정보이론 기반의 병합식 클러스터링 기법 연구

(Agglomerative Clustering Methods based on Information Theory)

김 병 희
서울대학교 컴퓨터공학부

Abstract

'Clustering' is an analyzing method which automatically grasps the features inherent in data. It can be implemented in diverse ways based on several distance-measures and algorithms. This thesis applied the 'double clustering' technique based on the 'information bottleneck (IB) method' to document and DNA microarray data. The IB method uses distance for probability distributions as the basis of distance-measurement. Compared with other bases, the accuracy of the IB method turned out to be higher in the experiments on document data. Although there exists no absolute answer for the clustering of microarray data, which made it difficult to compare the IB method with others, some meaningful results coincident to the prior biological knowledge have been extracted.

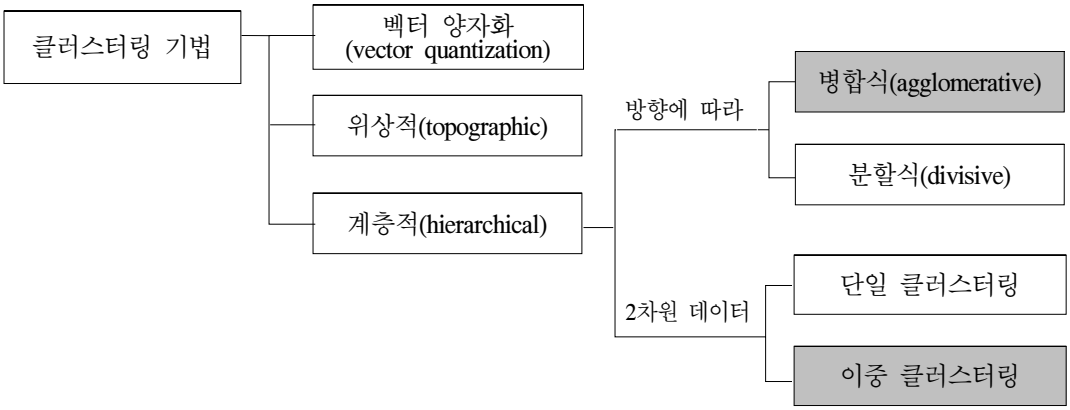
I. 서 론

정보화 및 기술의 발달에 따라 다양한 데이터가 엄청난 속도로 쏟아져 나오고 있다. 대표적인 예로는 인터넷 환경이 널리 확산되면서 기하급수적으로 증가하고 있는 각 분야의 웹 문서들과 생화학 실험기법의 발달에 따른 대량의 생물학적 데이터 등을 들 수 있다. 이러한 대량의 데이터에서 필요한 정보 및 지식을 산출하기 위해서는 적절한 분석 기법이 적용되어야 하며, 이는 과거의 소량의 데이터를 다루던 기법과는 다른 방법론들을 요구하고 있다. 대량의 데이터에 대한 분석법들은 통계학 및 인공지능의 기계학습 분야에서 연구되고 있다. 본 논문에서는 다양한 데이터 분석법 중 클러스터링(clustering) 기법을 다룬다. 클러스터링은 데이터에 내재된 특성을 자동으로 추출하는 기법으로 기계학습의 무감독학습(unsupervised learning)에 해당한다.

클러스터링은 주어진 데이터를 의미 있는 집단(subgroup)들로 분류하며, 데이터 분석, 시각화, 압축 및 전처리와 관련된 많은 분야에서 널리 응용되고 있다. 클러스터링을 통

해 데이터는 비슷한 항목들마다 별개의 클러스터를 형성하며, 이 과정에서 가장 중요한 것은 항목들 간의 유사도 또는 거리를 측정하는 기준이다. 이 기준 및 알고리즘을 바탕으로 클러스터링 기법은 [그림 1]과 같이 벡터 양자화(vector quantization), 위상적(topographic) 클러스터링, 계층적(hierarchical) 클러스터링 등으로 구분될 수 있다 [6].

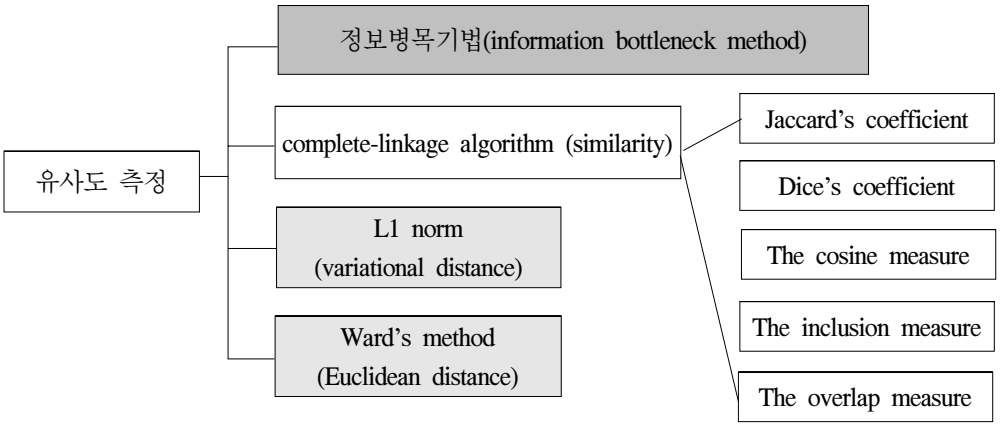
벡터 양자화는 클러스터링을 데이터 압축에서 보는 관점이다. 위상적 클러스터링은 분석 결과를 위상적 지도(topographic map)의 형태로 표현하며 대표적인 기법들로는 자기조직신경망(self-organizing map, SOM), GTM(generative topographic mapping) 등이 있다. 계층적 클러스터링은, 가장 상위에 모든 문서를 포함하는 하나의 클러스터로부터 출발하여 가장 하위에는 각 문서가 하나의 클러스터에 할당되는 트리 구조를 형성하는 모델이다. 트리를 구성하는 방법에 따라 상향식(bottom-up) 또는 병합식(agglomerative)과 하향식(top-down) 또는 분할식(divisive)으로 나눌 수 있으며, 이 논문에서 다룰 기법은 **병합식**에 해당한다.



[그림 1] 클러스터링 기법의 분류 (본 논문에서 적용할 기법은 음영표시가 되어 있다.)

기존의 클러스터링 기법에서는 대상이 되는 변수에 대해 직접 클러스터링을 실시하였으나, 최근에 소개된 이중 클러스터링(double clustering) 기법에 따르면, 관련 변수에 대한 클러스터링을 먼저 실시하여 정보를 압축한 상태에서 대상 변수에 대한 클러스터링을 실시하는 것이 더 효율적이다 [10]. 이 논문에서는 **이중 클러스터링 기법**의 이론적 기반을 살펴보고 그 효율성을 확인해볼 것이다.

병합식 클러스터링에서 가장 중요한 사항은, 병합할 클러스터를 선택하는 기준이다. 보통, 각 클러스터간의 유사도를 측정하여 가장 유사한 두 클러스터를 하나로 병합하게 된다. 유사도를 측정하는 다양한 방법이 제시되어 있으며, 이번 논문에서 다루게 될 방법을 [그림 2]로 구성하였다.



[그림 2] 클러스터링 기법에서의 유사도 측정 방법 분류

이번 논문에서는 '분포 **클러스터링(distributional clustering)**'이란 개념에 적절한 유사도 측정 기법으로서 제시된 **정보 병목 기법을 이용한** 클러스터링을 실시한다. 더불어 [그림 2]에 제시된 다른 기법(L1 norm 기법, Ward의 기법)을 병행하여 그 차이를 알아본다.

각 클러스터링의 결과를 판단하고 차이를 알기 위해서는 적절한 평가 기준이 필요하다. 가장 좋은 방법은 정답이 주어진 상태에서 실험 결과를 여기에 견주어 보는 것이다. 이 때, 정답으로서의 자료를 선택하는 기준이 필요한데, 이에 대해서는 본문에서 다루겠다. 정확도는 [10]에서 다루어진 '**contingency table**'을 이용한 측정 및 '**F_1-measure**'법을 적용하여 측정할 것이다.

본 논문에서는 정보 이론을 기반으로 정보 병목 기법, 이중 클러스터링 알고리즘이 구성되는 과정을 살펴볼 것이다. 각 이론과 관련된 문헌을 연구하고 정리하는 데 중점을 둔다. 기존의 여러 클러스터링 기법들 및 그 한계를 알아보고, 이론적으로 이중 클러스터링과 정보 병목 기법이 제시하는 개선된 해결책에 대해 살펴본다.

알고리즘을 구성한 후, C++를 이용해 이를 프로그램으로 구현한다. 그리고 클러스터링 결과를 신뢰할 수 있는 적절한 대상으로서 TREC-8 정보검색 분석대회에 사용된 문서-단어(document-word) 행렬을 이용해 문서 클러스터링을 실시한다. 이 때, 다양한 유사도 측정 기법에 대해 살펴보고 이들을 적용한 실험을 병행하여 그 결과를 비교, 분석한다. 다음으로, 같은 기법을 DNA 마이크로어레이(microarray) 데이터 분석에 적용해본다. 데이터는 CAMDA 2001 DNA 칩 데이터 분석 경진대회 문제에 사용된 자료를 사용한다.

논문의 구성은 다음과 같다. 먼저 2장에서 클러스터링에 대해 전반적으로 살펴보고, 문서 클러스터링과 DNA 마이크로어레이 데이터에 대해 간략하게 살펴본다. 3장에서는

클러스터링의 이론적 기반으로서 필요한 ‘정보 이론’의 주요 개념을 정리하고, 실험에 적용할 ‘정보 병목 기법’과 ‘이중 클러스터링’에 대해 이론적으로 살펴본다. 이를 바탕으로 알고리즘 및 프로그램을 구성한다. 4장에서는 클러스터링 실험 구성 및 결과를 다룬다. 결과를 신뢰할 수 있는 문서 및 마이크로어레이 자료의 선택, 클러스터링 결과에 대한 평가방법 선택을 거쳐 클러스터링을 실시한다. 단일 클러스터링과 이중 클러스터링, 여러 유사도 측정 기법 및 평가방법을 병행하여 그 결과를 비교한다. 5장에서는 연구 결과를 요약하고, 실험에 적용한 알고리즘의 단점에 대해 논의한다.

II. 관련연구

1. 기계 학습 개관

기계 학습은 크게 감독학습(supervised learning), 보강학습(reinforcement learning), 무감독학습(unsupervised learning)으로 분류할 수 있다. 감독 학습은 분류(classification) 및 회귀(regression)를 위한 학습법으로, '입력-목표(예를 들면 특정 클래스)출력 쌍'의 학습 데이터를 바탕으로, 임의의 입력과 목표값 사이의 관계를 추정하는 것을 목표로 한다. 보강 학습은 입력에 대한 행동에 대해 환경으로부터 보상(reward)이 주어지며, 보상을 최대로 받는 것을 목표로 하는 학습법이다.

