE has lower error rate starting from the first batch compared to the other algorithms. This indicates that TBE handles concept drift better than the other algorithms with a low variation throughout the batches. TBE also performs consistently by handling the occurrence of concept drifts for both small and large datasets. Overall, the performance of TBE is significantly better than AWE, Active Classifier and Learn++.

Table A.2 illustrate the comparison between the four algorithms (AWE, Learn++, Active Classifier, and TBE). The average ranks provide a fair comparison of the algorithms. The Friedman test checks whether the average ranks are significantly different from the mean rank, 2.5. The average rank is computed based on accuracy of the algorithms. Consequently, the test result indicates a significant difference between the four algorithms (with $p < 0.05$). A further post-hoc test is conducted for pairwise comparisons. Accordingly, the error rate of AWE and Active classifier is significantly lower than TBE and Learn++ (with $p < 0.05$). The error rate of AWE is insignificantly different to Active Classifier. Overall, AWE and Active classifier are significantly better than the other algorithms and TBE is found to be relatively less capable next to Learn++ in handling noisy data. Similarly, Figure B.2 illustrate the results of a Friedman test with the Nemenyi post hoc test for hip-replacement dataset generated using SEA. The result shows:

- The error rate of AWE is not significantly different than Active Classifier with p-value of 0.99
- The error rate of Active Classifier is significantly lower than Learn++ with p-value of 0.00
- The error rate of Active Classifier is significantly lower than TBE with p-value of 0.02
- The error rate of AWE is significantly lower than Learn++ with p-value of 0.00
- The error rate of AWE is significantly lower than TBE with p-value of 0.01.
- The error rate of TBE is significantly lower than Learn++ p-value with of 0.02

This result shows that the error rate of AWE is significantly lower than TBE and Learn++ with 99% and 99.9% confidence intervals respectively. Similarly, the error rate of Active Classifier is significantly lower than TBE and Learn++ with 98% and 99.9% confidence intervals respectively. The SEA generates concepts with recurrent, abrupt drifts. and 10% noise. Thus, TBE is found relatively less capability in handling noisy data. Moreover, as Figure 6.3 provides visual information about the four algorithms. AWE starts with a very high error rate but outperforms the other algorithms by adapting quickly to changes as the number of examples increases. Active classifier starts with a better performance than the other algorithms but the error rate increases starting from the 10-th batch. TBE starts with better performance next to the Active classifier but adapts gradually for both small and large datasets by maintaining the performance of existing classifiers.

Table A.3 illustrate a comparison of the algorithms (AWE, Learn++, Active Classifier, and TBE) based on average ranks. The Friedman test is again used to determine whether the average ranks are significantly different from the mean rank, 2.48. The average rank is computed based on the ranks of the classification accuracies. As a consequence, there is a significant difference between the four algorithms (with $p < 0.05$). Similarly, Figure B.3 presents the result of a Friedman test with the Nemenyi post-hoc test on the poker-hand dataset. The results of the test is used to evaluate the significance of the difference between the performances of the algorithms. The result shows:

- The error rate of Active Classifier is significantly lower than AWE with p-value of 0.00001
- The error rate of Learn++ and Active Classifier is not significantly different with p-value of 0.09931.
- The error rate of TBE and Active Classifier is not significantly different with p-value of 0.13161.
- The error rate of Learn++ and AWE is not significantly different with p-value of 0.05374.
- The error rate of TBE is significantly lower than AWE with p-value of 0.00000
- The error rate of TBE error rate is significantly lower than Learn++ with p-value of 0.00003

This result shows that TBE and Active Classifier have a similar performance. TBE has low significant error rate than AWE and Learn++ with 99.9% and 99.9% confidence intervals respectively. However, Active Classifier has similar performance with Learn ++ and low error rate than AWE with 99.9% confidence interval. Moreover, Figure 6.4 provides visual information about the four algorithms. TBE started with lower error rate and ended with lower error rate compared to other algorithms.

To summarize, the experimental results from Figure 6.2 to Figure 6.4 illustrate that TBE performs better on average compared to the other algorithms. TBE performed significantly better than the other algorithms on the hip-replacement data set simulated by STAGGER and on the Poker dataset.Moreover, the results show that TBE handles concept drift in a consistent manner for both small and large datasets. However, AWE and Active Classifier perform better on a noisy hip-replacement dataset that is simulated by SEA concept generator.

7.1 Validity Threats

The validity threats of the experimental results are discussed from different perspective, namely internal validity, external validity, construct validity and conclusion validity. Each validity threat of the research is discussed below.

7.1.1 Internal Validity

The research has high internal validity by manipulating the effects of the independent variable on the dependent variables of the experiment. The effect of the independent variables on the classification performance is clearly identified. Moreover, to avoid the instrumentation problem an identical data collection and recording technique is used for both the primary and the secondary data. However, since the sample size collected is small a synthetic data is generated based on the hospital real data to make the research valid and introduce the different types of concept drift from the sample data.

