Project Report
KNN algorithm for classification and prediction
COMP9417

Group member:
Zehao feng, z3435529

Introduction

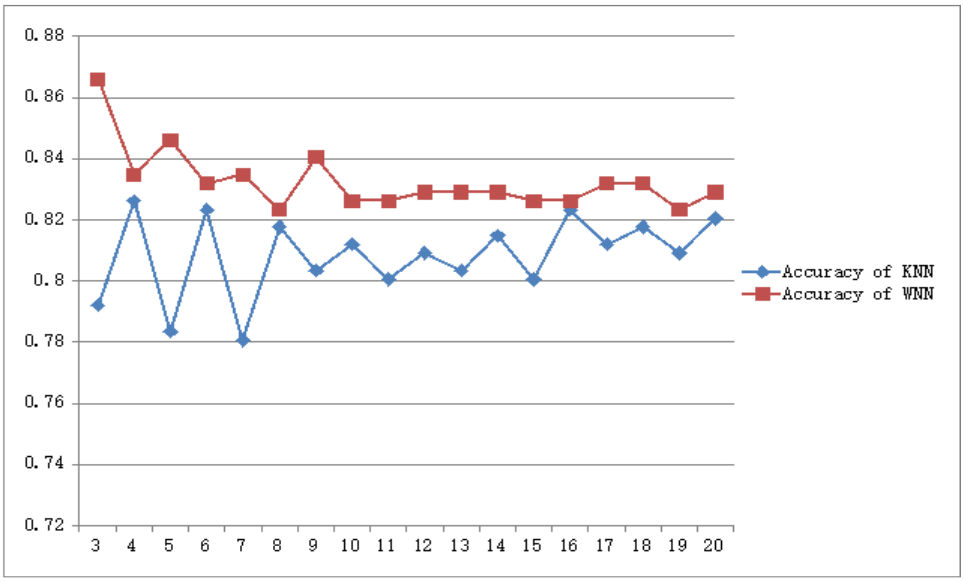
K near neighbor algorithm is a classic algorithm in machine learning. It is simple and easily understood. In this project, the goal was to achieve an implementation of KNN algorithm and a distant-weighted version as well, then evaluate performance by leave-one-out cross-validation. Additionally, an attempt was make to get graphical output showing how the system learns.

Implementation

Two data sets were used in the project from UCI Machine Learning Repository. Database 1 is ionosphere database for classification, and each entry contains 34 attributes and 1 class attribute. Database 2 is autos database for classification, and each entry contains 15 continuous attributes,10 nominal attributes and 1 integer attribute . In this project, our version of KNN and WNN algorithm will use 2 databases for numeric prediction, then evaluated by leave-one-out cross-validation and get the result with range of values of k. For the second database, we firstly only use 15 continuous attributes and 1 integer attribute to calculate, then encoding 10 nominal attributes and use them to calculate as well.
In each test, we calculate the distance of 2 instance with standardized euclidean distance. After get the input ,the program test accuracy of different k, it first select a leave_one_out instance, for each one, it get k-nearest instances. Then use these k-nearest instance to get the number or weigh of each predicted value and use the most one as the predicted result. At last, It compare the predicted result with real result in data and calculate the accuracy of the result for each k.It also calculate the average difference of the predicted value and the real value of each k. Then output the result.
In the WNN algorithm, we use 1/distance as the weigh of each instances in k-nearest to calculate which is the predicted value. Moreover, if the number or the weight of 2 value is the same in k-nearest instance, we set it as the first one.

Experiment

The KNN and WNN algorithm for classification on database ionosphere were evaluated by leave-one-out cross-validation. The output results (Fig.1) contains the average accuracy with different K value (3-20).



