BIG DATA PROCESSING ON EDUCATIONAL DATA MINING USING PYSPARK WITH JUPYTER NOTEBOOK

VINITHA A/P RAVICHANDRAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Computer Science

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2018

Dedicated to my beloved family and friends, without their understanding, supports and most of all love, the completion of this work would not have been possible.

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main supervisor, Professor Dr. Siti Mariyam Hj. Shamsudin for encouragement, guidance, critics and friendship. I am also very thankful to my co-supervisor, Dr. Shafaatunnur Hassan for her guidance, advises and motivations. Without their continuous support and interest, this thesis would not have been the same as presented here.

My fellow postgraduate students should also be recognized for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. I am truly grateful to all my family members too. Unfortunately, it is impossible to list all of them in this limited space.

ABSTRACT

The rapid advancement of the information technology brings new challenges and put new demands on our education system. The process of teaching and learning have moved from classroom to Computer Aided Learning (CAL) system. Big data technology and machine learning plays an important role in Computer Aided Learning (CAL) system due to the massive information or data generated by the system. This leads to the rapid development of data mining in education denote as Educational Data Mining (EDM). The abundance of data collected by the system can be used to analyse, predict and solve many societal issues in the education field such as improve the quality of education, predict as well as monitor educational outcomes. Effective analysing or predicting the future growth of students' performance can make the Computer Aided Learning (CAL) system a better platform for learning compared to traditional learning. Machine learning techniques were used to get reliable and accurate prediction on students' performance. Apache Hadoop has been the backbone for big data technology until the emergence of Apache Spark. However, only several researches are done on EDM using Apache Spark. In this dissertation, PySpark was be integrated with Jupyter Notebook to perform EDM on Educational Process Mining (EPM) data set. The Spark MLlib was used to compare four classification algorithms such as Logistic Regression, Naïve Bayes, Decision Tree and Random Forest to deal with EPM data set. Random Forest classifier outperformed other classifiers in Accuracy, Area Under the Precision-Recall(PR) and Area Under the Receiver Operating Characteristic (ROC) although with slightly slower Execution Time in this study. Random Forest classifier are the best classifier when dealing with EDM.

ABSTRAK

Perkembangan pesat teknologi maklumat menjurus kepada pembaharuan dan pembaikpulihan dalam sistem pendidikan. Proses pengajaran dan pembelajaran telah berubah dari sistem bilik darjah ke sistem pembelajaran daripada komputer (CAL). Data besar dan pembelajaran mesin memainkan peranan penting dalam sistem CAL disebabkan oleh maklumat besar-besaran yang dihasilkan oleh sistem. Ini membawa kepada perkembangan pesat penyiasatan data dalam sistem pendidikan yang dikenali sebagai Penyiasatan Data Pendidikan (EDM). Data yang dikumpul oleh sistem boleh digunakan untuk menganalisis, meramal dan menyelesaikan pelbagai isu berkaitan bidang pendidikan seperti meningkatkan kualiti pendidikan, meramal serta memantau hasil pendidikan. Analisis yang berkesan dapat meramal pertumbuhan masa depan prestasi pelajar dan menjadikan sistem pembelajaran daripada komputer (CAL) sebagai platform pembelajaran yang lebih baik berbanding pembelajaran tradisional. Teknik pembelajaran mesin digunakan untuk mendapatkan ramalan yang boleh dipercayai dan tepat terhadap prestasi pelajar. Apache Hadoop telah menjadi tulang belakang untuk teknologi data besar sehingga kemunculan Apache Spark. Walau bagaimanapun, hanya beberapa penyelidikan yang dilakukan untuk EDM menggunakan Apache Spark. Dalam disertasi ini, PySpark telah diintegrasikan dengan Jupyter Notebook untuk melaksanakan EDM pada set data Educational Process Mining (EPM). Spark MLlib digunakan untuk membandingkan empat algoritma klasifikasi iaitu Regresi Logistik, Naïve Bayes, Pohon Keputusan dan Random Forest apabila berurusan dengan set data EPM. Klasifikasi Random Forest memberi keputusan yang lebih baik berbanding klasifikasi lain dan ia amat sesuai digunakan apabila berurusan dengan EDM.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
AML	-	Amazon Machine Learning
ASF	-	Apache Software Foundation
CAL	-	Computer Aided Learning
DM	-	Data Mining
EDM	-	Educational Data Mining
EPM	-	Educational Process Mining
HDFS	-	Hadoop Distributed File System
ICT	-	Information and Communication Technologies
KDDs	-	Knowledge Discovery in Databases
ML	-	Machine Learning
MLlib	-	Machine Learning Library
PR	-	Precision-Recall
RDDs	-	Resilient Distributed Datasets
ROC	-	Receiver Operating Characteristic
SVM	-	Support Vector Machine
YARN	-	Yet Another Resource Negotiator

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

The rapid advancement of the information technology brings new challenges and put new demands on our education system. The process of teaching and learning have moved from classroom to Computer Aided Learning (CAL) system.

