Assignment 2

COMP3308

SID: 490503902

University of Sydney

Contents

1 Aim

The aim of the study is to examine the accuracies of the K-Nearest Neighbour and Naive Bayes classifiers in predicting whether a given Pima Indian has diabetes or not. Through implementation and evaluation of these classifiers on the pima-indian-diabetes.data dataset using the stratified cross validation method, this study demonstrates the importance of comparing different machine learning techniques, as demonstrated via Weka, and the impact of feature selection in classification.

Diabetes is a chronic disease that can potentially lead to impaired vision and damage to blood vessels of the heart, brain and legs if left unmanaged. The importance of this study is in evaluating the accuracies of different machine learning techniques, one gains a better understanding of the most effective classifiers for the given problem. Accurate classifiers can assist medical workers in diagnosis and prevention of diabetes, especially for populations at risk. It can give doctors insight into the key factors contributing to diabetes in different populations and can be used to make informed data-driven decisions to manage at risk population (e.g. encouraging personalised dietary lifestyles suitable to a specific culture).

2 Data

2.1 Description of Dataset

The Pima Indian Diabetes dataset can be originally sourced for the National Institute of Diabetes and Digestive and Kidney Diseases but for this study, the dataset was sourced from UCI Machine Learning Repository. The dataset was modified by replacing missing values with averages and changing class to nominal values for consistency. The dataset was donated to the National Institute of Diabetes and Digestive and Kidney with the date received being 9 May 1990. The dataset contains 768 patient records, where all patients were females aged 21 years or older of Pima Indian heritage. Each of the records contains 8 numeric attributes and a binary class; each attribute representative of a personal characteristics of a given female and the class indicates whether a patient has ”yes” or ”no” diabetes. Overall, the dataset contains 500 patients with no diabetes and 268 patients with diabetes.

For this study, the dataset attributes were normalised to be in the range of [0,1] using Weka. The class attribute was not normalised.

2.2 Summary of CFS and list of selected attributes

For attribute selection, the correlation-based feature selection (CFS) method was used. CFS is an algorithm that selects the best subset of features based on the hypothesis that a good subset of features are highly correlated with the class but are uncorrelated with other features. The CFS does this by taking multiple subsets of the features and scores each subset, thereby choosing the subset of features with the best score.

Note: The list below was taken from the dataset description.

Without CFS applied, the attributes (numeric-valued except for class) in pima.csv are:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (”yes” or ”no”)

With CFS applied, the attributes chosen in pima-CFS.csv are:

1. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
2. 2-Hour serum insulin (mu U/ml)
3. Body mass index (weight in kg/(height in m)^2)
4. Diabetes pedigree function
5. Age (years)
6. Class variable (”yes” or ”no”)

3 Results and Discussion

3.1 Results

Table 1: Accuracies of algorithms using Weka

	ZeroR	**1R**	**1NN**	**5NN**	**NB**	**DT**	**MLP**	**SVM**	**RF**
No feature selection (%)	65.1	70.8	67.8	74.5	75.1	71.8	75.4	76.3	74.9
CFS (%)	65.1	70.8	69.0	74.5	76.3	73.3	75.8	76.7	69.7

Table 2: Accuracies of algorithms using study’s implementation

	My1NN	**My5NN**	**MyNB**
No feature selection (%)	71.1	67.2	71.6
CFS (%)	67.1	75.0	73.7

3.2 Discussion

Performance of Classifiers

The **ZeroR** classifier is the simplest algorithm that ignores all attributes in the input table and focuses only on the class. The algorithm constructs a frequency table, counts all instances of "yes" and "no", and chooses the class with the majority. In the case of this study, the most frequent class was "yes", hence the algorithm classifies all new examples as "yes" and is correct 65.1% (derived from $\frac{500}{768}$) of the time, performing the worst of all classifiers. The **ZeroR** classifier is considered the baseline performance classifier whereby all other classifiers are compared to this classifier. Hence, under CFS or no feature selection (FS), accuracy remains the same since removing attributes has no impact on the class.

The **1R** classifier is an algorithm that generates one rule based on the attribute with the lowest number of errors in predicting the input table class. It does this by generating a rule for each attribute, thereby constructing a frequency table for each value of the attribute. For each value, the classifier chooses the majority value and then aggregates the total number of errors, assigning the attribute with a specific error number. The classifier then chooses the rule with the smallest number of errors. This rule can be understood easily by humans. It is simple and fast, but there can be problems with overfitting, especially when dealing with a numeric attribute with a large number of possible values.

