**<Programming Assignment #2>**

2017029716 박혜정

**1. Environment**

- OS : Windows

- Language : C++

- Tool : Visual Studio Community 2019

**2. 컴파일 및 실행방법**

**1) 컴파일 방법**

- Visual studio의 도구창에서 Debug->Release로 변환해주세요.



- Ctrl + F5 를 눌러 실행 시켜주시고, 실행된 파일은 종료 시켜주세요.

- 솔루션 폴더의 Release 폴더에 dt.exe가 생성된 걸 확인할 수 있습니다.

**2) 실행 방법**



소스코드(dt.cpp)은 /decisionTree/dt 폴더 안에,

실행파일(dt.exe)과 데이터셋(dt\_train.txt, dt\_train1.txt, dt\_test.txt, dt\_test1.txt)는 /decisionTree/Release 폴더 안에 있습니다.

따라서 /decisionTree/Release 폴더에서 cmd창에

[실행파일명] [training file name] [test file name] [result file name]

를 입력하여 프로그램을 실행시킬 수 있습니다.

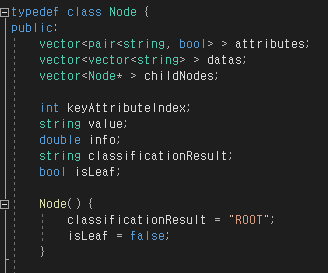
**3. 코드 설명**

**1) Summary**

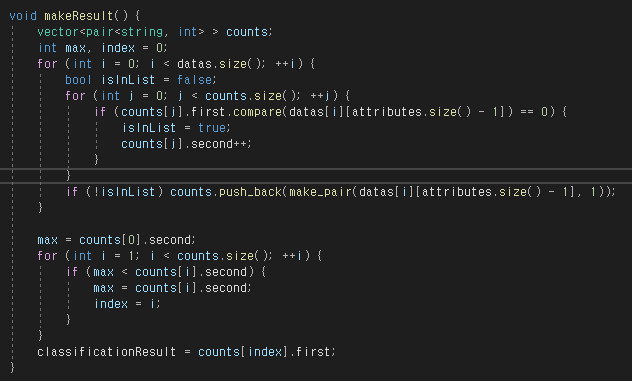
맨 처음 파일을 읽어 attribute List와 data들을 읽어온 후, 각 attribute로 분류 시 entropy가 얼마나 줄어드는지 확인합니다. 가장 많이 줄어드는 attribute로 분류를 진행하며, 분류된 집합에도 같은 동작을 반복하며 decision tree를 만들어나갑니다. (information gain 방법 사용)

Test의 경우 data를 읽어와 tree를 따라가며 결과값을 예측합니다.

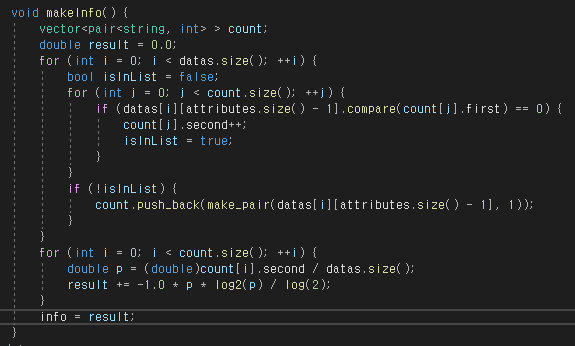
**2) 자료 구조**



Node는 다음과 같이 attribute List 와 데이터들을 담는 datas, 그리고 자식 노드를 가르키는 childNodes를 가진다. 또한 자식을 분류할 때 쓰는 attribute의 index를 저장하는 keyAttributeIndex, 자식 노드의 경우 분류될 때 사용된 attribute의 값을 저장하는 value, 노드의 information을 저장하는 info와 그 노드에서 멈출 경우 cateforical value와 말단 노드인지를 확인하는 isLeaf를 변수로 가지고 있다.

