논문 목차 (오프라인 구두)

11/18(금) 15:00~15:40 성남시니어산업혁신센터

OA : 15:00~15:40
좌장 : 김영철(홍익대학교)
발표장소 : 1F 오리엔테이션룸

기존 Covid-19와 유행성 인플루엔자 간의 질병 비교 예측 모델

Disease Comparative Prediction Model between the current Covid-19 and Pandemic influenza

정민용[*1], 이다윤[*2] 조나현[*3], 진조아[*4], 윤예동[**5],김영철[**6]

Min-Yong Jung[*], Da-Yun Lee[*], Na-Hyeon Cho[*], Jo-A Jin[*], Ye Dong Yoon[**], R. Young Chul Kim[***]

{pw050503[*1], dana0417, whskgus12, jinjoa1}@naver.com, yedong9477@gmail.com, bob@hongik.ac.kr

요 약

코로나 바이러스로 인한 전례가 없는 감염으로 팬데믹이 선언되는 등 전 세계적으로 심각한 문제가 발생하고 있다. 이에 대한 기존 해결책으로 의료 백신, 의료공학의 선제적 검사 도구를 발명했다. 하지만, 국가 및 지역 간의 격차로 인해 열악한 의료체계 및 의료기관의 접근이 어렵기도 하다. 미국의 경우, 높은 의료 비용 및 검사 비용으로 인해 병원에 가지 않아 피해가 더 커지기도 했다. 따라서, 코로나 바이러스의 감염여부를 식별하기 위한 증상 데이터 기반 소프트웨어 예측 모델을 한다. 특히 다양한 모델들(결정트리, Naive Bayes, KNN, 다중 퍼셉트론 신경망)을 분석 후, 가장 적합한 결정 트리 모델을 선정한다. 이를 위해 모델을 구축하는 과정에서는 설문조사 데이터 및 해외의 증상 데이터를 판단모델에 적용하여 도구화로 구축한다. 즉, 결정트리를 적용하여 데이터 수집 및 분석을 통해 의료적인 검사 및 진단 도구 없이도 코로나 바이러스 및 독감, 알러지, 감기의 판단이 가능하기를 기대한다.

Key Words : Decision Tree Classifier, Machine Learning, Naive Bayes, KNN

I. 서 론

본 논문은 2022년 1 · 2학기 홍익대학교 소프트웨어융합학과 종합설계 프로젝트 결과물로써, 현재 코로나 바이러스가 전 세계적으로 유행하면서 전례가 없는 감염으로 팬데믹이 선언되는 등 심각한 문제가 발생하고 있다. 코로나의 유행으로 인해 하늘길이 막히고 활동이 제한되면서 경제적, 사회적으로 큰 타격을 입게 되었다. 코로나로 인한 사망자 또한 전 세계에서 650만명이 넘게 발생했다.

세계보건기구(WHO)에 따르면, 코로나 바이러스는 독감, 알러지, 감기와 고열, 피로, 기침 등 공통점이 있다. 차이점으로는 전파속도, 전염력, 치료법, 위험도 등이 있다. 위 질병들은 감염 여부를 판단하기 위해서는 PCR검사나 자가진단키트를 이용한다. 하지만, 미국과 같이 높은 의료 비용 및 검사 비용을 부담해야 하는 경우 검사를 하지 못해 본인이 감염 모른 채 전염시킬 수 있다.

따라서, 본 논문에서는 증상 데이터 기반의 소프트웨어 예측 모델을 제안한다. 이는 결정트리 모델을 기반으로 데이터를 수집하고 분석하여 코로나, 독감, 알러지, 감기의 질병들을 판단한다. 모델을 통해 일반 사람들이 검사 도구의 사용 없이도 질병을 식별하여 보다 안전을 기대하려고 한다.

본 논문의 구성은 다음과 같다. 2장에서는 선행연구 및 확률형 머신러닝 알고리즘을 소개한다. 3장에서는 비교 예측 프로토타입 모델에 대해 설명한다. 4장은 결론 및 향후 연구를 언급한다.

