Health Data Classification through KNN and Decision Tree Models

Sarah Zhou
shi.q.zhou@mail.mcgill.ca
Ziqing Liang
ziqing.liang@mail.mcgill.ca
Zuchan Fang
zuchan.fang@mail.mcgill.ca

Abstract

The application of classification models in healthcare data is of great significance in the intelligentization of medical information. In this project, we investigated the performance of two machine learning models, K-Nearest Neighbour and Decision Trees, on two medical benchmark datasets,the Wisconsin Breast Cancer dataset (WBC) (uci) and the Hepatitis dataset (HP) (hep), to predict the target patient condition in a binary classification setting. We drew some general conclusions on the adaptability of the two models. In view of the advantages and disadvantages of the two models, the subsequent improvement directions of this experiment are also discussed at the end of this paper.

The results obtained from this experiment indicate that overall the Decision Tree model performs well slightly better than the KNN model. We found that the Decision Tree approach using Entropy of Gini index cost function achieved better accuracy than K - Nearest Neighbour on the Hepatitis dataset.

1 Introduction

Classification is the task of detecting a model or function. The function is extracted based on the analysis of a set of training data. Different classification techniques provide different ways to model decision boundaries between classes. The goal of this project is to implement two classification techniques from scratch —K-Nearest Neighbour and Decision Trees— and compare these two algorithms on two distinct health data sets.

In this study, we analyzed the two datasets from a statistical standpoint, and drew conclusions about the general characteristics of the two sets of data before processing. In terms of model implementation, both algorithms are implemented as Python classes, and the main functions are constructed for model training, output prediction and accuracy evaluation. In the later stage of this study, the two data sets are divided into training, validation and test set. We tuned the hyper-parameters of the two models on the validation set, and draw a general conclusion about the adaptation of the two models. At the end, we discussed how the two models can achieve more accurate goals based on the findings of the experimental data.

1.1 Related Work

Various researches have explored the application effects of classification models. Most of these studies have been published in recent years. Among all, *Comparative study between decision tree and knn of data mining classification technique* published back in 2018 (Mohanapriya and Jayabalan, 2018) is the closest to the topic of our study. This study starts from the respective classification principles of KNN and Decision Trees, and compared their adaptability, advantages and disadvantages from a very comprehensive perspective.

Moreover, the two medical datasets used have great value for classification learning. In the past two decades, these two sets of data have been repeatedly cited in a large number of research papers focusing on decision trees. As a typical example, the *Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm* published in 1995 (Turney, 1995). The Hepatitis dataset was cited to optimize cost-sensitive decision tree classification. Likewise, the Wisconsin Breast Cancer dataset was cited in *Heterogeneous Forests of Decision Trees* (Grabczewski and Wlodzislaw, 2002), which augmented the decision tree based on the Separability of Split Value(SSV) with the capability to generate heterogeneous forests of trees instead of single trees.

2 Dataset

2.1 Wisconsin Breast Cancer Dataset

The dataset presents a set of patients with breast cancer. There are 699 instances of 11 attributes, with 16 instances in Groups 1 to 6 that contain a single missing values, denoted “?”.

2.2 Hepatitis Dataset

The dataset labels whether patients with hepatitis live or die. The purpose of the dataset is to predict the presence/absence of hepatitis disease by using the results of medical tests of a patient. There are 155 instances with 20 attributes labeled to two different classes (32 for “die”, 123 for “live”). The 20 attributes include 14 binary ones and 6 numerical ones. Instances have missing values, such as 16th, 19th that have 29 and 67 missing values respectively.

2.3 Preprocessing

The two data sets were prepossessed by filtering out the invalid health records containing missing values in certain fields, converting each field to numeric values and moving the class labels to the first column for consistency across datasets. Upon further analysis, in the Wisconsin Breast Cancer dataset, the column ’Sample code number’ is dropped as it is not a measure but rather an assigned ID.

Furtheremore, the class labels are converted to with 0 and 1, with 0 serving as live/benign, and 1 serving as die/malignant. The rest of the categorical data also follows the conversion of the labels. We then normalized the data in preparation for fitting. Due to the small amount of available data, a 8:1:1 ratio was chosen to split train vs validation vs test set, so as to maximize training data.

