Predicting diabetes using machine learning classifiers

Lina Phaijit
COMP3308 Assignment 2
Faculty of Engineering
University of Sydney

***Abstract*—This paper analyses the performance of the different types of machine learning classifiers and the effects of feature selection on the Pima Indians Diabetes dataset.**

I. Aim

Diabetes Mellitus is a metabolic disease characterized by high levels of blood glucose which, if left untreated, can lead to severe damage to the body's systems. It is estimated that 415 million of the world's population have diabetes with 46% of them being undiagnosed. The number is expected to increase to 642 million by the year 2040. [1] Past research has shown that certain populations have a disproportionately high prevalence of diabetes due to genetic predisposition and unfavourable living conditions. [2] Past research has also established that diabetes occurrence is highly prevalent in patients of Pima Indians heritage. [3] [4] Patients of this ethnic group have unfavourable living conditions and hence already have difficult access to healthcare. Thus, it is even more crucial that accurate and fast diagnoses can be given. The purpose of this paper is to implement machine learning classifiers and perform analysis on the various types of classifiers for predicting whether the patient has diabetes. The patients selected in this study are women of at least 21 years of age and of Pima Indian heritage. The classifiers on Weka to be analysed are ZeroR, 1R, 1-Nearest Neighbour (1NN), 5-Nearest Neighbours (5NN), Naïve Bayes (NB), Decision Tree (DT), Multilayer perceptron (MLP), Support Vector Machines (SVM) and Random Forest (RF). The classifiers I will be implementing on Python which will also be analysed are 1NN, 5NN and NB.

II. Data

A. Dataset

The data used in this analysis was collected by the National Institute of Diabetes and Digestive and Kidney Diseases and was donated by Vincent Sigillito of The Johns Hopkins University on 9th May 1990. The data was modified for the unit COMP3308 on March 2015 to replace the missing values with averages and classes were changed to nominal values. There were several constraints placed on the selection of the data, such as the constraints of all patients being females of at least 21 years of age and of Pima Indian heritage. There are 768 instances included in the dataset, each with eight attributes and one class. The eight numeric attributes and their units are shown in TABLE I. The class variable is 'yes' or 'no' where 'yes' is interpreted as the patient being 'tested positive for diabetes'.

TABLE I. List of attributes in the Dataset

Attribute	***Unit***	***Selected by CFS***
Number of times pregnant	-	No
Plasma glucose concentration after 2 hours in an oral glucose tolerance test	-	Yes
Diastolic blood pressure	mm hg	No
Triceps skin fold thickness	mm	No
2-Hour serum insulin	mu U/ml	Yes
Body mass index	kg/m^2	Yes
Diabetes pedigree function	-	Yes
Age	years	Yes

Fig. 1. The attributes in the dataset for COMP3308 and the attribute selection by CFS

B. Attribute Selection

This paper also looks into the efficacy of correlation feature selection (CFS) on the provided dataset. CFS is a metric which evaluates the subsets of features based on the hypothesis that good feature subsets contain features which possess high correlation with the classification of the data and which are uncorrelated to each other. CFS will be used to select the subset of features for improving the prediction of the class of the data. Non-CFS data will also be used to analyse the effects of CFS on the classification for the different classifiers. Feature selection is done on Weka and the attributes which are discarded as a result are: number of times pregnant, diastolic blood pressure and triceps skin fold thickness. The remaining attributes are selected by CFS to be used in the building of classifiers (TABLE I).

III. Results and Discussion

TABLE II. Accuracy in Percentage

	ZeroR	***1R***	***1NN***	***5NN***
No feature selection	65.1%	70.8%	67.8%	74.5%
CFS	65.1%	70.8%	69.0%	74.5%

Fig. 2. The accuracies of the ZeroR, 1R, 1NN and 5NN classifiers on Weka using 10-fold stratified cross validation

TABLE III. ACCURACY IN PERCENTAGE

	NB	*DT*	*MLP*	*SVM*
No feature selection	75.1%	71.7%	75.4%	76.3%
CFS	76.3%	73.3%	75.8%	76.7%

Fig. 3. The accuracies of the NB, DT, MLP and SVM classifiers on Weka using 10-fold stratified cross validation

TABLE IV. ACCURACY IN PERCENTAGE

	RF	*My 1NN*	*My 5NN*	*My NB*
No feature selection	74.9%	68.4%	75.4%	75.3%
CFS	75.9%	68.2%	75.0%	76.0%

Fig. 4. The accuracies of the RF classifier on Weka and my 1NN, 5NN and NB classifiers using10-fold stratified cross validation

The results are as displayed in TABLES II-IV. ZeroR is the simplest classification algorithm and its accuracy is the lowest (65.1%) amongst all the classification algorithms, for both CFS and non-CFS methods. As predicted, its accuracy remains the same whether CFS is incorporated or not. This is because the classifier simply chooses the majority class ('yes' or 'no'), disregarding the attributes in the dataset. Hence, reducing the number of features has no effect on the accuracy of the ZeroR classifier.

