Cancer diagnosis prediction using multiple machine learning techniques

ABSTRACT

Cancer is a leading cause of death worldwide, accounting for an estimated 9.6 million deaths in 2018. Breast cancer is one of the most common cancer worldwide, contributing about one ninth of the total number of new cases diagnosed in 2018. Breast cancer survival rates have increased, Although the number of deaths associated with breast cancer is steadily declining largely because of earlier detection, breast cancer survival rates have increased. Recently, machine learning (ML) approaches have been applied in cancer prediction. With the help of ML, we can better understand whether the tumor is benign or malignant so as to help early diagnosis and reduce unnecessary surgical excision. However, different ML platforms and prediction methods may lead to varied performance outcome. In this study, in order to get better accuracy of prediction performance, we utilize multiple ML prediction models provided by Python and its machine learning library (scikit-learn), Apache Spark and its machine learning library (MLlib), Microsoft SQL server 2012 and its mining package to compare those approaches toward a better diagnostic prediction of breast cancer. In addition, how feature selection could contribute to the prediction accuracy is also explored in this study.

KEYWORDS

Machine learning, breast cancer, feature selection

1. Introduction

Artificial intelligence (AI) and machine learning technologies have the potential to transform health care by introducing new and important insights from the vast amount of data generated in health care industry every day. Among those, cancer prediction using ML technologies is increasingly studied [1]. The understanding of whether the tumor is benign or malignant in order to help early diagnosis and to reduce unnecessary surgical excision is an important application of cancer ML prediction.

Using ML techniques, researchers want to discover and identify patterns and relationships among multiple variables from complex datasets to effectively predict outcomes of a cancer type or subtypes [2].

Many studies have been reported that have focused on the classification of benign or malignant tumors. These studies have applied different approaches and achieved some good classification accuracies [3]. For instance, researchers classified the benign and malignant tumors based on digital image segmentation in the morphology of breast cancer cells [4]. Application of Convolutional Neural Networks (CNNs) model was able to discriminate benign cysts from malignant masses

in breast ultrasound images [5]. In order to investigate different approaches to combine the models such as bagging, boosting and other ensemble techniques, it has been reported recently that the combined model could obtain better performance and higher robustness [6].

However, different prediction methods may lead to varied performance outcome. In this study, in order to get better accuracy of prediction performance, we will utilize multiple machine leaning models and different algorithms/engines to compare different methods and optimize approaches toward a potential of a big volume of data in future. Therefore, it is urgent to compare the prediction performance and accuracy side by side among different platforms/tools and prediction models [7, 8]. To the best of our knowledge, this study is among the first to directly compare three widely-used ML platforms and multiple prediction models together with optimization of statistically selected features within the same dataset. Our long-term goal is to develop a ML model to predict benign or malignant tumors of breast tumor as a standard clinical decision support tool which can be used in the clinical practice.

2. Literature Review (Background)

2.1 Machine Learning Platforms/Tools

The Python programming language has already been one of the most popular languages for scientific computing. Scikit-learn is a Python module integrating a wide range of machine learning algorithms for medium-scale supervised and unsupervised problems [9]. Scikit-learn is built on Numpy, SciPy, and Matplotlib. It is an ease of use and efficient tool for data mining and data analysis.

In term of large-scale data processing, Apache Spark is a popular open-source platform. It uses data-parallelism to store and operate on data [10]. MLlib is Spark's open-source distributed machine learning library. Based on Spark, MLlib supports several languages including Scala, Java, R, Python, and SQL API. The library targets large-scale learning settings, and it benefit from data-parallelism or model-parallelism to store data and operate models. MLlib consists of fast and scalable implementations of widely used learning algorithms for common learning settings including classification, segmentation, regression, clustering etc. [11].