무감독 학습법은 데이터의 샘플 외에는 목표값도, 보상도 주어지지 않는 학습법이다. 학습의 목표는 데이터에 내재된 분포 특성에 대한 적절한 표현을 찾아내는 것이다. 구체적인 목표로는 중복성 절감, 정보량 최대화, 엔트로피 최소화, 재구성(reconstruction) 오차 최소화 등을 들 수 있다. 무감독 학습법의 응용은 다양한 분야에서 이루어지고 있으며, 감독 학습법을 위한 전처리, 데이터 압축, 밀도 측정, 원인(source) 분리, 데이터 가시화 등을 예로 들 수 있다. 무감독 학습법의 주된 접근법의 하나가 바로 클러스터링으로서 이 논문에서는 초점을 여기에 둘 것이다.

2. 클러스터링 개요

클러스터링은 고차원의 데이터를 시각화하거나 압축하는 방법 중 하나로서, 데이터 분석과 패턴 인식을 통해 내재된 클러스터 구조를 추출하려는 기술이다. 그러나, 클러스터링이란 말은 다소 '애매한(fuzzy)' 용어로, 많은 종류의 클러스터링 알고리즘이 존재하며, 클러스터링을 적용하는 문제 자체도 다양할 뿐만 아니라 대부분의 경우 여러 알고리즘을 다양하게 적용해 볼 수 있는 특성이 있다.

클러스터링과 관련된 여러 중요한 논점은 다음과 같다 [6][7].

(1) 데이터 종류

가장 상위 수준에서 클러스터링 알고리즘은 처리되는 데이터의 특성에 의해 구분할 수 있다. 가장 일반적인 경우는 각 데이터가 유클리드 공간에서의 벡터로 표현 가능한 경우이다. 즉, 절대 좌표계에서의 수치로 정의된 데이터라 할 수 있다.

또 다른 경우는 관계형(relational) 데이터로서, 데이터가 상대 좌표로, 즉 두 점 사이의 거리의 쌍으로서 주어지는 경우이다. 이 경우 거리의 쌍은 유사도(similarity)(또는 비유

사도)의 개념으로 이해되는데, 종종 거리의 세 가지 공리(양수, 대칭성, 삼각부등식)가 성립하지 않는다.

다루기 힘들어지는 경우는 데이터의 특성이 벡터 또는 관계형으로 표현할 수 없고, 근본적으로 정성적(qualitative)인 경우이다. 이런 데이터는 명사형(nominal) 데이터라 불리기도 한다.

(2) 유사도 / 거리 측정

다양한 형태의 계층적(hierarchical) 클러스터링을 위시한 여러 클러스터링 알고리즘의 시작점은 클러스터를 구분할 대상 간의 유사도(similarity) 또는 거리(distance)의 쌍에 대한 행렬이다. 때로는 이 거리쌍 대신 클래스의 중심과 데이터 위치 간의 이격도를 벡터 양자화(vector quantization) 방법으로 측정하기도 한다. 유사도, 거리 또는 이격도를 엄밀하게 정의하는 것은 결정적인 부분으로, 클러스터링 알고리즘의 결과에 상당한 영향을 미칠 수 있다. 어떤 경우든, 이런 과정을 통해 클러스터링 문제를 여러 방식의 최적화 문제로 변환하여, 명확하게 구분되고 내부적으로도 유사 자료로 이루어진 클래스를 찾아내는 것이 목적이다.

서론에서 다룬 ‘L1 norm’과 ‘Ward's method’는 유클리드식 거리를 응용한 대표적 예로 볼 수 있고, cosine measure 또는 Pearson correlation coefficient는 두 단위 벡터 간의 스칼라곱을 이용한다.

이 주제에 대한 내용은 3장에서 정보 병목 기법에 대해 논하면서 조금 더 다루도록 하겠다.

(3) 파라미터 / 비(非)파라미터 접근법

파라미터(parameter) 접근법은 자료상에 어떤 구조가 있다는 가정에서 시작한다. 일반적으로, 이런 가정은 구체적으로 전역적인 최적화 조건, 즉 비용 함수(cost function)로서 나타나며, 이 접근법의 목적은 비용 함수값을 최소화하는 것이다. 이 종류의 알고리즘은 다시 생성적(generative) 모델과 재구성(reconstructive) 모델로 구분할 수 있다.

비파라미터 접근법은 데이터 구조에 대해서는 거의 가정을 하지 않으며, 일반적으로 어떤 부분적 제약조건을 만족시키는 클러스터 구조를 찾아나간다. 비파라미터 접근법의 대표적인 예가 계층적 클러스터링 알고리즘으로, 병합식(agglomerative) 또는 분할식(divisive)으로 클러스터의 트리를 구성하게 된다.

비파라미터 클러스터링 기법의 중요한 장점은 데이터의 분포에 대한 어떠한 가정도

하지 않는다는 것이다. 또, 일반적으로 이들 기법은 데이터 아이템을 유클리드 공간 상의 벡터로 표현할 필요가 없으며 다만 유사도 또는 비유사도 쌍만 주어지면 된다. 단점으로는 클러스터가 겹치거나 형태, 밀도, 크기가 다른 경우 좋은 결과를 얻기 힘든 점을 들 수 있다.

3. 클러스터링의 응용

클러스터링은 패턴 인식, 통신, 생화학, 심리학, 경영 등 상당히 다양한 분야에서 적용이 되고 있다. 적용되는 문제의 성격은 크게 세 가지로 구분된다 : 데이터 분석 및 시각화, 데이터 압축 그리고 데이터 전처리(preprocessing)이다.

자질(feature) 벡터 또는 유사도 쌍의 값으로 이루어진 대량의 데이터 집합이 주어졌을 때, 데이터에서 클러스터 또는 그룹 구조를 먼저 찾아보는 것이 합리적인 전략이다. 이런 구조는 데이터의 뼈대로 볼 수 있으며 추가로 이루어질 탐색을 위한 기반이 될 수 있다. 유클리드 공간 상의 자질 벡터를 이용하는 경우, 산출된 클러스터의 중심은 데이터와 같은 공간에 존재하는 자질 벡터로서 각 데이터 그룹의 표준으로 해석될 수 있다. 클러스터링 알고리즘이 계층적 또는 위상적(topological)인 클러스터 구조에 관련된 정보까지도 제시하는 경우에는 데이터의 흥미로운 자질들이 더 잘 드러나게 된다.

통신 및 데이터 저장 문제에서는 데이터를 압축하는 것이 중요한 경우가 많다. 이는 통신 채널의 대역폭(bandwidth)이나 저장장치의 용량이 한정되어 있기 때문이다. 이런 문제에서 많이 사용되는 방법으로 벡터 양자화를 들 수 있는데, 데이터 벡터의 대표값으로 코드 벡터를 얻어내어 데이터를 코드 벡터에 대한 상대 벡터로서 표현하는 것이다. 코드 벡터를 잘 선택할수록 더 좋은 압축률을 얻을 수 있다.

클러스터링 기법은 감독(supervised) 학습법 등을 위한 전처리(preprocessing) 과정으로 사용될 수도 있다. 아이디어는 데이터 압축의 경우와 비슷하다. 즉, 감독 학습 기법을 효율적으로 적용할 수 있도록 데이터를 단순화하여 표현하는 것이다. 이러한 접근법이 적용된 예를 들어보면 로보틱스 응용분야에서의 텍스처 분할(texture segmentation)을 들 수 있다. 클러스터링 기법은 또 RBF망(radial basis function network)에서 베이시스 함수를 초기화하거나, 역전사(counterpropagation) 네트워크에서 데이터를 전처리하는 데 사용된다. 이런 전처리 과정을 통해 얻은 위상적(topographic) 매핑, 즉 클러스터 사이의 공간적 정보를 보존하는 매핑은 문자 인식 등과 같은 문제 해결에 중요한 정보를 제공할 수 있다.

4. 문서 클러스터링

문서 클러스터링이란 대용량의 문서 집합을 주제에 따라 분류하는 것으로 정보 추출을 위한 중요한 도구로 오래 전부터 다루어져왔으며, 문서 검색, 혹은 정보 검색에서 색인 후에 검색의 전처리(preprocessing) 단계에 많이 사용된다. 주어진 다량의 문서를 미리 분류해 두면 사용자의 특정 정보에 대한 검색 요구가 들어올 때 모든 문서를 검색하는 대신 사용자의 요구와 가장 가까운 주제의 클러스터 내의 문서만을 검색함으로써 탐색 시간을 절약할 수 있고 검색의 효율을 향상시킬 수 있다. 최근, 문서 클러스터링은 웹 탐색 엔진의 중요한 도구로 사용되고 있으며, 탐색 및 문서목록 검색과 분산화된 추출에서 유용하게 적용되고 있다.

문서 클러스터링 과정은 색인(indexing - stemming[1], stop word[2] 제거) 과정을 거쳐 검색의 대상이 되는 문서들을 [그림 3] 과 같은 문서-색인어 행렬로 만들어 둔 상태에서 이루어진다. 문서를 이와 같은 벡터로 표현한다는 것은 문서에 나타난 단어의 의미(context)나 단어의 순서 등을 고려하지 않는다는 것을 뜻한다. 이러한 접근 방법은 문서를 bag-of-words로 보는 관점으로 단어의 순서를 고려하는 *n*-gram 모델, 더 나아가 자연언어처리를 필요로 하는 언어 모델과 구별된다. Bag-of-words 접근법에서의 가정은 같은 클러스터, 즉 같은 주제에 속한 문서들은 주제와 연관된 단어들에 대해 비슷한 출현 빈도, 즉 비슷한 단어들의 출현 패턴을 보인다는 것이다.

	T_1	T_2	· · · · · ·	T_m
D_1	t_{11}	t_{12}	· · · · · ·	t_{1m}
D_2	t_{21}	t_{22}	· · · · · ·	t_{2m}
· · ·	· · ·	· · ·		· · ·
D_n	t_{n1}	t_{n2}	· · · · · ·	t_{nm}

n: 문서의 개수
m: 단어의 개수
$D_i(i=1,\cdots,n)$: i 번째 문서
$T_j(j=1,\cdots,m)$: j 번째 단어
t_{ij}: i 번째 문서에 나타난 j 번째 단어의 빈도수

[그림 3] 문서 클러스터링에 사용되는 데이터의 형태 : 문서-색인어 행렬

이런 관점은 “분포 클러스터링(distributional clustering)”이라 불리는 기법의 적용으로 볼 수 있다. 분포 클러스터링은 문서 데이터의 클러스터링의 중요한 패러다임이며, 이에 대해서는 3장에서 자세히 살펴보겠다.

