Journal of The Korea Society of Computer and Information
Vol. 20 No. 10, pp. 31-38, October 2015

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.10.031

Design and Implementation of the Ensemble-based Classification Model by Using *k*-means Clustering

Sung-Yeol Song*, A-Ra Khil **

Abstract

In this paper, we propose the ensemble-based classification model which extracts just new data patterns from the streaming-data by using clustering and generates new classification models to be added to the ensemble in order to reduce the number of data labeling while it keeps the accuracy of the existing system. The proposed technique performs clustering of similar patterned data from streaming data. It performs the data labeling to each cluster at the point when a certain amount of data has been gathered. The proposed technique applies the *K*-NN technique to the classification model unit in order to keep the accuracy of the existing system while it uses a small amount of data. The proposed technique is efficient as using about 3% less data comparing with the existing technique as shown the simulation results for benchmarks, thereby using clustering.

▸ Keyword : Streaming data, Ensemble, Clustering, Classification, *K*-NN

I. Introduction

데이터마이닝이란 저장된 데이터 안에서 유의미한 정보를 찾아내는 것을 말한다[1-2]. 대표적인 데이터마이닝 기법에는 학습데이터를 이용하여 생성한 분류모델을 기반으로 하여 새로운 데이터를 예측하는 분류기법이 있다. 분류기법은 서버 로그 데이터나 센서 데이터를 모니터링하여 사이버 범죄 여부 또는 사용자의 기호를 예측하는데 많이 사용한다[1]. 준비된 학습데이터를 이용하는 분류모델을 탑재한 모니터링 시스템은 새롭게 발생한 데이터를 이미 알려진 클래스를 기준으로 구분함으로써 얻을 수 있는 예측정보를 사용자에게 제공한다[3]. 사용자가 원하는 정보가 변경되거나 새로운 패턴의 데이터가 발생하는 경우, 이러한 모니터링 시스템은 기존의 분류기법 기반에 의하여 올바른 예측정보를 제공할 수 없다. 따라서 계속적으로 새로운 데이터 패턴으로의 변화가 요구되는 분야를 위한 분류모델은 적절한 시점에 새로이 생성 또는 갱신하는 작업이 반드시 필요하다.

스트리밍 데이터에 대하여 분류모델을 생성할 경우, 데이터 패턴의 개념변화(concept drift) 발생이 언제 어떻게 발생할 지에 대하여 예측할 수 없기 때문에 주기적으로, 저장되어있는 데이터들을 사용하여 분류모델을 갱신하는 방법을 사용한다[4-6]. 이와 같은 방법은 데이터 패턴의 개념변화가 발생하지 않은 경우에도 새로운 분류모델을 생성 또는 갱신하는 시도를 할 수 있다. 분류모델 생성 또는 갱신작업은 해당 분야의 전문가가 수작업으로 레이블링한 학습데이터의 이용이 반드시 필요하다. 따라서 분류모델 생성 또는 갱신작업의 횟수가 많아질수록 레이블링 횟수가 늘어나게 되며 결국은 많은 시간과 노력이 필요하게 된다[7]. 무한히 발생하는 스트리밍 데이터에 대하여 레이블링 작업을 수행하는 것은 현실적으로 불가능하기 때문에, 최근에는 부분 데이터 또는 입력 데이터의 분포를 비교, 분석하여 새로운 패턴의 데이터에 대해서만 분류모델을 생성 또는 갱신하는 방법에 대하여 연구가 진행되고 있다[7]. 그러나 이러한 방법 또한 새로운 패턴의 첫 번째 데이터를 비교 기준으로 삼기 때문에 첫 번째 데이터 의존도가 높다는 문제가 있다.

본 논문에서는 스트리밍 데이터에 대한 모니터링 시스템에서, 무한히 발생하는 새로운 데이터 패턴의 입력에 대한 레이블링 횟수를 줄이기 위하여 군집화를 기반으로 하는 앙상블 학습기반 분류모델 관리기법을 제안함으로써 과도한 레이블링에 의한 불필요한 시간과 노력을 절감한다.[1)]

본 논문에서 제안하는 앙상블 학습기반 분류모델 관리기법

• First Author: Sung-Yeol Song, Corresponding Author: A-Ra Khil
*Sung-Yeol Song(revwind@gmail.com), Dept. of Computer Science and Engineering, Soongsil University
**A-Ra Khil(ara@ssu.ac.kr), Dept. of Computer Science and Engineering, Soongsil University
• Received: 2015. 07. 15, Revised: 2015. 08. 03, Accepted: 2015. 09. 22.

은 데이터 수집 이전 별도의 기준을 설정하지 않고, 데이터들을 군집화하여 일정량이상 유사한 분포를 가지는 데이터로 묶일 때, 레이블링함으로써 불필요한 레이블링 횟수를 줄인다. 이러한 신규 학습데이터는 새로운 분류모델을 생성하는데 사용되며 기존의 앙상블에 추가함으로써 추가적인 레이블링 횟수를 최소화한다.

또한, 제안하는 앙상블 학습기반 분류모델 관리기법은 기존의 데이터 단위에 적용하는 K-NN기법[8]을 분류모델 단위로 확장, 적용한다. K-NN 기법을 분류모델 단위로 적용하는 방법에 의하여 앙상블 학습기반 분류모델은 기존의 방법에 비하여 레이블링 횟수를 감소시킴에도 불구하고 기존의 정확도를 유지할 수 있다.

