논문 2010-47SP-6-16

클러스터링 해쉬 테이블을 이용한 다차원 선박 USN 스트림 데이터의 효율적인 처리

(Efficient Processing of Multidimensional Vessel USN Stream Data using Clustering Hash Table)

송 병 호*, 오 일 환**, 이 성 로***

(Byoung-Ho Song, Il-Whan Oh, and Seong-Ro Lee)

요 약

디지털 선박에서는 선박 내의 각종 센서로부터 측정된 디지털 데이터에 대한 정확하고 에너지 효율적인 관리가 필요하다. 그러나, 센서 네트워크에서 대용량 스트림 데이터를 제한된 네트워크, 전력, 프로세서를 이용하여 모든 센서 데이터를 전송하고 분석하는 것은 어렵고 효율적이지 못하다. 그러므로, 연속적으로 입력되는 데이터를 사전에 분류하여 특성에 따라 선택적으로 데이터를 처리하는 데이터 분류 기법이 요구된다. 본 논문에서는 디지털 선박 내에 다수 개의 센서(온도, 습도, 조도, 음성 센서)를 배치하고 효율적인 입력 스트림 처리를 위해서 슬라이딩 윈도우 기반으로 다중 Support Vector Machine(SVM) 알고리즘을 이용하여 사전 분류(pre-clustering)한 후 요약된 정보를 해쉬 테이블로 관리하는 효율적인 처리 기법을 제안한다. 해쉬 테이블을 이용하여 다차원 스트림 데이터의 저장될 레코드 순서를 빠르게 찾아 저장 및 검색함으로서 처리 속도가 향상되고 메모리에 해쉬 테이블 만을 유지하면 되므로 메모리 사용량이 감소한다. 35,912개의 데이터 집합을 사용하여 실험한 결과 제안 기법의 정확도와 처리 성능이 향상되었다.

Abstract

Digital vessel have to accurate and efficient mange the digital data from various sensors in the digital vessel. But, In sensor network, it is difficult to transmit and analyze the entire stream data depending on limited networks, power and processor. Therefore it is suitable to use alternative stream data processing after classifying the continuous stream data. In this paper, We propose efficient processing method that arrange some sensors (temperature, humidity, lighting, voice) and process query based on sliding window for efficient input stream and pre-clustering using multiple Support Vector Machine(SVM) algorithm and manage hash table to summarized information. Processing performance improve as store and search and memory using hash table and usage reduced so maintain hash table in memory. We obtained to efficient result that accuracy rate and processing performance of proposal method using 35,912 data sets.

Keywords : Stream Data; USN; Digital Vessel; SVM; Hash Table.

* 정회원, 목포대학교 정보산업연구소
(Institute of Information Science and Engineering Research, Mokpo National University)
** 정회원, *** 정회원, 목포대학교 정보전자공학과
(Dept. of Information & Electronics, Mokpo National University)
※ 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 대학중점연구소지원사업으로 수행된 연구임(2009-0093828)
※ 본 논문은 2010학년도 목포대학교 학술(정책)연구비 지원에 의하여 연구되었음
접수일자: 2010년6월4일, 수정완료일: 2010년7월12일

Ⅰ. 서 론

디지털 선박에서 실시간으로 정보를 수집하기 위해서 많은 센서들이 필요하고 센서 네트워크를 통해서 수집된 데이터에 대한 효율적인 처리 방법이 요구되어 진다.

선박 Ubiquitous Sensor Network(USN)에서는 온도, 습도, 조도, 해양의 특성 등과 같은 것을 측정하는 여러 개의 센서들이 분산되어 있으며, 이러한 센서들에서 수집되는 데이터들이 계속해서 끊임없이 도착하는데, 이러한 데이터의 특징은 데이터 도착 속도가 가변적이며 데이터의 크기가 매우 커서 무한한 데이터 스트림을 형성하기 때문에 기존의 데이터베이스를 이용하는 것은 효율적이지 못하다.

이러한 스트림 데이터를 처리하기 위하여 새로운 스트림 데이터 모델을 제안하게 되었는데 이 모델에서는 데이터의 양이 아주 크거나 또는 무한하며 데이터에 대해 순차적인 액세스만이 가능하다고 가정하고 제한된 용량의 디스크나 메모리 등을 이용해서 질의 처리나 데이터 분석을 처리한다.

그동안 많은 관련 데이터 스트림 연구 분야에서는 주로 제한된 용량의 디스크나 메모리 등을 이용해서 질의 처리나 데이터 분석을 처리하는 분야에 대해 연구가 진행되어 왔다[1~2]. 그러나 스트림 데이터를 처리하기 위해서는 모든 입력 스트림을 하나하나 다 비교해야 하므로 입력 스트림 데이터 수만큼 순차적으로 비교를 해야 하는 문제점을 가지고 있다. 즉, 많은 양의 연속적인 데이터와 규칙을 처리하는 방법에 따라 스트림 데이터 처리의 효율성이 달라 질 수 있다. 따라서 다차원의 스트림 데이터를 효율적으로 처리하는 방법에 대해서 연구할 필요가 있다.