A combination of Integration Services, Reporting Services, SQL Server data mining tool is licensed platform that is particularly suitable for business intelligence and other purposes. It provides an integrated platform for predictive analytics that encompasses data cleansing and preparation, machine learning, and reporting. For instance, SQL Server 2012 provides nine algorithms to perform data mining tasks including Classification, Regression, Segmentation, Association, and Sequence analysis. All models have integrated visualizations to help develop, refine, and evaluate the models. A user can perform mining work directly with a database via SQL Server Data Tools. SQL Server can also work

with a connected database in order to provide on or off-premise facility [12].

2.2 Feature Selection

Machine learning uses features (variables or attributes) to generate predictive models. A dataset usually contains many possible variables or attributes for a predictive. Feature selection is one of the core concepts in machine learning which largely impacts the performance of your model. Feature Selection is the process where you automatically or manually select features which contribute most to your prediction output or variable. Careful feature selection schemes are of great importance for effective ML and subsequently for precise and accurate cancer predictions. Feature selection is essential for obtaining high precision and accuracy, reducing overfitting, and saving computer time and power. We generally want to restrict the features that are most relevant for the response variable we want to predict. It is very meaningful for reducing unnecessary testing and saving diagnosis duration in clinical piratical.

There are several ways to identify how much each feature contributes to the model and to restrict the number of selected features. Chi-square is a univariate statistical method that can be applied to rank features, whereas Recursive Feature Elimination (RFE) tests different subsets of features. Methods that use ensembles of decision trees including Random Forest and Extra Trees can also compute the relative importance of each input. In addition to those statistical approaches, Principal Component Analysis (PCA) is a dimensionality reduction algorithm where new features are created to represent the original feature dimensions in a lower dimension yet maintain the total information [13].

3. Design and Methods/Proposed Solutions

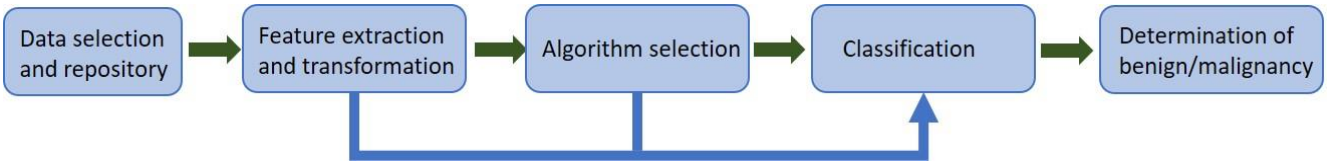

Fig.1. Procedure of data collection and analysis

3.1 Solution Design

In this study, the Wisconsin breast cancer (Diagnostic) dataset (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data) is chosen for entire analysis. The procedure of data collection and analysis is shown in Figure 1.

The dataset has 569 real records, each record has thirty-one columns (attributes) that includes one predictable (diagnosis) and 29 clinical variables as well as histological parameters that can be fed

as input to the prognostic procedure. The data has already been de-identified and the privacy of patients is fully protected. The size of data set is moderate. However, thirty-one attributes are very attractive for further development and establishment of data mining.

scikit-learn, MLlib, Microsoft SQL server analytic method etc. machine learning platforms will be utilized in this study. Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), K Nearest Neighbors (KN), Support Vector Machine (SVM), Neural Network (NN), and Clustering (Cl) will be used as predictive models.

3.2 Prediction using Scikit-learn

Scikit-learn package is applied. Jupyter notebook in Anaconda3 was used for computation and data description. 67% of dataset was used as training set, while the remaining 33% was used as test set. Using the Chi-square test, I was able to determine the best combination of one, two, three, four features. Different combinations of attributes were used as input. Diagnosis attribute was set as only predictable. I compared four different prediction models including SVM, KN, LR, RT and DT respectively.

3.3 Prediction using MLlib

In a VirtualBox environment, MLlib of Apache-Spark 2.3 is applied to perform ML. Diagnosis (M or B) was set as “Label”, and all thirty features was set as “Features/Terms”. 67% of dataset was used as training set, while the remaining 33% was used as test set. LR, DT, and RT models were used to predict the diagnosis.