본 논문에서는 제안하는 앙상블 학습기반 분류모델 관리기법의 타당성 및 효율성은 벤치마크 데이터에 대한 모의실험결과를 통하여 나타낸다. 모의실험결과에 의하여, 앙상블 학습기반 분류모델 관리기법은 첫 번째 데이터를 기준으로 분류모델을 생성, 갱신하는 기존의 방법과 비교하는 경우, 평균적으로 3%의 레이블링 횟수가 감소함을 알 수 있다. 또한, 본 논문에서 제안하는 앙상블 학습기반 분류모델 관리기법은 기존의 주기적으로 앙상블 모델을 생성하는 기법에서 사용하는 데이터양의 평균 9.8%의 데이터만을 사용하여도 기존의 정확도를 유지할 수 있음을 나타내 보인다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장은 스트리밍 데이터에서 분류모델을 갱신하는 방법론과 기존 연구들을 설명한다. 3장에서는 앙상블 학습기반 분류모델 관리기법을 기술한다. 4장은 제안한 방법의 타당성을 검증하기 위한 실험 결과들을 기술하고, 끝으로 5장에서는 본 논문의 결론과 함께 향후 연구를 제시한다.

II. Preliminaries

1. Related works

스트리밍 데이터에 대한 기존의 분류모델 관리기법들은 사전에 정의한 주기를 기준으로 분류모델을 갱신한다[4-6, 8]. 사전에 주기를 정의하여 분류모델을 갱신하는 방법은 점증적 학습방법[8]과 앙상블 접근 방법[4-7]으로 나눌 수 있다. 점증적 학습방법은 기존 분류모델 자체에 새로운 학습 데이터를 점증적으로 추가 적용함으로써 분류모델을 확장, 갱신하는 방법이다.

앙상블 접근 방법은 새로운 데이터에 적합한 새로운 분류모델을 생성하여 기존의 앙상블에 추가함으로써 기존 분류모델을 포함하는 앙상블을 갱신하는 방법이다.

그림 1의 (a)는 10개의 데이터가 저장될 때 마다 분류모델을 갱신하도록 정의하였을 때의 점증적 학습방법의 분류모델 갱신방법의 예이며 그림 1의 (b)는 같은 조건에서의 앙상블 접근방법의 분류모델 갱신 방법의 예이다.

이와 같이 점증적 학습방법은 기존의 분류모델의 오류율을 지속적으로 모니터링하여 오류율이 크게 떨어지는 시점에 분류모델을 갱신하도록 하는 방법으로 분류 능력을 향상시키는 반면, 앙상블 학습방법은 새로운 분류모델을 지정된 주기에 따라 주기적으로 생성하여 앙상블에 추가하는 방법을 사용한다. 점증적 학습방법으로 구축, 관리되는 분류기는 전문가의 레이블링 작업에 의하여 생성된 분류모델에 의해 모든 스트리밍 데이터의 분류가 가능하다고 가정한다. 그러나 앙상블 학습방법을 사용하는 분류기는 일정 시간 동안의 일부 데이터만을 사용하여 새로운 분류모델을 생성하기 때문에 스트리밍 데이터의 특화된 정확한 분류가 가능하므로 점증적 학습 방법에 비해 사용이 용이하다. 특히 데이터 분포가 변하는 스트리밍 데이터의 경우 앙상블 학습 방법을 이용한 다중 분류모델이 점증적 학습방법을 이용한 단일 분류모델보다 더 정확하게 데이터를 분류하는 것으로 알려져 있다[9].

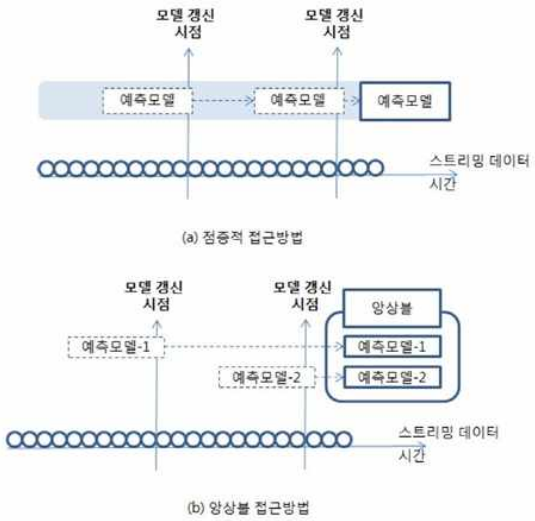


Fig. 1. Incremental and Ensemble Approach

스트리밍 데이터를 위한 앙상블 기반 분류모델 생성 방법에 대한 초창기 연구는 데이터 분포 변화를 처리할 수 있도록 앙상블을 구성하고 있는 분류모델들의 가중치를 보정하는방법에 초점이 맞춰졌다. Wang et al.[4]이 제안한 Accuracy Weighted Ensemble(AWE)은 주기적으로 가장 최근에 수집된 데이터를 이용하여 앙상블의 새로운 분류모델을 생성하는 동시에 동일한 데이터로 앙상블에 존재하는 기존 분류모델들을 평가한다. 이 평가 결과에 따라 해당 분류모델들의 가중치를 보정하고, 앙상블의 분류모델 개수가 사전에 정의된 최대 개수보다 큰 경우에는 가장 낮은 평가 결과를 보인 분류모델을 삭제한다.