7.1.2 External Validity

The external validity of the research is assessed from a different point of view, that is the ability to generalize the research to-and-across population, setting, time and outcome, and treatment variations. The sample data collected from Blekinge hospital is not large enough to drive generalization. Thus, additional real-world datasets and synthetic hospital datasets are used for the generalization purposes. Similarly, the experiment is conducted in different setting, i.e different domain data, to increase the validity of the research. The experiment has high temporal validity that the treatment works independent of the time. The experiment does not have a high validity in outcome because there are different types of performance evaluators in machine learning that was not checked in the experiment.

7.1.3 Construct Validity

The construct validity is concerned about the relation between the theoretical study and the observations. The research must confirm the relation between the cause and the effect is causal. First, the treatment reflects the construct of the cause well. Secondly, outcome reflects the construct of the effect well (Wohlin et al., 2000, Demšar, 2006). However, in the research all possible variables that may affect the dependent variable are not investigated deeply which may imply the experiment does not have high construct validity. The confounding factor of the research is the occurrence of noise in the sample data. Noise produces an effect on the outcome that makes it impossible to distinguish concept drift from noise. Thus, a mechanism to handle noise is proposed as future work.

7.1.4 Conclusion Validity

Statistical tests are used for hypotheses testing to determine whether the relationship between the independent and dependent variables is significant or not. As a result, the research drew high conclusion validity from the experiment by testing the significance of the null hypotheses tests. The highest power of the statistical test is used while evaluating the four algorithms in order to increase the probability of revealing a true pattern in data. Thus, the hypotheses test has a low risk to draw a wrong hypothesis.

8 Conclusion and Future Work

The research focused on the problem of prediction models losing their prediction performance in dynamic environments, such as surgery referral from primary to secondary health cares. The research began by hypothesizing that surgery prediction can be improved by managing concept drift properly.

The occurrence of concept drift in hip-replacement dataset caused a sudden decrease of prediction performance. On Figure 6.1-(a), the prediction performance starts to decrease at the 5-th batch. Similarly, the hypotheses test result shows that concept drift has a negative impact on classification performance of surgery prediction.

A classification performance is ensured when training and test data are drawn independently from an identical distribution. In other words, if the distribution of examples across the batches is similar, classifiers' performance should not decrease. However, the results show that the error rate of the algorithm increases as the number of examples increases. Furthermore, the hypotheses tests showed that a significant variation in classification performance occurs when the concept it learned is changed.

The negative effect of concept drift can be reduced through concept drift handling algorithms. Subsequently, a concept drift handling algorithm that manages the occurrence of concept drift in surgery prediction is investigated. The existing concept drift handling algorithms, single classifiers and ensemble approaches, are reviewed to mine a sequence of hospital data with concept drift. The investigation led to the development of an algorithm called Trigger Based Ensemble, which showed a comparatively better ability to detect concept drift and adapts to changes incrementally.

The Trigger Based Ensemble, actively detects the occurrence of changes on each incoming batches and adapts to the changes incrementally. It uses the current and past predictions of classifiers combined with dynamically updated voting weights. It is assumed that adding an active detector reduces the chance of ensemble classifiers suffering from outvoting, the growth of the number of incompetent classifiers. Moreover, the ensemble size is not extremely large. Thus the contribution of the research is twofold, improving the performance of surgery prediction and developing a generic algorithm that performed comparatively better, sometimes similar, than the existing concept drift handling algorithms.

The authors of this thesis conclude that with the help of automated decision support systems, which are capable of handling concept drift in surgery prediction, general practitioners will be able to make patient referrals of better performance. This will have a great contribution in the improvement of healthcare productivity.

Future Work

The future research has three main directions: (i) optimizing the performance of TBE and developing noise handling capabilities, (ii) performing additional experiments on datasets from other domains with different characteristics to validate and optimize the performance of the Trigger-based Ensemble method, and (iii) investigating how to properly distinguish real concept drift from noise for selected domains. The additional experiments could also include time and memory consumption measurements.

References

Alippi, C., Boracchi, G., and Roveri, M. (2011). An effective just-in-time adaptive classifier for gradual concept drifts. *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1675–1682.

Bach, S. H. and Maloof, M. A. (2008). Paired Learners for Concept Drift. In *2008 Eighth IEEE International Conference on Data Mining*, pages 23–32. IEEE.

Bifet, A., Holmes, G., Pfahringer, B., and Frank, E. (2010). Fast perceptron decision tree learning from evolving data streams. In *Proceedings of the 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part II*, PAKDD'10, pages 299–310, Berlin, Heidelberg. Springer-Verlag.

Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., and Gavaldà, R. (2009). New ensemble methods for evolving data streams. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 139–148, New York, NY, USA. ACM.

Carroll, R. and Schneider, H. (1985). A note on levene's tests for equality of variances. *Statistics and Probability Letters*, 3(4):191–194. cited By (since 1996) 22.