3.4 Prediction using Microsoft SQL Server 2012

Microsoft SQL server 2012 Microsoft machine learning platform was used to mine the breast cancer dataset. I choose top one, top four features or all features and that were selected by Chi-square as input, and I choose diagnosis (M=1, B=0) as only predictable. Four mining models including DT, Cl, LR, NN were chosen to predict.

4. Data Analysis Results

4.1 Feature selection and Prediction using Scikit-learn

Using the Chi-square test, I was able to determine the best combination of one, two, three, and four features. The best one feature is area_worst. The best two features are area_worst and area_mean". Area_worst, area_mean, and area_se are the top three features. While the top four features are area_worst, area_mean, perimeter_worst, and area_se. Five different models including SVM, KN, LR, RT and DT were examined. As show in figure 2 and table 1, in Scikit-learn platform, one feature is able to predict very well (85.42± 8.85% of five models). Among five algorithms, KN, LR,

RT, and DT obtained similar performances in all five feature combination selections. The prediction rate of SVM is relatively low (64.36±4.63% of five feature combination), and the low prediction is not improved when more features were selected.

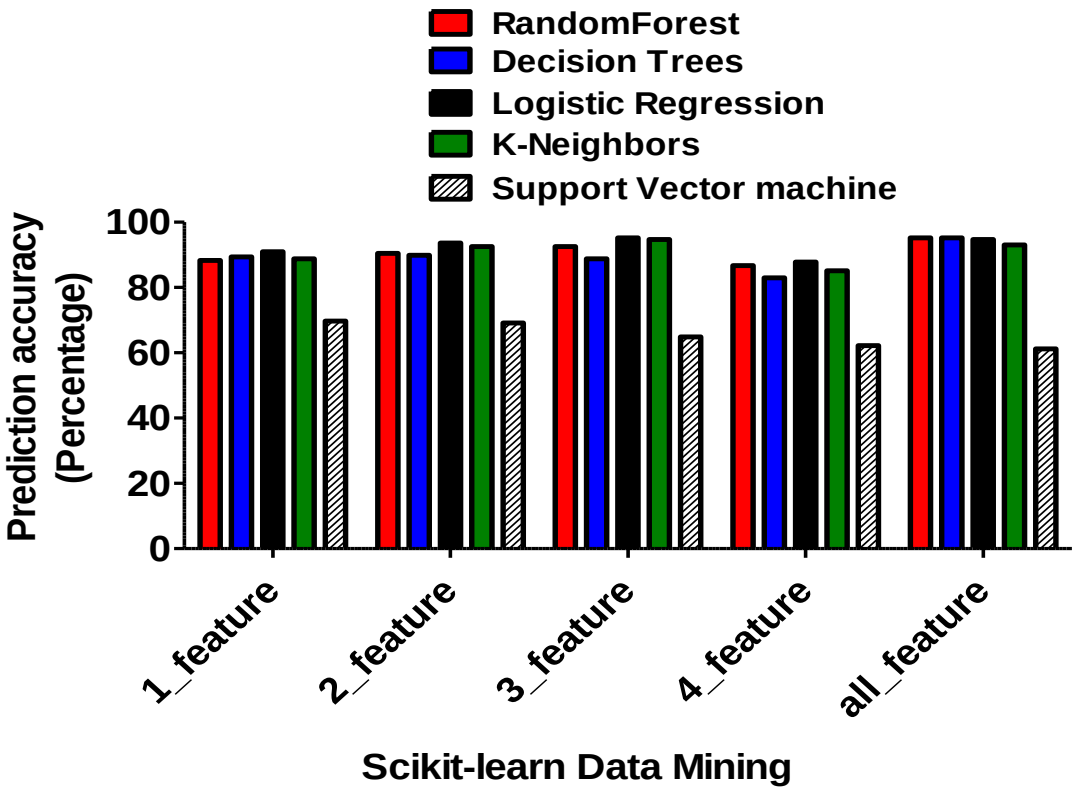

Fig.2. Prediction rate using Scikit-learn. The feature selection was determined by Chi-square computation in Scikit-learn. Support Vector machine (SVM), K Nearest Neighbors (KN), Logistic Regression (LR), Random Forest (FR), and Decision Tree (DT) were used as prediction models.