Dynamic Weighted Majority(DWM)[5]방법 역시 앙상블을 구성하고 있는 분류모델들의 가중치를 보정한다. 여기서는 분류 오류를 발생하는 분류모델의 가중치를 내리고, 올바르게 분류하는 분류모델의 가중치를 올린다. DWM은 주기적으로 각 분류모델의 가중치를 조사하여 사전에 정의한 임계값보다 작은 가중치를 갖는 분류모델을 삭제한다. 분류모델 가중치 조사 때,

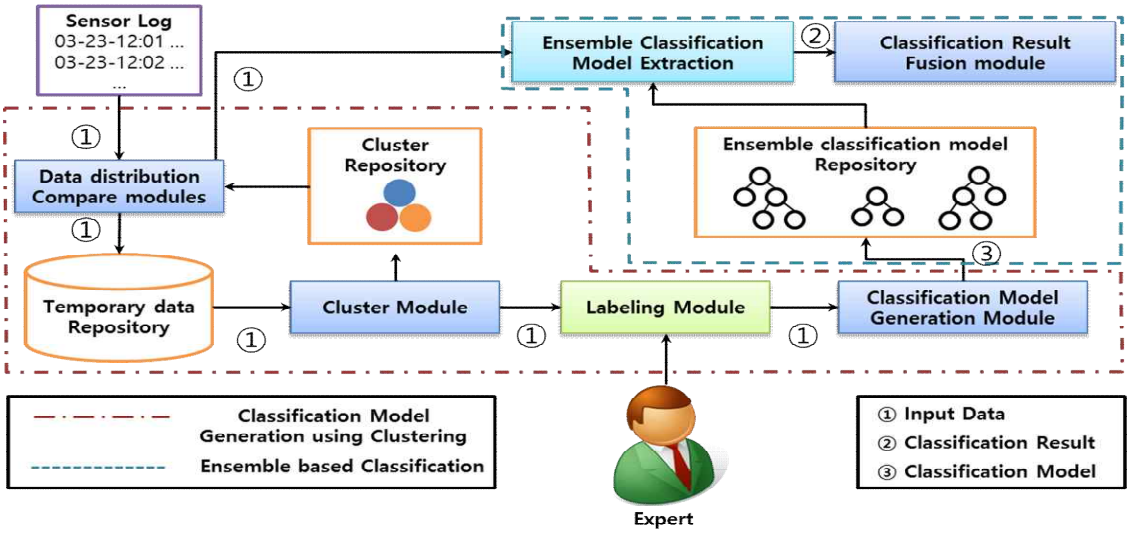


Fig. 2. System Architecture

입력 데이터데 대한 앙상블의 분류가 오류를 발생하게 되면 분류기는 새로운 앙상블 분류모델을 생성한다.

최근에는 정확성과 활용성을 높이는 연구가 진행되고 있는데, AWE 방법을 개선한 Accuray Updated Ensemble(AUE)[6]는 주기적으로 기존의 앙상블 분류모델들의 가중치를 보정할 뿐만 아니라 점증적 학습 방법을 이용하여 분류모델을 갱신한다.

이상의 앙상블 기반의 분류방법들은 새로운 분류모델을 만들기 위해서 모든 입력 데이터를 대상으로 주기적으로 전문가의 수작업으로 이루어지는 레이블링 작업이 필요하다. 특히, 스트리밍 데이터 대상의 분류기인 경우, 주기적으로 무한 반복하는 레이블링 작업으로 인하여 불필요한 시간과 노력의 낭비가 발생하지 않도록 하는 연구가 필요하다.

데이터 분포 기반 앙상블 분류모델 관리기법[7]은 현재 앙상블의 분류모델을 생성할 때 사용한 데이터 분포와 상이한 데이터분포가 입력되는 경우에만 새로운 분류모델을 생성하여 현재의 앙상블을 갱신하는 방법을 제안함으로써 불필요한 레이블링 수를 감소시키는 장점이 있다. 그러나 이 방법은 새로운 분류모델이 생성되기 이전에 입력되는 새로운 분포의 데이터에 대하여 분류 오류가 발생할 가능성이 높으며, 새로운 데이터 분포 패턴이 나타나는 첫 번째 데이터 패턴을 기준으로 새로운 분류모델을 생성하기 때문에 첫 번째 데이터에 대한 의존도가 지나치게 높다는 문제점이 있다.

III. The Proposed Scheme

스트리밍 데이터는 무한히 발생하기 때문에 분류 정확도를 유지하기 위해서 항상 새로운 데이터에 대한 추가 학습이 필요하다. 본 논문에서는 스트리밍 데이터 환경에서 추가 학습을 위해 전문가에 의한 레이블링을 필요로 하는 데이터양을 줄이는 방법을 연구한다. 이를 위해 군집화기법을 사용하여 입력되는 스트리밍 데이터를 분포가 유사한 데이터 군집으로 만들고, 각 군집에 해당하는 모델을 생성하여 앙상블에 추가한 후 데이터 분류를 수행한다. 앙상블 학습기반 분류모델 관리기법은 크게 군집화를 통한 분류모델 생성, 앙상블 기반 데이터 분류로 나뉘어진다.

그림 2는 본 논문에서 제안하는 앙상블 학습기반 분류모델 관리기법의 구조도이다. 그림에서와 같이 스트리밍 데이터가 입력되면 입력 데이터와 기존에 분류모델을 생성할 때 사용한 데이터 군집과 비교하여 입력 데이터가 새로운 분포의 데이터인지, 혹은 기존의 군집에 속하는 데이터인지를 확인한다. 만약 존재하는 군집 중 어느 영역에도 속하지 않는 새로운 데이터들, 즉 분포가 이전과 다른 데이터가 발생한다면 추가학습이 필요하기 때문에 임시저장소에 저장한다. 이와 반대로 분포가 이전과 같다면 앙상블을 이용한 분류결과를 생성하고 다음 데이터의 입력을 기다린다. 임시저장소에 저장된 데이터는 기존의 데이터 분포와 전혀 다른 데이터이기 때문에 군집화기법을 이용하여 군집을 생성한다. 본 논문에서는 K-means기법[10]을 이용하여 데이터 군집을 수행하였다. 군집으로 묶인 데이터는 추가 학습을 위하여 전문가가 각 데이터 클래스에 레이블링을 하고, 이를 학습데이터로 이용한 분류모델을 생성한다. 생성된 각 분류모델은 앙상블에 추가되어 스트리밍 입력데이터 분류에 사용된다.