*홍익대학교 소프트웨어융합학과 학부생, **담당 조교, ***교수

II. 관련연구

2.1 선행 연구

기존의 COVID19와 유행서 인플루엔자의 비교 예측 프로토타입 모델들을 참조 개발 진행했다[1]. 이 선행된 연구는 Naive Bayes와 KNN(최근접 이웃 분류) 모델들로 진행하였다. 그러나 분류 대상은 코로나 바이러스와 A형 독감을 대상으로, 증상 데이터로는 발열, 기침, 호흡곤란, 인후통 등의 7개의 증상을 선정하였다. 데이터의 경우 Kaggle 데이터를 이용했다. 비교 예측 프로토타입 모델의 성능은 Naive Bayes(약 94%), KNN(100%)의 정확도를 보였다.

2.2 Decision Tree Classifier

이는 머신러닝 지도 학습 알고리즘의 한 종류이며, 의사결정 트리를 사용하여 입력 변수와 목표 변수를 연결한다. 분류와 회귀가 모두 가능하며, 분류는 목표 값이 가질 수 있는 값에 대하여 유한한 경우에 적용한다. 트리는 가지와 노드로 구성되어 있으며 가지는 입력에 대한 참과 거짓을 가지는 논리연산으로 나타낸다[3]. 의사 결정 트리에서 불순도를 계산하는 방법은 대표적으로 지니 계수와 엔트로피 2가지가 있으며, 지니 계수의 경우 해당 수치를 통해 중요한 속성을 파악하여 분류의 기준으로 삼을 수 있고, 가장 중요한 속성

에 대해서만 결정트리를 학습시켜 학습 시간을 최소화 할 수 있다는 특징이 있다[4]. 우리는 불순도 계산법을 지니 계수로 설정하여 연구를 진행 한다.

2.3 Naive Baysian

이는 머신러닝 지도 학습 알고리즘의 또다른 종류이다. 사용되는 데이터의 모든 특성 값은 서로 독립임을 가정하며, 베이즈의 정리(Baye’s Theorem)를 사용하여 이전 사건과 현재 사건을 바탕으로 새로운 사건의 확률을 추론 및 분류한다[5].

```
[기존]
Naive Baysian Train Accuracy:  0.917
Naive Baysian Test Accuracy:  0.920

[개선]
Naive Baysian Train Accuracy : 0.918
Naive Baysian Test Accuracy:  0.922
0.9176028777564051
{'var_smoothing': 1}
```

그림1. GridSearchCV로 개선한 Naive Baysian

그림1은 우리의 데이터로 학습시킨 모델에 GridSearchCV를 적용하여 성능을 개선한 결과이다. 파라미터는 var_smoothing 하나로 진행하고, 0부터 10까지의 범위를 지정하여 GridSearchCV 알고리즘을 적용한 결과 var_smoothing = 1 일 때 정확도는 92%에서 92.2%로 개선되었다.

2.4 K-Nearest Neighbors Classifier

KNN은 머신러닝 지도 학습 알고리즘의 한 종류로서 K개의 최근접하는 이웃을 이용한다는 의미이다. 새로운 데이터가 들어오면, 해당 데이터에 유클리드 거리를 기반으로 가장 가까운 기존 데이터 K개의 이웃 데이터 정보를 통해 새로운 데이터의 정보를 예측한다[6].

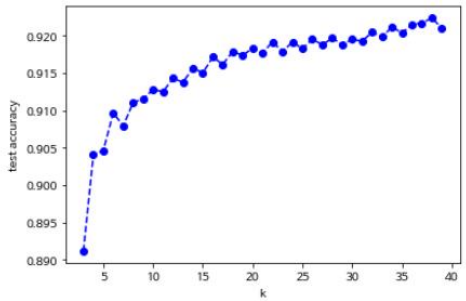

그림2. K에 따른 KNN모델 성능 측정 결과(Accuracy)

그림2는 KNN 알고리즘에 우리의 데이터를 적용하여 정확도를 측정한 결과이다. KNN의 경우 최근접 이웃의 수인 K에 따라 성능이 달라지므로 , 3≤k≤40 해당 범위 내에서 동일한 조건으로 정확도를 측정해보았다. 그 결과 최근접 이웃의 수가 38일 때, 약 92%의 정확도 이다.