The following are key metrics for the finalized datasets we worked with:

Wisconsin Breast Cancer dataset

683 valid records: 546 training set - 68 validation test - 69 test set.
Available features: 9 different health related metrics with integer value in the range 1-10.
Target classes: 2 (0: benign and 1: malignant)

Hepatitis dataset

80 valid records: 64 training set - 8 validation test - 8 test set.
Available features: 19 different health related metrics.
Target classes: 2 (0: live and 1: die)

2.4 Statistical Analysis

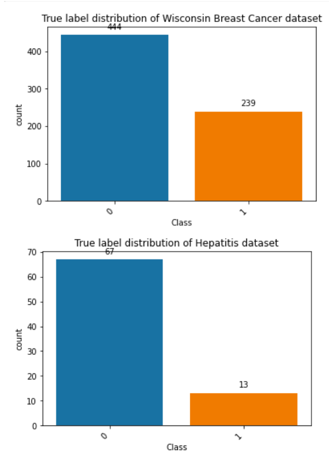
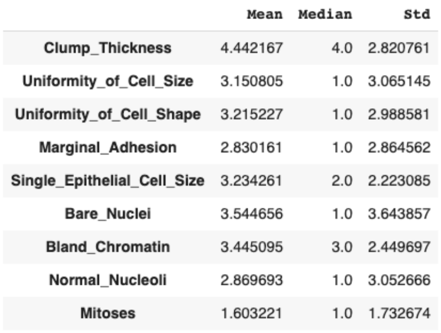


Figure 1: True labels distribution



	Mean	Median	Std
Clump_Thickness	4.442167	4.0	2.820761
Uniformity_of_Cell_Size	3.150805	1.0	3.065145
Uniformity_of_Cell_Shape	3.215227	1.0	2.988581
Marginal_Adhesion	2.830161	1.0	2.864562
Single_Epithelial_Cell_Size	3.234261	2.0	2.223085
Bare_Nuclei	3.544656	1.0	3.643857
Bland_Chromatin	3.445095	3.0	2.449697
Normal_Nucleoli	2.869693	1.0	3.052666
Mitoses	1.603221	1.0	1.732674

Figure 2: WBC statistics

As seen in Figure 1, the two datasets are small and unbalanced, with a significantly higher amount of live/benign labels. Namely, after prepossessing, the benign class has 444 instances and the malignant class has 13 samples.

For the WBC dataset, over all attributes, almost the standard deviations are between 2.0 and 3.7 (Figure 2). We find the minimum mean value in clump thickness.

Different attributes have different distributions. In summary, for WBC dataset, under benign condition, most of uniformity of cell size and cell shape, Marginal adhesion, Normal Nucleoli, Mitoses lie between 0 and 2 (Figure 3); under

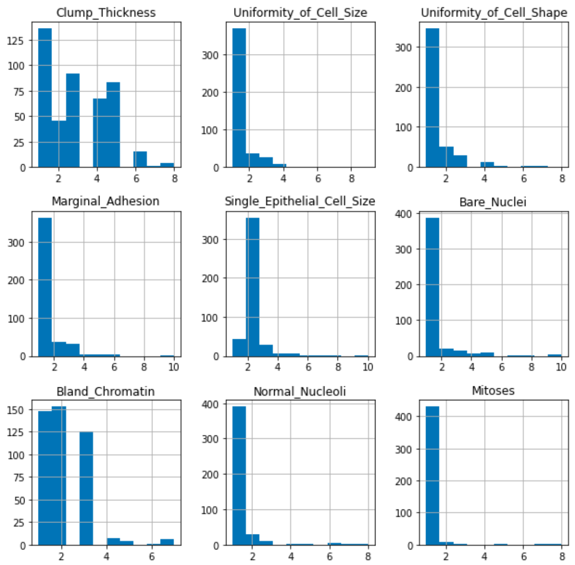


Figure 3: Histograms of WBC benign features

malignant condition, nearly all of the attributes are almost randomly distributed.

It is important to highlight the ethical concerns that reside for those datasets. There exists the potential for identifiable information to be deliberately collected about the patient, even without their permission. In many cases, cancer registries already have access to identifiable patient records without the patient realising. The data sets we used in this study is retrieved from public portal, therefore there is no such concern. However, in real data analysis, whether the channels for obtaining medical data are legal and compliant should be considered.

3 Models

The implementation of the two models largely follows the standard implementations given in the context of COMP 551: Applied Machine Learning class (Ravanbakhsh, 2020).

3.1 KNN Model

The KNN algorithm is a generalization algorithm for nearest neighbor rules. Its inductive offset is the class label of the k-sample with the class label to be tested most similar to the nearest one. Compared with the nearest neighbor, it differs in that it expands the nearest neighbor to k in the decision making phase. This extension allows the KNN algorithm to obtain and utilize more information. It omits the process of learning processing relative to other classification algorithms with distinct training phases.