The 1R classifier has an accuracy of 70.8% for both CFS and non-CFS methods. It has a higher accuracy than that of ZeroR. Since the 1R classifier uses the attribute with the smallest error, its higher accuracy than ZeroR implies that the attributes in the dataset do have relevancy towards the model prediction. From the results, CFS has no impact on the accuracy of 1R as expected. This is because CFS removes attributes that are redundant or irrelevant towards the classification. It is highly likely that the removed attributes do not have the smallest error amongst the rest of the attributes.

The 1NN algorithm yields an accuracy of 67.8% when no feature selection is employed. It is lower than that of the 1R classifier and only 2.7% higher than that of the ZeroR classifier. Since the 1NN algorithm takes into account all the attributes and only outputs a small increase in accuracy than the ZeroR method, this suggests that there may be attributes that are not that effective in the classification. In addition, the fact that the accuracy of 1R is higher strongly suggests the presence of attributes that contribute to the inaccuracy of the results. The 1NN classifier is sensitive to irrelevant features since its sample complexity grows exponentially with the amount of irrelevant features. [5] This is evident in the higher accuracy of the 1NN classifier when CFS is used, 69.0%, compared to when there is no feature selection, 67.8%.

In contrast, the 5NN algorithm yields a higher accuracy of 74.5% which is 6.7% higher than that of 1NN with no CFS. This is because the decision boundaries for 1NN are more highly variable and hence more unstable. If there is noise from the redundant or irrelevant data, the deviations in the values of the training set can cause large deviations in the classification. From the results, CFS on the 5NN classifier has shown to have no impact on the accuracy of the 5NN method. This suggests that using 5 as the k value for K-Nearest Neighbours is large enough that it sufficiently smoothens the decision boundaries and hence simplifies the model, lowering the variance which may be from redundant or irrelevant data.

The Naïve Bayes classifier on Weka produces an accuracy of 75.1% with no feature selection. It is higher than the accuracy obtained from 5NN. The Naïve Bayes algorithm is based on the assumption that attributes are independent given the class [5] and numerical values are assumed to have a normal probability distribution. Learning takes place and a model is built based on the training data. The prior probabilities of the classes and the probabilities of the different attributes for each class are learnt in Naïve Bayes. It is not a lazy method, unlike K-Nearest Neighbours which relies on local optimization and henceforth is susceptible to overfitting and noise. This, therefore, results in the accuracy of the Naïve Bayes classifier being higher. With CFS, the accuracy of the algorithm improves further. The irrelevant features which negatively impact the accuracy of the classification are removed in feature selection.

The classifier which has the highest accuracies is Support Vector Machines (SVM) with an accuracy of 76.3% with no CFS and 76.7% with CFS. The decision boundary for the classifier is defined by support vectors and is a maximum margin hyperplane. This prevents overfitting of data and produces enhanced generalization. CFS increases the accuracy as expected.

The DT (Decision Tree) classifier has higher accuracies than the simpler classification methods such as ZeroR and 1R. It produces a tree based on all of its attributes. Decision tree algorithms may therefore have a risk of overfitting training data and consequently produce trees that are unnecessarily large. The removal of redundant and irrelevant data in CFS reduces the size of the tree. [5] It reduces the complexity and also improves the accuracy from 71.7% to 73.3%.

Since the Random Forest (RF) classifier is comprised of a large number of uncorrelated individual decision trees, the individual errors from the trees shall 'cancel out' themselves. Hence RF outperforms a single Decision Tree. As expected, the accuracies of RF are higher than the accuracies of the DT for both CFS and non-CFS methods.

The Multilayered Perceptron (MLP) classifier with its adaptive learning also performs well in the experiment with relatively high accuracies of 75.4% for non-CFS and 75.8% with CFS.

