ACTA SCIENTIFIC MEDICAL SCIENCES (ISSN: 2582-0931)

Volume 3 Issue 12 December 2019

Review Article

An Empirical Study: Role of Prior Dataset Analysis for Disease Risk Prediction System

MM Faniqul Islam[1] and Rahatara Ferdousi[2]*

[1]Bpp University, London, United Kingdom

[2]Lecturer, Department of Cse, Metropolitan University, Sylhet, Bangladesh

***Corresponding Author:** Rahatara Ferdousi, Lecturer, Department of Cse, Metropolitan University, Sylhet, Bangladesh.

Received: September 30, 2019; **Published:** November 18, 2019

DOI: 10.31080/ASMS.2019.03.0478

Abstract

Health informatics systems have pivotal clinical impact on patients due to its ability to predict or diagnose diseases in early stage, applying machine learning techniques. Researchers from different domains have been conducting numerous contemporary studies to propose a novel data driven medical predictive system. However, the characteristic of dataset is important to consider before selecting the appropriate machine learning approach for disease risk prediction system. Therefore, prior to developing an ideal clinical prediction system, proper analysis of datasets is mandatory for better accuracy as well as cost-effectiveness. In this work, an effective analysis of two different diabetes datasets have been conducted; dataset-1 has 768 instances and 9 attributes, dataset-2 has 500 instances and 17 attributes. An open source data mining tool termed as WEKA has been used for our experiment on data mining techniques. Furthermore, the fluctuation of performance of the same algorithm for different datasets have been demonstrated by analyzing individual classification models.

Keywords: Data mining; Risk Prediction; WEKA; Diabetes

Introduction

Data mining is the technique of extracting hidden information from a large set of database. The ultimate goals of data mining are prediction and description of diseases. Data mining techniques play a pivotal role in healthcare analysis. The large amount of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-effectiveness and further decision making [1-16]. Data mining provides a set of tools and techniques that can be applied to the processed data to discover hidden patterns. Figure 1 depicts the basic data mining process model.

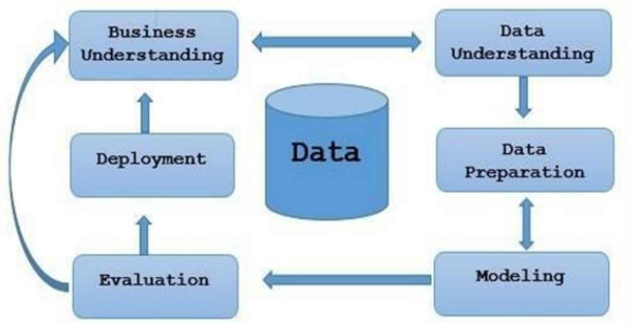

Figure 1: Data Mining Process Model.

This paper demonstrates the analysis of various data mining techniques which can assist medical analysts or practitioners to accurately select the right classifier model for disease risk prediction that rely on the dataset. For analysis, we use two datasets: Dataset-1 has 768 instances and 9 attributes, Dataset-2 has 500 instances and 17 attributes. The data sets are used to test and justify the performances of several classification algorithms. WEKA an open source data mining tool [17] has been used for the experimental analysis. Furthermore, comparative analysis has been performed and subsequent experimental results have been tabulated.

Common Data Mining Techniques for Predictive Systems

Naïve Bayes

The Naïve Bayesian classifier is based on Bayes theorem with independence assumptions between the predictors. Bayes Theorem-It works on conditional probability. Conditional probability is the probability that an occasion will happen, given that other occasion has just happened.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

P (A|B): the conditional probability that event A occurs, given that B has occurred. This is also known as the posterior probability.

P (A) and P (B): probability of A and B without regard of each other.

P (B|A): the conditional probability that event B occurs, given that A has occurred.

Advantages

1. A Naïve Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large data sets.
2. When the assumption of independence holds, a Naive Bayes classifier performs better compared to other models like logistic regression and need less training data.
3. It performs well in case of categorical input variables compared to numerical variable(s). For numerical variable, the normal distribution is assumed (bell curve, which is a strong assumption).