It started with CD-ROM that only provide a set of programmed instructions used for educational purpose. Unfortunately, students still have to depend on their teachers since there is no interaction between student and the system. Besides, there is no personalization since the system is meant for general educational purpose. An effective Computer Aided Learning (CAL) system should present material in stimulating way, give questions relevant to individual students' knowledge area, interest, background and skills. It should have also provided platform to ask questions and give response whether it is between students or student and the system.

Big data technology and machine learning plays an important role in Computer Aided Learning (CAL) system due to the massive information or data generated by the system. This leads to the rapid development of data mining in education denote as Educational Data Mining (EDM). The abundance of data collected by the system can be used to analyse, predict and solve many societal issues in the education field such as improve the quality of education, predict as well as monitor educational outcomes.

Education data from schools, colleges, universities or even through eLearning environment/simulator are used to perform research in EDM. EDM is performed by applying various data mining techniques available such as classification and clustering to do prediction, discover hidden patterns and adjust program actions accordingly. It is done even more easier with the help of machine learning. The discovered knowledge can then be used by the educational institutes to overcome the problem of low grades and poor performance of students. EDM is important so that measures can be taken to improve the overall academic performance.

Enhancing learner models and domain models, studying the pedagogical support provided by learning software and conducting scientific research on learning and learners are the suggested four key areas of EDM application based on Baker, *et* al. (2009 & 2010). While, Castro, *et* al., (2007) categorized EDM jobs into four distinctive areas which are applications that deal with the assessment of students learning execution, course adjustment and learning proposals to customize students learning based on individual student's behaviour, developing a strategy to assess materials in online courses, approaches that utilize input from leaners and tutors in e-learning courses, and detection models for revealing student learning behaviours.

There are plenty of open source machine learning frameworks to deal with big data such as Apache Hadoop, Apache Singa, Amazon Machine Learning (AML), Apache Spark and many more. Each one has its own advantages and disadvantages. Apache Spark is the new future of big data technology so for this dissertation Apache Spark MLlib integrated with Jupyter Notebook will be used to do the EDM on Educational Process Mining (EPM) data set that was built from the recordings of 115 subjects' activities through a logging application while learning with an educational simulator.

1.2 Background of the Problem

Educational simulator allows students to learn materials through specialized browsers by allowing them to solve various problems with different levels of difficulty. Learners get to verify their knowledge and test understanding on particular subject. It is crucial for learners to do self-assessment because it can develop the ability to identify their strengths and weaknesses on a specific topic. By engaging with different level of difficulty, students can focus their study efforts on the particular area they believe needs improvement.

Besides that, educational institutes need to improve their education quality. It can only be increased by providing learners with the appropriate learning materials, monitoring learners' performance, predicting the final grade of students and providing solutions to encounter dropout issues before it is too late. This is when Educational Data Mining emerges to improve education field.

The growth and development of nation together with young generation are influenced by the effectiveness of educational system and quality of education provided by the educational institutions. These gains the interest of researchers and institutions to invest money as well as effort on Education Data Mining (EDM).

A study was conducted on student academic performance of the Dr. R.M.L. Awadh University, Faizabad, India where a total number of 300 (226 males, 74 females) students of BCA (Bachelor of Computer Applications) course from five colleges in the 2009-2010 examination were selected (Bhardwaj & Pal, 2011). Students' performance was predicted to identify the difference between fast learners and slow learners using the Bayesian Classification. The result can be helpful for the teachers to improve the students' performance by giving special attention on the slow learners.

Besides that, a study on student academic performance at S.G.R. Education Foundation’s College of Engineering and Management was conducted by using sampling data taken for first year engineering in the year 2009-10 and 2010-11.

Engineering students’ past performance data were applied to the classification technique and decision tree algorithms to predict their performance. Confusion matrix was generated and analysed to identify the students’ fail records. The accuracy of the model improves their attributes for better performance.

Analysing or predicting the future growth of student’s performances using various approaches and techniques is one of the rising trends in Educational Data Mining (EDM). Enhancing the student’s performance in their academic will directly improve the educational process but there is where the challenge lies within, especially when have to deal with massive data.