또한, 선박 USN 시스템에서는 하나의 데이터가 아니라 동시에 다차원 데이터를 처리해야하므로 데이타 스트림에 대한 보다 빠른 처리가 요구되는데 적은 용량의 메모리를 사용하는 요약 정보에 대한 해쉬 테이블을 이용하는 것이 이러한 환경에 보다 빠른 처리를 가능하게 해줄 것이다.

이에 본 논문에서는 사전 클러스터링을 통해 전체 데이터에 대한 정보를 요약한 후 해쉬 테이블에 저장하여 데이터를 처리한다. 입력 데이터 크기 n이 커지면 이들에 대한 다중 I/O 스캔으로 병목 현상이 일어나고, 비선형 시간 복잡도로 인한 처리비용이 급격히 증가된다는 제약점을 극복하기 위해 먼저 전체 데이터를 스캔해내는 사전-클러스터링(pre-clustering)단계를 수행한 후 가능한 메모리에 맞는 부클러스터에 대해 요약정보를 갖고 있는 해쉬 테이블을 검색함으로써 효율적인 데이터 처리를 수행한다.

사전 클러스터링은 입력되는 스트림 데이터가 다차원 비선형 데이터이므로 다중 Support Vector Machine(SVM) 알고리즘을 이용하여 모델링 하였고 비선형 데이터에 대한 클러스터링 알고리즘인 K-nearest neighbor(Knn) 알고리즘과 비교하여 분석하였다.

해쉬 테이블은 사전 클러스터링 후 요약된 정보의 인덱스와 레코드의 위치를 관리한다. 해쉬 테이블을 이용하여 다차원 데이터의 저장될 레코드 순서를 빠르게 찾아 저장함으로서 데이터 생성 속도가 향상된다. 또한 해쉬 테이블 만을 유지하면 되므로 메모리 사용량이 감소한다. 따라서 해쉬 테이블의 사용으로 데이터의 빠른 검색과 효율적인 데이터 처리가 가능하다.

본 논문은 다음과 같이 구성되었다. Ⅱ장에서는 관련 연구, Ⅲ장은 시스템 구성 및 설계, Ⅳ장에서는 실험 및 구현 결과를 보여주고 Ⅴ장에서는 결론과 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

1. 스트림 데이터 관리 시스템

스트림 데이터 처리에 있어서 중요한 부분은 센서 네트워크 환경에 흩어져 있는 센싱 정보들에 대하여 센서 노드들의 제한된 리소스들을 고려하면서 효율적으로 센싱 데이터의 질의를 만족시켜 주는 것이다[3]. 즉, 각 센서 노드들에서의 에너지 소모율을 최소화 시키면서 질의에 대한 정확성 및 신속성을 최대화 시킬 수 있는 질의 처리기(Query Processor)를 만드는 것이다. 그림 1은 데이터 스트림 환경에서의 질의 처리에 대한 설명을

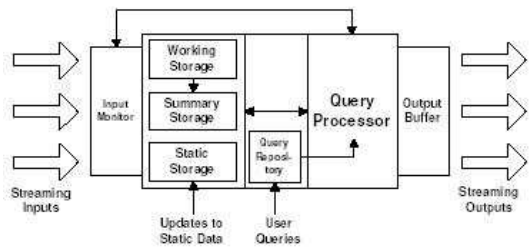

그림 1. 스트림 데이터 관리 시스템
Fig. 1. Data Stream Management System.

단계적으로 나타낸 데이터 스트림 관리 시스템(DSMS)의 구조도이다. DSMS는 다중 데이터 스트림에 대하여 다중 연속 질의 처리를 수행한다. 여기서 입력 데이터의 과부하로 인하여 시스템의 처리 용량을 초과하는 현상이 발생 할 수도 있다.

그림 1에서 데이터의 저장은 세 부분으로 나눌 수 있다. 첫번째, 임시 작업 저장소(temporary working storage)와 두 번째, 스트림 시놉시스 처리를 위한 요약 저장소(summary storage) 세 번째, 메타 데이터 처리를 위한 정적 저장소(static storage)이다. 실행되는 질의는 질의 저장소(query repository)에 저장되고 질의 처리기(query processor)는 입력 데이터양의 상황에 따라 질의 처리에 대한 최적화 작업을 실행한다.