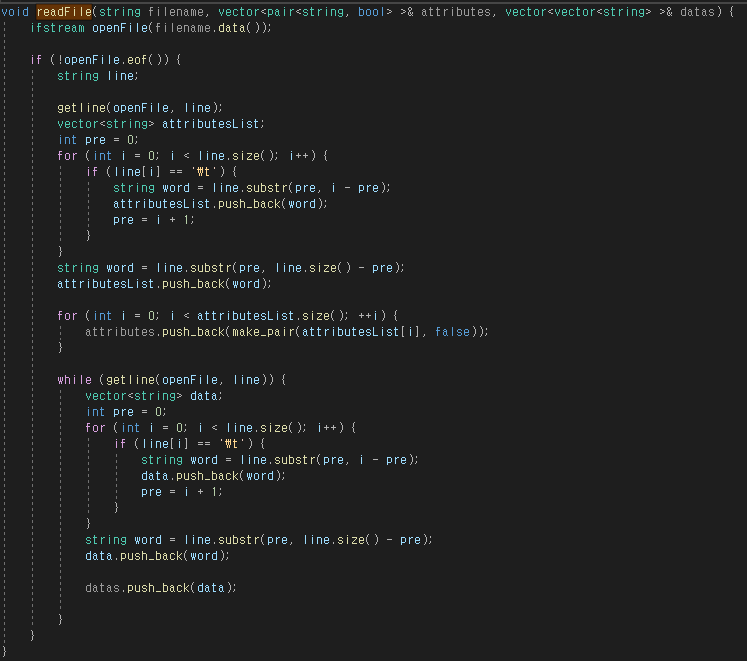


또한 categorical value를 찾는 makeResult 메소드가 있으며,



Information을 계산하는 makeInfo 메소드도 가지고 있다.

**3)파일 읽기**



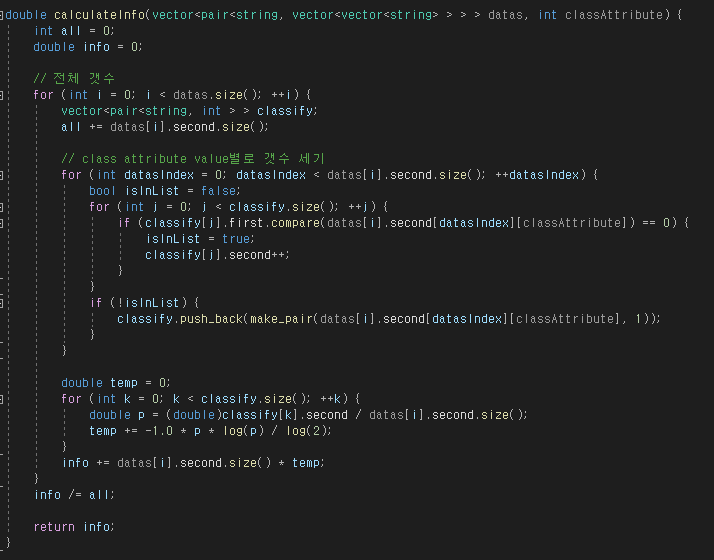
첫 줄을 읽어와 attribute List를 만들고, 그 다음줄부터 파일이 끝날 때까지 읽어서 데이터를 형성합니다. 데이터는 ‘\t’을 기준으로 끊어서 attribute의 값으로 저장합니다.