입력데이터의 분류를 수행할 때에는 K-NN 알고리즘을 데이터에서 모델 개념으로 확장한 방식으로 분류를 수행한다. 즉 입력 데이터와 최근접한 k개의 데이터 대신 k개의 모델들을 이용하여 분류를 수행한다. 앙상블 학습기반 분류모델의 각 앙상블에 속하는 모든 단일 분류모델에는 해당 분류모델이 담당하는 분류영역이 주어진다. 분류영역이란 해당 분류모델이 학습될 때 사용된 군집을 이용하여 정의되며 각 분류모델은 담당하는 분류영역에서 학습이 되어 있기 때문에 분류정확도를 높일 수 있다.

그림 3은 입력데이터가 주어졌을 때 새로운 분류모델을 앙상블 분류모델에 추가하는 과정의 알고리즘이다. 그림 3에서와 같이 앙상블 내에 저장된 군집들을 확인하여 입력데이터 *d*가 새로운 패턴의 데이터인지를 확인한다. 만약 해당하는 군집이 있는 기존과 유사한 패턴의 데이터라면 기존의 앙상블 분류모델을 유지한 채 종료한다. 만약 입력데이터가 속한 군집이 없는 새로운 패턴의 데이터라면, 해당 데이터를 임시저장소에 저장한다. 그리고 임시저장소에 저장된 데이터양이 일정량이상 되었을 경우 run_clustering 함수를 이용하여 군집화를 수행한다. 여기서 일정량 이상이란 학습 데이터 크기를 의미하기 때문에 Millan-giraldo et al.에서 학습데이터 개수를 결정했단 방식과 같이 "속성 개수 X 클래스 개수 X 10"을 이용하여 결정한다. 군집화 수행 후 한 개 이상의 군집에서 일정량 데이터를 가진 군집이 존재한다면, 해당군집의 데이터를 전문가에게 클래스 정보 입력을 요청(*input_category_information*)하고 전문가는 레이블링 작업을 수행한다. 전문가에 의해 레이블링되어 생성된 학습 데이터를 이용하여 *make_classification_model* 함수는 해당하는 군집의 분류모델 *m*을 생성한다. 생성된 분류모델 m은 *input_ensemble_module*를 통해 앙상블 분류모델에 추가되고 갱신된 앙상블 분류모델을 반환한다. 위의 알고리즘과 같이 기존 단일 분류모델들의 분류영역에 속하지 않는 데이터에 알맞은 새로운 분류모델을 생성함으로써 앙상블 분류모델이 항상 새로운 데이터에 적응할 수 있도록 한다.

```
input: Streaming Data d, Ensemble Classification Model e
output: New Ensemble Classification Model

initialize: i = 0, clustering_start_size = x, classification_start_size = y
{
    if(d is new pattern data AND x >= size of temporary_storage) {
        Set of temporary Cluster T = run_clustering();
        i = 0;
        while(i< number of T) {
            if(y >= size of t_i){
                input_category_information(t_i );
                classification_model m =
                    make_classification_model(t_i );
                e = input_ensemble_module(m);
            }
        }
    }
    return e;
}
```

Fig. 3. Algorithm for Building Classification Model

본 논문에서 제안하는 앙상블 학습기반 분류에서 분류결과를 생성하는 방법은 널리 사용되는 *K*-NN알고리즘의 확장이라고 볼 수 있다. 본래의 *K*-NN알고리즘에서 각 올바른 범주를 갖는 *k*개의 데이터 포인트를 이용하여 분류를 수행한다면, 앙상블 학습기반 분류모델은 데이터 포인트를 분류모델로 확장하여 *k*개의 분류모델을 이용하여 분류를 수행한다. 입력 데이터가 주어지게 되면 앙상블에 속한 분류영역들과 입력데이터를 비교하여 *k*개의 최근접 군집들의 분류모델을 이용하여 분류를 수행한다. 분류 수행 후 분류모델들의 결과를 종합하여 최종적으로 입력데이터의 범주를 예측한다. 이러한 방법은 입력데이터가 속하는 분류영역이 앙상블에 존재하지 않을 경우에도 신뢰성있는 분류결과를 예측할 수 있다는 장점이 있다. 또한 *K*-NN 알고리즘의 장점인 새로운 데이터에 쉽게 적응할 수 있는 높은 적응성을 가질 뿐만 아니라, *K*-NN알고리즘의 단점이라고 할 수 있는 낮은 일반화 성능을 데이터 공간상의 지역화된 분류모델의 학습을 통하여 전체적으로 일반화 성능을 향상시킬 수 있다. 그림 4는 본 논문에서 제안하는 앙상블 분류결과를 융합하는 모듈의 알고리즘이다.

```
input: Streaming Data d, Set of Cluster C_i
output: Classification Result

initialize: i = 0, k = max number of close cluster
{
    while(i < number of C_i) {
        if(d is in area of c_i) {
            model m = find_closest_cluster(C, d);
            result of classification r_i=
                make_result_classification(m, d);
            exit;
        }
    }
    k set of cluster C_k= find_close_cluster(k);
    i = 0;
    Set of model M = get_model(C_k );
    Set of result R;
    while(i < number of M){
        result of classification r_i= make_result_classification(m_i ,d);
    }
    return combine_result(R);
}
```

Fig. 4. Algorithm for Combining Classification Result

앙상블 분류모델에 입력데이터가 입력되면 해당하는 데이터가 속하는 군집이 있는지 여부를 검사한다. 만약 속하는 군집이 있다면, *find_closest_cluster*를 이용하여 가장 가까운 군집의 모델을 추출하고 분류결과를 생성(*make_result_classification*)한다. 속하는 군집이 없다면, *find_close_cluster*를 이용하여 *k*개의 최근접 군집을 탐색한다. 최근접 군집 탐색 후 해당하는 *k*개의 군집의 모델을 이용하여 분류결과를 생성하고 이를 융합(*combine_result*)하여 최종 분류 결과를 생성한다. 위의 과정을 통해 기존의 데이터와 분포가 유사한 데이터는 해당하는 분류모델을 이용하고, 새로운 분포의 데이터는 분포가 유사한 K개의 분류모델을 이용하여 분류결과를 생성함으로써 정확도를 높이거나 유지할 수 있다는 장점이 있다.