2. 사전 클러스터링

클러스터링은 사전 정보 없이 데이터의 분석을 통한 정보를 취득해야 하는 문제이므로 탐험적인 데이터 분석(exploratory data analysis)으로 불리기도 한다. 클러스터링 문제는 데이터의 패턴을 분석하여 패턴간의 유사도를 측정하고 그룹을 형성하는 것이다. 실세계의 문제를 다루는 데이터에서 가지는 고차원의 특징은 동일한 그룹의 특징을 가지지 않는다. 클러스터링의 결과는 다음 단계의 연구 진행을 위한 기본 자료로서 이용이 되기 때문에 보다 의미 있고 안정적인 정보를 얻을 수 있는 방법에 대한 연구가 진행되고 있다. 본 실험에 사용된 데이터는 선박에 설치된 온도, 습도, 음성, 조도 센서 값으로써 연속적이지만 데이터가 정규분포를 따르지 않는 비선형 데이터이다. 본 논문에서는 다중 SVM 알고리즘을 이용하여 사전 클러스터링을 수행하고 비선형 데이터 분류 문제에 대표적으로 많이 사용되어온 기법인 Knn을 적용한 모델과 성능을 비교 하였다.

2.1 다중 SVM 알고리즘

학습 이론에 기반을 둔 SVM은 주어진 문제를 항상 전역적 최적해가 보장되는 convex quadratic problem으로 변환하여 해를 구하기 때문에 패턴인식 분야에서 매우 우수한 성능을 보여주고 있다[4]. SVM은 이진 분류를 위해 개발되었기 때문에 실제 환경에서 여러 클래스를 가지는 문제들을 해결하기에는 많은 어려움이 있다. 때문에 이러한 문제점들을 해결하기 위해 많은 전략들이 제시되었는데 그 중에 대표적인 것이 One-against-all 기법과 One-against-one 기법이다. One-against-all 기법은 입력된 클래스의 개수만큼 SVM을 학습하는 방법이다. k개의 클래스가 입력되었을 때 k-1개의 SVM이 필요하며, 클래스 수가 늘어날수록 성능이 저하되고 각각의 클래스에 속하지 않는 입력이 주어지면 의미 없는 출력을 하게 된다. One-against-one 기법은 k개의 클래스가 입력되었을 때 k(k-1)/2개의 SVM으로 구성되며 각각의 학습데이터는 두개의 소속을 나타내는 데이터로만 구성된다. 각 학습에 사용되는 학습 데이터의 수가 적기 때문에 학습이 빠르다[5~6].

2.2 Knn 알고리즘

본 논문에서는 다중 SVM 알고리즘의 유용성 확인을 위해서 비선형 데이터 분류 문제에 대표적으로 많이 사용되어온 기법인 Knn을 적용한 모델과 성능을 비교 하였다. KNN 알고리즘은 새로운 데이터를 분류함에 있어 입력 데이터로부터 유사성을 계산하여 유사성이 가장 높은 데이터를 선택한다[7]. 그림 2는 KNN 알고리즘의 수행 과정을 나타낸다. KNN 알고리즘에 대한 선행 연구를 살펴보면 KNN 알고리즘은 최초로 Cover 와 Har에 의하여 각각 독립적으로 제안되었다[8]. 이 후 Smith와 Medin 등에 의하여 KNN 알고리즘은 논리학적으로 그 타당성을 인정받았지만 실제 알고리즘을 위한 모델은 개발되지 않은 상태였다[9]. 이후 Aha, Kibler and Albert에 의하여 몇 개의 개체중심 학습(instance- ased learning,IBL) 알고리즘이 개발되었다[10].

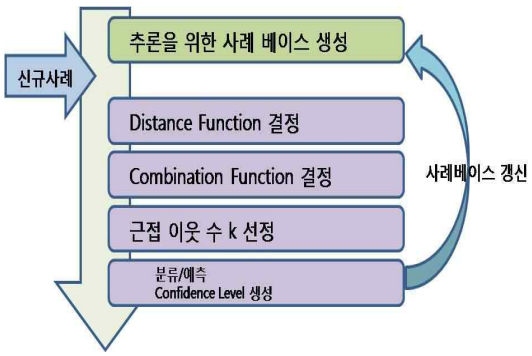

그림 2. kNN 알고리즘 수행 과정
Fig. 2. kNN Algorithm Execute Process.

Ⅲ. 시스템 구성 및 설계

본 논문에서 프로세서 보드는 Telos 플랫폼 계열이며, MSP430의 MCU와 CC2420 Radio Chip을 사용한다. 그리고 온도, 습도, 음성, 조도 센서가 통합된 센서 모듈

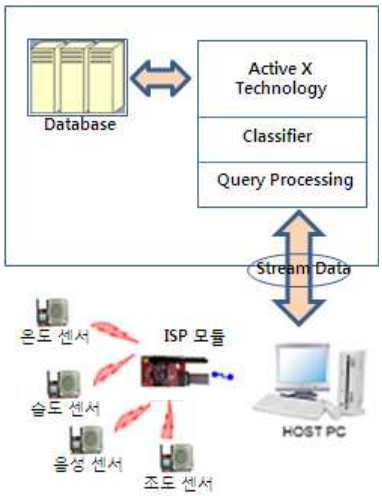

그림 3. 시스템 구성도
Fig. 3. System block diagram.