1) stemming: 문서나 쿼리(query)의 term들(즉, 단어)에서 접두사(prefix)와 접미사(suffix)를 제거하는 작업, 영문서를 대상으로 할 경우에 가장 많이 사용되는 알고리즘으로 PORTER STEMMER가 있다.
2) stop word: 문서의 내용에 영향을 주지 않는 관사나 전치사와 같은 단어를 말한다.

5. DNA Microarray

(1) 기능유전체학(functional genomics)

생물학적(biological), 생의학적(biomedical) 연구는 두 가지 주된 요인으로 인해 중대한 변환기를 맞고 있다. 한 요인은 DNA 서열 정보의 엄청난 증가이며, 이 정보를 이용하기 위한 기술의 발전이 또 다른 요인이다. 인간 유전자의 염기서열은 거의 모두 밝혀졌으며[3], 쥐나 다른 동식물의 완벽한 염기서열도 가까운 시일 내에 완성될 것이다. 그러나, 이들 수십억 개의 DNA 염기서열만으로는 각 유전자들의 역할이 무엇이고, 세포는 어떻게 동작하며, 세포가 어떻게 생물체(organism)를 형성하는지, 질병을 유발시키는 문제점은 무엇인지, 어떻게 의약품을 개발할지에 대한 답변이 되지 않는다. 그래서, 기능유전체학(functional genomics)이 점자 중요한 과학적 연구분야(discipline)로 자리잡고 있다 [8].

기능유전체학이란 유전자와 유전자 산물의 기능에 관한 연구를 하는 학문 분야로서 생물학적·생화학적 접근법을 통해 유전자 발굴과 유전자의 발현 양상을 조사한다. 유전자 발굴의 목표는 정상조직과 암조직, 성인과 발달과정에 있는 뇌의 여러 부위에서 발현하는 모든 유전자의 목록을 만드는 것이다. 그 다음으로 유전자의 발현 양상을 결정하고 발생·전이 및 다른 병리 상태에서의 유전자 발현 윤곽의 변화를 추적한다.

학자들은 이러한 연구를 통하여 3~6만개(추정)에 달하는 인간 유전자 각각의 기능을 상세히 알 수는 없어도 화학의 주기율표와 같이 인간 유전자의 분류가 가능할 것으로 예측하고 있다. 또한 이들 분류는 질병의 예후를 예측하거나 치료방법을 선택하는 데 유익한 정보를 제공할 것으로 추정되고 있다 [2].

(2) 마이크로어레이(Microarray) 기술

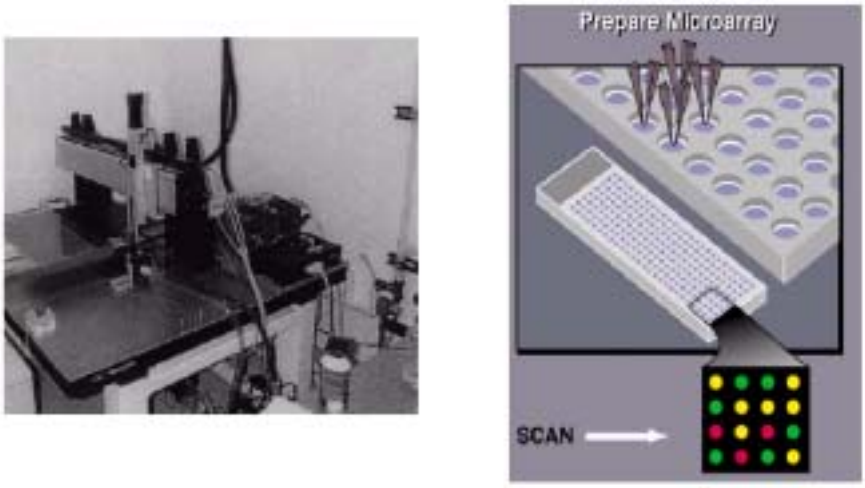
방대하고 빠른 속도로 증가하는 염기서열 관련 정보를 제대로 활용하기 위해서는 새로운 기술들이 필요하다. 유전체학에서 가장 강력하고 다양한 활용성을 가지는 기술 중 하나가 고밀도의 올리고뉴클레오티드(oligonucleotide, 핵산의 짧은 서열) 또는 상보적(complementary) DNA의 배열(array)이다.

DNA 배열은 칩(chip)의 표면에 특별한 목적을 위해 배치한 DNA 분자들(probe)과, 세포에서 뽑아내어 형광처리한 RNA나 DNA의 용액을 혼합(hybridization)시키는 방식으로 동작한다. 혼합반응은 전체 배열에서 동시에 병렬적으로 발생한다. 즉, 샘플을 배열에

3) 미국·영국 등 6개국 공동 연구팀인 HGP(Human Genome Project: 인간게놈프로젝트)와 미국의 생명공학회사인 셀레라 지노믹스가 2001년 2월 12일에 인간의 유전자에 대한 비밀을 푼 연구결과를 공동으로 발표하고 인터넷에도 공개하였다. 이 때 게놈지도의 99% 완성본이 공개되었다.

혼합하는 과정에서 각 분자가 배열상에서 매치가 되는 파트너를 찾는 고차원적인 병렬 탐색이 이루어진다. 스캐닝(scanning)을 통해 혼합된 배열로부터 형광 이미지를 얻으며, 형광의 강도가 수치적으로 변환되어 2차원 배열 형태의 마이크로어레이(microarray) 데이터가 생성된다. 이 데이터는 실험의 목적에 따라 다양한 방법으로 처리할 수 있다.

실험의 목적은 특정 발현 과정에 연관된 새로운 유전자를 구분하고, 잠재적인 의약 처리 대상을 찾아내며 질병 예측이나 진단에 사용될 수 있는 발현 표시자(expression marker)를 밝히는 것이다.



[그림 4] Microarray 실험기기 및 실험의 간략한 개요

마이크로어레이 데이터를 분석하고, 분류하며 분할하는 방법은 다양한데, 여러 가지 클러스터링 분석 기법이 중요한 도구로 널리 쓰이고 있다 [3]. 이 기법들은 특정 발현 과정에서 동시에 활성화 또는 상호작용하는 유전자들을 찾아내는 데 유용하다.

이 논문에서는 문서 클러스터링에 사용되는 데이터와 마이크로어레이 데이터가 모두 2차원의 배열로서 비슷한 방식으로 처리할 수 있다는 점에 착안하고, '정보 병목 기법'을 이용한 이중 클러스터링 기법을 문서 클러스터링 과정을 통해 효율성을 확인한 후 마이크로어레이 데이터 분석에 적용할 것이다.

III. 정보이론 기반의 이중 클러스터링

1. 정보 이론(Information Theory)

(1) Introduction

정보(information)는 한 마디로 정의하기는 힘들다. 그러나, 어떤 확률 분포에 대해 엔트로피(entropy)라 불리는 양을 정의할 수 있는데, 이 양은 정보의 측정이 어떤 식으로 이루어져야 하는지에 대한 직관적인 개념과 일치하는 많은 특성들을 가지고 있다 [5]. 이 개념을 확장하여 상호 정보량(mutual information)을 정의하며, 이는 한 변수가 다른 변수에 대해 담고 있는 정보의 양을 측정하는 척도가 된다. 상호 정보량의 개념에 따르면 엔트로피는 랜덤 변수의 자기 정보(self-information)로 해석된다. 상호 정보량은 보다 일반적인 상대적 엔트로피(relative entropy)의 특수한 경우이다. 상대적 엔트로피는 두 확률 분포 사이의 거리를 측정하는 양이다. 이 값들은 서로 긴밀한 연관을 가지며 몇 가지 간단한 속성들을 공유한다.

(2) 엔트로피(entropy)

엔트로피는 랜덤 변수(random variable)의 불확실성에 대한 척도로서 정의는 다음과 같다 [5].

정의> 이산(discrete) 랜덤 변수 X의 엔트로피 $H(X)$는 다음과 같이 정의된다. (이하에서 X가 취할 수 있는 모든 값의 집합을 편의상 같은 기호 X로 표현하기로 한다)

$$H(X) = -\sum_{x \in X} p(x) \log p(x) . \qquad (1)$$

때로는 $H(p)$라고 표현하기도 한다. log는 밑이 2이며 엔트로피는 비트(bit) 단위로 표현된다. 예를 들면, 동전 던지기의 엔트로피는 1 bit이다.

- $0\log 0 = 0$이다. 따라서, 확률 0인 항을 더하는 것은 엔트로피를 바꾸지 않는다.
- 엔트로피는 X의 분포에 대한 범함수(functional)이다. 따라서, 랜덤 변수 X가 가지는 값과는 무관하며 오직 확률값에 의해서만 정의된다.
- $H(X) \geqq 0$이다. ($\because 0 \leqq p(x) \leqq 1, \ \log(1/p(x)) \geqq 0$)

(3) 결합(joint) 엔트로피와 조건부 엔트로피

랜덤 변수 쌍에 대한 엔트로피, 즉 결합 엔트로피는 다음과 같이 정의한다. [5]

정의> 이산 랜덤 변수쌍 (X, Y) 및 이들의 결합 확률 분포 $p(x, y)$가 주어졌을 때의 *결합 엔트로피* $H(X, Y)$는 다음과 같다.

$$H(X, Y) = -\sum_{x \in X}\sum_{y \in Y} p(x, y)\log p(x, y). \tag{2}$$

한 랜덤 변수가 주어졌을 때의 다른 랜덤 변수의 조건부 엔트로피는 다음과 같이 정의한다 [5].

정의> $(X, Y) \sim p(x, y)$이면, *조건부 엔트로피* $H(Y|X)$는 다음과 같다.

$$H(Y|X) = \sum_{x \in X} p(x)H(Y|X = x) \tag{3}$$

$$= -\sum_{x \in X} p(x)\sum_{y \in Y} p(y|x)\log p(y|x) \tag{4}$$

$$= -\sum_{x \in X}\sum_{y \in Y} p(x, y)\log p(y|x). \tag{5}$$

결합 엔트로피와 조건부 엔트로피의 정의는 다음의 연쇄 규칙(chain rule)을 통해 자연스럽게 엔트로피와 연결된다.

정리> (연쇄 규칙)

$$H(X, Y) = H(X) + H(Y|X). \tag{6}$$

추론>

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z). \tag{7}$$

증명은 생략한다.

(4) 상대적(relative) 엔트로피와 상호 정보량(mutual information)

랜덤 변수의 엔트로피는 랜덤 변수의 불확실성에 대한 척도이다. 즉 이 랜덤 변수를 표현하는 데 평균적으로 필요한 정보의 양에 대한 척도이다.