Apache Hadoop has been the backbone for big data technology by offering distributed storage, superior scalability, ideal performance and fault tolerance. However, it requires a lot of processing time thereby increase latency. It also only support batch processing and no real-time data processing. Hadoop does not fit for small data. In security concern, Hadoop is missing encryption at storage and network levels. These leads to the emergence of Apache Spark and numerous researches done to conduct EDM with Apache Spark due to its advantages over Apache Hadoop.

1.3 Problem Statement

A wide range of researches have been conducted on Education Data Mining (EDM) but there are only a few done on EDM using big data technology. This is the biggest challenge because there are very less resources to refer as a guide for this dissertation. On the other hand, there is no EDM research conducted using Educational Process Mining (EPM) data set with Apache Spark or Apache Spark MLlib. The prior researchers who worked on Apache Spark concentrated on its framework and did not integrate web applications for better User Interaction (UI). This dissertation presents the use of PySpark with Jupyter Notebook to do EDM on Educational Process Mining (EPM) data set.

1.4 Dissertation Goal and Objectives

Effective analysing or predicting the future growth of students' performance can make the Computer Aided Learning (CAL) system a better platform for learning compared to traditional learning. Four machine learning techniques will be used to perform classification on the EPM data set and students' performance will be predicted. Based on the dissertation goal mentioned above, the objectives of this research are as follows:

1. To integrate Spark Python API (PySpark) with Jupyter Notebook and provide User Interface (UI) to assist non-technical people.
2. To perform Educational Data Mining (EDM) on Educational Process Mining (EPM) data set by doing classification using four different Apache Spark Machine Learning (Spark MLlib) algorithms. The classification results will be used to do prediction on students' performance.
3. To evaluate the effectiveness of Apache Spark Machine Learning (Spark MLlib) algorithms on handling large data set in Educational Data Mining (EDM).

1.5 Dissertation Scope and Assumptions

Researches in EDM are growing drastically over the decade since there are still many improvements and initiatives need to be taken care for better education system.

However, EDM with big data tool such as Apache Spark is still new in the research field. Therefore, this research is set within certain scopes that focuses on:

1. PySpark will be used for this dissertation, that means Python programming language will be used even though Apache Spark support multiple languages such as Scala, R and Java.
2. Spark MLlib will be used to perform EDM on Educational Process Mining (EPM) data set.
3. A web application called Jupyter Notebook will be integrated with Pyspark to perform the EDM.

1.6 Significance of Dissertation

Using PySpark in this dissertation will significantly reduce the execution time and storage to perform EDM on Educational Process Mining (EPM) data set compared to using Apache Hadoop or any other data mining tools. It is the first study to conduct EDM on Educational Process Mining (EPM) data set using PySpark integrated with Jupyter Notebook.

1.7 Organization of the Thesis

The thesis contains five chapters. Chapter 1 starts with introduction, problem background and statement, dissertation goal and objectives, scope, importance of the dissertation. Then, chapter 2 gives the literature review on big data technology, Education Data Mining (EDM), Deeds (Digital Electronics Education and Design Suite), machine learning concepts, big data tools focused on Apache Hadoop and Apache Spark, the comparison of both big data tools in EDM in prior researches, web application that will be used for this dissertation and other related issues that help to clarify, understand and solve the problems in this research. The methodologies of the

EDM on Educational Process Mining (EPM) data set using PySpark integrated with Jupyter Notebook are discussed in Chapter 3. Discussion based on the dissertation outcomes are done in Chapter 4. Lastly, dissertation conclusion and future work recommendations are presented in Chapter 5.

REFERENCES

Amra, I. A., and Maghari, A. Y. (2017). Students performance prediction using KNN and Naïve Bayesian. *2017 8th International Conference on Information Technology (ICIT).* doi:10.1109/icitech.2017.8079967

Andersson, L. (2016). Natural Language Processing In A Distributed Environment : A comparative performance analysis of Apache Spark and Hadoop MapReduce (Dissertation). Retrieved September 15, 2017, from http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-126865

Apache Hadoop 2.8.0 – Apache Hadoop YARN. (n.d.). Retrieved September 15, 2017, from https://hadoop.apache.org/docs/r2.8.0/hadoop-yarn/hadoop-yarn-site/YARN.html

Baker, R.S.J.d. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining,* vol. 1, no. 1, pp. 3–17.

Baker, R.S.J.d. (2010). “Data Mining for Education, In McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopedia of Education (3rd edition),* vol. 7, pages 112-118.