을 사용한다. 그리고 습도, 온도, 조도, 음성 센서가 통합된 센서 모듈을 사용하며 싱크 노드는 1개를 포함 총 10개의 센서 노드를 사용하였다. 그림 3은 제안한 시스템의 전체 구성도 이다.

시스템 구성도를 보면 각각의 센서로부터 입력 스트림 데이터를 처리하기 위해서 전체 데이터베이스를 한 번 스캔한 후 사전-클러스터링을 수행한다. 가능한 메모리에 맞는 분류된 요약 정보에 따른 레코드 순서와 위치값을 갖는 해쉬 테이블을 검색함으로써 효율적인 데이터 처리를 수행한다.

1. 센서 처리 및 데이터 구조

구현된 시스템은 스트림 데이터(온도, 습도, 조도, 음성)를 획득하기 위해 다수 개의 센서를 사용한다. 분석에 사용될 데이터는 동일한 환경에서의 데이터이므로 하나의 패킷으로 묶어서 전송한다. 각각의 패킷에 담게 되면 추가적인 트래픽 발생 및 에너지 소모가 일어남으로 단일 패킷으로 처리하여 질의를 처리한다. 그림 4는 센싱된 데이터의 패킷 구성을 나타낸다. 패킷의 총길이는 36바이트이며, 고정 헤더는 10바이트, 센서 노드 ID 및 채널은 6바이트, 버퍼 20바이트 부분으로 구성된다. 이중에서 버퍼는 앞에서부터 8바이트를 각각 2바이트씩 헥사값으로 습도, 온도, 조도 순으로 실제 센싱값이 들

헤더(10)	센서노드ID 및 채널(10)	습도(2)	온도(2)	음성(2)	조도(2)	(12)

그림 4. 패킷구성
Fig. 4. Packet Configuration.

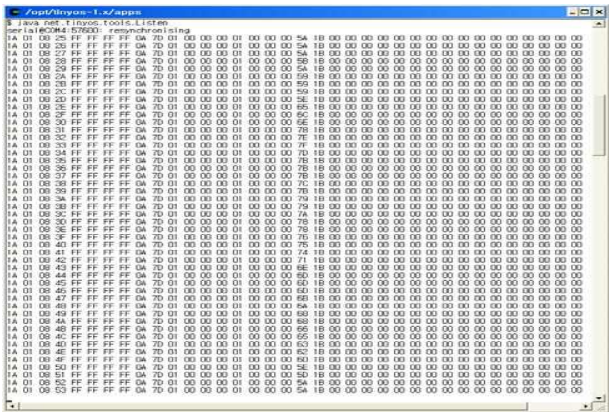

그림 5. 센싱된 데이터의 구조
Fig. 5. Structure of sensed data.

어오도록 설계하였다.

데이터의 구조는 그림 5와 같다. 각각의 묶음은 1바이트를 나타내고 있으며, 좌측에서부터 7, 8번째 값은 통신 방식, 15, 16번째 값은 채널을 알려준다. 17~24 번째는 습도, 온도, 조도값을 나타낸다.

2. 다중 SVM 사전 클러스터링

본 논문에서 적용한 알고리즘은 다중 SVM 분류로서 입력된 데이터를 특정 범주로 분류해주는 역할을 한다. 특정 범주에 해당하는 요약 정보를 해쉬 테이블에 저장하여 데이터베이스의 효율성을 높이고자 한다. SVM 분류는 두 그룹을 잘 분리시키는 분류 초평면을 찾는 방법이다[11]. SVM은 기존의 선형 분류방법보다 확장성이 좋고 학습 시마다 성능이 달라지는 신경망 분류방법과는 달리 항상 일정하게 우수한 성능을 보여준다[12]. SVM의 기본 원리는 선형 분리가 가능한 문제에서부터 출발한다. d-차원에서 입력데이터 X_i가 주어졌을 때 학습데이터의 출력으로 -1과 +1처럼 이진 값으로 구분되는 문제를 고려한다. 두 집합을 분류하기 위한 모델을 정의하기 위하여 그림 6과 같은 선형 식별함수인 초평면 (hyperplane)을 정의할 수 있다. 여기에서 Support Vector란 분류 규칙을 결정 짓는 경계와 밀접한 연관이 있는 표본을 의미한다. 본 실험 데이터처럼 선형 분리가 불가능한 데이터인 경우에는 비선형 사상 Φ를 이용하여 입력 벡터의 차원보다 높은 선형분류가 가능한 차원으로 변형한 후 선형 분류를 하게 된다. 비선형 사상은 kernel 함수를 이용하여 N차원의 입력공간의 데이터를 고차원의 특징 공간(Q차원)으로 변환함으로써 선형적으로 구별할 수 있으며 식(1)은 kernel함수와 결정함