**4) 분류 함수**



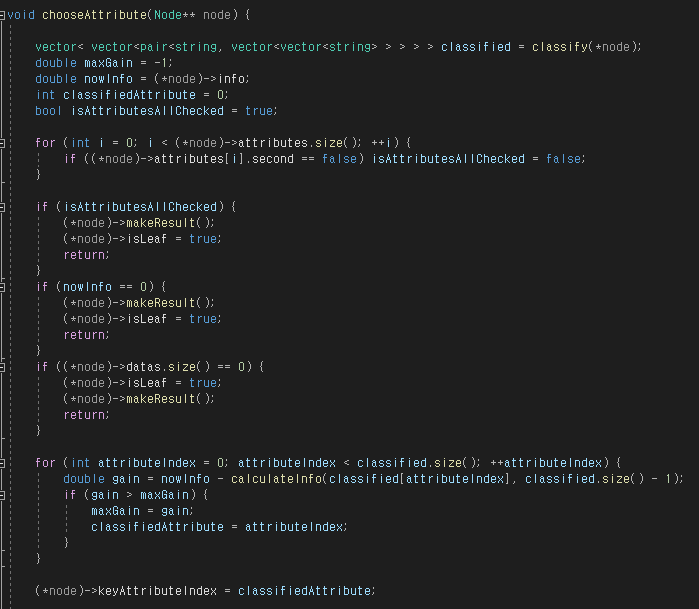
주어진 데이터 셋을 특정 attribute의 값을 기준으로 하여 분류하는 함수입니다.

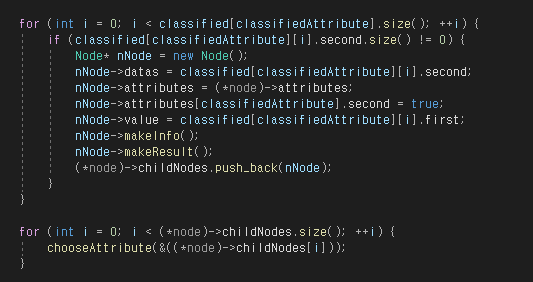
**5) information 계산 함수**



주어진 데이터 셋의 information을 계산하는 함수입니다.

**6) decision tree 만들기**

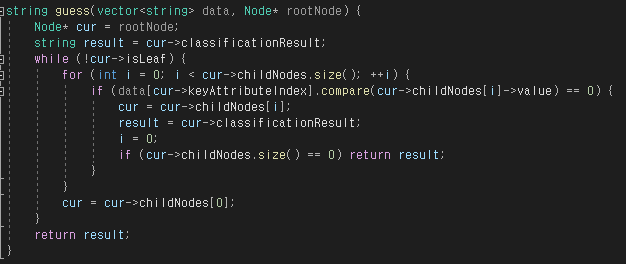




위의 분류함수와 information 계산 함수를 이용하여, 각 attribute를 분류 기준으로 삼았을 때의 information gain을 측정하고, information gain이 가장 큰 attribute로 partition을 진행합니다.

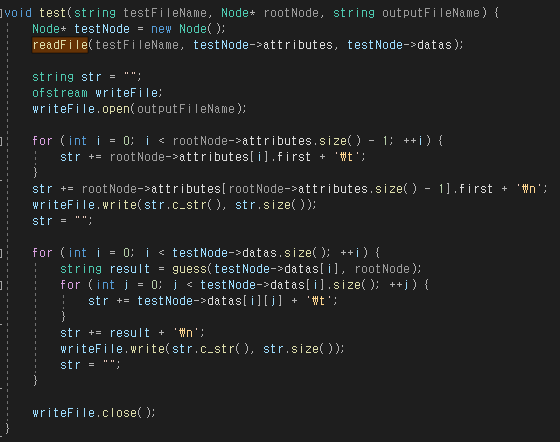
자식노드들에도 재귀적으로 이 과정을 반복하며, 분류된 데이터가 모두 같은 class에 속하거나, 더 이상 데이터가 없거나, 모든 attribute를 분류 기준으로 사용한 경우 분류과정을 멈춥니다.

**7) 결과 예측 함수**



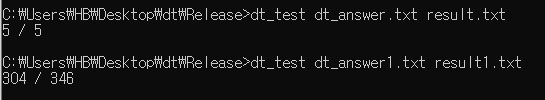
만들어진 decision tree를 따라가며 주어진 data의 categorical value를 예측합니다.

**8) 테스트 함수**



테스트 파일을 읽어와 그 데이터의 결과값을 예측하여 파일로 저장합니다.

**4. dt\_test.exe 결과**



다음과 같은 결과를 확인할 수 있었습니다.