4.2. Prediction using MLlib

All clinical variables and histological parameters (29 total), top one ("area_worst") or four features ("area_worst", "area_mean", "perimeter_worst", and "area_se") were inputted for prediction. I compared three different models including LR, DT, and RF respectively. As show in figure 3 and table 1, in MLlib platform, all three feature combinations are able to predict very well, and all_feature group performed best in MLlib as the predictive rate is as high as 97.33±1.81%. Among three models in all three feature sections, RF model obtained highest accuracy (95.27±5.29%). All three models have high performance with different feature selection.

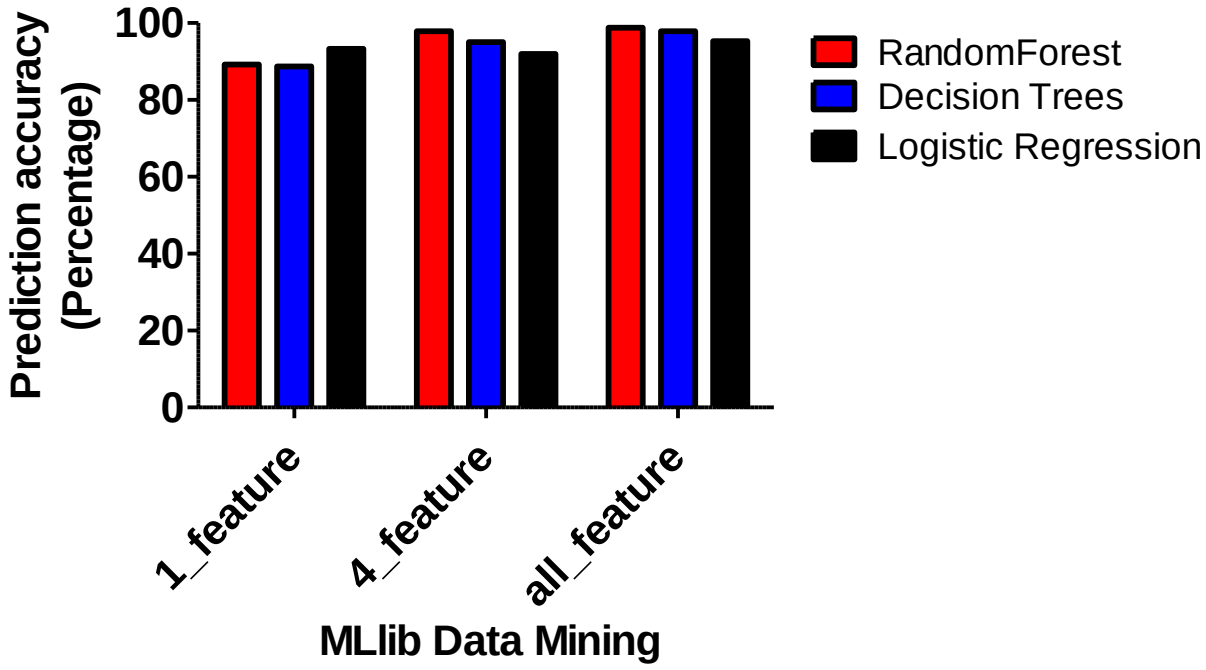

Fig.3. Prediction rate using MLlib. All 29 features (all_feature), four top features (4_feature), or 1 top feature (1_feature) were utilized for input. Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) were used as prediction models.