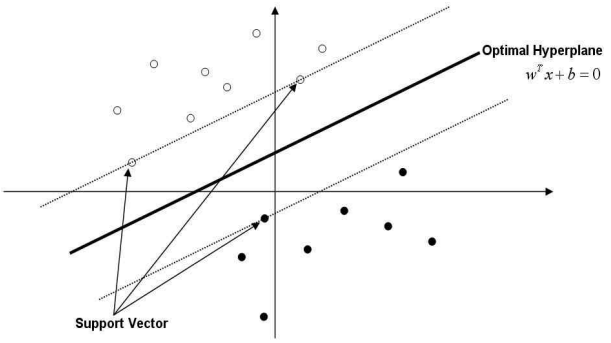

그림 6. 최적화 초평면과 서포트 벡터
Fig. 6. Optimal Hyperplane and Support Vector.

수이다.

$$K(x,y)=\phi(x)\cdot\phi(y)$$

$$f(x)=\sum_{i=1}^{n}a_i y_i K(x,x_i)+b \quad (1)$$

SVM은 단순히 분류 평면을 찾는다거나 표본 에러를 최소화하는 작업을 하는 것이 아니라 분류 여백(Margin)을 최대화함으로써 학습데이터가 아닌 새로운 데이터에 대해서도 올바르게 분류할 가능성이 높다.

SVM은 이진 분류를 위해 개발되었기 때문에 실제 환경에서 여러 클래스를 가지는 문제들을 해결하기에는 많은 어려움이 있다. 때문에 이러한 문제점들을 해결하기 위해 One-against-all기법과 One-against-one 기법이 제시 되었다.

One-against-one 기법은 k개의 클래스가 입력되었을 때 k(k-1)/2개의 SVM으로 구성되며 각각의 학습데이터는 두 개의 소속을 나타내는 데이터로만 구성되고 각 학습에 사용되는 학습 데이터의 수가 적기 때문에 학습 속도가 빠른 것으로 나타났다. 본 논문에서는 학습의 성능 향상을 위해서 One- against-one 기법을 이용하여 실험을 수행하였다.

Algorithm : SVM
학습을 위한 데이터의 개수 : N
Inputs: sample x to classify 데이터 셋 : I_i
I_{i1} : 온도, I_{i2} : 조도, I_{i3} : 습도, I_{i4} : 음성
Output: decision $y \in \{-1, 1\}$
Classify using SVM, get the result in the form of a real number.

그림 7. 알고리즘 구성
Fig. 7. Algorithm Configuration.

본 논문에서는 그림 7과 같이 SVM 알고리즘을 구성하였다.

제안한 SVM 알고리즘은 주어진 트레이닝 데이타의 각 Feature에 대해 최대 여분(margin)이 많게 생성되는 hyper-plane을 생성한다. 테스트 단계에서는 트레이닝 단계에서 생성된 hyperplane에 의해 분할된 다차원 공간에 매핑하여 새로운 데이타를 분류한다.

3. 클러스터링 주변점 결정

다차원 스트림 데이터의 클러스터링에서 전체적인 클러스터링 패턴에 대해 중요하지 않은 주변점들이 있을 수 있다. 만일 클러스터링 과정이 종료된 후에 생성된 클러스터 중에서 극히 소수의 점만을 갖는 클러스터가 있다면 그 클러스터에 있는 점들은 주변점으로 간주된다. 예를 들어, 클러스터링 후에 두세 개의 점만을 갖는 클러스터가 다른 클러스터와 떨어진 위치에 존재한다면 그 클러스터는 주변점을 포함하고 있을 확률이 높게 된다. 특정 클러스터가 주변점을 포함하고 있는 클러스터인 지를 판정하기 위하여 몇 개의 점을 기준으로 할 것인가는 응용 분야에 따라 경험적으로 결정될 수 있다. 주변점 값이 너무 작으면 중요하지 않은 클러스터가 인덱스에 포함될 수 있으며, 이는 메모리의 효율을 저하시킨다. 본 실험에서는 극히 작은 값들은 의미있는 정보가 아니므로 인덱스에 포함되지 않고 삭제한다. 반면에 데이터 값이 너무 큰 경우는 의미있는 이벤트 발생으로 인식하여 데이터베이스에 저장하여 모니터링 시스템에 전송한다.

4. 자료의 표준화

질의 처리 후 센싱된 데이터 값을 표준화된 데이터로 변환하기 위하여 데이터 변환 알고리즘을 통해 실제 데이터 값으로 변환한다. 온도센서 모듈은 스플릿이라는 한 단계의 계산 과정을 더 거쳐 센싱값을 출력한다. 실제온도는 그림 6의 변환 식에 대입하여 계산된다.