Cesa-Bianchi, N., Gentile, C., and Zaniboni, L. (2006). Worst-case analysis of selective sampling for linear classification. *J. Mach. Learn. Res.*, 7:1205–1230.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Elwell, R. and Polikar, R. (2011). Incremental learning of concept drift in nonstationary environments. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 22(10):1517–31.

Fan, W. and Streamminer, W. D. (2004). Streamminer: A classifier ensemble-based engine to mine concept-drifting data streams.

Forman, G. (2006). Tackling concept drift by temporal inductive transfer. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 252–259, New York, NY, USA. ACM.

Furuhata, M., Mizuta, T., and So, J. (2010). Paired Evaluators Method to Track Concept Drift: An Application for Hedge Funds Operations. *2010 IEEE International Conference on Data Mining Workshops*, pages 808–815.

Gama, J., Medas, P., Castillo, G., and Rodrigues, P. (2004a). Learning with drift detection. In *In SBIA Brazilian Symposium on Artificial Intelligence*, pages 286–295. Springer Verlag.

Gama, J., Medas, P., Castillo, G., and Rodrigues, P. (2004b). Learning with drift detection. In Bazzan, A. and Labidi, S., editors, *Advances in Artificial Intelligence – SBIA 2004*, volume 3171 of *Lecture Notes in Computer Science*, pages 286–295. Springer Berlin Heidelberg.

Goodacre, R., Vaidyanathan, S., Dunn, W., Harrigan, G., and D.B.a, K. (2004). Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends in Biotechnology*, 22(5):245–252. cited By (since 1996) 396.

Hulten, G., Spencer, L., and Domingos, P. (2001). Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 97–106, New York, NY, USA. ACM.

Jose, M. B.-G., Campo-Avila, J. D., Fidalgo, R., Bifet, A., Gavalda, R., and Morales-bueno, R. (2006). Early drift detection method. In *ECML/PKDD 2006, Workshop on Knowledge Discovery from Data Streams*, KDD '06, pages 77–86, New York, NY, USA. ACM.

Karnick, M., Ahiskali, M., Muhlbaier, M., and Polikar, R. (2008). Learning concept drift in nonstationary environments using an ensemble of classifiers based approach. In *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 3455 –3462.

Katakis, I., Tsoumakas, G., and Vlahavas, I. (2009). Tracking recurring contexts using ensemble classifiers: an application to email filtering. *Knowledge and Information Systems*, 22(3):371–391.

Kelly, M. G., Hand, D. J., and Adams, N. M. (1999). The impact of changing populations on classifier performance. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 367–371, New York, NY, USA. ACM.

Khaing, H. W. (2011). Data mining based fragmentation and prediction of medical data. In *Computer Research and Development (ICCRD), 2011 3rd International Conference on*, volume 2, pages 480 –485.

Klenner, M. and Hahn, U. (1994). Concept versioning: A methodology for tracking evolutionary concept drift in dynamic concept systems. In *In Proc. of ECAI 1994*, pages 473–477. Wiley.

Klinkenberg, R. (2004). Learning drifting concepts: Example selection vs. example weighting. *Intell. Data Anal.*, 8(3):281–300.

Klinkenberg, R. and Joachims, T. (2000). Detecting concept drift with support vector machines. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 487–494, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Klinkenberg, R. and Renz, I. (1998). Adaptive information filtering: Learning in the presence of concept drifts. In *Workshop Notes of the ICML/AAAI-98 Workshop Learning for Text Categorization*, pages 33–40. AAAI Press.

Kolter, J. Z. and Maloof, M. A. (2003). Dynamic Weighted Majority : A New Ensemble Method for Tracking Concept Drift. *Learning*.

Kolter, J. Z. and Maloof, M. A. (2007). Dynamic weighted majority: An ensemble method for drifting concepts. *J. Mach. Learn. Res.*, 8:2755–2790.

Kothari, C. (2008). *Research Methodology: Methods and Techniques*. New Age International Limited.

Law, Y.-N. and Zaniolo, C. (2005). An adaptive nearest neighbor classification algorithm for data streams. In Jorge, A., Torgo, L., Brazdil, P., Camacho, R., and Gama, J., editors, *Knowledge Discovery in Databases: PKDD 2005*, volume 3721 of *Lecture Notes in Computer Science*, pages 108–120. Springer Berlin / Heidelberg.

Magoulas, G. D. and Prentza, A. (2001). Machine learning in medical applications. In *Machine Learning and Its Applications, Advanced Lectures*, pages 300–307, London, UK. Springer-Verlag.

Markou, M. and Singh, S. (2003). Novelty detection: A review - part 1: Statistical approaches. *Signal Processing*, 83:2003.

Masud, M. M., Chen, Q., Khan, L., Aggarwal, C., Gao, J., Han, J., and Thuraisingham, B. (2010). Addressing Concept-Evolution in Concept-Drifting Data

Streams. *2010 IEEE International Conference on Data Mining*, pages 929–934.