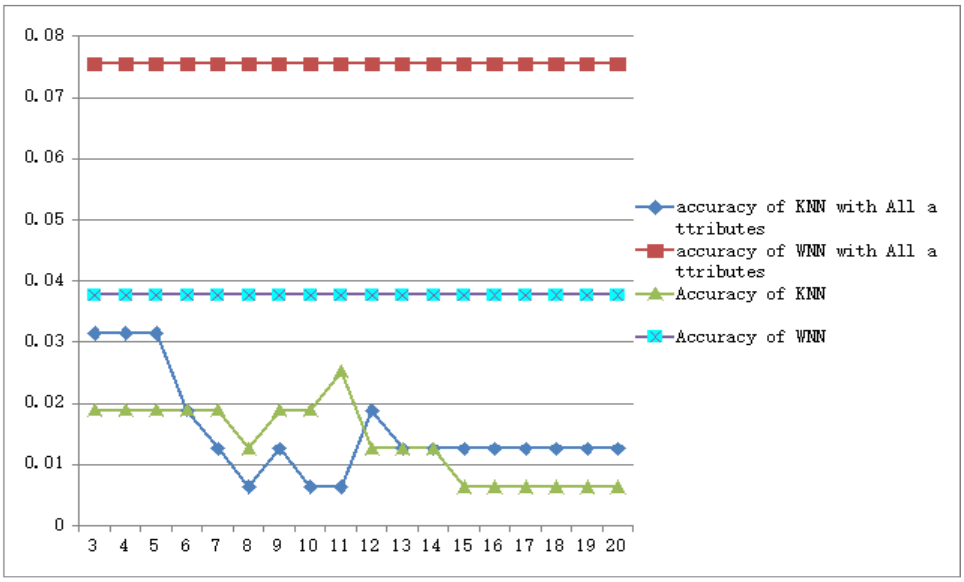
(Fig.1 Accuracy of KNN and WNN algorithm with different K)

The figure 2 compare the accuracy of KNN and WNN algorithm for numeric prediction on database autos with all attributes and database autos without nominal attributes evaluated by leave-one-out cross-validation.

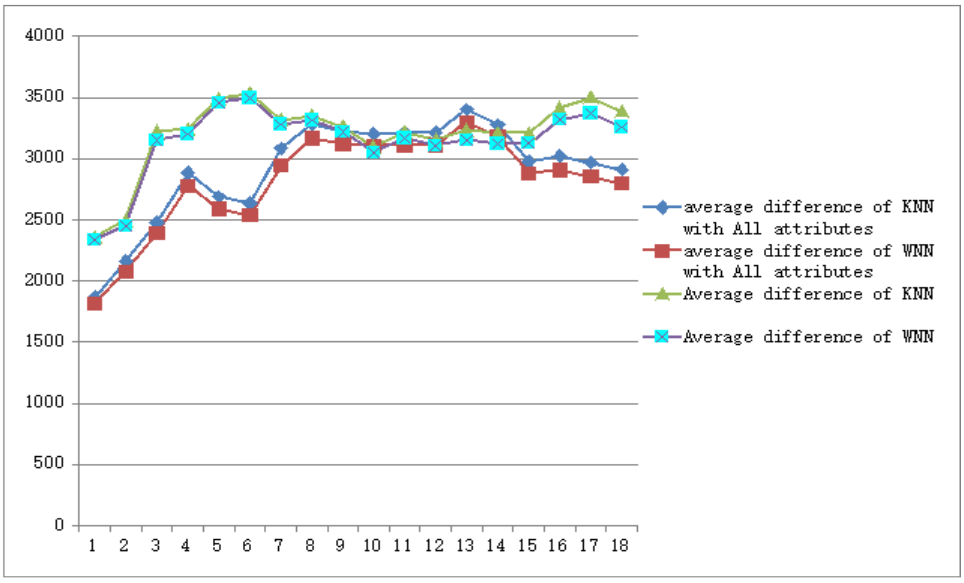
The dark blue line shows the accuracy of KNN with all attributes, the red line shows the accuracy of WNN with all attributes, the green line shows the accuracy of KNN with without nominal attributes and the light blue line shows the accuracy of WNN without nominal attributes.

The figure 3 compare the price difference from the predicted price to real price in database of KNN and WNN algorithm for numeric prediction on database autos with all attributes and database autos without nominal attributes evaluated by leave-one-out cross-validation.

The dark blue line shows price difference from the predicted price to real price of KNN with all attributes, the red line shows the price difference from the predicted price to real price of WNN with all attributes, the green line shows price difference from the predicted price to real price of KNN with without nominal attributes and the light blue line shows price difference from the predicted price to real price of WNN without nominal attributes.



