Internet Application

<HW #5>

산업공학과 황영석

2010-12086

1. (100점) **MovieReview** 데이터를 활용하여, 아래의 작업을 수행하시오. *(1-1*의 경우 문제에서 요구하는 *arff* 파일을 함께 제출하시오*. 1-2, 1-3, 1-4*의 경우*,* 결과 비교 및 의견 서술 내용만을 *MS-Word* 문서 파일 형태로 작성하여 제출하시 오*.* 모든 문제에서*,* 명시한 것 이외의 옵션들은 기본값으로 유지하고 진행하시오*.)*

1-1. (10점) Weka에서 StringToWordVector filter를 사용하여, 원본 데이터를 word vector 형태로 변환하고 이를 **word-vector-data.arff** 파일로 저장하시오. 이 과정에서 아래의 옵션을 설정하고 진행하시오.

 - **outputWordCounts: True, stemmer: SnowballStemmer,**  **stopwordsHandler: Rainbow, tokenizer: WordTokenizer**

**“word-vector-data.arff” file submitted**

1-2. (30점) 1-1에서 설정한 옵션 조합을 아래와 같이 변경하면서, 최소 4가지 조합 이상의 옵션들을 각각 적용하여 데이터를 다시 변환한 뒤, Naïve Bayes classifier를 사용하여 sentiment classification을 수행하고 분류 정확도를 서로 비교 분석하시오. 테스트 시에는, 아래의 테스트 옵션을 설정하고 진행하시오.

**- outputWordCounts: {True, False}, stemmer: {SnowballStemmer, NullStemmer}, stopwordsHandler: {Rainbow, Null}, tokenizer: {WordTokenizer, NGramTokenizer}**

- **Test options: Percentage split (%70)**

**1) 1-1 설정한 옵션 (True, SnowballStemmer, Rainbow, WordTokenizer)**

Correctly Classified Instances 403 67.1667 %

**2) 1-1에서 outputWordCounts: True -> False**

Correctly Classified Instances 477 79.5 %

**3) 1-1에서 stemmer: SnowballStemmer -> NullStemmer**

Correctly Classified Instances 403 67.1667 %

**4) 1-1에서 tokenizer: WordTokenizer -> NGramTokenizer**

Correctly Classified Instances 412 68.6667 %

**outputWordCounts: True -> False** 로 한 것이 결과값이 높게 나왔다. 나머지는 비슷비슷한 수준이었다. 다른 조합을 찾아서 사용한다면 더 성능을 높일 수도 있겠다.

1-3. (30점) 1-1에서 변환하여 얻은 데이터에 대하여, 커널을 변경하면서 SVM classifier를 사용하여 sentiment classification을 수행하고, 분류 정확도를 서로 비교 분석하시오. 이 과정에서, 아래의 커널 옵션 및 테스트 옵션을 설정하고 진행하시오.

**- kernel: {PolyKernel(exponent=1.0), PolyKernel(exponent=2.0), RBFKernel}**

- **Test options: Percentage split (%70)**

1. **PolyKernel(exponent=1.0)**

Correctly Classified Instances 465 77.5 %

1. **PolyKernel(exponent=2.0)**

Correctly Classified Instances 482 80.3333 %

1. **RBFKernel**

Correctly Classified Instances 472 78.6667 %

다항 커널로 exponent=2.0 한 것이 분류정확도가 제일 좋았다. RBF, 다항커널 (exponent=1.0)이 그 다음으로 좋았다.

1-4. (30점) 1-1에서 변환하여 얻은 데이터에 대하여, K값을 2~10 사이의 값으로 조정하면서 K-means clustering을 적용하여 sentiment classification을 수행하고, 분류 정확도를 서로 비교 분석하시오. - **Cluster mode: Classes to clusters evaluation**

k=2 Incorrectly clustered instances : 998.0 49.9 %

k=3 Incorrectly clustered instances : 998.0 49.9 %

k=4 Incorrectly clustered instances : 998.0 49.9 %

k=5 Incorrectly clustered instances : 928.0 46.4 %

k=6 Incorrectly clustered instances : 927.0 46.35 %

k=7 Incorrectly clustered instances : 927.0 46.35 %

k=8 Incorrectly clustered instances : 928.0 46.4 %

k=9 Incorrectly clustered instances : 932.0 46.6 %

k=10 Incorrectly clustered instances : 932.0 46.6 %

k 증가시킬 때마다 컴퓨터 수행 시간 증가 -> 계산량 많아지므로

k가 조금 커져야 결과가 조금씩 향상되긴 했으나 어쨌든 그닥 좋은 분류율을 나타내진 않음.

k=6,7이 제일 좋았다.

Distance Measure나 maxIteration 등을 바꾸면 달라질 수 있을 것이다.