McAllester, D. A. (1998). Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, pages 230–234, New York, NY, USA. ACM.

McKenzie, J., Pinger, R., and Kotecki, J. (2011). *An Introduction to Community Health*. Jones & Bartlett Learning.

Minku, L., White, A., and Yao, X. (2010). The impact of diversity on online ensemble learning in the presence of concept drift. *Knowledge and Data Engineering, IEEE Transactions on*, 22(5):730 –742.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York, 3 edition.

Molin, R. and Johansson, L. (2005). *Swedish Health Care in an International Context: A Comparison of Care Needs, Costs, and Outcomes*. Swedish Association of Local Authorities and Regions.

Nishida, K. and Yamauchi, K. (2007a). Adaptive classifiers-ensemble system for tracking concept drift. In *Machine Learning and Cybernetics, 2007 International Conference on*, volume 6, pages 3607 –3612.

Nishida, K. and Yamauchi, K. (2007b). Detecting concept drift using statistical testing. In *Proceedings of the 10th international conference on Discovery science*, DS'07, pages 264–269, Berlin, Heidelberg. Springer-Verlag.

Nunez, M., Fidalgo, R., and Morales, R. (2007). Learning in environments with unknown dynamics: Towards more robust concept learners.

Ouyang, Z. Z. Z. G. Y. W. T. (2011). Study on the classification of data streams with concept drift. *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, pages 1673–1677.

Patist, J. P. (2007). Optimal window change detection. In *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*, pages 557 –562.

Perner, P. (2009). Concepts for novelty detection and handling based on a case-based reasoning process scheme. *Engineering Applications of Artificial Intelligence*, 22(1):86 – 91.

Persson, M. and Lavesson, N. (2009). Identification of surgery indicators by mining hospital data: A preliminary study. In *Database and Expert Systems Application, 2009. DEXA '09. 20th International Workshop on*, pages 323 –327.

Persson, M., Lavesson, N., and Magnusson, M. (2010). Automatic identification of surgery indicators. In *Database Technology for Life Sciences and Medicine*, pages 295–318.

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.*, 1(1):81–106.

Rodríguez, J. J. and Kuncheva, L. I. (2008). Combining online classification approaches for changing environments. In *Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, SSPR & SPR '08, pages 520–529, Berlin, Heidelberg. Springer-Verlag.

Schlimmer, J. and Granger, R. (1986a). Beyond incremental processing: Tracking concept drift. In Kehler, T. and Rosenschein, S., editors, *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 502–507. San Francisco, CA: Morgan Kaufmann.

Schlimmer, J. C. and Granger, R. H. (1986b). Incremental learning from noisy data. *Machine Learning*, 1:317–354. 10.1023/A:1022810614389.

Scholz, M. and Klinkenberg, R. (2007). Boosting classifiers for drifting concepts. *Intell. Data Anal.*, 11(1):3–28.

Stiglic, G. and Kokol, P. (2011). Interpretability of sudden concept drift in medical informatics domain. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pages 609–613, Washington, DC, USA. IEEE Computer Society.

Street, W. N. and Kim, Y. (2001). A streaming ensemble algorithm (sea) for large-scale classification. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 377–382, New York, NY, USA. ACM.

Su, B., Shen, Y.-D., and Xu, W. (2008). Modeling concept drift from the perspective of classifiers. In *Cybernetics and Intelligent Systems, 2008 IEEE Conference on*, pages 1055 –1060.

Syed, N. A., Liu, H., and Sung, K. K. (1999). Handling concept drifts in incremental learning with support vector machines. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 317–321, New York, NY, USA. ACM.

Tsymbal, A. (2004). The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*.

Tsymbal, A., Pechenizkiy, M., Cunningham, P., and Puuronen, S. (2006). Handling local concept drift with dynamic integration of classifiers: Domain of antibiotic resistance in nosocomial infections. In *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems*, CBMS '06, pages 679–684, Washington, DC, USA. IEEE Computer Society.

Wang, H., Fan, W., Yu, P. S., and Han, J. (2003). Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 226–235, New York, NY, USA. ACM.

Wang, H., Yin, J., Pei, J., Yu, P. S., and Yu, J. X. (2006). Suppressing model overfitting in mining concept-drifting data streams. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, page 736.

Wang, S., Schlobach, S., and Klein, M. (2011). Concept drift and how to identify it. *Web Semantics Science Services and Agents on the World Wide Web*, 9(3):247–265.

Wenerstrom, B. and Giraud-Carrier, C. (2006). Temporal data mining in dynamic feature spaces. In *Data Mining, 2006. ICDM '06. Sixth International Conference on*, pages 1141 –1145.

Widmer, G. and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23:69–101. 10.1023/A:1018046501280.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington, MA, 3 edition.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2000). *Experimentation in software engineering: an introduction*. Kluwer Academic Publishers, Norwell, MA, USA.