습도센서는 온도의 변화에 따른 습도, 즉 상대 습도를 계산하기 위한 센서로서 이는 습기의 흡수 또는 흡착에 따른 전기 저항의 변화 특성을 이슬점 또는 서릿점 때의 응결 상태 감지 원리를 이용한다. 상대 습도는

* 실제온도 = 실제 센싱값 * 0.01 - 40

그림 8. 온도의 실제값 변환식
Fig. 8. Conversion Formula of Temperature

현재의 수증기량과 그 온도에 있어서의 포화수증기량의 비로 나타내며 습도의 변화는 주로 기온 변화에 의하여 발생된다. 조도 센서는 빛의 밝기에 따라 출력 저항의 값이 변하는 성질을 이용한 센서로서 출력 전압을 A/D 컨버팅한 후 그 값을 읽어오는 방식이다. 센서 출력 전압을 256등분하여 읽어올 수 있으므로 외부 회로의 구성없이 디지털 출력이 가능하다.

5. 해쉬 테이블 구성

해쉬 테이블은 입력 스트림 데이터에 대한 요약정보의 저장할 레코드의 순서를 계산하기 위한 것으로 각 필드마다 중복 없는 레코드 값을 이용해 각각의 독립적인 해쉬 테이블로 만든다. 해쉬 테이블에 입력된 레코드 값에는 레코드 순서 계산을 위한 레벨 값으로 순차적인 값이 할당되며, 이러한 레벨 값을 이용해서 저장 레코드 순서를 계산한다. 그림 9의 첫 번째 테이블은 해쉬 테이블에 '0'부터 순차적인 값을 할당하여, n까지 레벨 값이 저장된 것을 보여준다. 또한 레코드에 대한 레벨 값을 할당한 후에는 'Total' 레코드의 레벨 값을 추가한다. 해쉬 테이블에 'Total'레코드를 포함하는 이유는 다차원 집계 연산에서 일반화에 필요하기 때문이다. 이러한 해쉬 테이블은 연산 결과 순서를 맞추기 위해 필드를 이름순으로 정렬해야한다. 이름순으로 정렬된 필드는 해쉬 테이블 생성 시 중복되지 않는 레코드 값을 구분하는 비용을 줄여준다.

해쉬 테이블은 저장하기 위한 레코드 순서를 계산하는데 있어, 각 레코드의 값을 빠르게 찾아 계산할 수 있다. 해쉬 테이블이 생성되면, 다음으로 해쉬 테이블의 레벨 값으로 계산된 레코드 순서에 대해 실제 레코드 저장 위치를 관리한다. 그림 9와 같이 테이블에 저장될 순서와 임시 레코드의 주소를 저장하는 구조로 구성된다. 그림 9의 해쉬 테이블을 보면 특정 레코드에 대해 저장할 레코드 순서 값에 실제 저장위치가 매핑되어 있는 것을 볼 수 있다. 레코드의 실제 저장은 계산된 레코드들 순서로 하기 때문에 해쉬 테이블의 순서 값으로 결과의 순서를 보장한다. 따라서 해쉬 테이블은 저장 공간을 효율적으로 사용할 수 있다.

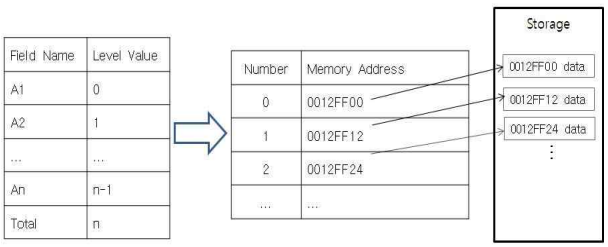

그림 9. 해쉬 테이블의 구성
Fig. 9. Hash Table Composition.

Ⅳ. 실험 및 구현 결과

본 논문에서 실험을 위해 사용된 프로세서 보드는 Telos 플랫폼 계열이며, MSP430의 MCU와 CC2420 Radio Chip을 이용하여 실험을 수행한다. 1개의 Sink 노드와 9개의 중간노드 총 10개의 노드를 사용하여 5초마다 한 번씩 온도, 습도, 조도, 음성 값에 대하여 해쉬 테이블을 통해서 데이터베이스에 저장한다. 사전-클러스터링을 수행하기 위해서 총 35,912개의 데이터 집합을 사용하여 모델링하였고, 다중 SVM 알고리즘의 유용성 확인을 위해서 비선형 데이터 분류 문제에 대표적으로 많이 사용되어온 기법인 Knn을 적용한 모델과 성능을 비교 하였다. 실험 데이터는 선형적인 관계가 아닌 실세계를 반영한 불규칙한 데이터를 사용했기 때문에 본 실험에서는 슬라이딩 윈도우의 크기 변화에 따른 오차율을 측정했다. 실험의 오차율 측정을 위해 식 (1)과 같이 RMSE를 사용하였다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2} \quad (1)$$

표 1은 윈도우 크기에 따른 오차율 측정 결과이다. 측정 결과 윈도우의 크기가 더 클수록 낮은 오차율을 보였다.