Disadvantages

1. If the categorical variable has a category (in the test data set), which was not observed in the training data set, then the model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
2. Naive Bayes is also known as a bad estimator, so the probability outputs are not to be taken too seriously.
3. Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

Decision Tree (J48)

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller subsets with increase in depth of tree. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor is called root node. Decision trees can handle both categorical and numerical data.

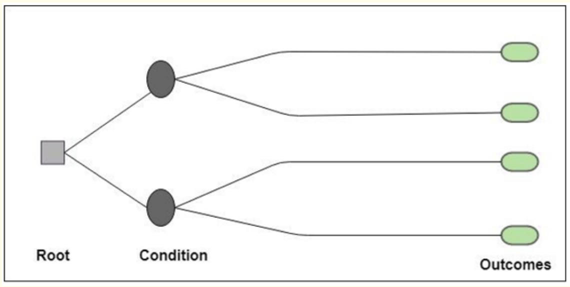

Figure 2: Decision tree.

Advantages

1. Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.
2. The models are inexpensive to construct, easy to interpret, easy to integrate with database system.
3. A decision tree does not require normalization of data.

Disadvantages

1. A small change in the data can cause a large change in the structure of the decision tree causing instability.
2. For a decision tree sometimes calculation can go far more complex compared to other algorithms.
3. Decision Tree algorithm is inadequate for applying regression and predicting continuous values.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm characterized by a separating hyperplane. The hyperplane is a line that partitions a plane in two sections where each class lay in either side.

There are 2 sorts of SVM classifiers:

1. Linear SVM Classifier.
2. Non-Linear SVM Classifier.

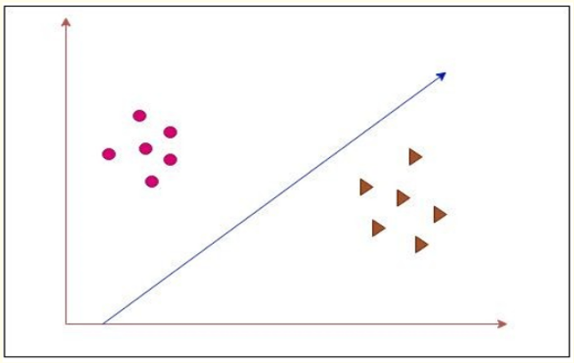

Figure 3: Support Vector Machine.

Advantages

1. SVM classifiers are very good when we have no idea on the data.
2. Works well with even unstructured and semi structured data like text, images and trees.
3. It scales relatively well to high dimensional data.
4. SVM models have generalization in practice; the risk of overfitting is less in SVM.

Disadvantages

1. Long training time for large datasets.
2. Choosing a "good" kernel function is not easy.
3. Difficult to understand and interpret the final model, variable weights and individual impact.

K-Nearest Neighbour

This classifier is considered as a statistical learning algorithm and it is extremely simple to implement and leaves itself open to a wide variety of variations. In brief, the training portion of nearest-neighbour classifier finds the closest training-point to the unknown point and predicts the category of that training point according to some distance metric. The distance metric used in nearest neighbor methods for numerical attributes can be simple Euclidean distance.

Advantages

1. Simple to implement.
2. Flexible to feature/distance choices.
3. Naturally handles multi-class cases.
4. Can do well in practice with enough representative data.

Disadvantages

1. Need to determine the value of parameter K (number of nearest neighbors).
2. Computation cost is quite high because of the need to compute the distance of each query instance to all training samples.
3. Storage of data.
4. Prior information of meaningful distance function is required.

Random Forest

Random forest algorithm is a supervised classification algorithm. This algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives high accuracy results.

Advantages

1. The same Random Forest algorithm or the Random Forest classifier can use for both classification and regression tasks.
2. Random Forest classifier will handle the missing values.
3. When we have more trees in the forest, Random Forest classifier won't fit the model.
4. Random Forest is comparatively less impacted by noise.