2.5 다층 퍼셉트론 신경망 모델

다층퍼셉트론의 신경망은 입력계층, 1개 이상의 은닉계층, 마지막으로 출력계층으로 이뤄진 계층구조 이다. 이 계층 네트워크는 입력, 은닉, 출력 방향으로 연결되고, 각 계층 내부의 연결과 입력계층과 출력계층 사이의 직접적인 연결은 존재하지 않는다. 이런 계층 구조에서 각 계층에 속한 뉴런들은 가중치 연산과 활성화 함수를 통해 분류를 수행한다. 이 모델의 장점은 은닉계층에서 비선형의 input이라도 선형분류가 가능하게끔 매핑시켜주기 때문에 분류 모델의 지도 학습 환경에 효율적인 학습이 가능하다[7].

2.6 모델간의 비교

표1. 분류알고리즘의 성능 비교

항목	Accuracy
Naive Baysian	92.22 %
KNN	92.11 %
Decision Tree	92.02 %
Perceptron	92.07 %

코로나 예측 판단 시스템에 적합한 모델을 선정하기 위해 연구를 진행한 알고리즘의 성능을 비교이다. 비교 지표는 테스트 데이터의 분류 정확도로 설정하고, 테스트 데이터는 총 데이터에서 랜덤(Random_state = 2)으로 30%를 추출한 13352개의 데이터이다. 각 모델들은 GridSearchCV 알고리즘을 통해 찾은 최적화 된 하이퍼 파라미터 값을 할당한 후 생성된다. 성능 비교를 진행한 결과, 모든 알고리즘이 예측에 대해 동일한 성능을 보인다. 따라서 우리는 추후 개선 가능성을 고려하여, 알고리즘의 논리구조에 대해 시각화가 가능한 Decision Tree Classifier를 앞서 제안한 프로토타입 모델에 적용한다.

III. 비교 예측 프로토타입 모델

3.1 데이터 수집

실험 데이터로 사용을 위해 코로나19와 독감을 걸린 적이 있는 홍익대 소프트웨어융합학과 동료 학생들을 대상으로, 설문조사는 2022년 5월 24일부터 2022년 5월 28일, 2022년 8월 21일부터 2022년 8월 26일까지로 두 차례동안 총 11일에 걸쳐 진행되었다. 감염된 변이 바이러스의 종류 및 연령대와 발열감, 기침, 오한, 근육통, 구토, 인후통, 복통, 설사, 콧물, 코막힘, 미각 상실, 후각 상실, 재채기 등 25여 가지의 질병 발병 후 증상을 수집하는 목적으로 진행했으며, 코로나 163건, 독감 25건의 데이터를 수집하였다.

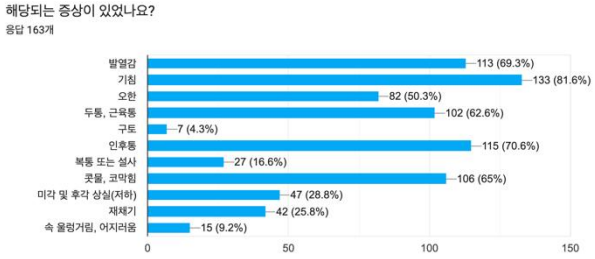

그림3. 코로나 19 감염 후 증상 설문응답

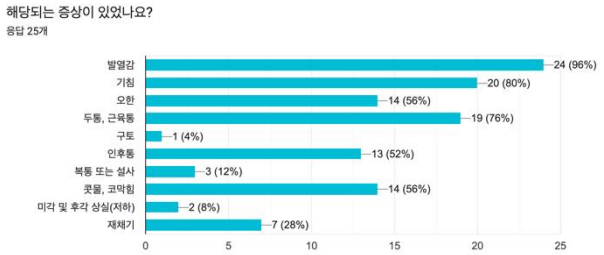

그림4. 독감 감염 후 증상 설문응답

설문 응답 결과, 코로나와 독감 모두 발열감, 기침, 오한이 가장 많이 증상으로 나왔으며, 코로나의 경우 콧물, 코막힘과 미각 및 후각 상실, 어지러움 증상 등이 독감보다 높게 나타났다.

설문 조사로 수집한 증상 데이터들로 모델링을 진행한 결과, 모델에 데이터 부족으로 인해 학습 데이터에 대해서는 정확도가 높지만 새로 들어오는 실제 데이터에 대한 오차가 커지는 과적합 현상이 일어났다. 과적합 현상을 제거하기 위해 GitHub에서 앞서 진행한 설문조사의 증상데이터들과 비슷한 증상 컬럼을 가진 데이터셋을 발견하여 추가로 사용하였다. 데이터셋에는 기침, 근육통, 피로, 목아픔, 콧물, 코막힘, 열, 미각손실 등 20개의 증상 컬럼과 심각한 코로나, 경증 코로나, 알레르기, 감기, 일반 독감으로 구분되는 1개의 병명 컬럼을 가지며 총 44506개의 행의 데이터가 포함되어 있다[2].