상대적 엔트로피(relative entropy)는 두 확률분포 사이의 거리에 대한 척도이다. 좀 더 명확하게 표현하면, 상대적 엔트로피 $D_{KL}[p \parallel q]$는 실제 확률분포가 p일 때 확률 분포를 q라고 가정하는 경우의 비효율성을 측정하는 것이다 [5]. 예를 들면, 랜덤 변수의 실제 분포를 알고 있다면, 평균 표현길이(description length) $H(p)$로서 코드를 구성할 수 있다. 대신, 확률 분포 q의 코드를 사용한다면 랜덤 변수를 표현하는 데는 평균적으로

$H(p) + D_{KL}[p \parallel q]$ bit가 필요하다.

정의> 두 확률질량함수(probability mass function) $p(x)$와 $q(x)$ 사이의 *상대적 엔트로피* 또는 ***Kullback-Leibler divergence***는 다음과 같이 정의한다 [5].

$$D_{KL}[p \parallel q] = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}. \tag{8}$$

- $0 \log \frac{0}{q} = 0,\ p \log \frac{p}{0} = \infty$이다.
- $D_{KL}[p \parallel q] \geq 0$이며, 등식은 $p = q$일 때에만 성립한다.

상대적 엔트로피는 분포 사이의 실제 거리는 아니다. 이유는 p와 q에 대해서 대칭적이지 않으며, 삼각 부등식을 만족하지 않기 때문이다. 그럼에도 불구하고, 상대적 엔트로피를 분포 사이의 "거리"로 생각하는 것은 보통 유용하다.

상호 정보량은 한 랜덤 변수가 다른 랜덤 변수에 대해 담고 있는 정보량에 대한 척도이다. 또는 다른 랜덤 변수에 대한 지식으로 인해 한 랜덤 변수에 대한 불확실성이 줄어드는 정도이다.

정의> 두 랜덤 변수 X와 Y 및 이들의 결합확률질량함수 $p(x, y)$, 주변(marginal) 확률질량함수 $p(x)$와 $p(y)$가 주어져 있다고 하자. *상호 정보량* $I(X;Y)$은 결합확률분포와 확률분포의 곱 $p(x)p(y)$ 사이의 상대적 엔트로피이다.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{9}$$

$$= D_{KL}[p(x, y) \parallel p(x)p(y)]. \tag{10}$$

(5) 엔트로피와 상호 정보량 사이의 관계

상호 정보량 $I(X;Y)$의 정의를 다시 쓰면 다음과 같다.

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{11}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \tag{12}$$

$$= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \tag{13}$$

$$= -\sum_{x} p(x) \log p(x) - (-\sum_{x,y} p(x,y) \log p(x|y)) \tag{14}$$

$$= H(X) - H(X|Y). \tag{15}$$

따라서 상호 정보량 $I(X;Y)$는 Y에 관련된 지식으로 인해 X의 불확실성이 줄어드는 정도이다.

대칭성에 의해 다음 식도 성립한다.

$$I(X;Y) = H(Y) - H(Y|X). \tag{16}$$

따라서 X가 Y에 대해 언급하는 만큼 Y도 X에 대해 언급한다.

(2)절에서 살펴보았듯이 $H(X,Y) = H(X) + H(Y|X)$이므로

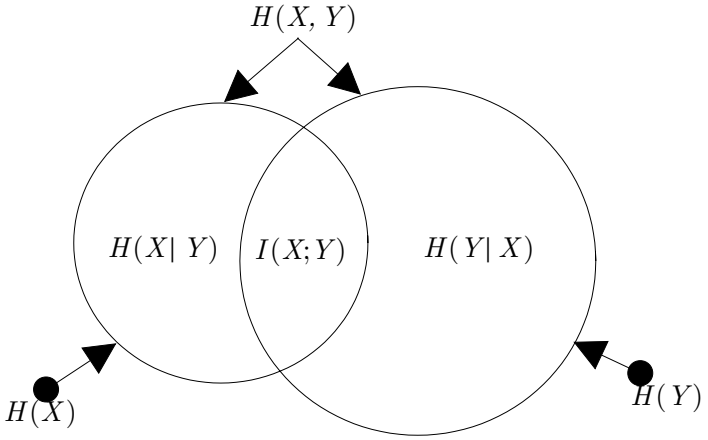
$$I(X;Y) = H(X) + H(Y) - H(X,Y), \tag{17}$$

그리고,

$$I(X;X) = H(X) - H(X|X) = H(X). \tag{18}$$

따라서 한 랜덤 변수 자신에 대한 상호 정보량은 랜덤 변수의 엔트로피이다. 이런 이유로 엔트로피는 때때로 *자기정보(self-information)*라고 불리기도 한다.

지금까지의 결과를 종합하면 다음의 벤 다이어그램으로 표현된다 [5].



[그림 5] 엔트로피와 상호 정보량 사이의 관계

2. 정보 병목 기법

대부분의 클러스터링 알고리즘은 두 점 사이의 '거리'의 쌍(pairwise 클러스터링) 또는 클래스의 중심점과 데이터 위치 사이의 이격도 측정(벡터 양자화)에서 시작한다. 일단 거리의 행렬이나 이격도 측정값이 주어지면, 클러스터링 작업은 다양한 방법을 통해 최적화 문제를 푸는 작업이 된다. 즉, 클래스 내부적으로 이격도가 작은 또는 연결성이 큰 소수의 클래스를 찾아내는 것이 목적이 된다.

이러한 접근 방법에서 가장 큰 문제가 되는 것은 거리 또는 이격도 측정의 선택 문제이다. 거의 대부분의 경우, 선택은 임의로 이루어지고 데이터의 표현 방식에 상당히 민감하게 되는데, 이런 과정에서는 고차원에서 표현되는 여러 다양한 요소들의 구조를 정확히 반영한다는 것을 확신할 수 없다.

문서 클러스터링의 경우, 두 문서 간 유사도를 측정하는 자연스러운 기준은 각각에서의 단어의 조건부 확률분포의 유사도이다. X를 문서의 집합, Y를 단어의 집합이라 하면, 모든 문서에 대해 다음과 같은 조건부 확률을 정의할 수 있다.

$$p(y|\,x) = \frac{n(y|\,x)}{\sum_{y \in Y} n(y|\,x)}. \tag{19}$$

$n(y|\,x)$는 문서 x에서의 단어 y의 빈도수이다. 대략적으로, 우리가 바라는 바는 단어에 대한 조건부 분포가 유사한 문서들이 같은 클러스터에 속하도록 하는 것이다. 이런 식으로 한 집합(예를 들면 문서)의 원소들의 클러스터 구조를 다른 집합(예를 들면 단어)의 집합의 원소에 대해 비슷한 조건부 분포를 기준으로 하여 찾아내는 방법은 "분포(distributional) 클러스터링"이라 부르며 [8]에서 처음으로 소개가 되었다.

각 분포간의 '정확한' 거리를 어떻게 측정할지의 문제에 대한 해법으로서 최근에 Tishby, Pereira 및 Bialek [11]은 이격도 또는 거리 측정에서의 임의 선택에 따른 문제를 피할 수 있는 '정보 병목 기법(information bottleneck method)'을 제시하였다. 이 새로운 접근 방법에서는 두 확률변수 간에 실험적으로 얻어진 결합 확률분포 $p(x, y)$가 주어졌을 때, 우리가 할 일은 관련 변수 Y의 정보를 최대한으로 보유하고 있는 X의 축약된 표현을 찾는 것이다. 이런 직관적인 개념을 정보 이론 방식으로 자연스럽게 논리화하면 다음과 같은 문제로 요약할 수 있다:

문제> *집합 X 구성원의 클러스터 집합 $\tilde{X}$를 찾아라. $\tilde{X}$는 X에서 추출한 정보에 대한 제한조건 $I(\tilde{X};X)$을 만족하면서, 상호 정보량(mutual information) $I(\tilde{X};Y)$가 최대가 되도록 해야 한다.*

두 변수 X와 Y 간의 상호 정보량 $I(X;Y)$는 변수 X가 변수 Y에 대해 담고 있는 정보에 대한 모순되지 않은(consistent) 유일한 통계적 측정이다. 표현의 압축률은 $I(\tilde{X};X)$에 의해 결정되며, 클러스터 집합 $\tilde{X}$의 질(quality)은 이 집합이 Y와 관련해 정보를 보유하는 비율, 즉 $I(\tilde{X};Y)/I(X;Y)$로 정의된다.

이 일반적인 문제에 대해, 정확하고 최적화된 해가 존재하며, 이 해는 결합 확률분포 $p(x,y)$를 어떻게 만들어냈는지에 대한 아무런 가정을 필요로 하지 않는다.

해는 각 클러스터 $\tilde{x} \in \tilde{X}$의 특징을 반영하는 다음의 세 가지 확률분포의 조합으로 주어진다.

- $p(\tilde{x})$: 해당 클러스터의 prior probability
- $p(\tilde{x}|\, x)$: 해당 클러스터의 membership probabilities
- $p(y|\, \tilde{x})$: 해당 클러스터의 관련 변수에 대한 분포도(distribution)

일반적으로 membership probabilities $p(\tilde{x}|\, x)$는 'soft'하다. 즉, X의 모든 원소 x는 모든 클러스터에 특정(정규화된) 확률로 할당될 수 있다.

정보 병목 원리에서는 x와 $\tilde{x}$간의 이격도 측정법을 다음과 같이 조건부 분포 $p(y|\, x)$와 $p(y|\, \tilde{x})$ 간의 Kullback-Leibler divergence로 정의한다.

$$D_{KL}[p(y|\, x) \parallel p(y|\, \tilde{x})] = \sum_{x,\, y} p(y|\, x) log \frac{p(y|\, x)}{p(y|\, \tilde{x})}. \qquad (20)$$

※ $0log\frac{0}{q} = 0$, $plog\frac{p}{0} = \infty$ 로 정의한다.

문제의 해는 다음의 세 개의 분포(distribution) 방정식으로 주어지며, 이들은 함께 풀어야 한다.

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(\beta, x)} exp(-\beta D_{KL}[p(y|x) \parallel p(y|\tilde{x})]) \quad (21)$$

$$p(y|\tilde{x}) = \frac{1}{p(\tilde{x})} \sum_{x} p(\tilde{x}|x) p(x) p(y|x) \quad (22)$$

$$p(\tilde{x}) = \sum_{x} p(\tilde{x}|x) p(x). \quad (23)$$

※ $Z(\beta, x)$는 정규화 인자이며, 라그랑지 인자(Lagrange multiplier) β는 분류의 'softness'를 결정한다.

전체 과정을 직관적으로 살펴보면, X에 담겨 있는 Y 관련 정보는, 압축된 클러스터 집합 $\tilde{X}$(bottleneck)을 통해 '짜내어'지며, $\tilde{X}$는 X에서 Y에 대해 '관련된' 부분을 표현하도록 만들어진다.