Compared to **ZeroR** and all other classifiers, the **1R** classifier performs quite well with 70.8% for both no FS and CFS. This is due to **1R** chooses the attribute, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, which exists in both no FS and CFS datasets. A possible reason for the relatively good performance is that the underlying problem of diabetes has a very strong correlation with blood sugar levels, hence a single attribute can perform well.

The **K-Nearest Neighbour (KNN)** classifier is a similarity measure algorithm that takes in a test instance, calculates the Euclidean distance using each instance of the training data against the test instance, assigns each training instance with a Euclidean distance score, selects k number of training instances with smallest Euclidean distance scores and determines the majority class. This majority class is then assigned to the test instance. The classifier keeps working until all test instances in the test data file have been assigned a class. As **KNN** is a lazy learner, it does not build a classifier until a new example needs to be classified, making the classifier very fast in training but slow when testing. In this study, the numerical data was normalised to prevent one attribute from dominating and ties were broken with the output being "yes".

From table 1, the increase in K number has increased the accuracy where under no FS, the accuracy increased from 67.8% (**1NN**) to 74.5% (**5NN**) on **Weka**. For table 2, under no FS, the accuracy decreased from 71.1% (**My1NN**) to 67.2% (**My5NN**). This is incorrect and appears to be a result of constructing the folds (pima-folds.csv). Despite each fold having the ratio of 50 (no) :27 (yes), some folds have attributes where the mean and standard deviation of the attribute is significantly different across folds. For example, let us assume that the mean body measurement index (**BMI**) is 0.2 and standard deviation (**std**) is 0.1 for fold 1 and for fold 2, the mean of BMI is 0.3 and std is 0.25. This affects the euclidean distances calculation across different folds, impacting the selection of **KNN**.

Moreover, under CFS, using **Weka**, the accuracy for **1NN** increased by 1.2% from 67.8% (**no FS**) to 69.0% (**CFS**) but for **5NN**, the accuracy did not increase. This shows that CFS is effective in increasing or equaling the accuracy of no FS. According to table 2, due to the issue with folds, the accuracy decreased for **My1NN** from 71.1% (**no FS**) to 67.1% (**CFS**). For **My5NN**, the CFS increased from 67.2% to 75%, a significant increase demonstrating that **KNN** is very sensitive to irrelevant attributes and noise. This suggests that compared to **1R**, an increase in the

K variable can improve accuracies and is likely a result of a trend in the given dataset.

The **Naive Bayes (NB)** algorithm is a statistical classifier based on Bayes theorem that, given a specific test instance, generates probabilities of the test instance belonging to each class value, then chooses the class value with the highest probability. In this study, as the data given was numerical, a probability density function must be used to evaluate the probability of each attribute, a normal distribution of data was assumed and classification ties were broken with the default output of "yes". **NB** assumes that the attributes are independent of each other and are of equal importance. This assumption can be problematic, especially in the context of diabetes, a person's age can impair their serum insulin levels and affect their body mass index. Hence, the importance of choosing the most relevant attributes is demonstrated in **Weka**'s implementation and the study's implementation of **NB** where using CFS compared to no FS, the accuracy increased by 1.2% (**table 1: NB**) and by 2.1% (**table 2: MyNB**). Furthermore, the study's implementation and **Weka**'s implementation of **NB** were very similar and so the study's implementation appears to be correct. So far, the **NB** appears to be the best performer compared to the previous classifiers and is due to being robust to isolated noise.

The **Decision Tree (DT)** algorithm is a supervised learning algorithm that builds a decision tree by choosing the best attribute as the root, creating the tree, and generating leaf nodes that correspond to a class value. It does this by maximising information gain. In table 1, **DT** performs slightly worse than **NB** and **5NN**, but better than **ZeroR**, **1R**, **1NN** because simple **DTs** are more prone to overfitting compared to other classifiers. The accuracy of **DT** increased by 1.5% using CFS, which highlights that pruning the attributes of a **DT** can lead to better performance.

The **multi-layer perceptron(MLP)**, **support vector machine (SVM)**, and **random forest (RF)**'s performance, under no FS, were 75.4%, 76.3%, and 74.9%, respectively, with **SVM** achieving the highest accuracy of all classifiers tested. Even when using CFS, **SVM** is the most accurate classifier. A possible reason for **SVM**'s success is the use of the maximum margin hyperplane enables great tolerance to noise, enabling accurate generalisation to new examples. Despite **SVM**'s success, **SVM** has a costly disadvantage of long training times for large datasets, highlighting the issue of scalability.