Disadvantages

1. Quite slow to create predictions once trained. More accurate ensembles require more trees, which means using the model becomes slower.
2. Random Forest require much more time to train as compared to decision trees as it generates a lot of trees and makes decision on the majority of votes.
3. Results of learning are incomprehensible. Compared to a single decision tree, or to a set of rules, they don't give you a lot of insight.

Multilayer Perceptron (MLP)

Multilayer perceptron which makes use of multiple layers of the neural network is created by using the set of various parameters which are selected to adjust the models with the help of correlation between parameters and prediction of the disease. MLP utilizes a supervised learning technique called back propagation for training the network.

Advantages

1. Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.
2. One of the preferred techniques for gesture recognition.
3. Multi-layered neural networks are basically used to manage data sets that have an extensive number of features, especially non-linear ones.
4. A two layer backpropagation network with sufficient hidden nodes has been proven to be a universal approximator.

Disadvantages

1. Convergence can be slow.
2. MLP needs long training time.
3. Local minima can affect the training process.
4. Hard to scale.

Stochastic Gradient Descent (SGD)

The stochastic gradient method is a gradient descent method optimized by the rate of convergence. The difference between the traditional gradient methods is that the elements are considered separately. Stochastic gradient descent (SGD) approximates the gradient using only one data point. So, evaluating gradient saves a lot of time compared to summing over all data. This is very useful while specifically working with big data sets.

Advantages

1. It is easier to fit into memory due to a single training sample being processed by the network.
2. It is computationally fast as only one sample is processed at a time.
3. For larger datasets it can converge faster as it causes updates to the parameters more frequently.

Disadvantages

1. Due to frequent updates the steps taken towards the minima are very noisy. This can often lead the gradient descent into other directions.
2. Also, due to noisy steps it may take longer to achieve convergence to the minima of the loss function.
3. Frequent updates are computationally expensive due to using all resources for processing one training sample at a time.

Table 1 depicts the comparative analysis of various classification algorithms. Different data sets with different kinds of variables and the number of instances determine the type of algorithm that will perform well.

Classifier	Decision Tree (J48)	Support Vector Machine	K-Nearest Neighbour	Random Forest	Stochastic Gradient Descent	Multilayer Perceptron	Naïve Bayes
Accuracy in general	**	****	**	***	***	***	*
Speed of learning with respect to the number of attributes and the number of instances	***	*	****	**	****	***	****
Speed of classification	****	****	*	**	****	**	****
Tolerance for missing values	***	**	*	****	**	*	****
Tolerance to irrelevant attributes	***	****	**	***	*	**	**
Tolerance to highly interdependent attributes	**	***	*	**	***	***	*
Tolerance to noise	**	**	*	***	*	**	*
Dealing with the danger of overfitting	**	**	***	****	**	*	***
Attempts for incremental learning	**	**	****	**	**	***	****
Explanation ability/ transparency of knowledge/classification	****	*	**	***	***	***	****
Applications	Emotion recognition , Verbal column pathologie s, Churn Analysis, Investment Solutions, High Customer	Face detection,text & hypertext categorization, Bioinformatics, handwriting recognition.	Text mining, Agriculture , Finance, Medicine[3 0].	Machine learning, Genetic algorithm, Fault diagnosis, Rotating Machiner y [27].	Surgeries, health information system(HIS) analyzing, medical image analyzing. and processing.	Speech recognition, Image recognition, Machine translation software [27].	Text classification, Spam filtering, Online Application, Hybrid recommender system.

Table 1: Comparison of seven classification algorithms (**** stars represent the best and * star the worst performance) [29-34].

Dataset Description

We have carried out performance analysis on two different diabetes datasets. Dataset-1 is Pima Indian diabetes dataset, collected from data hub [16]. It contains 768 instances with 9 parameters and Dataset-2 is an experimental diabetes dataset, collected from several hospitals. It contains 500 instances with 17 attributes.

The attributes of both datasets are described in table 2 and table 3.