4.3. Prediction using Microsoft SQL Server 2012

All clinical variables and histological parameters (29 total) or top one (area_worst) and four features (area_worst, area_mean, perimeter_worst, and area_se) were inputted for prediction. I compared four different models including LR, DT, NN, Cl respectively. As show in figure 4 and table 1, in SQL server data mining platform, all features (60.08% ± 4.70%) and one-feature (65.15% ± 8.81%) are not able to predict very accurately. However, four top features combination is able to give a good prediction as the average accuracy is 96.71% ± 45%.

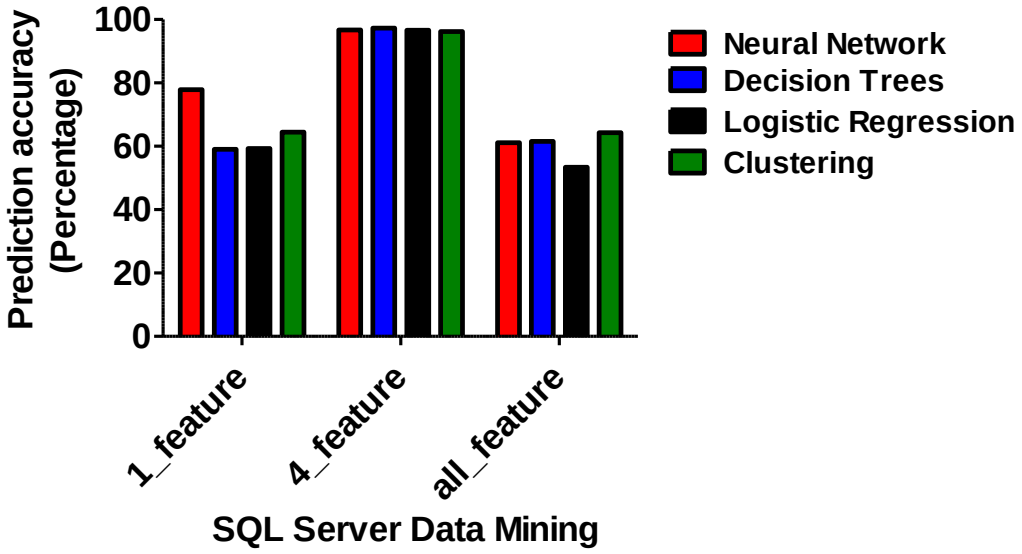

Fig.4. Prediction accuracy rate using SQL Server Data Mining platform. All 29 features (all_feature), four top features (4_feature), or 1 top feature (1_feature) were utilized for input. Logistic Regression (LR), Decision Tree (DT), Neural Network (NN) and Clustering (Cl) were used as prediction models.

Table 1. Summary of direct comparison of three machine leaning platforms

	scikit-learn					MLlib			SQL Server 2012			
	LR	DT	RF	KN	SVM	LR	DT	RF	LR	DT	NN	Cl
Top 1_feature	90.95	89.36	88.29	88.82	69.68	93.29	88.70	89.18	59.27	59.05	77.84	64.45
Top 4_feature	87.76	82.97	86.70	85.10	62.23	91.97	95.02	97.84	96.66	97.30	96.70	96.20
All_feature	94.68	95.21	95.21	93.08	61.17	95.30	97.90	98.80	53.36	61.56	61.11	64.32
Mean	91.13	89.18	90.06	89.00	64.36	93.52	93.87	95.27	69.76	72.63	78.55	74.99
SD	3.46	6.12	4.52	3.99	4.63	1.67	4.70	5.29	23.47	21.39	17.80	18.36

Table1. SD: Standard Deviation. LR: Logistic Regression, DT: Decision Tree, RF: Random Forest, KN: K Nearest Neighbors, SVM: Support Vector Machine, NN: Neural Network, and Clustering (Cl)