3.2 Decision Tree Model

Decision trees is like a hierarchical model. It classifies objects by sorting all objects based on attribute value. Typical construction of decision trees contain node – root node, leaf node, and branch. Each node in the tree represents the object based on attributes. Branches denotes objects value. From the root node the objects are classified. After that based on the attribute value the objects are sorted. A decision tree is a classifier which provides some rules in a tree structure.

3.3 Evaluation metrics

For all the models, we used training, validation and test accuracy to evaluate the performance on the task of headline generation:

1. Accuracy - ranging from 0 to 1, 1 being the optimal score.

Accuracy is an efficient and computationally inexpensive way to check the ratio of matches of target labels against the true labels. As a result, we will use it as our evaluation method.

The training accuracy provides a benchmark on how the models perform compared to seen data. The validation accuracy serves in hyper-parameter tuning. The test accuracy applies to unseen data.

4 Results

The results that indicate the scoring of these two models can be found in Table 1.

As shown in Table 1, both of the KNN model and the Decision Tree model performed well on the two datasets, achieving very high accuracy. For WBC data set, the two models are similarly well performed. When it comes to the data set with higher feature dimensions (HP dataset), the performance of the Decision Tree is significantly better than KNN.

This conclusion is consistent with the theoretical knowledge learned in class, that KNN is sensitive to feature scaling while Decision Tree is not. Overall, the Decision Tree model performed slightly better

Model	Dataset	K/Max-depth	Cost function	Accuracy score		
				Train	Validation	Test
KNN	WBC	3	Euclidean	97.8.	97.1.	98.6.
	WBC	1	Manhattan	100.0.	97.1.	98.6.
	HP	4	Euclidean	90.6.	87.5.	87.5.
	HP	4	Manhattan	90.6	87.5.	87.5.
Decision Tree	WBC	5	Miss-classification	98.0.	97.1.	98.6.
	WBC	3	Entropy	97.1.	94.1.	97.1.
	WBC	7	Gini index	99.1	94.1.	94.2.
	HP	3	Miss-classification	96.9.	87.5	75.0.
	HP	2	Entropy	93.8.	87.5.	100.0.
	HP	2	Gini index	93.8.	87.5.	100.0.
KNN (with feature selection)	HP	1	Euclidean	100.0.	100.0.	100.0
	HP	4	Manhattan	89.1.	87.5.	87.5.
Decision Tree (with feature selection)	HP	2	Miss-classification	92.2.	62.5.	75.0.
	HP	2	Entropy	85.9.	87.5.	100.0.
	HP	2	Gini index	92.2..	87.5.	75.0.

Table 1: Reported accuracy for model variants.

than the KNN model if we choose misclassification cost for the WBC dataset and Entropy/Gini index cost for the HP dataset.

4.1 KNN Model

Changing the value of k, we found that, as the value of k increases, the training & testing data accuracy of the model quickly peak reaching the optimal accuracy and gradually decrease afterwards. Take the performance of the KNN model on the WBC data set when the Euclidean distance is used as an example, the best accuracy is achieved when the data result is k=3 (Figure 4).

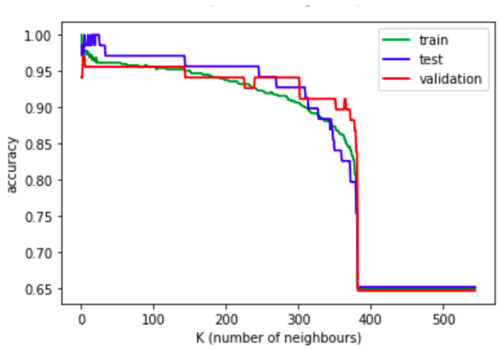


Figure 4: k values vs. accuracy on WBC

We also compared two distance formulas: Euclidean distance and Manhattan distance. Compared to Euclidean distance, using Manhattan distance gives the model a higher training data accuracy on the WBC data set. Otherwise, the difference between the two distance formulas can hardly be observed on the two datasets (Table 1).

4.2 Decision Tree Model

Changing the value of max depth, we found that, as the value of the max depth increases, the training data accuracy of the model gradually increases as the max depth increases. Furthermore, we observe that as the max depth increases, the test data accuracy will have a trend of rising and falling in one of the intervals, and tends to be flat at other times. Take the performance of the Decision Tree model on the WBC dataset when the Misclassification cost is used as an example, the best accuracy is achieved when max depth = 3 (Figure 5).