3. 이중(double) 클러스터링 기법

정보 추출의 관점에서 두 가지 방식의 클러스터링이 연구되어 왔다. 하나는 문서를 공통으로 포함하는 단어의 분포를 기준으로 클러스터링하는 방법이며, 다른 하나는, 단어를 클러스터링하되 단어가 포함된 문서의 분포를 기준으로 하는 것이다.

[10]에서는 이 두 가지 방식을 하나의 정보이론적 프레임웍, 즉 정보 병목 기법을 바탕으로 조합을 하였다. 구체적인 내용은 다음 절에서 다룰 것이며 여기에서 간단히 그 아이디어를 살펴보면 다음과 같다. 실험을 통해 얻은 두 변수의 조합 분포가 주어져 있을 때, 한 변수를 압축하되 다른 변수에 관련된 상호 정보량(mutual information)은 최대한 보존하도록 하는 것이다. 문서 클러스터링에서는 두 변수는 각각 문서집합과 단어집합이다. 그러므로, 문서 집합의 정보를 최대한 담고 있는 단어 클러스터를 찾을 수도 있고, 또는 문서에 나타나는 단어에 관련된 정보를 최대한 담고 있는 문서 클러스터를 찾을 수도 있을 것이다. 이중 클러스터링에서는 이 두 가지 경우를 하나로 조합하여 두 단계의 클러스터링을 실시한다. 즉, 먼저 문서 관련 정보를 최대한 담고 있는 단어 클러스터를 얻어낸다. 다음 단계에서는 문서와 단어의 co-occurrence 행렬 대신 문서 내의 단어 클러스터의 co-occurrence를 기반으로 한 보다 압축된 표현으로 대체한다. 이 새로운 문서 표현을 이용해, 앞 단계와 같은 과정을 통해 문서 클러스터를 얻어낸다.

이중 클러스터링의 주된 이점은, co-occurrence 행렬이 상당히 고차원의 데이터이기 때문에 필수불가결하게 나타나는 노이즈를 상당히 줄일 수 있다는 것이다. 단어 클러스터를 기반으로 한 축약된 행렬은 더 조밀하고 엄정하며, 문서 집합에 내재된 구조를 더

잘 반영한다.

4. 실험 구성

(1) Clustering의 방향 :

이번 실험에서 구성할 클러스터는 'hard' 클러스터이다. 즉, 모든 $x \in X$는 정확히 하나의 클러스터 $\tilde{x} \in \tilde{X}$에 포함된다. 이는 정보 병목 기법의 해(solution)에서 $\beta \to \infty$인 경우로서, 정보 병목 기법 구현이 간단해지며, 각 분포간의 자연스러운 거리 측정을 할 수 있다. 'hard' 클러스터를 사용할 경우, $\tilde{x} \in \tilde{X}$를 특정 클러스터라고 하면 각 분포식은 다음과 같이 단순해진다.

$$\begin{cases} p(\tilde{x}|x) = \begin{cases} 1 & \text{if } x \in \tilde{x} \\ 0 & otherwise \end{cases} & (24) \\ p(y|\tilde{x}) = \dfrac{1}{p(\tilde{x})} \sum_{x \in \tilde{x}} p(x)p(y|x) & (25) \\ p(\tilde{x}) = \sum_{x \in \tilde{x}} p(x). & (26) \end{cases}$$

이 분포식을 이용하면 클러스터 집합 $\tilde{X}$와 Y간의 상호 정보 $I(\tilde{X};Y)$를 쉽게 측정할 수 있다.

(2) 이중 클러스터링

정보 병목 기법의 목적은 변환 $p(\tilde{x}|x)$에 의해 정의되는 X의 파티션을 얻어내되, 이 파티션이 상호 정보 함수의 최대값이 되도록 하는 것이다. 이 과정을 자세히 살펴보면 ((11), (19)), X와 Y의 역할이 서로 바뀔 수 있음을 알 수 있다. 즉, X와 Y는 대칭적인 관계를 가지며, 이를 클러스터링에 적용한다면, 문서와 관련하여 최대의 상호 정보량을 가지는 단어의 클러스터를 얻을 수도 있고, 또는 단어에 관해 최대 상호 정보량을 보유하는 문서 클러스터를 찾아낼 수도 있다. 이 두 과정을 합친 것이 '이중 클러스터링(double clustering)'이다.

이 실험에서는 '단어'에 대한 클러스터링을 통해 얻은 단어 클러스터 집합을 이용해 '문서' 클러스터링을 실시한다.

5. 프로그램 구현

(1) 알고리즘

먼저, 한 단계의 알고리즘을 구성한다. 여기에 적용되는 일반적인 프레임웍은 agglomerative greedy hierarchical 클러스터링 알고리즘이다. 이 알고리즘은 X의 원소 각각을 하나의 클러스터로 나눈 상태, 즉 $|X|$ 개의 클러스터에서 시작한다.

각 단계에서 현재 파티션 상에 있는 두 성분(component)을 병합하여(merge) 하나의 새로운 성분으로 만드는데, 이 때 제한조건은 국부적으로(locally) 상호 정보량 $I(\tilde{X};Y)$의 손실을 최소화하는 것이다. 모든 병합과정 $(\tilde{x_i}, \tilde{x_j}) \rightarrow \tilde{x_*}$ 은 다음의 방정식에 의해 명확하게 정의된다.

$$\begin{cases} p(\tilde{x_*}|x) = \begin{cases} 1 & \text{if } x \in \tilde{x_i} \text{ or } x \in \tilde{x_j} \\ 0 & otherwise \end{cases} & (27) \\ p(y|\tilde{x_*}) = \frac{p(\tilde{x_i})}{p(\tilde{x_*})} p(y|\tilde{x_i}) + \frac{p(\tilde{x_j})}{p(\tilde{x_*})} p(y|\tilde{x_j}) & (28) \\ p(\tilde{x_*}) = p(\tilde{x_i}) + p(\tilde{x_j}). & (29) \end{cases}$$

이 병합 과정에서의 상호 정보량 $I(\tilde{X};Y)$의 감소 정도를 계산하면 다음과 같다. 이 항목을 "병합에 필요한 비용"으로 해석할 수 있다.

$$\delta I(\tilde{x_i}, \tilde{x_j}) \equiv (p(\tilde{x_i}) + p(\tilde{x_j})) \cdot D_{JS}[p(y|\tilde{x_i}), p(y|\tilde{x_j})]. \quad (30)$$

함수 D_{JS}는 Jensen-Shannon (JS) divergence로서 다음과 같이 정의된다.

$$D_{JS}[p_i, p_j] = \pi_i D_{KL}[p_i \| \bar{p}] + \pi_j D_{KL}[p_j \| \bar{p}], \quad (31)$$

이 실험에서 각 항목은 다음과 같이 구할 수 있다.

$$\begin{cases} p_i, p_j \equiv p(y|\tilde{x_i}), p(y|\tilde{x_j}) & (32) \\ \pi_i, \pi_j \equiv \frac{p(\tilde{x_i})}{p(\tilde{x_*})}, \frac{p(\tilde{x_j})}{p(\tilde{x_*})} & (33) \\ \bar{p} = \pi_i p(y|\tilde{x_i}) + \pi_j p(y|\tilde{x_j}) \quad . & (34) \end{cases}$$

JS-divergence는 0 이상이며, 상한값은 1이고, 두 인자가 동일한 경우에만 0이 된다.

정보의 최적화라는 기준을 둠으로써 유사도 측정법을 얻을 수 있었다. 이제 정리된 한 단계의 클러스터링 알고리즘은 아주 간단하다. 매 단계마다 "가장 좋은 병합", 즉 $\delta I(\tilde{x_i}, \tilde{x_j})$을 최소화하도록 두 클러스터 $\{\tilde{x_i}, \tilde{x_j}\}$를 선택해 병합하는 것이다. 전체 과정에

대한 pseudo code는 [표 Ⅰ]과 같다 [10].

[표 Ⅰ] 병합식 정보병목 기법 알고리즘의 pseudo-code

입력 : 결합 확률분포 $p(x, y)$

출력 : X를 분할하여 생성한 m개의 클러스터, $^{\forall}m \in \{1...|X|\}$

초기화 :

- 클러스터 초기화. $\tilde{X} \equiv X$
- merging cost matrix 초기화

$^{\forall}i, j = 1...|X|$, $i < j$ 에 대해 다음을 계산

$$d_{i,j} = (p(\tilde{x_i}) + p(\tilde{x_j})) \cdot D_{JS}[p(y|\tilde{x_i}), p(y|\tilde{x_j})]$$

반복 :

- For $m = |X| - 1...1$
  - $d_{i,j}$를 최소로 하는 인덱스 쌍 $\{i, j\}$를 찾는다.
  - 병합. $(\tilde{x_i}, \tilde{x_j}) \rightarrow \tilde{x_*}$
  - 클러스터 갱신. $\tilde{X} = \{\tilde{X} - \{\tilde{x_i}, \tilde{x_j}\}\} \cup \{\tilde{x_*}\}$
  - merging cost matrix 갱신. $\tilde{x_*}$와 관련된 $d_{i,j}$를 갱신.
- End For

(2) 알고리즘 분석

X는 단어의 집합, Y는 문서의 집합, $|X| = n$, $|Y| = m$ 이라 하면

- Time complexity : $O(n^2 m)$ - D_{KL} 계산시
- Space complexity : $O(n^2 + nm + m^2) = O((n+m)^2)$

  $O(n^2)$ - merging cost matrix, word cluster list 저장시 필요한 공간

  O(nm) - n×m 크기를 가지는 기본 data matrix 저장시 필요

  $O(m^2)$ - document cluster list 저장시 필요한 공간

(3) 다른 클러스터링 기법

병합식 클러스터링 알고리즘의 일반적인 프레임웍을 유지하면서, 다른 유사도 측정법을 적용하여 '정보 병목 기법'의 결과와 비교를 하고 이 방법의 타당성을 확인해본다.

첫 번째 비교 대상은 L1 norm(일명 변동(variational) 거리)을 적용하는 경우로서, 정의는 다음과 같다.

$$L_1(p_i, p_j) \equiv \sum_{y \in Y} |p_i(y) - p_j(y)| . \qquad (35)$$

JS-divergence와는 달리, L1 norm은 삼각 부등식을 포함해 모든 해석적(metric) 속성을 만족하는 '거리'측정방법이다. 이 경우 (30)에서 *JS*-divergence 대신 L1 norm을 사용하게 된다.

$$d_{i,j} \equiv (p(\tilde{x_i}) + p(\tilde{x_j})) \cdot L_1(p(y|\tilde{x_i}), p(y|\tilde{x_j})). \quad (36)$$

두 번째로는 유클리드식 거리에 바탕을 두고 있는 Ward 기법을 적용한다. 이 경우의 유사도 측정은 (37)과 같이 이루어진다.

$$d_{i,j} \equiv \frac{p(\tilde{x_i})p(\tilde{x_j})}{p(\tilde{x_i}) + p(\tilde{x_j})} \sum_{y \in Y} (p(y|\tilde{x_i}) - p(y|\tilde{x_j}))^2. \quad (37)$$

(4) 이중 클러스터링으로 확장

식 (30), (36), (37)의 세 가지 기준식은 근본적으로 X와 Y의 역할이 대칭적이다. 즉, 어떤 변수를 압축해야 되는지에 대한 아무런 전제조건이 없다. 여기에서는 두 가지 선택적 클러스터링을 조합을 하려 하며, 이를 위해 두 단계의 클러스터링을 실시한다.