Effect of feature selection

Overall, feature selection was beneficial. Using the "Best-First-Search" method, the CFS reduced the attributes from 8 attributes + class to 5 attributes + class. The subset of original features chosen by CFS makes intuitive sense as a person who has high plasma glucose concentration, serum insulin levels, is overweight, has diabetic parents, and is of great age is more prone to be diabetic.

In general, CFS increases accuracy for all algorithms. As with all classifiers, when the dimensionality of data increases, this can lead to overfitting. CFS reduces dimensionality, making the algorithms more accurate. For example, this is demonstrated in table 1, where **Weka**'s implementation consistently highlights that CFS either increases or equals accuracy from no CFS results, except for the **random forest**. The issue with the random forest is that it is a collection of random decision trees and so when CFS is performed, the individual decision trees are smaller, therefore decreasing the **random forest** accuracy. Hence, the results in table 1:**RF** accuracy drop from 74.9% (**no FS**) to 69.7% (**CFS**). Moreover, CFS has the added benefit of on a large scale reducing computational complexity, thereby increasing the speed of processing. This is highly important when doctors have limited time and computational power.

Further comments

In this study, 10-fold stratified cross validation was conducted to reduce the variability of results. The best performing algorithm was the **SVM** but this classifier has the issue of significant training time for large datasets. Similarly, **NN** has a memory requirement of O(mn) where m is the number of training examples with the dimensionality of n. This makes it impractical for using **NN** on a large dataset.

Overall, the researcher of this study would recommend either the **1R** and **DT** classifiers for use by non-technical individuals. This is due to the output of each classifier is easily understood by humans as a decision tree, enabling insights to be interpreted immediately, thereby assisting informed decision making.

4 Conclusion

4.1 Summary of main findings

To conclude, this study demonstrates the importance of understanding different machine learning algorithms, the original dataset, and the prescribed problem before drawing conclusions that influence decision making. The study highlights that it is good practice to try simple ML algorithms before trying more complex ones. This is further elaborated through choosing the "best" ML algorithm is dependent on the problem and needs to consider not only accuracy but also computational complexity. Though the simple algorithms such as **1R** and **Nearest Neighbour** are less accurate than the complex algorithms such as **random forest** and **multi-layer perceptron**, the simple algorithms can execute quicker on large datasets and produce output that is easily understood by non-technical humans. Moreover, higher accuracies can be a result of overfitting, meaning the algorithms learn the training data too well and are not good at generalising. Machine learning algorithms are a tool, which means a doctor still needs to have a deciding role in diagnosis.

4.2 Future Work

A suggestion for future work is to investigate further the attributes that influence plasma concentration of glucose. The attributes discovered can inform medical scientists into the key foods commonly consumed and reveal specific causes of diabetes. It could also shine light on outlier cases such as young teenagers who have diabetes but do not have any diabetic family members. This future work will enable doctors to prescribe personalised plans for patients on the verge of becoming diabetic.

Another suggestion would be further research into scalable machine learning models that produce rules/decision trees and have higher accuracies than the ones presented in this study. Computationally efficient machine learning algorithms that work well on large datasets can reveal insights into a particular community at risk of diabetes, which can prompt lawmakers to initiate public reform such as a sugar tax. Further research into how machine learning can complement diabetic decision making is suggested.

Further, there should be research into software that after running a dataset, reveal why success rates differ among the different classifiers and if this discrepancy is significant.

5 Reflection

From this study, I have discovered the importance of choosing the right classifier for a given problem is contextual and dependent on accuracy, issues of overfitting, data noise, and computational speed. To elaborate further on this discovery, I have gained a greater appreciation of complete, consistent, and reliable data in the machine learning process, which is necessary for effective classification. The importance of good data cannot be understated, giving greater credit to the phrase ”Garbage in, garbage out”. This has prompted me to be more vigilant of the data used in my projects and be more critical of other conclusions drawn from unknown data sources. Moreover, choosing the right classifier can accelerate scientific discoveries and help researchers to better respond to changing communities and emerging diseases. This has increased my belief that machine learning enables better informed medical decisions, which can lead to early intervention in fatal diseases such as terminal cancer, saving potentially millions of lives.