Wu, J., Ding, D., Hua, X.-S., and Zhang, B. (2005). Tracking concept drifting with an online-optimized incremental learning framework. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, MIR '05, pages 33–40, New York, NY, USA. ACM.

Xiang, C., Chen, M., and Wang, H. (2009). An ensemble method for medicine best selling prediction. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on*, volume 1, pages 100 –103.

Yang, Y., Webb, G. I., and Wu, X. (2010). Discretization methods. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 101–116. Springer US.

Zhang, P., Zhu, X., and Shi, Y. (2008). Categorizing and mining concept drifting data streams. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 812–820, New York, NY, USA. ACM.

Zhenzheng, O., Zipeng, Z., Yuhai, G., and Tao, W. (2011). Study on the classification of data streams with concept drift. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, volume 3, pages 1673 –1677.

Zhu, Q., Hu, X., Zhang, Y., Li, P., and Wu, X. (2010). A double-window-based classification algorithm for concept drifting data streams. In *Granular Computing (GrC), 2010 IEEE International Conference on*, pages 639 –644.

Zliobaite, I., Bifet, A., Pfahringer, B., and Holmes, G. (2011). Active learning with evolving streaming data. In *Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases - Volume Part III*, ECML PKDD'11, pages 597–612, Berlin, Heidelberg. Springer-Verlag.

Appendix A Experimental Results

Table A.1: Comparison of accuracy(%) between concept drift handling algorithms on hip-replacement data set simulated by STAGGER concept generator

Batch	Instances	TBE(Rank[a])	AWE(Rank)	Learn++(Rank)	Active Classifier(Rank)
1	500	**80.00**(1)	14.00(3.5)	14.00(3.5)	46.00(2)
2	1,000	**87.80**(1)	50.40(3)	13.60(4)	72.80(2)
3	1,500	**90.80**(1)	65.20(3)	12.93(4)	81.40(2)
4	2,000	**92.15**(1)	73.90(3)	12.35(4)	84.30(2)
5	2,500	**92.92**(1)	79.12(3)	11.96(4)	85.36(2)
6	3,000	**93.50**(1)	82.60(3)	12.30(4)	85.47(2)
7	3,500	**93.89**(1)	85.09(3)	12.20(4)	85.89(2)
8	4,000	**94.13**(1)	86.95(2)	12.28(4)	86.05(3)
9	4,500	**94.42**(1)	88.40(2)	11.69(4)	86.82(3)
10	5,000	**94.60**(1)	89.56(2)	11.58(4)	87.08(3)
11	5,500	**94.71**(1)	90.51(2)	11.47(4)	87.31(3)
12	6,000	**94.88**(1)	91.30(2)	11.37(4)	87.52(3)
13	6,500	**95.06**(1)	91.97(2)	11.45(4)	87.52(3)
14	7,000	**95.17**(1)	92.54(2)	11.31(4)	87.73(3)
15	7,500	**95.23**(1)	93.04(2)	11.16(4)	87.95(3)
16	8,000	**95.34**(1)	93.48(2)	11.14(4)	88.03(3)
17	8,500	**95.45**(1)	93.86(2)	11.04(4)	88.18(3)
18	9,000	**95.47**(1)	94.20(2)	11.04(4)	88.21(3)
19	9,500	**95.51**(1)	94.51(2)	11.04(4)	88.25(3)
20	10,000	**95.56**(1)	94.78(2)	11.08(4)	88.25(3)
Average Rank		**1**	**3.02**	**3.98**	**2.65**

[a] The rank is computed based on accuracy

Table A.2: Comparison of accuracy(%) between concept drift handling algorithms on hip-replacement data set simulated by SEA concept generator

Batch	Instances	TBE(Rank[a])	AWE(Rank)	Learn++(Rank)	Active Classifier(Rank)
1	5,00	58.60(2)	36.60(3.5)	36.60(3.5)	**60.40**(1)
2	1,000	**62.10**(1)	52.50(3)	35.60(4)	60.70(2)
3	1,500	61.87(3)	62.93(2)	36.53(4)	**67.93**(1)
4	2,000	67.25(3)	69.40(2)	36.55(4)	**72.85**(1)
5	2,500	70.32(3)	72.72(2)	36.44(4)	**74.80**(1)
6	3,000	73.40(3)	75.70(2)	35.93(4)	**76.77**(1)
7	3,500	75.09(3)	77.09(2)	36.34(4)	**77.80**(1)
8	4,000	76.28(3)	78.45(2)	36.53(4)	**78.83**(1)
9	4,500	77.38(3)	79.64(2)	36.00(4)	**79.67**(1)
10	5,000	78.16(3)	**80.46**(1)	35.92(4)	80.38(2[b])
11	5,500	78.69(3)	**81.07**(1)	36.07(4)	79.67(2)
12	6,000	79.05(3)	**81.70**(1)	36.25(4)	79.33(2)
13	6,500	79.54(2)	**82.18**(1)	36.18(4)	79.52(3)
14	7,000	79.87(3)	**82.60**(1)	36.40(4)	79.99(2)
15	7,500	80.28(3)	**82.96**(1)	36.28(4)	80.32(2)
16	8,000	80.70(3)	**83.40**(1)	36.24(4)	80.84(2)
17	8,500	80.98(3)	**83.66**(1)	36.31(4)	81.14(2)
18	9,000	81.17(3)	**83.83**(1)	36.36(4)	81.20(2)
19	9,500	81.37(3)	**84.04**(1)	36.26(4)	81.44(2)
20	10,000	81.60(3)	**84.23**(1)	36.28(4)	81.74(2)
Average Rank [c]		**2.80**	**1.58**	**3.98**	**1.65**