3.2 데이터 전처리

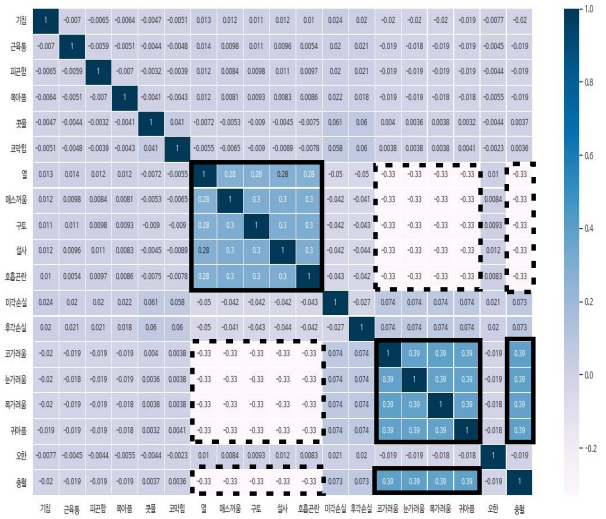

그림5. 변수간의 상관계수

그림5는 데이터 속성 값 사이의 상관성을 보여주는 HeatMap이다. 굵은 선은 양적 상관관계이며, 점선은 음적 상관관계이다. 열과 매스꺼움, 구토, 설사는 양의 상관관계를 보이고 있으며, 코가려움, 눈가려움, 목가려움, 귀아픔의 경우는 음적 상관관계를 보이고 있다. ±0.3 정도는 약한 양적, 음적 선형관계이므로 ,우리의 속성 값 들은 강한 관계성이 보이지 않는다. 따라서 독립변수의 일부가 다른 독립변수의 조합으로 표현이 되는 다중공선성의 문제점이 발생하지 않을 것이라 판단된다.

표2. 라벨 인코딩 데이터 변환방법 테이블

Label Encoder	병명
0	ALLERGY
1	COLD
2	COMMON FLU
3	COVID

수집 데이터는 MILD COVID(경증), SEVERVE COVID(중증), COLD(감기), ALLERGY(알러지), COMMON FLU(독감) 총 5개의 병명이 존재하며, 데이터의 원활한 학습을 위하여 중증 코로나와 경증 코로나 두 데이터를 결합하여 하나의 결합된 코로나 증상 데이터를 구성하였다. 구성 이후 데이터의 전처리를 위하여 4가지 병명 ALLERGY, COLD, COMMON FLU, COVID에 대한 라벨 인코더를 진행하였다.

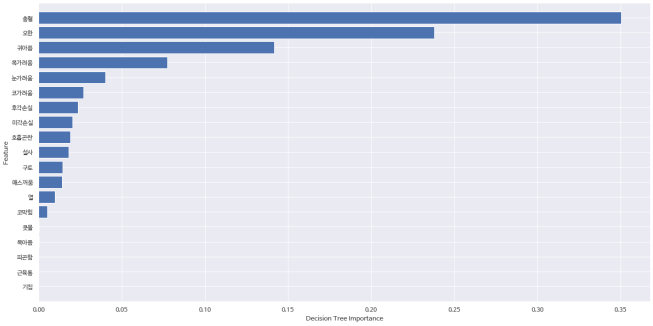

그림6. Decision Tree 변수 중요도

이후 결정트리 중요도를 그래프화 시켰다. 각 Feature 별 중요도는 충혈, 오한, 귀아픔, 목가려움, 눈가려움, 코가려움, 후각손실, 미각손실, 호흡곤란, 설사, 구토, 매스꺼움, 열, 코막힘, 콧물, 목아픔, 피곤함, 근육통, 기침으로 구성되어있으며, 이 중 충혈이 중요도가 가장 높고, 오한, 귀아픔, 목가려움 등이 순서대로 높았다.