첫 번째 단계에서 각 단어 y의 문서 집합에 대한 조건부 확률분포 $p(x|y)$를 구한다. 다음으로, 앞 절에서 설명한 클러스터링 알고리즘을 이용해 단어 클러스터 $\tilde{Y}$를 얻는다. 이 때 $|\tilde{Y}| \ll |Y|$ 가 되도록 한다.

두 번째 단계에서는 이 단어 클러스터를 이용해 기존의 문서 표현 방식을 바꾼다. 문서를 단어의 조건부 확률 분포 $p(y|x)$를 이용해 표현하는 대신, 다음과 같이 정의되는 단어 클러스터에 대한 조건부 확률 분포 $p(\tilde{y}|x)$를 이용해 표현한다.

$$p(\tilde{y}|x) = \frac{n(\tilde{y}|x)}{\sum_{\tilde{y} \in \tilde{Y}} n(\tilde{y}|x)} = \frac{\sum_{y \in \tilde{y}} n(y|x)}{\sum_{\tilde{y} \in \tilde{Y}} \sum_{y \in \tilde{y}} n(y|x)}. \quad (38)$$

이 축약된 표현을 이용해, 앞 절의 클러스터링 알고리즘을 다시 적용하여 원하는 문서 클러스터 $\tilde{X}$를 얻는다. 이 과정을 pseudo-code로 표현하면 다음과 같다 [10].

[표 II] 이중 클러스터링 절차

입력 : 결합확률분포 $p(x, y)$

1단계 :

- $\{p(x|\, y)\}$를 이용해 클러스터 집합 $\tilde{Y}$를 찾는다.

2단계 :

- 모든 $x \in \tilde{X}$에 대해, $p(y|\, x)$를 보다 축약된 표현 $p(\tilde{y}|\, x)$로 대체한다.
- $\{p(\tilde{y}|\, x)\}$를 이용해 클러스터 집합 $\tilde{X}$를 찾는다.

이 이중 클러스터링 프레임웍 기반에서 정보 병목 기법을 적용하여, 생성되는 결과 클러스터의 본질에 대해 명확한 정보를 얻을 수 있다. 첫 번째 단계에서 주어진 문서와 관련된 가장 의미 있는 정보를 담고 있는 단어 클러스터를 추출한다. 좀 더 정연하게 표현하자면 첫 번째 단계에서 $I(X;\tilde{Y}) \lesssim I(X;Y)$를 만족하는 클러스터 집합 $\tilde{Y}$를 찾는다. 두 번째 단계에서는 단어 클러스터와 관련된 가장 의미 있는 정보를 포함하는 문서 클러스터 집합 $\tilde{X}$를 찾는다. 이 과정을 통해 변수간의 상호 정보량의 손실은 크지 않으면서도($I(\tilde{X};\tilde{Y}) \lesssim I(X;\tilde{Y}) \lesssim I(X;Y)$) 기존 변수들의 차원(dimension)을 둘 모두 상당히 줄일 수 있다.

IV. 실험 및 결과

1. 문서 데이터 선택

실험의 주요 목적은 클러스터링의 정확도 측정이므로, 클래스가 명확히 구분되는 문서집합 데이터가 필요하다. 문서 클러스터링을 적용하기 위해 선택한 데이터는 [그림 3]과 같은 2차원의 문서-단어 행렬 형태로서, (1,069 문서×5,286 단어)의 크기를 가진다. 데이터는 다음과 같은 과정을 통해 생성되었다.

- TREC-8 ad-hoc task data 중에서 4가지 주제, 즉 'Foreign minorities, Germany' (ID 401), 'Estonia, economy' (ID 434), 'inventions, scientific discoveries' (ID 439), 'King Husayn, peace' (ID 450)에 포함된 1,069 개의 문서를 선택한다.
- 각 문서에서 stop word를 제거하고, 출현문서 빈도가 5 이상인 단어를 선택

2. 실험 상세구성

(1) 주요 인자 설정

- 클러스터링의 방향 : 데이터 작성시 사용된 4가지의 주제를 문서의 클래스로 삼는다. 문서를 4개의 클러스터로 압축하여 이 클래스를 기준으로 정확도를 측정한다.
- Single/Double : 문서만의 단일(single) 클러스터링, 단어-문서 순서로 이중(double) 클러스터링을 실시한다. 이중 클러스터링에서는 실험 결과가 단어 클러스터의 수에 따라 달라질 수 있음을 고려하여, 20, 40, 60, 80, 100 개의 단어 클러스터를 구성하였다.
- 유사도 측정 방법 : 정보 병목 기법 외에, L1 norm 기법, Ward 기법을 병행하여 비교
- 클러스터링 결과 평가 : 정확도 측정

(가) Contingency Table

Confusion matrix와 유사한 개념으로서 [표 Ⅲ]과 같은 테이블을 구성한다. 이 때, 클러스터의 순서는 클러스터 내에서 가장 높은 빈도를 가지는 것으로 판단된 클래스에 맞춰 배열하였다. 대각선에 있는 값들이 정확하게 클러스터링된 문서의 수이다.

[표 Ⅲ] Contingency Table. '단일 클러스터링 - 정보 병목기법' 실행결과.
정확도는 (340+288+293+122)/1069 ≒ 0.976 이다.

	Cluster1	Cluster2	Cluster3	Cluster4	RowSum
Class1	**340**	1	3	3	347
Class2	0	**288**	5	0	293
Class3	6	0	**293**	1	300
Class4	6	1	0	**122**	129
ColSum	352	290	301	126	**1069**

(나) F_1-measure

먼저 정확률(precision)과 재현률(recall)을 정의한다.

$$Precision(i,j) = n_{ij}/n_i \qquad Recall(i,j) = n_{ij}/n_j . \qquad (39,\ 40)$$

n_{ij} : 실제로 클래스 i에 속한 문서인데 클러스터 j에 속한다고 판단된 문서의 수
n_i : 클래스 i에 포함된 문서의 수
n_j : 클러스터 j에 포함된 문서의 수

정확률은 어떤 클러스터에 속한다고 판정된 문서 중 제대로 클러스터링이 된 문서의 비율을 말하고, 재현률은 실제로 어떤 클래스에 속하는 문서 중 제대로 클러스터링이 된 문서의 비율을 말한다. n_{ij}, n_i, n_j는 contingency table에서 바로 구할 수 있다.

클러스터 j와 클래스 i에 대한 F_1-measure는 (41)과 같다. $n_{ij}=0$이면 $F_1(i,j)=0$으로 처리한다. 전체 클러스터링 결과는 (42)와 같이 측정한다.

$$F_1(i,j) = \frac{2*Recall(i,j)*Precision(i,j)}{Precision(i,j)+Recall(i,j)} . \qquad (41)$$

$$F_1 = \sum_i MaxF_1(i,j)\frac{n_i}{n} . \qquad (n=\sum_i n_i) \qquad (42)$$

3. Clustering 결과 및 분석

(1) 결과데이터 명칭 구분

IB : 정보병목기법, L1 : L1 norm 기법, Ward : Ward 기법

(2) 실험 결과

(가) 단일 클러스터링

[표 Ⅳ] 문서에 대한 단일 클러스터링 결과

정확도 측정법	IB	L1	Ward
Contingency table	0.976	0.711	0.864
F_1-measure	0.976	0.755	0.859

(나) 이중 클러스터링

- Contingency Table로 정확도 측정

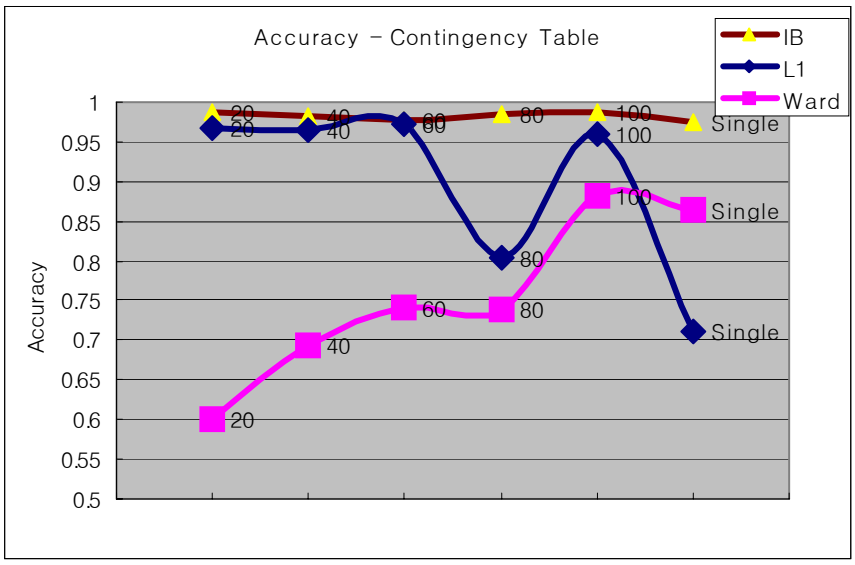
[표 Ⅴ] 문서에 대한 이중 클러스터링 결과 (정확도 측정법 : contingency table)

단어 클러스터 수	20	40	60	80	100	평균	표준편차
IB	0.988	0.983	0.977	0.986	0.987	0.982	0.004
L1	0.968	0.964	0.972	0.804	0.961	0.934	0.073
Ward	0.601	0.693	0.741	0.738	0.881	0.731	0.101

● F_1-measure로 정확도 측정

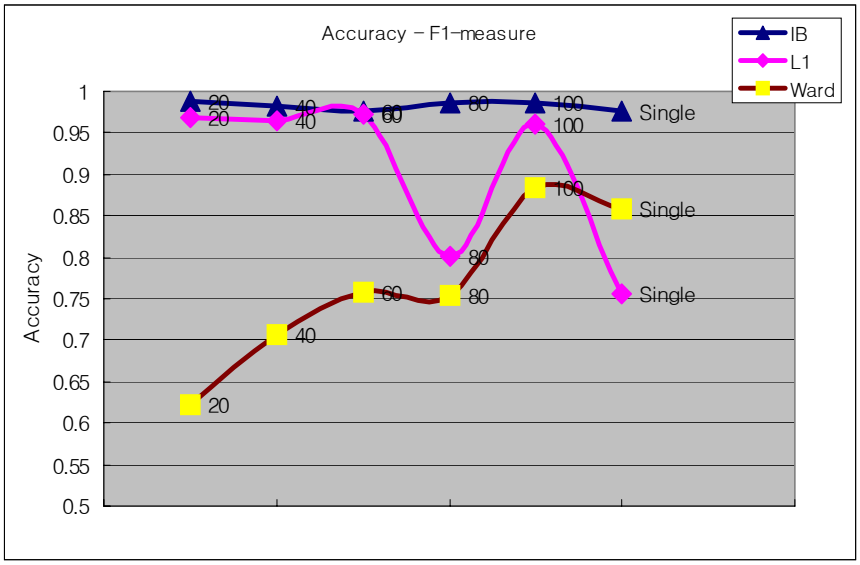
[표 VI] 문서에 대한 이중 클러스터링 결과 (정확도 측정법 : F_1-measure)

단어 클러스터 수	20	40	60	80	100	평균	표준편차
IB	0.988	0.983	0.977	0.986	0.987	0.984	0.004
L1	0.968	0.965	0.972	0.801	0.961	0.933	0.074
Ward	0.622	0.706	0.757	0.753	0.883	0.744	0.095



[그림 6] 문서 클러스터링 결과 - contingency table로 정확도 측정

※ 각 포인트의 첨자는 사용된 단어 클러스터의 수를 표시하며, single은 단일 클러스터링을 의미한다.