[a] The rank is assigned based classification accuracy

[b] Accuracy starts to decrease

[c] The average rank of the algorithms.

Table A.3: Comparison of accuracy(%) between Concept Drift Handling Algorithms on Poker Data set.

Batch	Instances	TBE(Rank[a])	AWE(Rank)	Learn++(Rank)	Active Classifier(Rank)
1	500	32.40(2.5)	32.40(2.5)	32.40(2.5)	**38.00**(1)
2	1,000	**52.80**(1.5)	51.50(4)	**52.80**(1.5)	52.10(3)
3	1,500	**56.26**(1)	48.19(4)	51.06(3)	52.86(2)
4	2,000	**62.00**(1)	46.45(4)	49.05(3)	50.55(2)
5	2,500	**65.64**(1)	41.76(4)	55.27(3)	56.59(2)
6	3,000	**65.26**(1)	43.46(4)	53.63(3)	59.66(2)
7	3,500	**68.25**(1)	42.17(4)	54.17(3)	58.68(2)
8	4,000	**70.35**(1)	43.10(4)	58.02(3)	61.30(2)
9	4,500	**72.00**(1)	44.24(4)	54.95(3)	63.62(2)
10	5,000	**74.02**(1)	41.12(4)	58.06(3)	64.74(2)
11	5,500	**74.61**(1)	40.80(4)	60.10(3)	65.52(2)
12	6,000	**74.35**(1)	39.46(4)	61.05(3)	65.68(2)
13	6,500	**73.87**(1)	40.98(4)	60.06(3)	66.01(2)
14	7,000	**72.64**(1)	40.34(4)	60.00(3)	65.85(2)
15	7,500	**72.65**(1)	41.08(4)	60.68(3)	66.25(2)
16	8,000	**72.33**(1)	41.80(4)	59.47(3)	67.41(2)
17	8,500	**72.81**(1)	43.77(4)	60.50(3)	67.80(2)
18	9,000	**73.38**(1)	44.33(4)	59.64(3)	67.43(2)
19	9,500	**73.64**(1)	45.23(4)	60.66(3)	68.09(2)
20	10,000	**73.74**(1)	43.84(4)	60.95(3)	67.64(2)
Average Rank[b]		**1.10**	**3.93**	**2.90**	**2.00**

[a] The rank is assigned based on accuracy.

[b] The average rank of the algorithms.