No. of Datasets	Dataset Name	No. of Attributes	No. of Instances
1	Pima Indian diabetes dataset	9	768
2	Experimental diabetes dataset	17	500

Table 2: Characteristics of the datasets.

Attribute	Value
Age	Numeric
Pregnancies	Numeric
Glucose	Numeric
Blood Pressure	Numeric
Skin Thickness	Numeric
Insulin	Numeric
BMI	Numeric
Diabetes Pedigree function	Numeric
Class	{1=Yes, 0=No}

Table 3: Attribute Description of Dataset-1.

Age	Value
Age	Numeric
Gender	{M,F}
Polyuria	Numeric
Polydipsia	Numeric
Sudden weightloss	Numeric
Weakness	Numeric
Polyphagia	Numeric
Genital thrush	Numeric
Visual blurring	Numeric
Itching	Numeric
Irritability	Numeric
Delayed healing	Numeric
Partial paresis	Numeric
Muscle stiffness	Numeric
Alopecia	Numeric
Obesity	Numeric
Class	{1=Yes, 0=No}

Table 4: Attribute Description of Dataset-2.

Experimental analysis of classifiers

Following popular classification algorithms for prediction systems have been shown with their convenient WEKA name: (Table 5)

Generic ClassifiersName	WEKA Name
C4.5 Decision Tree	J48
Support Vector Machine	SMO
K-Nearest Neighbour	IBk
Random Forest	Random Forest
Stochastic Gradient Descent	SGD
Multilayer Perceptron	MLP
Bayesian Network	Naïve Bayes (NB)

Table 5: Seven Classification Algorithms.

In table 6 and table7 details of used datasets have been depicted. Table 8 and table 9 provides information about performance accuracy factors for both datasets.

Dataset	No. of training data	No. of test data	Total
Dataset-1	66 %	34 %	768

Table 6: Number of instances with Dataset-1.

It is evident from Figure 4 that the performance of the classifiers is not similar across the data sets. Performance of IBk, MLP, J48 and Random Forest is lower in the dataset-1. J48 and Random forest result in 76.2452%, while Naïve Bayes and SMO are comparatively higher accurate, i.e. 77.0115% and 79.3103% accuracy respectively. IBk and MLP perform, poorly among seven classifiers.

Data Mining Techniques	Accuracy	Mean Absolute Error	Model Construction Time
Naïve Bayes	77.0115 %	0.266	0s
SMO	79.3103 %	0.2069	0.02s
J48	76.2452 %	0.3125	0.02s
IBk	72.7969 %	0.2729	0.05s
Random Forest	76.2452 %	0.3058	0.06s
SGD	80.8429 %	0.1916	0.02s
MLP	74.3295 %	0.3186	0.57s

Table 7: Shows the Performance metrics with Dataset-1.

Dataset	No. of training data	No. of test data	Total
Dataset-2	66 %	34 %	500

Table 8: Number of instances with Dataset-2.

Data Mining Techniques	Accuracy	Mean Absolute Error	Model Construction Time
Naïve Bayes	86.4706 %	0.147	0s
SMO	91.7647 %	0.0824	0.11s
J48	91.7647 %	0.0921	0.03s
IBk	97.0588 %	0.0312	0s
Random Forest	96.4706 %	0.0789	0.24s
SGD	91.1765 %	0.0882	0.08s
MLP	94.1176 %	0.0644	0.76s

Table 9: Shows the Performance metrics with Dataset-2.

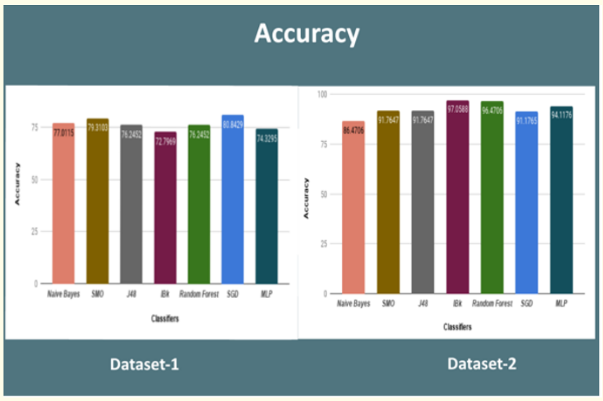

Figure 4: Classifiers Performance with both data sets.