(Fig.2 Accuracy of algorithm on different range of k for autos without nominal attributes)



(Fig.3 Average differences of autos on different range of k for autos with all attributes)

Discussion and Conclusion

The result of the first database ionosphere is better than the result of second database autos. It means KNN algorithm performance better when attributes are continues number and predicted class have less value.

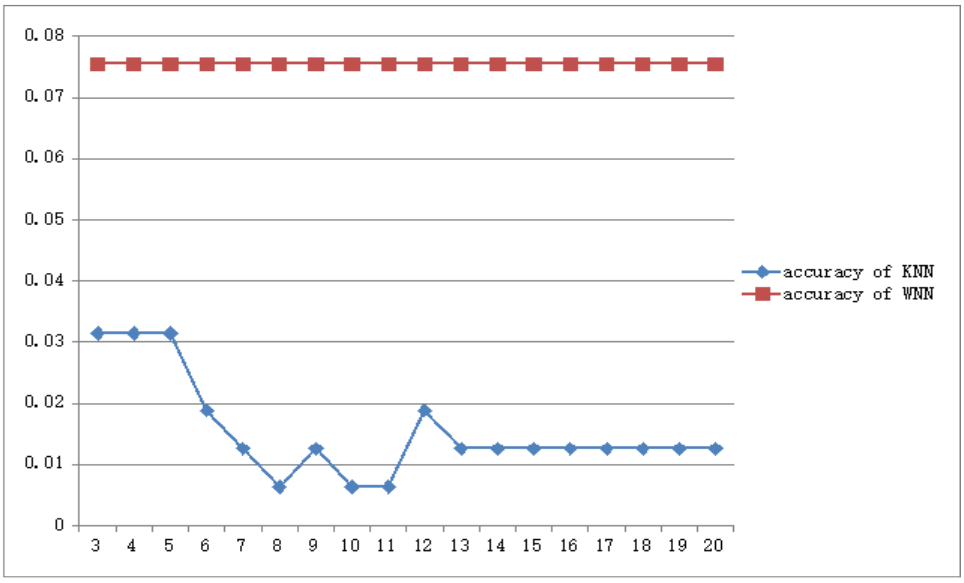
From all results it can be concluded that the distance weighted version of KNN(WNN) algorithm has a better performance than normal KNN.This is due to the more similar instance usually more similar with the real result than the less similar instance, so ,use distance weighted could increase the accuracy of the result.

From comparison between algorithms with different range of attributes, we can find that more attribute does not increase the accuracy of the result. In some range of k, it increased the accuracy, but in some range of k, it does not. It means not more attribute, the result is better. However, better correlation attributes may give better result.

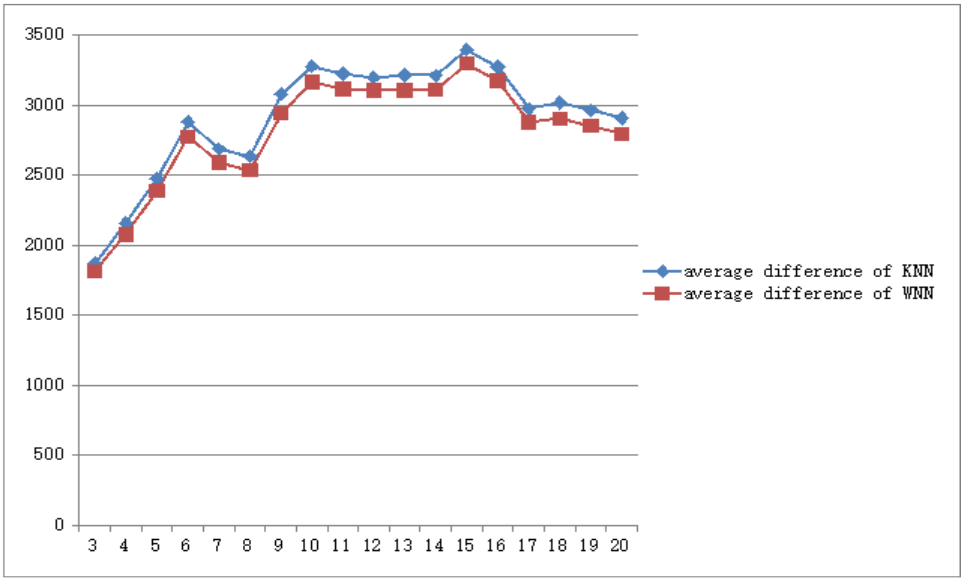
Reference

Tom M. Mitchell, 1997. *Machine Learning*. 1 Edition. McGraw-Hill Education.