표 2는 35,912개의 데이터 집합을 사용하여 튜플의 개수에 따라 윈도우의 크기를 1000, 3000, 5000, 7000, 10,000개로 분할하여 정확도를 측정한 결과이다. 실험한 결과 윈도우의 크기를 5000으로 분할했을 때 정확도가 0.882로 가장 높았고 평균 정확도는 0.863으로 SVM 알

표 1. 윈도우 크기 변화에 따른 오차율 측정 결과
Table 1. Measure Result of Error Rate by Window Size Change.

윈도우 크기	오차율
1000	2.43%
3000	2.32%
5000	2.03%
7000	1.98%
10000	1.76%

표 2. 윈도우 크기 변화에 따른 정확도 측정 결과
Table 2. Measure Result of Accuracy Rate by Window Size Change.

윈도우 크기	Knn	SVM
1000	0.817	0.842
3000	0.822	0.851
5000	0.825	0.882
7000	0.839	0.863
10000	0.854	0.877

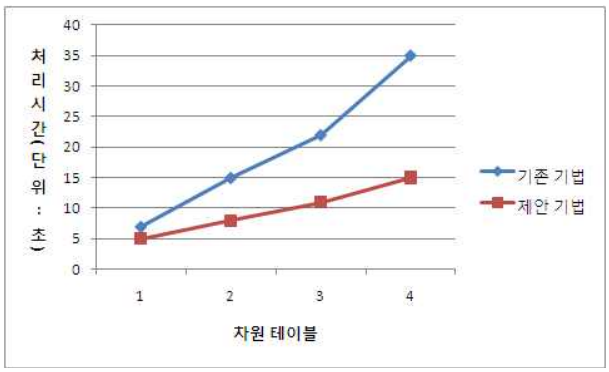

그림 10. 처리 시간 측정 결과
Fig. 10. Result of access time.

고리즘의 성능이 더 좋은 것으로 나타났다.

또한, 본 논문에서의 사용 데이터는 온도, 조도, 습도, 음성 4차원의 데이터 이므로, 처리 성능 평가를 위해 테스트셋을 기반으로 1개부터 4개까지의 차원 테이블을 이용한 처리 시간의 변화를 측정하였다. 그림 10을 보면, 제안 기법을 사용하면 49.3%의 성능이 향상된 것을 볼 수 있다.

본 논문에서 처리시간에 대한 평가는 크게 기존 기법과 제안 기법, 두 가지 경우로 나누어 비교 평가하였다. 여기에서 기존 기법이란 사전-클러스터링을 거치지 않고 센서를 통해 입력되는 모든 튜플을 적재하여 다차원 데이터를 처리하는 것을 의미한다.

메모리에 모든 튜플을 적재하여 연산을 수행하면 차원 테이블의 개수가 증가할수록 많은 메모리를 요구하여 트리 구축비용과 검색 비용이 증가한다. 제안 기법의 경우 사전-클러스터링 후 해쉬 테이블만 메모리에 유지하면 되므로 메모리 사용량이 적고 빠르게 연산이 가능함을 알 수 있다.

본 실험에 사용된 데이터는 델파이를 이용하여 디지털 선박 모니터링 시스템을 구현하였다. 표 3은 시스템 구현 환경을 나타낸다.

구현 결과는 센싱된 데이터의 분류 결과와 수치데이

표 3. 시스템 구현 환경
Table 3. system implement environment.

	항목	종류
소프트웨어	운영체제	Windows XP
	사용언어	Delphi
	DBMS	MSSQL
하드웨어	DB서버	Sqlserver 2000
	서버	Pentium(R) Quad Core 2.66
	메인메모리 용량	4GHz DDR3 PC3-12800

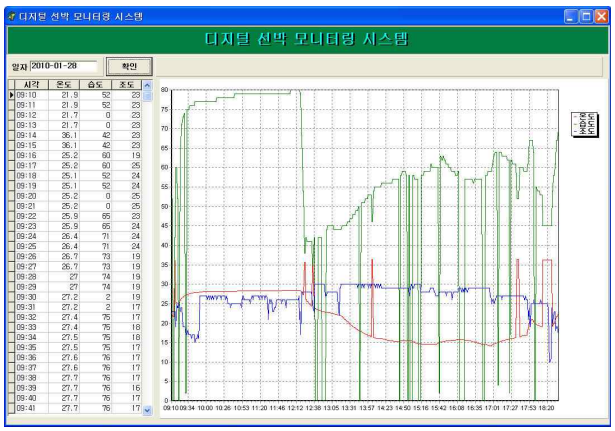

그림 11. 시스템 구현 결과
Fig. 11. Result of System Implementation.

터를 시간별, 일자별로 모니터링 할 수 있는 항목과 각각 온도, 습도, 조도, 음성의 변화를 볼 수 있는 그래프 항목으로 나뉘어진다. 그림 11은 시스템 구현 결과 화면이다.