Performance of SGD is the best in the dataset-1. Similarly, The performance of Naïve Bayes, J48, SMO, Random Forest, MLP and SGD is higher in the dataset-2. Surprisingly, IBk which performs the least in the dataset-1 among all classifiers and displays the best performance in the dataset-2. This may be attributed to the relevance of features which correctly classify the instances in the data sets.

Figure 5 shows, the error rate of the classifiers with different data sets. It is found that SGD shows the least error in the dataset-1 and IBK shows the least error in the dataset-2. Overall classifiers are performing less errors in the dataset-2.

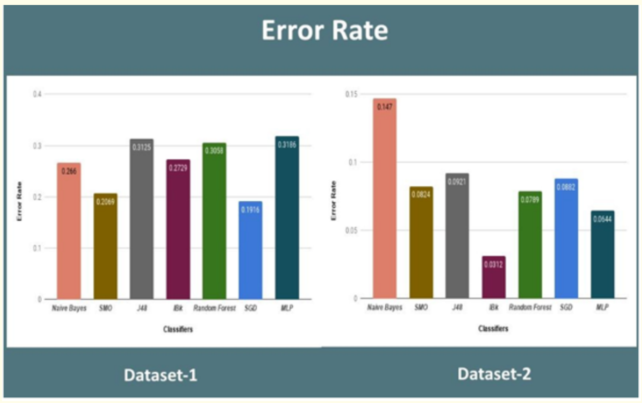

Figure 5: Classifiers error rate with both data sets.

Discussion

This experiment has been conducted to find the predictive performance of classifier models based for different datasets. Seven widely accepted machine learning classification algorithms have been selected considering their qualitative performance of the experiment. Two datasets were used for this experiment. For dataset-1, SGD's performance was found to be the best with 80.8429% accuracy and mean absolute error 0.1916. For dataset-2, IBk was the best in performance, i.e. 97.0588% accuracy and mean absolute error was 0.0312. Thus, this work also concludes that IBk classifier was the best as compared to other classifiers with the lowest error rate. Therefore, it can be stated that characteristic of dataset is a significant factor to analyse prior to design and develop a disease risk prediction system for the end users [18-34].

Conclusion

Data mining methods and tools are becoming more promising to predict risk of diseases depending on the characteristics of the dataset used. It's role in reducing the healthcare cost and burden through early risk prediction systems is undeniable. From this analysis, it has been observed that performance accuracy may vary for disease risk prediction for different datasets. For designing and developing a risk prediction system with better accuracy, a prior analysis of the dataset with number of potential data mining algorithms is essential.

Bibliography

1. M Mirmozaffari., *et al.* "Data Mining Classification Algorithms for Heart Disease Prediction". *International Journal of Computing, Communication and Instrumentation Engineering* 4.1 (2017).

2. M Nikhil Kumar., *et al.* "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools". *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 3.3 (2018): 2456-3307.

3. S Kodati and Dr R Vivekanandam. "Analysis of Heart Disease using Data Mining Tools Orange and Weka". *Global Journal of Computer Science and Technology: C Software and Data Engineering* 18.1 (2018): 0975-4350.

4. D Kinge and SK Gaikwad. "Survey on data mining techniques for disease prediction". *International Research Journal of Engineering and Technology (IRJET*) 5.1 (2018): 2395-0072.

5. K Gomathi and Dr Shanmugapriyaa. "Heart Disease Prediction Using Data Mining Classification". *International Journal for Research in Applied Science and Engineering Technology (IJRASET)* 4.2 (2016): 2321-9653.

6. K Ara Shekel., *et al.* "Dengue Disease Prediction Using Weka Data Mining Tool". *Proceedings of IIRAJ International Conference (ICCI-SEM-2K17), GIFT, Bhubaneswar, India, ISBN* (2018): 978-93-86352-38-559.