IV. Experimental Results

군집화를 이용한 스트리밍 데이터에서 앙상블 방법의 타당성을 검증하기 위하여 앙상블의 단일 분류모델을 주기적으로 생성하는 방법과, 군집화를 사용하지 않고 데이터 분포를 이용하여 앙상블 모델을 생성하는 방법과 비교한다. 모의실험에서는 Weka(www.cs.waikato.ac.nz/ml/weka/)에서 제공하고 있는 의사결정트리 알고리즘인 J48을 분류모델로 사용하고 군집화 알고리즘은 역시 Weka에서 제공하고 있는 군집화 알고

리즘 중 k-means인 SimpleKmeans를 사용한다.

1. Experimental Data and Measure

UCI 데이터 저장소에서 제공하고 있는 벤치마크 데이터 중 데이터 개수가 5,000개 이상인 4개의 벤치마크 데이터와 스트리밍 데이터 벤치마크 데이터 3개를 모의실험데이터로 사용하였다. 표 1에서 상위 4개의 데이터가 일반 벤치마크 데이터이고, 하위 3개의 데이터가 스트리밍 데이터이다. 스트리밍 데이터 중 EM(Electricity Market)은 2년 6개월 간의 전력시장에서의 가격의 변동을 나타낸 데이터이다. PAKDD2009 데이터는 5년동안 수집된 고객들의 신용평가 데이터이며, KDDCup99는 군 네트워크환경에서 여러 유형의 네트워크 침입탐지를 시뮬레이션한 데이터이다.

Table 1. Benchmark Data

Data	Number of Attribute	Number of Data	Rate of Class(%)
Mushroom	23	8,124	C_1(50), C_2(50)
MAGIC	11	19,020	C_1(65), C_2(35)
Adult	15	48,842	C_1(24), C_2(76)
MiniBoone	51	129,596	C_1(28), C_2(72)
EM	7	27,409	C_1(40), C_2(60)
PAKDD2009	24	50,000	C_1(80), C_2(20)
KDDCup99	42	494,021	C_1(20), C_2(80)

모의실험에서는 실험 데이터를 순차적으로 시스템에 입력하고, 조건에 만족하면 새로운 분류모델을 생성 및 앙상블에 추가한다. 최소 군집의 크기는 Millan-giraldo등의 연구에서 제안한 바와 같이 "속성 개수 X 클래스 개수 X 10"으로 결정하였다. 군집화 이후 생성되는 분류모델의 개수는 각 군집의 크기가 최소 군집의 크기 이상이 되면 생성하므로 복수의 분류모델이 생성될 수 있다. 또한 실험에서 사용되는 모든 앙상블의 공정한 성능 비교를 위해 이미 생성된 분류모델은 삭제되지 않았다. 앙상블 분류 결과를 융합할 때에는 입력데이터가 속하는 군집이 있을 경우에는 해당 군집의 분류모델을 사용하였으며, 속하지 않는 데이터였을 경우 K-NN기법과 유사하게 1~3개의 최근접 군집의 분류모델을 사용한 simple voting을 이용하여 분류결과를 생성하였다. 즉, 3개 이하의 분류모델을 가지는 초기의 앙상블에는 모든 분류모델을 사용하고, 3개를 초과할 경우 3개의 최근접 군집에 해당하는 모델을 이용하여 분류결과를 생성하였다.

본 실험에서는 스트리밍 데이터에서 앙상블 모델의 정확도를 유지하면서 전문가가 데이터를 레이블링해야하는 양을 효율적으로 줄이는가를 측정해야한다. 따라서 다음 세 가지의 평가척도를 사용한다.

(1) WSF(Weighted Sum F-measure): 가중합을 적용한 F-measure 값

(2) TC(the Total number of Classifiers): 앙상블 모델에서 생성된 단일 분류모델의 수

(3) RL(Rate of Labeled Data): 전체 테스트 스트리밍 데이터에서 전문가의 레이블링을 요구량

일정 기간 동안 많은 양의 단일 분류모델을 생성하였다면 그것은 빈번하게 전문가의 개입을 필요로 하고 있다는 것을 의미한다. 또한 RL수치가 높다는 것은 전문가가 레이블링해야 되는 데이터량이 그만큼 많다는 것을 의미한다. 따라서 정확도를 기존방법과 어느정도 비슷한 수준으로 유지하면서 TC, RL수치가 낮을수록 본 논문에서 제안한 방법이 스트리밍 데이터에서 실제로 유용하다고 할 수 있다.