3.3 모델 구조화

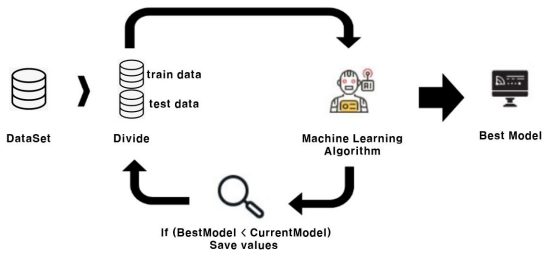

그림7. 모델 구성 방식

그림7은 적용할 모델 구성 방식이다. 모델의 성능을 최대한으로 끌어내기 위해 훈련데이터셋과 검증데이터셋의 비율을 1:9에서 3:7까지 할당했으며, 과적합을 방지하기 위해 랜덤으로 뽑을 수 있도록 시드값을 난수로 생성해 넣을 수 있도록 했다.

본 논문은 데이터셋의 비율 뿐만 아니라 학습 알고리즘 마다 존

재하는 파라미터들을 파이썬 sklearn패키지의 GreedsearchCV를 사용해서 그리디 알고리즘에 기반 모델학습 결과 비교를 통해 최적의 값을 찾아 최종모델에 적용한다.

3.4 학습모델 - Decision Tree Classifier

우리는 총 44506개의 데이터로 모델 학습을 진행하였으며, Train 데이터와 Test 데이터는 최종적으로 7:3의 비율로 분할하였다. 그림4의 첫 번째 모델은 Default값으로만 생성한 Decision Tree 모델이며, 정확도는 Test 데이터 기준 88.3%이다. 그림4의 두 번째 모델은 성능 개선을 위해 GridSearchCV 알고리즘을 적용한 모델이다. 해당 알고리즘을 통해 찾은 최적의 파라미터 값은 max_depth=13, min_samples_leaf=2 이며, Test기준 정확도는 88.3%에서 92%로 약 4% 상승하였다.

```
[ hyperparameter tuning before ]
1. Decision Tree Train Accuracy:  0.942
1. Decision Tree Test Accuracy:  0.883

[ hyperparameter tuning after ]
2. Decision Tree Train Accuracy::  0.928
2. Decision Tree Test Accuracy: :  0.920
best_params : {'max_depth': 13, 'min_samples_leaf': 2}
```

그림8. Decision Tree Classifier 파라미터 조정 전 후 성능 비교

```
[학습 데이터 속성 값]
Index(['기침', '근육통', '피곤함', '목아픔', '콧물', '코막힘', '열', '매스꺼움', '구토', '설사', '호흡곤란',
       '미각손실', '후각손실', '코가려움', '눈가려움', '목가려움', '귀아픔', '오한', '충혈', '병명'],
      dtype='object')

[설문조사 데이터 속성 값]
Index(['발열감', '기침', '오한', '두통', '근육통', '인후통', '콧물', '미각 및 후각상실(저하)', '코막힘',
       '재채기', '목통 또는 설사', '피로감 혹은 무기력함', '흉통(호흡곤란 포함)', '어지러움', '소화불량', '식욕부진',
       '피부질환(발진 가려움 건조함 등)', '시야흐림 등 시각문제', '가래(피가래 포함)', '귀 통증 혹은 청각저하', '탈모',
       '체력저하', '건망증', '코로나'],
      dtype='object')
```

그림9. 학습데이터와 설문조사 데이터 속성 값

```
[ Decision Tree Classifier ]
Decision Tree Train Accuracy:  0.928
Decision Tree Survey Accuracy:  0.366
```

그림10. 설문조사 데이터에 대한 프로토타입 모델 성능

프로토타입 모델이 추후에 추가될 데이터에도 잘 적용되는지 확인하기 위해, 우리가 수집한 설문조사 데이터에 적용해보았다. 예측 결과는 134개의 데이터 중 49개를 맞췄고, 약 37%의 정확도이다. 그림10에서 보이듯이, 학습 데이터와 기존 데이터의 속성 값이 다른 것을 감안한다면, 추후에 현재 학습 데이터 속성과 일치하는 데이터가 추가되었을 때, 성능이 개선될 여지는 충분해 보인다.