V. 결 론

디지털 선박에서는 선박 내의 각종 센서로부터 측정된 디지털 데이터에 대한 정확하고 에너지 효율적인 관리가 필요하다. 이에 따라 본 논문에서는 디지털 선박 내에 다수 개의 센서(온도, 습도, 조도, 음성 센서)를 배치하고 효율적인 입력 스트림 처리를 위해서 슬라이딩 윈도우 기반으로 다중 SVM 알고리즘을 이용하여 사전 분류(pre-clustering)한 후 요약된 정보를 해쉬 테이블로 관리하는 효율적인 처리 기법을 제안한다. 유효한 데이터는 디지털 선박 모니터링 시스템에 이용하였다. 35,912개의 데이터 집합을 사용하여 실험한 결과 윈도우의 크기를 5000으로 분할했을 때 정확도가 0.882로 가장 높았고 평균 정확도는 0.863으로 SVM 알고리즘

의 성능이 더 좋은 것으로 나타났다. 또한 처리 성능 평가를 위해 1개부터 4개까지의 차원 테이블을 이용한 처리 시간의 변화를 측정한 결과 49.3%의 성능이 향상 되었다. 향후 연구 방향으로는 처리 시간을 고려한 보다 효율적인 알고리즘을 개발하고 시간의 흐름에 영향을 받는 데이터들의 처리를 위해 시간 기반 슬라이딩 윈도우 질의 처리에 대해 연구한다.

참 고 문 헌

[1] L. Golab and M. T. Ozsu, “Issues in Data Stream Management,” SIGMOD Record, vol.32, no. 2, June 2003.
[2] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, “Models and Issues in Data Stream Systems,” In Proc. of ACM SIGACT-SIGMOD-SIGART Sym. on Principles of Database Systems, pp. 1-16, Wisconsin, USA, June 2002.
[3] 이수안외 3명 , “유비쿼터스 센서 네트워크에서 스트림 데이터를 효율적으로 관리하는 저장 관리자 구현”, 전자공학회논문지, 제46권 CI편, 제3호, 24-33쪽, 2009년 5월.
[4] Burges, C., “A Tutorial on Support Vector Machines for Pattern Recognition”, Data Mining and Knowledge Discovery, 1998.
[5] Hyunchul Ahn, Kyoung-jae Kim, “Multiclass SVM, Model with Order Information”, International Journal of Fuzzy Logic and Intelligent Systems, Vol.6, No.4, pp.331-334, December 2006.
[6] 고재필, “Support Vector Machines을 이용한 다중 클래스 문제 해결”, 정보과학회논문지:소프트웨어 및응용, 제32권, 제12호, pp.1260-1270, 2005.12
[7] 이희성외 2명 , “KNN 규칙과 새로운 특징 가중치 알고리즘을 결합한 패턴 인식 시스템”, 전자공학회 논문지, 제42권 CI편, 제4호, 43-50쪽, 2005년 7월.
[8] T. M. Cover and P. E. Hart, “Nearest Neighbor Pattern Classification,” IEEE Transactions on Information Theory, Vol. 13, 1967.
[9] E. E. Smith and D. L. Medin, “Categories and Concepts,” Cambridge, MA: Harvard University Press, 1981.
[10] D. Aha, D. Kibler and M. Albert, “Instance-based Learning Algorithms,” Machine Learning, 6(1) pp.37-66, 1991.
[11] Zhuang, D., Zhang, B., Yang, Q., Yan, J., Chen, Z., & Chen, Y. 2005. “Efficient Text Classification by Weighted Proximal SVM.” Proceedings of the Fifth IEEE International Conference on Data Mining: 538-545.
[12] Y. Liu, R. Wang, H. Huang, Y. Zeng, and H. He,“Applying support vector machine to P2P traffic identification with smooth processing,” IEEE Int. Conf. on Signal Processing, Vol. 3, pp. 16-20, 2006.

저 자 소 개

송 병 호(정회원)

1998년 조선대학교 전산통계 학사 졸업.
2000년 조선대학교 전산통계 석사 졸업.
2008년 조선대학교 전산통계 박사 졸업.
2008년～2009년 Murdoch University Post.Doc.
2009년～현재 목포대학교 정보산업중점연구소 연구전임교원
<주관심분야 : 인공지능, USN, 신호처리>

이 성 로(정회원)-교신저자

1987년 고려대학교 전자공학 학사 졸업.
1990년 한국과학기술원 전기및 전자공학 석사 졸업.
1990년 한국과학기술원 전기및 전자공학 박사 졸업.
2009년～현재 목포대학교 정보전자공학과 교수.
<주관심분야 : 디지털통신, 위성통신, 해양텔레매틱스, USN>

오 일 환(정회원)

1982년 연세대학교 전기학과 학사 졸업
1985년 12월 텍사스대학교 전기 전자공학과 석사
1988년 12월 Rhode Island대학교 전기전자공학과 박사
1990년 3월~현재 목포대학교 정보전자공학과 교수
<관심분야> 광통신, 센서네트워크