표1에서 보는 바와 같이 실험 데이터에서 각 클래스이 비율이 균등하지 않기 때문에 정확도를 측정하는 방법은 아래의 식과 같은 F-measure의 가중합 방법을 사용하였다. C는 데이터의 클래스의 집합을 나타내고 $F(c_i)$는 i번째 클래스에

Table 2. Result of simulation

Data	The proposed method			Ensemble method using a data distribution			CEA(Chunk-based Ensemble Approach)		
	WSF	TC	RL(%)	WSF	TC	RL(%)	WSF	TC	RL(%)
Mushroom	0.810	10	33	0.824	7	42	0.795	16	96
MAGIC	0.628	5	3	0.739	8	9	0.774	85	99
Adult	0.517	6	2	0.532	7	4	0.673	161	99
MiniBoone	0.713	5	2	0.782	4	3	0.862	125	99
EM	0.639	10	3	0.667	13	6	0.68	194	99
PAKDD2009	0.334	4	2	0.24	8	7	0.18	102	98
KDDCup99	0.886	17	2	0.971	16	2	0.644	586	99
평균	0.644	8.1	7	0.6	8.2	10	0.6	160.2	98.2

대한 정확도(precision)와 재현율(recall)의 조합평균인 F-measure값을 의미한다. w_i는 i번째 클래스의 가중치이다.

$$WSF = \sum_{c_i \in C} w_i F(c_i) \quad (1)$$

여기서 가중치 w_i는 i번째 클래스가 테스트 스트리밍에서 차지하는 비율이 클수록 작은 값을 갖는다. 본 실험에서는 각 클래스의 가중치를 아래의 식과 같이 테스트 스트리밍 데이터에서의 클래스 비율에 따라 결정한 후 앙상블 방법의 정확성을 계산한다.

$$w_i = 1 - \frac{n_i}{N} \quad (2)$$

실험에서 사용한 데이터는 클래스가 2분화 되어있는 데이터이기 때문에 가중합 비율을 계산하는 부분은 1이 되므로 생략하였다. WSF 값은 1.0에 가까울수록 높은 정확성을 의미한다.

2. Experimental Results

스트리밍데이터에서 분류모델을 갱신할 때 주기적으로 갱신하는 것은 가장 보편적인 접근이다. 따라서 제안한 앙상블 학습 방법과 데이터 분포를 이용하여 분류모델을 갱신하는 방법, 주기적으로 갱신하는 방법과 결과들을 비교한다. 주기적으로 단일 분류모델을 갱신하는 방법 CEA(Chunk-based Ensemble Approach) 중 입력 데이터의 클래스를 예측할 때 결과를 결합하는 방법에는 여러 가지가 있는데, 실험에는 simple voting을 사용하여 결합하는 방법을 사용하였다. simple voting은 결과들 중 가장 높은 빈도를 갖는 클래스로 입력 데이터를 분류한다.

표 2는 7개의 실험 데이터에 대해서 제안한 앙상블 방법과 데이터 분포를 이용한 앙상블방법, CEA 방법의 결과를 보여준다. 제안한 방법은 실험 데이터에 대해서 CEA와 비교하여 평균 91.2% 포인터 적은 데이터를 가지고 단일 분류모델을 생성하였으며, 데이터 분포를 이용한 앙상블 방법보다 3% 포인트 적은 데이터를 사용하였다. 또한 최종적으로 생성된 분류모델은 CEA방법에 비해 94.9% 적은 단일 분류모델을 사용하였으며 이는 데이터 분포를 이용한 앙상블 방법과 비슷하다. 적은 양의 데이터를 이용하여 비슷한 양의 분류모델을 생성한 것은 기존 데이터 분포를 이용하여 분류모델을 생성한 방법은 한번에 하나의 단일 분류모델을 생성하지만, 본 논문에서 제안하는 방법은 군집화 이후 복수의 분류모델을 생성할 수 있기 때문으로 보인다. 평균 정확도의 경우 기존의 방법들에 비해 0.04의 평균정확도가 향상된 것을 볼 수 있지만 이는 오차범위 이내이므로 비슷한 정확도를 보인다고 할 수 있다. 따라서 본 논문에서 제안하는 방법이 기존의 방법들과 비