3.5 학습모델 가시화

생성한 모델의 판단과정을 시각화한 자료가 그림 11 이다. 트리의 깊이는 13층이며, 한 노드 당 최소한 2개의 가지가 나오도록 설정하였다. Root Node는 '귀아픔' 이며, 각 노드의 증상에 해당되면 오른쪽 노드로, 해당되지 않으면 왼쪽 노드로 뻗어나가는 구조이다.

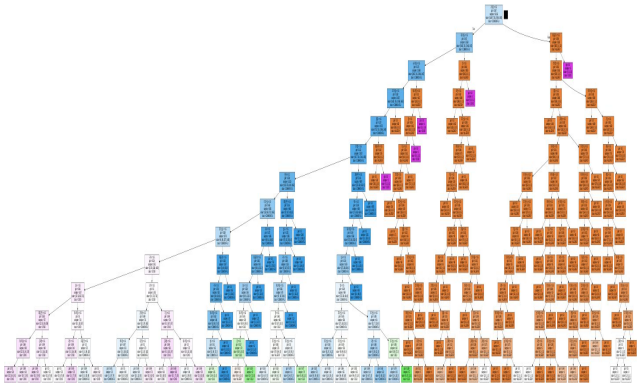

그림 11. 프로토 타입 모델 Decision Tree 알고리즘 시각화

그림 11은 우리가 생성한 모델의 판단과정을 시각화한 자료이다. 트리의 깊이는 13층이며, 한 노드 당 최소한 2개의 가지가 나오도록 설정하였다. Root Node는 '귀아픔' 이며, 각 노드의 증상에 해당되면 오른쪽 노드로, 해당되지 않으면 왼쪽 노드로 뻗어나가는 구조이다. 시각화 결과 현 프로토 타입 모델은 알레르기와 독감에 대한 분류가 우세한 모델이며, 해당 결과를 토대로 논리구조의 취약점을 개선할 수 있을 것이라 생각한다.

IV. 결 론

본 논문은 4가지의 질병(코로나, 독감, 알러지, 감기)을 판단하는 증상 데이터 기반 소프트웨어 예측 모델을 제안한다. 코로나 바이러스로 인해 큰 피해를 겪고 위드 코로나가 계속되는 가운데, 본 논문이 제안하는 모델은 결정트리 모델을 적용하여 위 질병들을 판단한다. 본 연구를 통해, 일반 사람들 및 의료접근이 어려운 지역의 사람들도 보다 쉽게 검사도구 없이 코로나 감염여부를 판단할 수 있을 것으로 기대한다.

향후, 데이터 수집을 지속적으로 수집하여 증상데이터를 추가하고 지속적으로 모델을 학습시켜 성능을 향상시키기를 기대한다.

코로나 바이러스 외에도 다른 질병에 대한 신뢰성 있는 데이터를 수집한다면, 이 또한 적용하여 모델을 확장시킬 수 있을 것이다. 이후 제안한 모델의 서비스 제공을 위해 웹페이지 및 어플리케이션을 개발을 진행하여 한다.

참 고 문 헌

[1] 유승빈, 김영희, 김장환, 김영철, "COVID19와 유행성 인플루엔자의 비교 예측 프로토타입 모델", 한국정보과학회, pp. 1295-1297, 2020.12

[2] Covid19_Prediction_Using_Symptoms, https://github.com/GuduruAishwarya/Covid19_Prediction_Using_Symptoms

[3] 진민우.염성관, "트래픽 속성 개수를 고려한 의사 결정 트리 DDoS 기반 분석" 한국정보통신학회논문지,Vol.25,.Jan.2021.

[4] 염성관.박상윤.신광성, "중요도를 고려한 의사 결정 트리 기반 DDoS 공격 분석, 한국정보통신학회 2021년 춘계 종합학술대회

논문집
[5] 정하림.김홍회.박상민.한음.김경헌.윤일수,나이브 베이즈 빅데이터 분류기를 이용한 렌터카 교통사고 심각도 예측’,한국ITS학회논문지,Vol.16 No.4(2017),August, 2017
[6] 최은선.박남제, K-NN 알고리즘 이해를 기반한 머신러닝교육프로그램의 개발 및 적용,Journal of The Korean Association of Information Education Vol. 25, No. 1,February 2021
[7] 김민하. "다층 퍼셉트론과 부분 투영을 이용한 차선 색상 및 형태 인식." 국내석사학위논문 부산대학교 대학원, 2014. 부산