Appendix B Hypothesis Test Results

Graphs

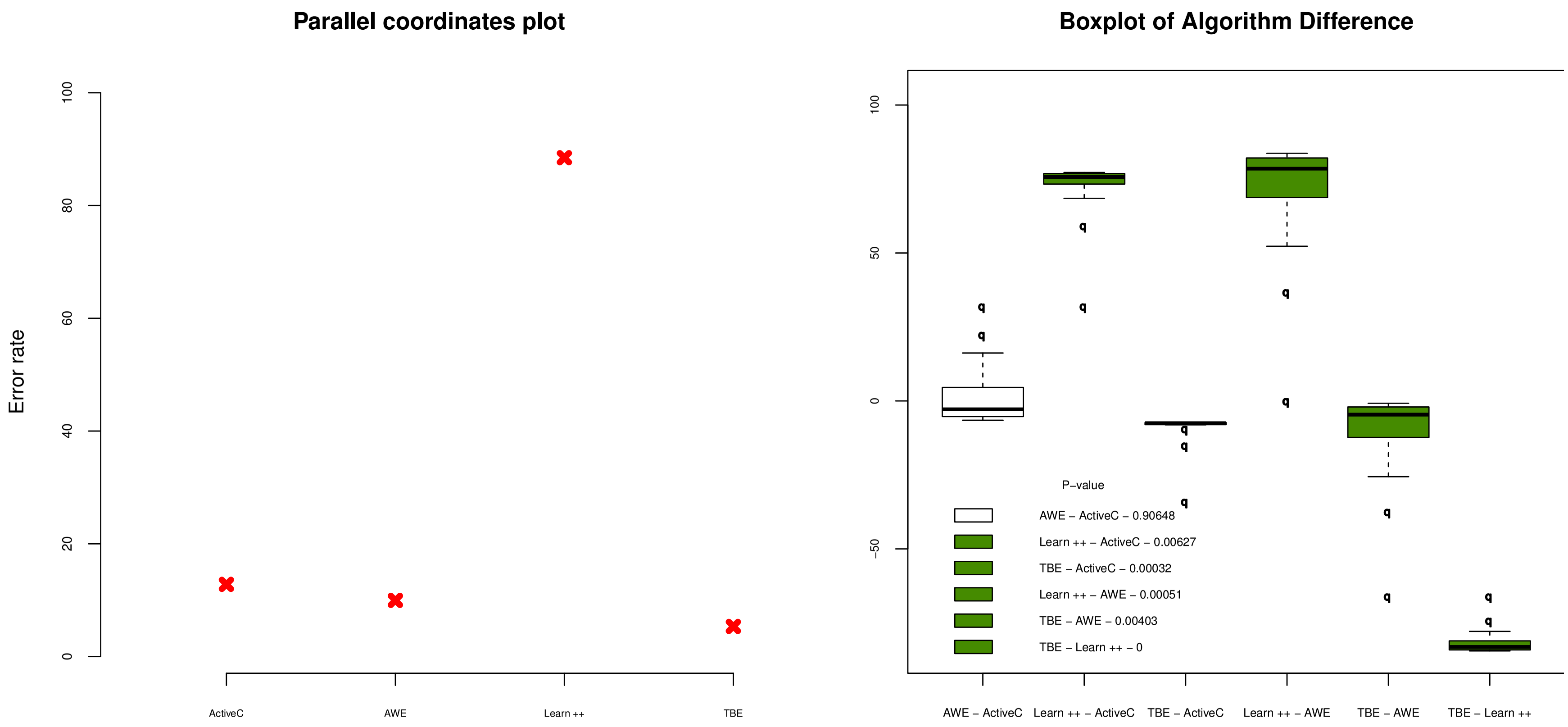


Figure B.1: Hip-replacement Dataset based on STAGGER concept generator

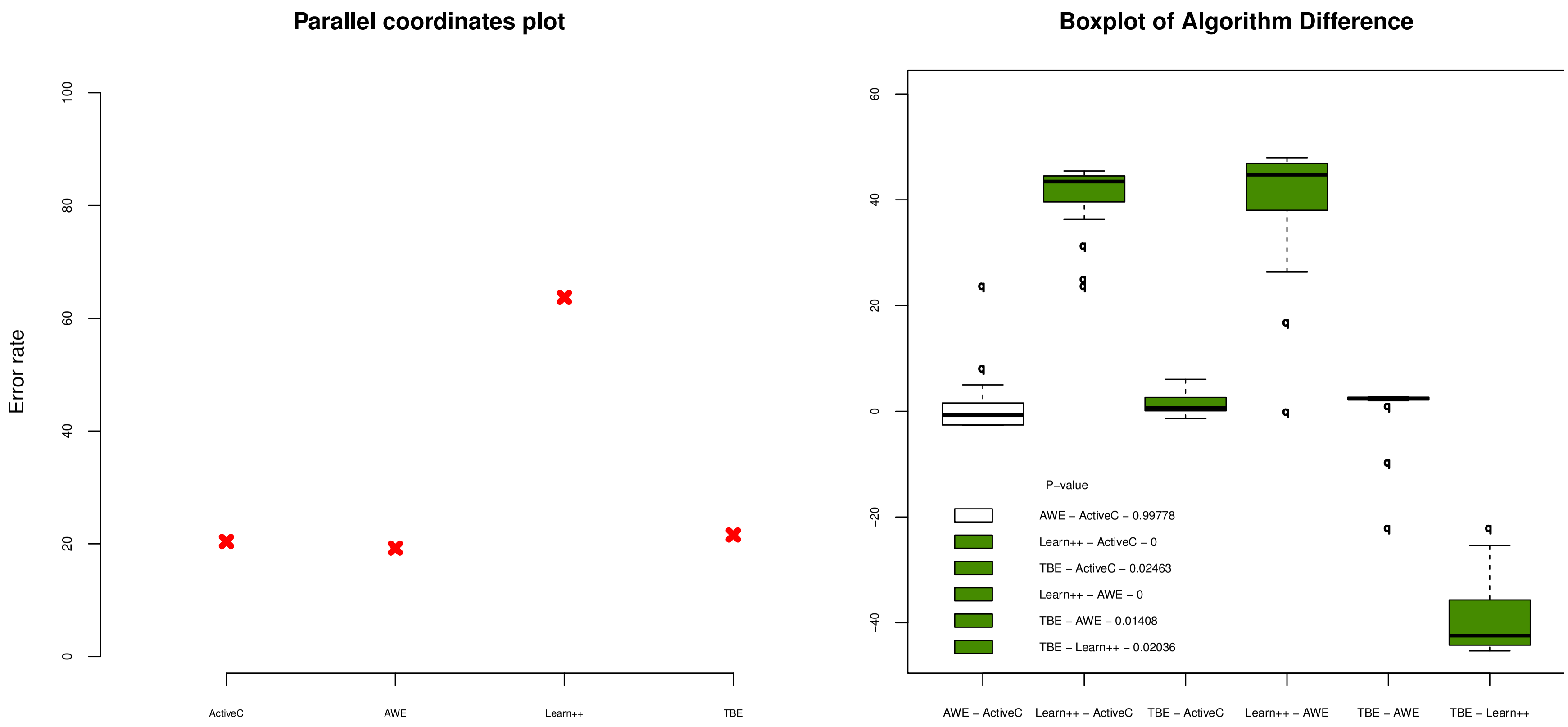


Figure B.2: Hip-replacement Dataset based on SEA concept generator

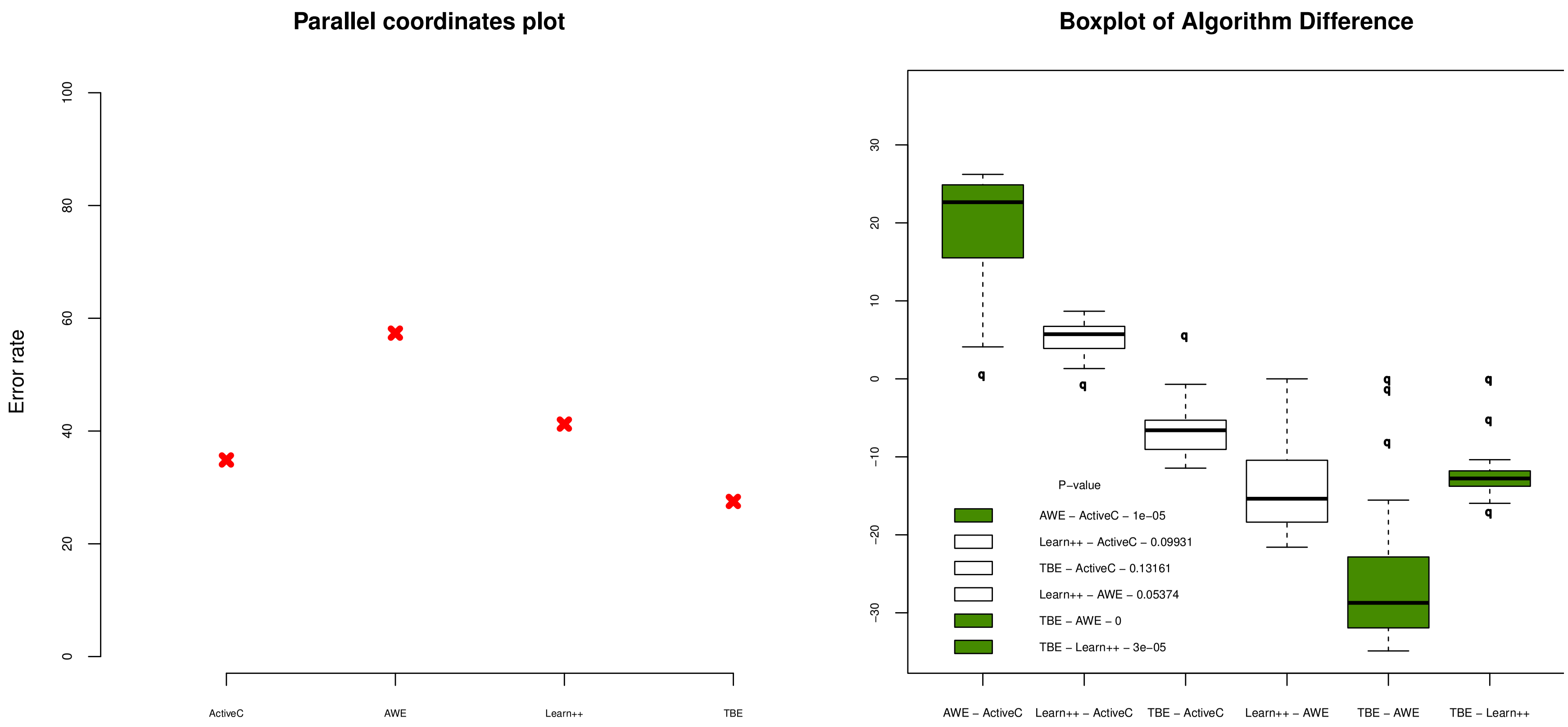


Figure B.3: Poker Dataset

Appendix C Acronyms

AWE	Accuracy-Weighted Ensemble
CD	Concept Drift
DDM	Drift Detection Method
EDDM	Early Drift Detection Method
GP	General Practitioner
PAC	Probably Approximately Correct
MOA	Massive Online Analysis
SEA	Streaming Ensemble Algorithm
TBE	Trigger Based Ensemble