Baker, R.S.J.d., Hershkovitz, A., Rossi, L.M., Goldstein, A.B., and Gowda, S.M. (2013). Predicting Robust Learning With the Visual Form of the Moment-by-Moment Learning Curve. *Journal of the Learning Sciences*, 22(4), pages 639-666.

Beck, J. E., and Woolf, B. P. (2000). High-Level Student Modeling with Machine Learning. *Intelligent Tutoring Systems Lecture Notes in Computer Science,* pages 584-593. doi:10.1007/3-540-45108-0_62

Bienkowski, M., Feng, M., and Means, B. (2012). Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief, Washington, D.C., 2012. *U.S. Department of Education, Office of Educational Technology*.

Bhardwaj, B. K., and Pal, S. (2011). Data Mining: A prediction for performance improvement using classification. *(IJCSIS) International Journal of Computer Science and Information Security*, Vol. 9, No.4.

Brundin, M. (2016). Data Stream Queries to Apache SPARK (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-301326

Brusilovsky, P. and Peylo, C. (2003). Adaptive and intelligent Web-based educational systems, *Int. J. Artif. Intell. Edu.* , vol. 13, nos. 2–4, pages 159–172.

Castro, F., Vellido, A., Nebot, A., and Mugica, F. (2007). Applying data mining techniques to e-learning problems. *Evolution of teaching and learning paradigms in intelligent environment*, Springer, pages 183–221.

Dean, J. and Ghemawat, S. (n.d.). MapReduce: Simplified Data Processing on Large Clusters. Static.googleusercontent.com. Retrieved September 15, 2017, from http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009). Predicting students drop out: a case study. In T. Barnes, M. Desmarais, C. Romero, & S.Ventura (Eds.), *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009*, July 1-3, 2009. Cordoba, Spain, pages 41-50.

Donzellini, G. (2018). Digital Electronics Deeds. Retrieved from https://www.digitalelectronicsdeeds.com/deeds.html#

García, E., Romero, C., Ventura, S. and de Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education*, 14(2), pages 77-88.

GraphX | Apache Spark. (n.d.). Retrieved September 15, 2017, from https://spark.apache.org/graphx/

Gopalani, S. and Arora, R. (2015). Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means. *International Journal of Computer Applications*, 113(1), pages 8-11.

Gu, L., and Li, H. (2013). Memory or Time: Performance Evaluation for Iterative Operation on Hadoop and Spark. *2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE*

International Conference on Embedded and Ubiquitous Computing. doi:10.1109/hpcc.and.euc.2013.106

Guo, B., Zhang, R., Xu, G., Shi, C., and Yang, L. (2015). Predicting Students Performance in Educational Data Mining. *2015 International Symposium on Educational Technology*.

Gupta, A., Shaikh, F., Singh, G., N., and Blog, G. (2016, October 14). Comprehensive Introduction - Apache Spark, RDDs & Dataframes (PySpark). Retrieved September 01, 2017, from https://www.analyticsvidhya.com/blog/2016/09/comprehensive-introduction-to-apache-spark-rdds-dataframes-using-pyspark/

Harvey, C. (n.d.). 50 Top Open Source Tools for Big Data - Datamation. Datamation.com. Retrieved September 15, 2017, from https://www.datamation.com/data-center/50-top-open-source-tools-for-big-data-1.html

HDFS Architecture Guide. (n.d.). Retrieved September 15, 2017, from https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Hess, K. (2016). New Hadoop survey makes big data predictions for 2016. Retrieved October 30, 2017, from http://www.zdnet.com/article/new- hadoop-survey-makes-big-data-predictions-for-2016/

Hogo, M. (2010). Evaluation of e-learning systems based on fuzzy clustering models and statistical tools. *Expert Systems with Applications,* 37(10), pages 6891-6903.

Jia, J., and Mareboyana, M. (2014). Predictive Models for Undergraduate Student Retention Using Machine Learning Algorithms. *Transactions on Engineering Technologies,* pages 315-329. doi:10.1007/978-94-017-9115-1_24

Kabra, R.R., and Bichkar, S. (2011). Performance Prediction of Engineering Students using Decision Trees. *International Journal of Computer Applications*, Volume 36– No.11, pages 0975– 8887.

Karau, H., Konwinski, A., Wendell, P., and Zaharia, M. (2015). Learning Spark. *Beijing: O'Reilly*.

Laney, D. (2001). 3D Data Management:Controlling Data Volume, Velocity, and Variety. *Application Delivery Strategies Metagroup*.