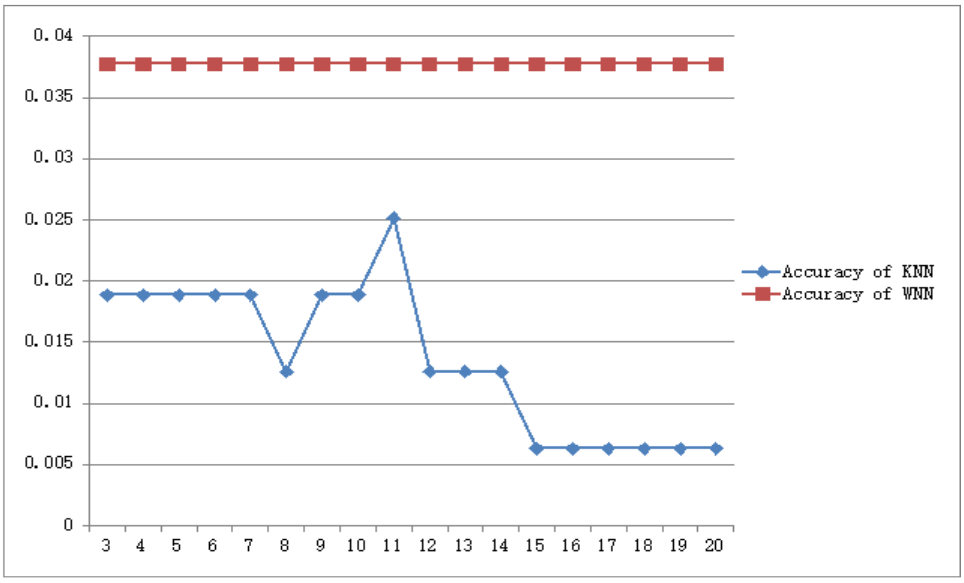
Appendix



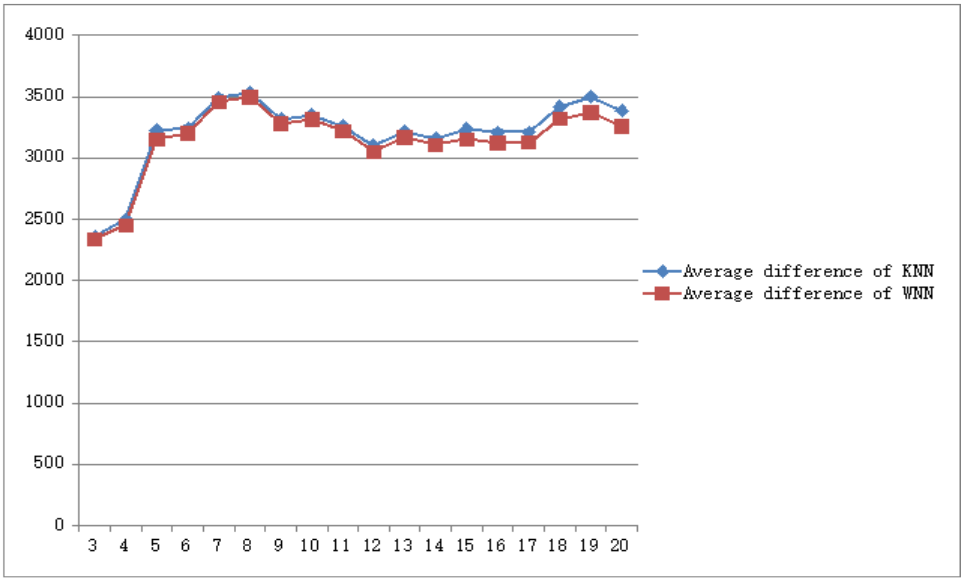
(Fig.4 Accuracies of KNN and WNN algorithm with different K without nominal attributes)



(Fig.5 Average differences of KNN and WNN algorithm with different K without nominal attributes)



(Fig.4 Accuracies of KNN and WNN algorithm with different K with all attributes)



(Fig.5 Average differences of KNN and WNN algorithm with different K with all attributes)