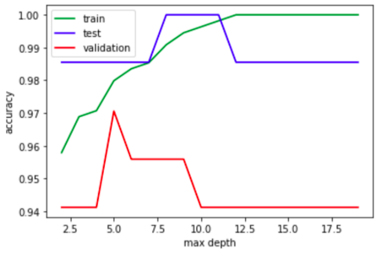


Figure 5: max depth vs. accuracy on WBC

We compared three cost functions: Misclassification, Entropy and Gini index. The impacts of the three cost functions on the training/test data accuracy of the model are very similar: the model achieved very high training data accuracy on both datasets. For the test data accuracy, the performance differed slightly on the datasets. For HP, we observe better accuracy when using Entropy and Gini Index functions (Table 1).

4.3 Decision Boundary Analysis

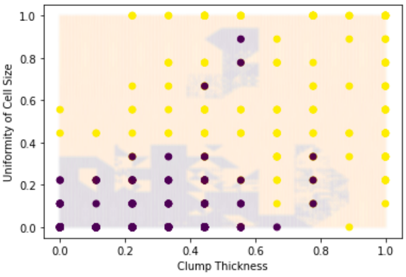


Figure 6: KNN Decision Boundary Plot

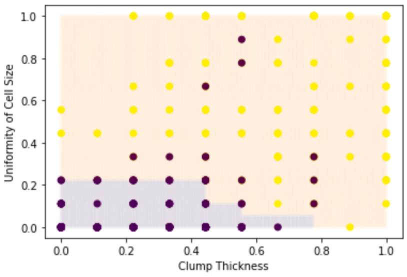


Figure 7: DT Decision Boundary Plot

From the decision boundary plots of the two models (Figure 6 & 7), we can conclude that the decision boundary of the KNN model is more complex and smoother. In contrast, the boundary of the decision tree is defined in lower dimension, sharply dividing classification areas.

4.4 Experimental Models

Through the analysis of the datasets, we noticed that the Hepatitis dataset is highly volatile. To potentially increase the accuracy and reduce the cost of algorithm, we performed feature reduction. Relief filtering, one of the classical feature selection algorithm, (Nancy.P, 2017) is able to select 6 important features - ascites, albumin, histology, bilirubin, spiders and steroids – over the 20 attributes. They are essential in diagnosis of hepatitis based on practical experience.(Sartakhti.J, 2011) Therefore, we selected them and tested our models on this reduced dataset.

Surprisingly, the obtained results show that the optimized model accuracy results did not bring the expected improvement (Table 1). We speculate that the reason behind it may be that the toy datasets used in the experiment that may have many deviants, therefore, leading to results that deviate from the theory.

5 Discussion and Conclusions

We believe this experiment consolidated our study of classification problems through an implementation from scratch in the context of disease condition categorization.

Potential extension of this project include exploring new cost functions, such as the Minkowski distance which is a combination of Euclidean and Manhattan distance (NIST, 2017). One very important improvement is to perform k-fold cross-validation as seen in class. Since the datasets are particularly small and unbalanced, it will benefit from such technique.

In addition, the methodology adopted for the missing values in data points was deletion. Since the proportion of missing data is small, directly deleting the missing values does not cause excessive accuracy errors. However, this processing method may not work for large datasets. We propose to use the KNN model in future learning to train the data points near the missing value for model fitting, and fill the missing values through prediction, to ensure a coherent value.

Overall, through the study of two classification models on two small health datasets, we conclude that both of the Decision Tree Model and KNN model on the given datasets performs well, and the Decision Tree Model performs even better with higer dimension datasets (less prone to curse of dimensionality). The accuracy is very high for all model variants explored. This, in turn, shows that simple classification models can perform very well given a small amount of data and correlated features.

6 Statement of Contributions

We all contributed equally to the report component of this assignment. In particular, Author 1 contributed to most of the coding component, All authors refined the written components of the report and helped with its structure.

References

Uci machine learning repository: Breast cancer wisconsin (diagnostic) data set.

Uci machine learning repository: Hepatitis data set.

Krzysztof Grabczewski and Duch Wlodzislaw. 2002. Heterogeneous forests of decision trees. pages 504–509.

M Mohanapriya and Lekha Jayabalan. 2018. Comparative study between decision tree and knn of data mining classification technique. *Journal of Physics: Conference Series*, 1142:012011.

Akiladevi.R Nancy.P, Sudha.V. 2017. R analysis of feature selection and classification algorithms on hepatitis data.

NIST. 2017. Minkowski distance.

Siamak Ravanbakhsh. 2020. Comp551: Applied machine learning (fall 2020).

Mozafari Sartakhti.J, Zangooei.M. 2011. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (svm-sa).

Peter D. Turney. 1995. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *J. Artif. Int. Res.*, 2(1):369–409.