Lin, J. and Dyer, C. (2010). Data-Intensive Text Processing with MapReduce. Lintool.github.io. Retrieved September 25, 2017, from https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf

Michalski, R., Carbonell, J., and Mitchell, T. (2013). Machine Learning An Artificial Intelligence Approach. *Berlin: Springer Berlin*.

MLlib | Apache Spark. (n.d.). Retrieved September 15, 2017, from https://spark.apache.org/mllib/

Moore, J. L., Dickson-Deane, C., & Galyen, K. (2011). E-Learning, online learning, and distance learning environments: Are they the same?. *The Internet and Higher Education*, 14(2), pages 129-135.

Musso, M., Kyndt, E., Cascallar, E. and Dochy, F. (2013). Predicting general academic performance and identifying the differential contribution of participating variables using artificial neural networks. *Frontline Learning Research*, 1(1).

Ogunde A. O. and Ajibade D. A. (2014). A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm. *Journal of Computer Science and Information Technology*, vol.2, no. 1, pages 21–46.

Pan, S. (2016). The Performance Comparison of Hadoop and Spark. *Culminating Projects in Computer Science and Information Technology*. Paper 7.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41(4), pages 1432-1462.

Project Jupyter. (n.d.). Retrieved October 27, 2017, from http://jupyter.org/

Python Programming Guide. (n.d.). Retrieved September 01, 2017, from https://spark.apache.org/docs/0.9.0/python-programming-guide.html

Sin, K., and Muthu, L. (2015). Application Of Big Data In Education Data Mining And Learning Analytics – A Literature Review. *ICTACT Journal on Soft Computing*, 05(04), pages 1035-1049. doi:10.21917/ijsc.2015.0145

Spark Streaming | Apache Spark. (n.d.). Retrieved September 15, 2017, from https://spark.apache.org/streaming/

Spark Streaming - Spark 2.2.0 Documentation. (n.d.). Retrieved September 15, 2017, from https://spark.apache.org/docs/latest/streaming-programming-guide.html

Ranjan, J. and Malik, K. (2007). Effective educational process: a data-mining approach. *VINE*, 37(4), pages 502-515.

Ratnapala, I. P., Ragel, R. G., and Deegalla, S. (2014). Students behavioural analysis in an online learning environment using data mining. *7th International Conference on Information and Automation for Sustainability*. doi:10.1109/iciafs.2014.7069609

Romero, C. and Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pages 601–618.

Romero, C., López, M., Luna, J. and Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68, pages 458-472.

Taylor, B. (2016). Apache Spark rises to become most active open source project in big data. Retrieved August 20, 2017, from https://www.techrepublic.com/article/apache-spark-rises-to-become-most-active-open-source-project-in-big-data/

UCI Machine Learning Repository: Data Set. (n.d.). Educational Process Mining (EPM): A Learning Analytics Data Set.Retrieved from https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+(EPM)%3A+A+Learning+Analytics+Data+Set

Vavilapalli, V. K., Seth, S., Saha, B., Curino, C., Omalley, O., Radia, S., . . . Shah, H. (2013). Apache Hadoop YARN. *Proceedings of the 4th annual Symposium on Cloud Computing - SOCC 13*. doi:10.1145/2523616.2523633

Welcome to Apache™ Hadoop®!. (n.d.). Retrieved September 15, 2017, from ` http://hadoop.apache.org/

What is Hadoop?. (n.d.). Retrieved September 15, 2017, from https://www.sas.com/en_my/insights/big-data/hadoop.html

What is Apache Spark? (n.d.). Retrieved September 01, 2017, from https://databricks.com/spark/about

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., Mccauley, M., J. Franklin, M., Shenker, S., and Stoica, I. (2008). Fast and interactive analytics over hadoop data with spark. *The Functional Approach to Data Management*, 51(1), pages 107-113.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., J. Franklin, M., Shenker, S., and Stoica. I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in- memory cluster computing. *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 2-2.

Zaharia, M., Chowdhury, M., J. Franklin, M., Shenker, S., and Stoica, I. (June 2010). Spark: Cluster computing with working sets. *Proceedings of the USENIX Conference on Hot Topics in Cloud Computing (HotCloud '10)*, page 10, 2010.

Zaharia, M., Das, T., Li, H., Hunter, T., Shenker, S., and Stoica, I. (2013). Discretized streams: Fault-tolerant streaming computation 49 at scale. *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, pages 423-438.

Zorrilla, M. E., Menasalvas, E., Marín, D., Mora, E., & Segovia, J. (2015). Web usage mining project for improving Web-based learning sites. *Computer Aided Systems Theory—EUROCAST*. Springer, pp. 205–210.

Zweig, M.H., and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry,* 39, pages 561-577.