My implementation of the 1NN and 5NN classifiers yields higher accuracies than the 1NN and 5NN classifiers on Weka by 0.6% and 1.1% respectively for non-CFS methods. I suspect that these deviations may be due to the variations in the algorithms and how folds are generated. Moreover, the way in which Weka may break ties may be different from my methods. However, it remains consistent that my 5NN classifier is more accurate than my 1NN classifier. What is interesting to note is that CFS lowers the accuracies of my 1NN and 5NN classifiers by 0.2% and 0.4% respectively. This suggests to me that my algorithm of kNN utilizes the least relevant attributes in such a way that they still help with the accuracies. Although three attributes have been removed in CFS, the accuracies decrease only slightly. This implies that the removed attributes are not as relevant as the remaining five attributes.

Lastly, the Naïve Bayes algorithm that I implemented has very similar accuracies of being greater by only 0.2% for non-

CFS methods and lower by only 0.3% for CFS-methods to the Naïve Bayes classifier in Weka. The values are very similar and the slight differences may be due to the small differences in the algorithm and equations, and in the way in which the folds were generated. The accuracy improves with the use of CFS as predicted.

CFS-included methods reduce the training times due to the reduced number of features in the data. The complexity of the algorithm is reduced as a result of the decreased number of data points. Redundant information acts as noise to the results. Hence, its impact is eliminated when redundant features are discarded from the data. Consequently, overfitting of data can be reduced. The accuracy of the classification can be improved by CFS due to the decreased number of misleading attributes present in the data.

IV. Conclusion

The results from this study have shown ZeroR to be the least accurate classifier and SVM as the most accurate. It is evident that some classifiers are more accurate than others. SVM and ZeroR differ in accuracy by over 10%. Thus, the type of classifier employed in the prediction of diabetes is important. The results amongst the classifiers also reveal that there are attributes present which are highly suggestive of the class and there are attributes which are relatively redundant. Namely, plasma glucose concentration after 2 hours in an oral glucose tolerance test, 2-hour serum insulin level, BMI, diabetes pedigree function and age are attributes which are suggestive of the class.

From this study, Correlation Feature Selection (CFS) has proven to be effective in increasing accuracy in many cases. It can be employed in data analysis to eliminate redundant and irrelevant features to improve model interpretability, reduce training times and reduce over-fitting. [6] It has been shown to also slightly reduce the accuracy in certain cases. However, this reduction is small enough that the benefits from the reduction in training times and complexity outweigh the slight decrease in accuracy. CFS has proven to correctly identify redundant and irrelevant information.

For future work, it would be beneficial to explore other parameters that may also have impact on the classification such as the level of physical activity. Moreover, there may be other parameters which lead to the features in this study. Since diabetes can be on the basis of genetics, it may be useful to also conduct studies and build classifiers for children and adults of under 21 years of age. Furthermore, to effectively compare the performance of my implementation and that of Weka's classifiers, additional datasets shall be used. The generation of the folds for the cross-validation shall also be kept the same.

V. Reflection

The analyses taken in this assignment have shown that artificial intelligence can be used to help humans predict crucial information such as the diagnosis of diabetes at a fast rate. The slightest details such as feature selection and normalization can be used to help with the classification and it is important to keep exploring these techniques. More work is required to ensure the classification is sufficiently accurate for diagnosis to be delivered with very high confidence. It is quite interesting to me that CFS does not improve the accuracy under all circumstances and that although it may lower the accuracy, it may even be beneficial in other ways. This assignment has taught me that the techniques are not quite black-and-white as they may seem. There are many complications to classification.

References

[1] [1]"Since 1996, the number of people with diabetes in the UK has risen from 1.4 million to 3.5 million. Diabetes prevalence is estimated to rise to 5 million by 2025.", *Diabetes*, 2020. [Online]. Available: https://www.diabetes.co.uk/diabetes-prevalence.html. [Accessed: 11- May- 2020].

[2] Schulz, L. O., Bennett, P. H., Ravussin, E., Kidd, J. R., Kidd, K. K., Esparza, J., & Valencia, M. E. (2006). Effects of traditional and western environments on prevalence of type 2 diabetes in Pima Indians in Mexico and the US. *Diabetes care*, *29*(8), 1866-1871.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] Knowler, W. C., Pettitt, D. J., Saad, M. F., & Bennett, P. H. (1990). Diabetes mellitus in the Pima Indians: incidence, risk factors and pathogenesis. *Diabetes/metabolism reviews*, *6*(1), 1-27. R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] Bennett, P., Burch, T., & Miller, M. (1971). Diabetes mellitus in American (Pima) indians. *The Lancet*, *298*(7716), 125-128. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[5] Hall, M. A. (1999). Correlation-based feature selection for machine learning.

[6] Zhang, M. (Ed.). (2016). *Applied Artificial Higher Order Neural Networks for Control and Recognition*. IGI Global.