[그림 7] 문서 클러스터링 결과 - F_1-measure로 정확도 측정

※ 각 포인트의 첨자는 사용된 단어 클러스터의 수를 표시하며, single은 단일 클러스터링을 의미한다.

(3) 결과 분석 및 논의

실험 결과 유사도 측정 방법의 정확도는 IB > L1 > Ward 순이며, 특히 IB는 단어 클러스터의 수에 관계없이 고른 결과를 보인다. 이는 (31), (34)의 식에서처럼, 중앙값을 지정하여 이에 대한 거리의 가중치를 고려한 결과라고 생각된다. L1 norm 및 Ward 기법에서도 이 개념을 적용한다면 결과값이 고르게 나올 것이라 예상된다.

단일 클러스터링과 이중 클러스터링 결과를 비교해볼 때, IB와 L1에서는 단일(single) 클러스터링보다는 이중(double) 클러스터링의 정확도가 더 높다. Ward 에서는 [표 Ⅷ]에서와 같이 특정 클러스터에 항목이 거의 포함되지 않는 현상이 있었으며, 단어 클러스터의 수가 적을수록 정확도가 떨어졌다.

세 경우를 종합해볼 때 이중 클러스터링에서 단어 클러스터의 수가 100일 때, 즉 $\frac{100}{5286} \approx 1.9\%$의 비율로 압축했을 때 좋은 결과를 얻을 수 있었다.

한 가지의 data set에 대해서만 실험을 했기 때문에 일반화를 하기는 힘들지만, [10]에서 제시한 결과, 즉, '정보병목기법'의 우수성, 이중 클러스터링이 단일 클러스터링보다 정확함을 확인할 수 있었다.

이 실험에서는 정확도 측정 방법에 따른 차이는 거의 없었다. 즉, 클러스터링이 잘 되는 경우의 정확도는 contingency table이나 F_1-measure 모두 비슷한 값을 가진다. Ward 기법의 경우는 [표 Ⅷ]과 같이 같은 클래스에 속한 원소들이 잘 모이긴 하였으나, 한 클러스터에 집중된 경우 contingency table은 대각선 값만을 고려하므로 정확도가 많이 떨어지게 되지만, F_1-measure로 측정시, 비슷한 원소들이 모여 있다는 점을 반영하기 때문에 상대적으로 더 정확도가 높게 측정된다.

[표 Ⅷ] 60 개의 단어 클러스터를 사용한 이중 클러스터링 결과: Ward 기법으로 유사도 측정

	Cluster0	Cluster1	Cluster2	Cluster3	RowSum
Class0	220	0	127	0	347
Class1	11	276	6	0	293
Class2	1	4	295	0	300
Class3	128	0	0	1	129
ColSum:	360	280	428	1	1069

- Contingency table 방식의 정확도 : 0.741
- F_1-measure 방식의 정확도 : 0.757

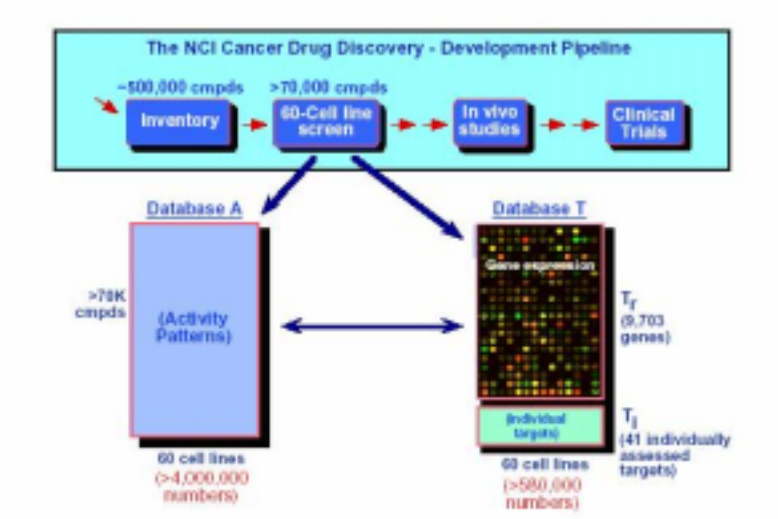
극단적인 경우로서, 대부분의 원소가 한 클러스터에 집중된 경우, contingency table로 측정한 정확도는 아주 낮아지지만, F_1-measure로 측정한 정확도는 높은 값을 가질 것이다.

이중 클러스터링에서, 먼저 클러스터링을 하는 대상(이 실험의 경우 단어)에 대해 몇 %의 클러스터를 구성하는 것이 좋은지는 추후 연구해볼 만한 과제이다.

4. Microarray data 분석에의 응용

(1) 데이터 선택

미국 국립 암센터(national cancer institute, NCI)는 암 치료제 개발을 위한 프로그램을 구성하였다. 이 프로그램의 주요 과정은 인간의 암과 유전자의 발현(expression) 그리고 암세포에 대한 약품의 활성(activity) 간의 관계를 밝히는 것으로서, [그림 8]에서와 같이 두 개의 데이터베이스(Database A and T)를 구축하였다 [9].



[그림 8] NCI Drug Discovery Program에서의 DB구축 개관

이 DB의 명칭은 'NCI60 Cell Lines Data Set'이며, 9개 조직 60여종의 인간 암세포의 유전자 발현정도(Database T), 약품의 활성도(Database A)에 대한 수치를 담고 있다. Database T가 바로 마이크로어레이를 이용해 생성한 자료로서, 60여 샘플에 대해 1,376개 유전자의 발현 정도를 수치화한 2차원 행렬 형태(1376×60)의 데이터이다.

이 실험에 사용할 자료는 1,376개의 유전자 중에서, 불완전한 데이터를 제거해 805개의 유전자만을 걸러낸 805×60 행렬 데이터로서, [4]에 사용된 것을 얻었다.

(2) 실험 상세구성

유전자의 발현 정도는 세포를 얻어낸 조직의 특성을 반영한다고 한다. 따라서 각 샘플의 클래스는 각 조직별로 9개로 구분되며, 60여개의 샘플을 9개의 클러스터로 구분하여 정확도를 측정할 수 있다. 유전자에 대한 클러스터링은 생물학적인 분석이 수반되어야 하며, 이 논문의 범위를 넘어서는 내용이므로, 실험은 60여개 샘플에 대한 일차 클러스터링만을 실시한다. 유사도 측정 방법은 문서 클러스터링 결과 정확도를 확인한 '정보 병목 기법'을 사용한다.

(3) 결과 및 분석

(가) 클래스의 명칭

BR : Breast (가슴)
CO : Colon (결장(結腸))
OV : Ovarian (난소)
RE : Renal (신장)
ME : Melanoma (피부암)
CNS : Central Nervous System (중추신경계)
LC : Lung (폐)
PR : Prostate (전립선)
LE : Leukemia (백혈병)

(나) 클래스와 클러스터 대응

이 실험에서는 어느 클래스에 해당하는지 결정하기 어려운 클러스터가 있었으며, 다음과 같이 대응이 잘 되는 클래스-클러스터를 먼저 짝지은 후 잔여처리를 하였다.

BR, CNS, CO, LE, OV에 대해서는 잘 대응이 되는 클러스터가 존재한다. RE에 대응되는 클러스터는 RE 샘플을 가장 많이 보유하고 있는 클러스터로 결정하였다. 클래스 LC, ME에 해당하는 클러스터는 이들의 샘플을 포함하는 클러스터로 결정하였으며, 마지막으로 남은 PR 클래스에 해당하는 클러스터는 이미 OV에 할당이 되어 있어서, 임의로 하나 남은 클러스터를 대응시켰다.

(다) Contingency Table

[표 Ⅷ] NCI60 Data Set의 마이크로어레이 데이터(Database T)에서,
60개 샘플에 대한 클러스터링 결과

Class \ Cluster	#1	#2	#3	#4	#5	#6	#7	#8	#9	계
1:BR	8									8
2:CNS		6								6
3:CO		1	5				1			7
4:LE				6						6
5:OV			1		5					6
6:RE			2			4		2		8
7:LC	1	3	2	1			2			9
8:PR					2					2
9:ME			4	2					2	8
합계	9	10	14	9	7	4	3	2	2	60

정확도는 0.63이며, 클래스-클러스터 매치가 잘 되지 않는 7~9번 데이터를 제외했을 때의 정확도는 0.57이다.

(라) F_1-measure : 0.62, 7~9번 데이터를 제외한 경우 0.50

(마) 결과 분석 및 논의

앞의 문서 클러스터링의 경우와는 달리, 클러스터간 구분이 명확하지 않았다. 원인으로서 고려할 수 있는 사항은, 상대적으로 적은 샘플의 수, 문서의 경우와는 달리, 유전자(단어에 해당)의 본질이 아직 완전히 밝혀지지 않았다는 점 등이다.

암세포가 발생한 조직에 따라서도 클러스터링 정도에 차이가 있음을 보인다. BR, CNS, LE의 경우는 완벽하게 클래스와 클러스터가 일치하였으나, LC, RE, ME의 경우 여러 클러스터에 샘플이 분산되는 것을 볼 수 있다.