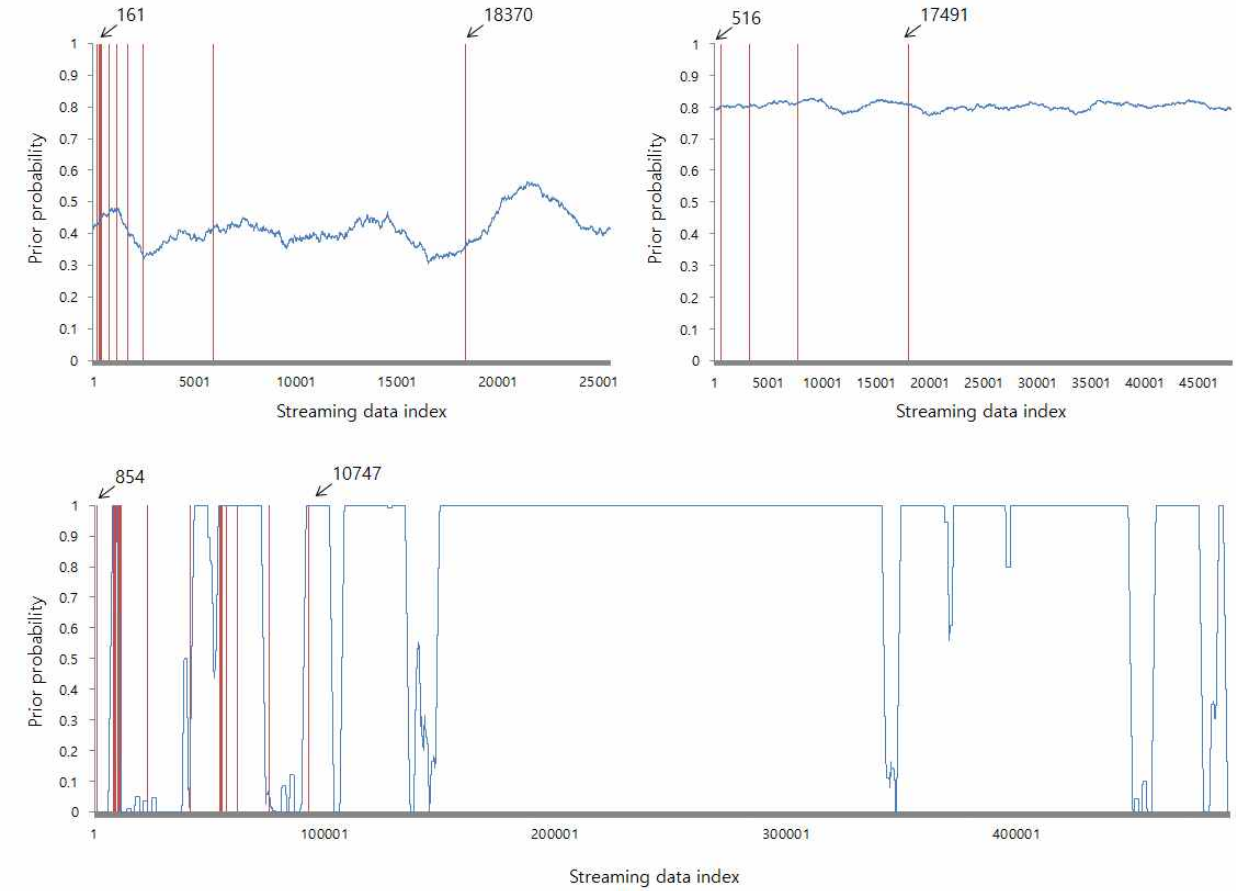
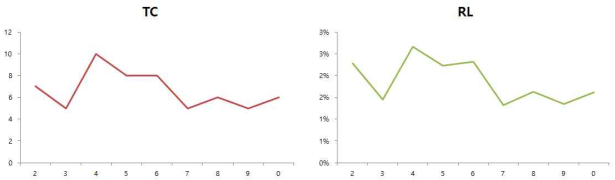


Fig. 5. Generation Point of Classification Model

슷한 정확도를 유지하면서 기존의 방법들에 비해 레이블링에 적은 양의 데이터를 사용한다고 할 수 있다. 이는 기존 방법이 첫 번째 입력데이터를 중심으로 하기 때문에 큰 의존도를 가지는 반면 본 논문에서 제안하는 방법은 군집화를 이용하여 새로운 분류영역의 중심데이터(Centroid)를 최적화하기 때문이다.

3. Change of Results according to *k* Value



	2	3	4	5	6	7	8	9	10
WSF	0.654	0.603	0.639	0.620	0.646	0.630	0.658	0.627	0.631
TC	7	5	10	8	8	5	6	5	6
RL	2%	1%	3%	2%	2%	1%	2%	1%	2%

Fig. 6. Change of TC, RL, WSF according to k value

k-means 알고리즘을 군집화에 사용하였기 때문에 *k*개의 변화에 따라 군집화 성능이 변하게 된다. 그림 6은 *k*개의 변화에 따른 앙상블 분류 실험결과이다. 실험에 사용한 데이터는 스트리밍 데이터 중 EM데이터를 사용하였다.

군집 개수가 2~3개일 때에는 3개의 항목의 변동치가 크지만 4개를 기점으로 군집의 수가 낮아짐에 따라 TC값과 RL값이 하락하고 있음을 알 수 있다. 이는 군집화과정에서 군집수가 많아짐에 따라 한번에 앙상블에 추가되는 모델의 수가 늘어났기 때문으로 분석된다. 앙상블에 모델이 추가되는 시점을 줄임에 따라 데이터 분포가 겹쳐지게 되는 경우가 점점 줄어들어 앙상블 모델이 전체 데이터를 다루는데 필요한 데이터가 적어지기 때문이다.

실험 데이터 중에서 벤치마크 스트리밍 데이터인 3개를 이용하여 제안한 앙상블 방법에 의해 생성된 단일 분류모델의 생성 시점들을 분석하였다. 아래 그림은 특정 시점에서 특정 클래스가 발생할 사전확률(prior probability)이 변하는 것을 보여주고, 수직선은 새로운 분류모델이 생성되는 시점을 나타낸다. 예를 들어 그림의 첫 번째 수직선은 161번째 “EM” 스트리밍 데이터에서 앙상블의 첫 번째 분류리가 생성되었음을 의미한다. 특정 클래스의 사전확률을 계산하는 방법은 Zhang 등의 연구를 참조하였다[9]. 실험 결과 초기에는 대체로 사전확률이 변하는 시점에 분류모델이 생성되었음을 볼 수 있고, 일정 수준이상의 분류모델이 생성된 이후는 거의 생성되지 않음을 볼 수 있다. 사전확률이 변함에도 새로운 분류모델이 생성되지 않는 것은 이미 이전에 생성된 분류모델의 분류영역이 변화하는 데이터 분포를 포함하고 있기 때문으로 보인다.

모의실험 결과 본 논문에서 제안하는 군집화를 이용한 앙상블 학습기반 분류모델방법은 기존의 분류 정확도를 유지하면서 레이블링양을 줄이는데 유용한 것으로 나타났다.