5. Summary and Conclusion

To analyze medical data, many machine learning platforms and prediction methods are available. An important challenge in the field is to build accurate and computationally efficient approaches for clinical applications. In this study, we compared scikit-learn, MLlib, and Microsoft SQL server 2012, three widely-used ML platforms, which are suitable for small to large scale of data analysis. MLlib has the highest accuracy rata, and all three tested prediction models including Logistic Regression, Decision Trees, and Random Forest performed very well with

more than 93.5% accuracy rate respectively. It is worthy noting that the high performance of MLlib was achieved even only top one feature was inputted for modeling. SVM in scikit-learn was not able to give a very accurate prediction perhaps due to the difference of training and test sets and selection of features [14]. Since Spark is a Java Virtual Machine (JVM)-based distributed data processing engine. It scales and it is fast compared to many other data processing frameworks with up to 100 times faster than Hadoop MapReduce in memory and 10 times faster on disk [11]. Therefore, MLlib is the most accurate and faster ML tool to computer this dataset.

With appropriate 4_feature combination, SQL machine learning tool was able to give good accuracy, but not with one and four features. However, SQL machine learning tools are very friendly to use, and many automatic functions could be useful for other analytic tasks [12]. In term of prediction models, we observed that Logistic Regression, Decision Trees and Random Forest performed reliably in both MLlib and scilit-learn. In this study, we discovered that the MLlib is the best tool to predict malignancy or benign of breast tumor.

6. Limitations and Future Work

The dataset does not have a large volume. This could limit the prediction accuracy using SQL server. The computation time could not be efficiently estimated based on this scale of the current dataset. It needs further study to validate the outcome of this srudy in other independent datasets before it could be used as a support tool in furfure clinical practice.

REFERENCES

[1] K. Kourou, T.P. Exarchos. K.P. Exarchos (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology*, 13: 8-17.

[2] Y. Sun, S. Goodison, J. Li, et al. (2007). Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*, 23 (2007), pp. 30-37.

[3] G.V. Sherbet, W.L. Woo, S. Dlay. (2018). Application of Artificial Intelligence-based Technologyin Cancer Management: A Commentary on the Deployment of Artificial Neural Networks. *Anticancer Res*, 38(12):6607-6613.

[4] A. Toprak. (2018). Extreme Learning Machine (ELM)-Based Classification of Benign and Malignant Cells in Breast Cancer, *Med Sci Monit*. 17(24):6537-6543.

[5] B.B. Ehteshami, M. Mullooly, R.M. Pfeiffer, et al. (2018). Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Mod Pathol*, 31(10):1502-1512.

[6] S. Boughorbel, R. Al-Ali, N. Elkum. (2016). Model Comparison for Breast Cancer Prognosis Based on Clinical Data. *PLoS One*, Jan 15;11(1):e0146413. doi: 10.1371/journal.pone.0146413. eCollection 2016.

[7]S.M. Domchek, A. Eisen, K. Calzone, J. Stopfer, A. Blackwood, B.L. Weber (2003). Application of breast cancer risk prediction models in clinical practice. *J Clin Oncol*, 21:593-601.

[8]J. Listgarten, S. Damaraju, B. Poulin, L. Cook, J. Dufour, A. Driga, *et al*.(2004). Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clin Cancer Res*, 10: 2725-2737.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, et al (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825−2830.

[10] X. Meng, J. Bradley, B. Yavuz, E. Sparks (2016). Mllib: Machine learning in apache spark. *Journal of Machine Learning Research*, 17: 1-7.

[11] E.R. Sparks, A. Talwalkar, V. Smith, et al (2013). MLI: An API for Distributed Machine Learning. In International Conference on Data Mining.

[12] Brian Larson (Ed.). 2012. *Delivering Business Intelligence with Microsoft SQL Server 2012* (3rd. ed.) McGraw-Hill Press, New York, Chapter 1-2.

[13] J. Li, W.K. Rich, R.D. Buhl-Brown. (2015). Texture analysis of remote sensing imagery with clustering and Bayesian inference. *International Journal of Image, Graphics, and Signal Processing*, 7(9): 1-10.

[14] M.F. Akay. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst Appl*, 36: 3240-3247.