(바) 다른 기법과의 결과 비교

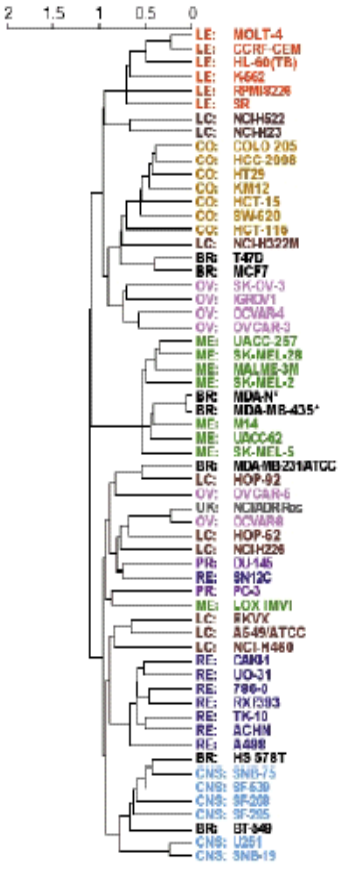
1) STVQ

[4]에서는 동일한 데이터를 벡터 양자화 기법의 일종인 STVQ(soft topographic vector quantization)을 이용해 클러스터링을 실시하였으며 결과는 [표 Ⅸ]와 같았다. [표 Ⅷ]과 [표 Ⅸ]를 비교한 결과는 [표 Ⅹ]과 같다.

STVQ는 위상적인 클러스터링 기법으로서, 행렬상의 위치가 비슷할수록 유사도가 큰 것이다. 각 클래스별로 비교한 결과는 [표 Ⅹ]과 같다. 편의상 [표 Ⅷ]을 행렬 A로, 원소를 a_{ij}로 표현하겠다. 이웃하는 경우는 []로 묶었다.

[표 Ⅸ] [4]에서 실시한 STVQ를 이용한 NCI60 T-matrix data set의 클러스터링 결과. [표 Ⅹ]을 구성하기 위한 자료

CNS:SNB-19 CNS:U251 CNS:SF-295	RE:A498 RE:ACHN RE:TK-10 RE:RXF-393 RE:UO-31 RE:CAKI-1	OV:OVCAR-3 OV:OVCAR-4 OV:IGROW1 OV:SK-OV-3	LC:NCI-H460 LC:A549/ATCC LC:EKVX
CNS:SF-268 CNS:SF-539 CNS:SNB-75 BR:BT-549 BR:HS578T LC:NCI-H226	ME:LOXIMVI PR:PC-3 LC:HOP-62	ME:M14 ME:MALME-3M ME:SK-MEL-28	BR:MDA-MB-435 BR:MDA-N
RE:SN12C LC:HOP-92 BR:MDA-MB-231 /ATCC	LC:NC-H23 LC:NC-H522 BR:MCF7 BR:T-47D	LE:SR LE:RPMI-8226 LE:K-562 LE:HL-60 LE:CCRF-CEM LE:MOLT-4	ME:SK-MEL-5 ME:UACC-62 ME:UACC-257 ME:SK-MEL-2
LC:NCI-H322M	CO:KM12 CO:HT29 CO:HCC-2998 CO:COLO205	CO:HCT-116 CO:SW-620 CO:HCT-15 OV:OVCAR-5	PR:DU-145 OV:OVCAR-8 BR:MCF7/ADF-RES



[그림 9] [9]의 결과

2) 다른 계층적 클러스터링 : average-linkage

[9]에서는 average-linkage 방식의 계층적 클러스터링을 실시하였으며 그 결과는 [그림 9]와 같다.

[표 X] 다른 클러스터링 기법과의 결과 비교 - 클러스터의 항목 수를 기준으로

Class	본 논문의 결과 (병합식, 정보병목기법)	[4]의 결과 (STVQ)	비교	[9]의 결과	비교
BR	8	[a21(2), a31(1), a32(2)], a24(2), a44(1)	상이	most heterogeneous	상이
CNS	6	[a11(3), a21(3)]	유사	6	유사
CO	5, 1, 1	[a42(4), a43(3)]에 분포	유사	7	유사
LC	1, 3, 2, 1, 2	a14(3), [a21(1), a22(1), a31(1), a32(2)], a41(1)	유사	3,2,2,1,1	유사
OV	1, 5	a13(4), [a43(1), a44(1)]	유사	4, 2	유사
PR	2	a22(1), a44(1)	유사	2	일치
RE	2, 4, 2	a12(7), a31(1)	상이	7, 1	상이
LE	6	a33(6)	일치	6	일치
ME	4, 2, 2	[a22(1), a23(3)], a34(4)	유사	7, 1	유사

전체적으로 클러스터링 결과가 유사하다고 판단할 수 있다. 각 조직별 암세포가 반드시 유전자의 발현 정도가 유사한지는 아직 연구가 진행 중인 부분으로서, '정답'은 없으며, 다만 기존의 다른 연구결과와 비교해볼 때 유사한 결과를 얻었으며, BR과 같이 기법마다 차이가 나는 부분은 생물학적 연구의 대상이 될 수 있을 것이다. 정리하면 '정보병목기법' 역시 DNA 마이크로어레이 데이터의 초기 분석 단계에서 좋은 도구로서 사용될 수 있다고 할 수 있다.

V. 결론 및 논의

이 논문에서는 클러스터링 기법의 전반적인 내용 및 관련 이론을 살펴보고, 여러 클러스터링 기법 중에서 계층적-병합식 클러스터링 기법에 초점을 두었다. 일반적인 클러스터링에서 많이 사용하는 2차원 행렬 형태의 데이터에 대한 클러스터링이 전체적인 윤곽이며, 최근에 소개된 두 가지 원칙(principle)을 직접 적용해보았다. 즉, 새로운 클러스터링 패러다임으로서 이중 클러스터링 기법을 선택하였고, 클러스터를 만들어나가는 데 가장 중요한 '유사도 측정' 기법으로서 최근에 소개된 '정보 병목 기법(information bottleneck method)'을 선택하여, 문서 클러스터링 및 DNA 마이크로어레이 데이터 클러스터링을 실시하였다.

실험 결과 단일 클러스터링보다는 이중 클러스터링이, 유사도 측정 기법에서는 '정보

병목 기법'을 적용했을 때 정확한 클러스터링이 이루어짐을 확인하였다. 동종의 여러 데이터 및 다양한 종류의 데이터에 적용하여 통계적인 결과를 얻지 못했다는 한계가 있으나, 문서 클러스터링에 적합한 '분포 클러스터링'개념에서의 이론적인 '유사도 측정법' 해답으로서 제시된 '정보 병목 기법'의 성능을 확인하기에는 충분하였다고 본다.

문서 클러스터링에 비해 DNA 마이크로어레이 데이터의 경우는 상대적으로 명확한 클러스터링이 되지 않았다. 클러스터링이 데이터의 전처리단계에 많이 사용된다는 점을 고려해볼 때, 다른 처리방법이 병행되어야 한다고 생각된다.

정확도 이외에 클러스터링 알고리즘의 복잡도를 짚고 넘어가야 할 것 같다. Ⅲ장 5-(2)에서 밝혔듯이, 이 실험에 사용한 병합식 알고리즘은 X를 단어의 집합이라고 하면, $O(|X|^3)$ 정도의 시간 복잡도를 가진다. 또, 공간 복잡도도 $O(|X|^2)$ 정도로서, 이는 데이터가 커질수록 알고리즘을 적용하는 데 상당한 비용이 든다는 것을 의미한다. 실험에 사용한 주요 컴퓨팅 환경은 [표 XI]과 같으며,

[표 XI] 실험에 사용한 컴퓨팅 환경

CPU	AMD Athlon 1.33G
RAM	SDRAM 133MHz 512M
OS	Windows 2000 Professional

문서 이중 클러스터링의 경우($|X| = 5286$) 1회 실험에 2시간 정도의 시간이 소요되었다. [10]에 따르면 이를 해결하기 위한 여러 가지 기법들이 제안되고 있으나, 아직 획기적으로 속도가 개선된 알고리즘은 없다.

정리하면, 병합식 클러스터링 기법으로서 '정보병목기법'을 적용한 이중 클러스터링을 통해 안정되고 정확한 결과를 얻을 수 있으며, 문서뿐만 아니라 DNA 마이크로어레이 데이터의 클러스터링에도 적용할 수 있다. 또, 데이터가 두 변수간의 결합확률분포 형태로 표현할 수만 있다면 데이터의 종류와 생성 방법에 관계 없이, 일괄적으로 '정보 병목 기법'을 적용할 수 있으므로, 이 기법은 좋은 도구로서 널리 활용될 수 있을 것이다.

다만, 많은 비용(속도, 메모리)이 필요하다는 단점이 있다. 기하급수적으로 빨라지는 컴퓨팅 속도와, 빠르게 감소하는 단위 용량당 메모리의 가격이 이 단점을 어느 정도 보충해줄 수 있으리라 생각된다. 또, 한편으로는 폭발적으로 증가하는 정보량에 따른 대량 데이터에 대한 클러스터링 기법의 수요를 충족시키기 위해서는, 근본적으로 속도를 개선하는 기법이 나와야 할 것이다.

Acknowledgement

이 논문은 서울대 컴퓨터공학부 바이오지능(bioinformatics) 연구실의 장병탁 교수님의 지도를 받아 작성하였습니다.

참고문헌

[1] 김영택, *자연언어처리*, 생능출판사, 2001, pp. 387-395

[2] 네이버 백과사전(http://100.naver.com), "기능유전체학" 항목

[3] Baldi, P. and Hatfield, G.W., *DNA Microarrays and Gene Expression*, Cambridge Press, 2002, pp.78-87

[4] Chang, J.-H, Hwang, K.-B, and Zhang, B.-T., "Analysis of gene expression profiles and drug activity patterns by clustering and bayesian network learning", *Methods of Microarray Data Analysis II*, pp.169-184, 2002.

[5] Cover, T.M. and Thomas, J.A., *Elements of Information Theory*. John Wiley & Sons, 1991, pp. 1-23

[6] Graepel T., "Statistical physics of clustering algorithms", Master thesis, Technical University of Berlin, 1998.

[7] Lin, S.M. and Johnson, K.F. (eds.), *Methods of Microarray Data Analysis II*. Kluwer Academic Publishers, 2002, pp. 9-17

[8] Pereira, F.C., Tishby, N., and Lee, L., "Distributional clustering of English words", In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics,* pp.183-190, 1993.

[9] Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., Scudiero, D.A., Eisen, M.B., Sausville, E.A., Pommier, Y., Botstein, D., Brown, P.O., and Weinstein, J.N., "A gene expression database for the molecular pharmacology of cancer", *Nature Genetics*, Vol. 24, pp. 236-244, 2000.

[10] Slonim, N. and Tishby, N., "Document clustering using word clusters via the information bottleneck method", In *Proceedings of SIGIR-2000*, pp.208-215, 2000.

[11] Tishby, N., Pereira, F.C., and Bialek, W., "The Information bottleneck method", In *Proceedings of the 37th Allerton Conference on Communication and Computation*, pp.368-377, 1999.