7. N Bhatla and K Jyoti. "An Analysis of Heart Disease Prediction using Different Data Mining Techniques". *International Journal of Engineering Research and Technology (IJERT)* 8.1 ISSN (2012): 2278-0181.

8. V Kirubha and S Manju Priya. "Survey on Data Mining Algorithms in Disease Prediction". *International Journal of Computer Trends and Technology (IJCTT)* 38.3 (2016): 2231-2803.

9. R Manimaran and Dr M Vanitha. "Prediction of Diabetes Disease Using Classification Data Mining Techniques". *International Journal of Engineering and Technology (IJET)* 9.5 ISSN (Print) (2017): 2319-8613.

10. Azrar., *et al.* "Data Mining Models Comparison for Diabetes Prediction". *International Journal of Advanced Computer Science and Applications (IJACSA)* 9.8 (2018).

11. Sivagowry S., *et al.* "An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease". *International Conference on Information Communication and Embedded Systems (ICICES)* (2013).

12. H Kaur and SK Wasan. "Empirical study on applications of data mining techniques in healthcare". *Journal of Computer Science* 2.2 (2006): 194-200.

13. R Deo Sah and Dr J Sheetalani. "Review of Medical Disease Symptoms Prediction Using Data Mining Technique". *IOSR Journal of Computer Engineering (IOSR-JCE)* e-ISSN: 19.3 (2017): 2278-0661.

14. Subhashri K., *et al.* "Analysis on Data Mining Techniques For Heart Disease Dataset". *International Research Journal of Engineering and Technology (IRJET)* 4.9 (2017): 2395-0072.

15. SK Sen and Dr S Dash. "Empirical Evaluation of Classifiers' Performance Using Data Mining Algorithms". *International Journal of Computer Trends and Technology (IJCTT)* 21.3 (2015): 2231-2803.

16. Pima Indian Diabetes data set.

17. WEKA tool.

18. Stochastic Gradient Descent (SGD).

19. https://medium.com/@jorgesleonel/multilayer-perceptron-6c5db6a8dfa3

20. DecisionTree.

21. https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf

22. https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589

23. SB Wankhede. "Analytical Study of Neural Network Techniques: SOM, MLP and Classifier-A Survey". *IOSR Journal of Computer Engineering (IOSR-JCE)* 16.3 (2014): 2278-8727.

24. SP Karkhanis and SS Dumbre. "A Study of Application of Data Mining and Analytics in Education Domain". *International Journal of Computer Applications* 120.22 (2015): 0975-8887.

25. FZ Maksood and G Achuthan. "Analysis of Data Mining Techniques and its Applications". *International Journal of Computer Applications* 140.3 (2016): 0975-8887.

26. R Arora Suman. "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA". *International Journal of Computer Applications* 54.13 (2012): 0975-8887.

27. S Saini., *et al.* "Comparative Analysis of Classification Algorithms Using Weka". *IOSR Journal of Engineering (IOSRJEN)* 8.10 (2018): 2250-3021.

28. S Gupta and N Verma. "Comparative Analysis of classification Algorithms using WEKA tool". *International Journal of Scientific and Engineering Research* 7.8 (2016): 2229-5518.

29. Osisanwo FY., *et al.* "Supervised Machine Learning Algorithms: Classification and Comparison". *International Journal of Computer Trends and Technology (IJCTT)* 48.3 (2017).

30. SB Imandoust And M Bolandraftar. "Application of K-nearest neighbor (KNN) approach for predicting economic events theoretical background". *International Journal of Engineering Research and Applications* 3.5 (2013): 605-610.

31. https://en.wikipedia.org/wiki/Multilayer_perceptron #Applications

32. N Kumar Choudhary., *et al.* "Impact of attribute selection on the accuracy of Multilayer Perceptron". 7.2 (2014): 32-36.

33. https://ieeexplore.ieee.org/document/4809024

34. https://www.sciencedirect.com/science/article/pii/S0925231200003052

Volume 3 Issue 12 December 2019