SVM을 적용한 선박 스트림 데이터 처리 기법

양진호*, 프라시스 포우델*, 시리 크리스나 아차레*, 서군 수베디*, 정민아**, 이성로*
*목포대학교 전자공학과
**목포대학교 컴퓨터공학과
*didwlsgh789@naver.com

Ship Stream Data Processing Techniques To Which The SVM

Yang Jin Ho*, Prasis Poudel*, Shree Krishna Acharya*, Sagun Subedi*, Min-A Jeong**, Seong-Ro Lee*
*Dept of Electronics Engineering, Mokpo National University
*Dept of Computer Engineering, Mokpo University

요 약

디지털 선박에서는 선박 내의 각종 센서로부터 측정된 디지털 데이터에 대한 정확하고 에너지 효율적인 관리가 필요하다. 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 집함을 사용하여 실험한 결과 제안 기법의 정확도와 처리 성능이 향상되었다.

1. 서론

본 논문에서는 사전 클러스터링을 통해 전체 데이터에 대한 정보를 요약 한 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.

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 테이블만을 유지하면 되므로 메모리 사용량이 감소한다. 따라서 해쉬 테이블의 사용으로 데이터의 빠른 검색과 효율적인 데이터 처리가 가능하다.

2. 관련 연구

1. 다중 SVM 알고리즘

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 것들이 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 의미없는 출력을 하게 된다. One-against-one 기법은 k개의 클래스가 입력되었을 때 k(k-1)/2개의 SVM으로 구성되며 각각의 학습데이터는 두 개의 소속을 나타내는 데이터로만 구성된다. 각 학습에 사용되는 학습 데이터의 수가 적기 때문에 학습이 빠르다[2][3].

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. SVM분류는 두 그룹을 잘 분리시키는 분류 초평면을 찾는 방법이다[4]. SVM은 기본 원리는 선형 분리가 가능한 문제에서부터 출발한다. $d-$차원에서 입력데이터 X_i가 주어졌을 때 학습데이터의 출력으로 -1과 +1처럼 이진 값으로 구분되는 문제를 고려한다. 본 실험 데이터처럼 선형 분리가 불가능한 데이터인 경우에는 비선형 사상 Φ를 이용하여 입

력 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 결정함수이다.

$$K(x,y) = \varnothing(x) \bullet \varnothing(y)$$

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

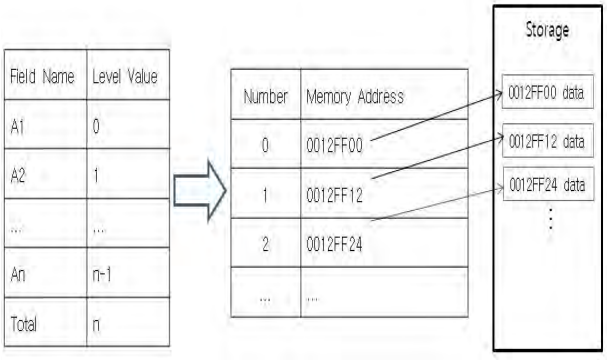
본 논문에서는 그림1과 같이 SVM 알고리즘을 구성하였다.

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 : decisin $y \in \{-1,1\}$
Classify using SVM, get the result in the form of a number.

(그림 1) 알고리즘 구성

3. 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 수에는 '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 생성시 중복되지 않는 레코드 값을 구분하는 비용을 줄여준다.



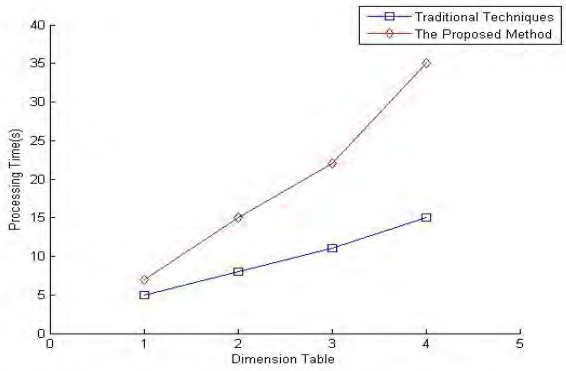
(그림 2) 해쉬 테이블의 구성

3. 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 5초 마다 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 (2)와 같이 RMSE를 사용하였다.

$$SE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2} \qquad (2)$$

본 논문에서의 사용 데이터는 온도, 조도, 습도, 음성 4차원의 데이터 이므로, 처리 성능평가를 위해 테스트셋을 기반으로 1개부터 4개까지의 차원테이블을 이용한 처리 시간의 변화를 측정하였다. 그림 3을 보면, 제안 기법을 사용하면 49.3%의 성능이 향상된 것을 볼 수 있다.



(그림 3) 처리 시간 측정 결과

4. 결 론

디지털 선박에서는 선박 내의 각종 센서로부터 측정된 디지털 데이터에 대한 정확하고 에너지 효율적인 관리가 필요하다. 이에 따라서 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 디

지털 선박 모니터링 시스템에 이용하였다. 35,912개의 데이터 집합을 사용하여 실험한 결과 윈도우 크기를 5000으로 분할했을 때 정확도가 0.882로 가장 높았고 평균 정확도는 0.863으로 SVM 알고리즘의 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 사간을 고려한 보다 효율적인 알고리즘을 개발하고 시간의 흐름에 영향을 받는 데이터들의 처리를 위해 시간 기반 슬라이딩 윈도우 질의 처리에 대해 연구한다.

ACKNOWLEDGMENT

본 연구는 2015년도 정부 (교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2009-0093828)와 미래창조과학부 및 정보통신기술진흥센터의 ICT융합 고급인력과정지원사업(IITP-2015-H8601-15-1006)의 연구결과로 수행되었음.

참고문헌

[1] Davies R. W.″ The Data Encryption standard in perspective,″Computer Security and the Data Encryption Standard, pp. 129-132.

[2] Miles E. Smid, ″From DES to AES,″ 2000, (http//www.nist.gov/aes).

[3] Shamir, A. ″On the security of DES,″ Advances in Cryptology, Proc.Crypto ′85, pp. 280-285, Aug. 1985.

[4] NIST, ″Announcing the Advanced Encryption Standard(AES),″FIPS PUB ZZZ, 2001, (http://www.nist.gov/aes).

[5] Daemen, J., and Rijmen, V. “AES Proposal: Rijndael, Version2.,” Submission to NIST, March 1999.