IV. Conclusion

무한히 발생하는 스트리밍 데이터를 효과적으로 분류하기 위해서는 데이터 분포가 바뀌는 시점에서 적절하게 분류모델을 갱신해주어야만 한다. 그러나 실시간으로 데이터 분포를 모니터링하고 분류모델을 생성하기 위해 매번 학습데이터를 생성하는 것은 현실적으로 불가능하다. 보통 스트리밍 데이터에서의 분류방법은 주기적으로 전문가가 학습데이터를 생성하고 분류모델을 갱신하는 방법을 사용한다. 하지만 기존의 접근방법은 주기적으로 분류모델을 생성하기 때문에 데이터 분포가 바뀌지 않아도 분류모델을 생성하거나 데이터 분포가 바뀌어도 분류모델을 생성하지 않는 경우가 발생할 수 있다. 이러한 문제를 해결하기 위하여 데이터 분포를 이용한 앙상블 분류방법이 연구되었지만, 이 방법 또한 데이터 분포가 다른 첫 번째 데이터를 기준으로 분류영역을 지정하기 때문에 이에 대한 의존도가 크다.

따라서 본 논문에서는 스트리밍 데이터에서 강인하면서도 첫 번째 데이터의 의존도를 줄이기 위해 군집화를 사용한 앙상블 분류방법을 제안한다. 즉 기존의 방법에서 거리값을 이용한 데이터 분포를 이용하는 대신 군집화 기법을 이용하여 분류영역을 설정함으로써 첫 번째 데이터에 의존하지 않으면서도 데이터 분포가 변하는 시점을 찾아내어 분류모델을 생성할 수 있다. 기존의 데이터와 분포가 다른 데이터가 입력되어 일정량 이상 모이게 되면 군집화를 수행하고 군집의 크기가 일정량 이상이 되면 해당하는 군집을 이용하여 분류모델을 생성한다. 데이터 분포가 변하는 시점을 자동으로 판단하면서도 기준의 방법에서 문제가 되었던 첫 번째 데이터에 대한 의존도가 없어졌기 때문에 더 효율적으로 앙상블 분류모델을 생성할 수 있다. 또한 주기적으로 분류모델을 생성하는 방법에 비해 전문가가 개입하여 레이블링하는 데이터량을 줄이면서도 분류모델 성능을 유지할 수 있다.

앙상블 분류모델에서는 다수의 모델에서 생성된 결과를 융합하기 위하여 *K*-NN방법을 모델 개념으로 확장한 방식을 사용하여 새로운 데이터에 쉽게 적응할 수 있게 하며, *k*개의 데이터로만 분류결과가 생성되는 데이터 단위가 아닌 모델 개념의 *K*-NN방법을 사용함으로써 일반화 성능을 향상시킬 수 있다는 장점이 있다.

본 논문에서 제안하는 방법의 유용성을 검증하기 위해 다양한 벤치마크 데이터를 사용한 실험결과 정확도를 유지하면서도 전문가가 개입해서 입력해야할 데이터량은 전체 데이터를 100%로 정의할 때 주기적으로 갱신하는 방법에 비해 평균 91%, 데이터 분포를 사용한 방법에 비해 평균 3%가 적은 데이터를 사용하는 것으로 나타나 본 논문에서 제안하는 방법이 유용한 것으로 나타났다.

제안한 방법은 군집화 방법에서 *k*-means, 모델을 융합할 때 *K*-NN방법을 사용함으로써 *K*값에 대한 의존도가 높은데

이를 보완 혹은 일반화할 방법에 대해 향후연구가 필요하다. 또한 앙상블에서 노후화된 분류모델을 삭제하는 방법에 대한 연구하고자 한다.

REFERENCE

[1] Hebah H. O. Nasereddin, "Stream Data Mining,"International Journal of Web Applications, vol.1,no.4, pp.183-190, 2009.

[2] Kantardzic, Mehmed. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons, 2011.

[3] Tsymbal, Alexey. "The problem of concept drift: definitions and related work." Computer Science Department, Trinity College Dublin 106 (2004).

[4] Wang, Haixun, et al. "Mining concept-drifting data streams using ensemble classifiers." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[5] Kolter, Jeremy Z., and M. Maloof. "Dynamic weighted majority: A new ensemble method for tracking concept drift." Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.

[6] Brzeziński, Dariusz, and Jerzy Stefanowski. "Accuracy updated ensemble for data streams with concept drift." Hybrid Artificial Intelligent Systems. Springer Berlin Heidelberg, 2011. 155-163.

[7] Joung-Woo Ryu and Myung-Won Kim, "An Ensemble Model based on Data Distribution for Streaming Data Classification," Journal of KIISE : Database Research, vol.40, no.2, 2013, 89-98.

[8] Altman, Naomi S. "An introduction to kernel and nearest-neighbor nonparametric regression." The American Statistician 46.3 (1992): 175-185.

[9] Domeniconi, Carlotta, and Dimitrios Gunopulos. "Incremental support vector machine construction." Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, 2001.

[10] Bock, Hans-Hermann. "Clustering methods: a history of k-means algorithms." Selected contributions in data analysis and classification. Springer Berlin Heidelberg, 2007. 161-172.

Authors

Sung-Yeol Song received the B.S. and degree in Computer Science and Engineering from Soongsil University, Seoul, Republic of Korea, in 2008 and 2010. Since 2010 he has been studied in Computer Science and Engineering from Soongsil University, Seoul, Republic of Korea. His current research interests include artificial intelligence, data mining, ubiquitous communications. Contact him at revwind@gmail.com

She received the B.S. degree in computer science from Ewha Women's University in 1987, the M.S. and Ph.D degrees in computer science from Korea Advanced Institute of Science and Technology(KAIST), Daejeon, Republic of Korea, in 1990 and 1997, respectively. Since 1995, she has been a senior engineer in Serome Ltd., Republic of Korea. Since 2003 she has been a member of board of directors in Dialpad Ltd., USA. Since 1997, she is with School of Computer Science and Engineering, Soongsil University, Seoul, Korea. Her current research interests include ubiquitous communications, wireless sensor networks, and embedded operating system. Contact her at ara@ssu.ac.kr.