



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2008년01월03일
(11) 등록번호 10-0791037
(24) 등록일자 2007년12월26일

(51) Int. Cl.

G06F 17/40 (2006.01)

(21) 출원번호 10-2007-0038262

(22) 출원일자 2007년04월19일

심사청구일자 2007년04월19일

(56) 선행기술조사문헌

KR1020010095973 A

(뒷면에 계속)

(73) 특허권자

인하대학교 산학협력단

인천 남구 용현동 253 인하대학교

(72) 발명자

이중식

인천 남구 용현3동 인하대학교 하이테크관 1409호

조규철

인천 남구 용현3동 인하대학교 하이테크관 1407호

(74) 대리인

이원희

전체 청구항 수 : 총 11 항

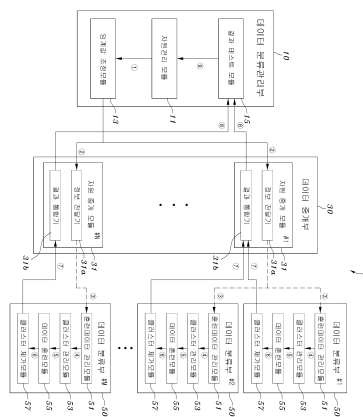
심사관 : 손준영

(54) 그리드 기반 대용량 데이터 분류 장치 및 방법

(57) 요약

본 발명은 그리드 기반 대용량 데이터 분류 장치 및 방법에 관한 것으로서, 대용량 데이터가 분류되도록 클러스터를 생성시켜 훈련하고, 생성된 클러스터 중 데이터를 대표할 수 없거나 해당 클러스터에 소속된 데이터가 일정 수 이하인 비중이 작은 이상치 클러스터를 제거함으로써, 이에 따른 클러스터의 감소는 데이터 분류의 처리 시간을 단축시키는 효과가 있으며, 분류가 완료된 결과를 기반으로 최적의 임계값을 유도하도록 조정함으로써, 최적의 인식률에 근접할 수 있으며, 동적으로 그리드 자원에 데이터 분류 작업을 할당함으로써, 대용량의 데이터 운용 및 처리가 가능해짐과 동시에 중앙 서버의 병목 현상을 방지하여 안정성을 증가시킬 수 있는 그리드 기반 대용량 데이터 분류 장치 및 방법을 제공하기 위한 것으로서, 그 기술적 구성은 대용량의 데이터 분류에 사용될 이용가능한 그리드 자원을 분류하는 자원 관리 모듈과, 상기 데이터 분류를 위한 임계값을 초기화하여 데이터를 분류하되, 분류가 완료된 데이터의 결과에 따라 임계값을 조정하는 임계값 조정 모듈과, 상기 임계값 조정 모듈로 분류가 완료된 데이터의 결과를 검사하여 결과를 전달하는 결과 테스트 모듈을 포함하는 데이터 분류 관리부; 상기 그리드 자원 및 임계값을 전달하고, 분류가 완료된 데이터를 병합하여 상기 결과 테스트 모듈로 전송하는 다수의 자원 중개 모듈을 포함하는 데이터 중개부; 및 상기 그리드 자원 및 임계값을 수신하여 대용량 데이터를 클러스터로 분류하여 훈련시키되, 이상치 클러스터는 제거하여 상기 데이터 중개부로 전송하는 다수의 데이터 분류부;를 포함한다.

대표도 - 도1



(56) 선행기술조사문헌

KR1020040052886 A

KR1020050055240 A

KR1020050078395 A

KR1020060119115 A

특허청구의 범위

청구항 1

대용량의 데이터 분류에 사용될 이용가능한 그리드 자원을 분류하는 자원 관리 모듈과, 상기 데이터 분류를 위한 임계값을 초기화하여 데이터를 분류하되, 분류가 완료된 데이터의 결과에 따라 임계값을 조정하는 임계값 조정 모듈과, 상기 임계값 조정 모듈로 분류가 완료된 데이터의 결과를 검사하여 결과를 전달하는 결과 테스트 모듈을 포함하는 데이터 분류 관리부;

상기 그리드 자원 및 임계값을 전달하고, 분류가 완료된 데이터를 병합하여 상기 결과 테스트 모듈로 전송하는 다수의 자원 중개 모듈을 포함하는 데이터 중개부; 및

상기 그리드 자원 및 임계값을 수신하여 대용량 데이터를 클러스터로 분류하여 훈련시키되, 이상치 클러스터는 제거하여 상기 데이터 중개부로 전송하는 다수의 데이터 분류부;

를 포함하는 그리드 기반 대용량 데이터 분류 장치.

청구항 2

제1항에 있어서,

상기 자원 관리 모듈은 이용 가능한 그리드 자원을 수집하여 저장 공간 및 작업 처리량에 따라 분류하는 것을 특징으로 하는 그리드 기반 대용량 데이터 분류 장치.

청구항 3

제1항에 있어서,

상기 이상치 클러스터는 생성된 클러스터 중에서 데이터를 대표할 수 없거나 또는 포함된 데이터의 비중이 크지 않은 클러스터인 것을 특징으로 하는 그리드 기반 대용량 데이터 분류 장치.

청구항 4

제1항에 있어서,

상기 자원 중개 모듈은

상기 그리드 자원 및 임계값을 데이터 분류부에 전달하는 정보 전달기; 및

다수개의 데이터 분류부로부터 출력된 데이터 훈련, 분류 결과 및 클러스터 정보를 병합하여 상기 결과 테스트 모듈로 전송하는 결과 통합기;

를 포함하는 것을 특징으로 하는 그리드 기반 대용량 데이터 분류 장치.

청구항 5

제1항에 있어서,

상기 데이터 분류부는

상기 데이터를 입력 및 관리하는 훈련 데이터 관리 모듈;

상기 그리드 자원을 사용하고, 임계값을 기반으로 데이터가 분류되도록 클러스터를 생성시켜 훈련하는 데이터 훈련 모듈;

상기 클러스터를 형성하는 대표 데이터를 관리 및 갱신하는 클러스터 관리 모듈; 및

생성된 클러스터 중 이상치 클러스터를 제거하는 클러스터 제거 모듈;

을 포함하는 것을 특징으로 하는 그리드 기반 대용량 데이터 분류 장치.

청구항 6

제5항에 있어서,

상기 클러스터 제거 모듈은 생성된 전체 클러스터를 검사하되, 각 클러스터에 소속된 데이터의 수가 일정 기준 이하 또는 해당 클러스터가 데이터를 대표할 수 없는 이상치 클러스터를 제거하는 것을 특징으로 하는 그리드 기반 대용량 데이터 분류 장치.

청구항 7

대용량 데이터 분류를 위해 수집된 그리드 자원을 저장 공간 및 작업 처리량에 따라 분류하고, 임계값을 초기화하는 단계;

상기 분류된 그리드 자원으로 동적으로 작업을 할당받아 임계값 기반으로 데이터가 분류되도록 클러스터를 생성하여 훈련하고, 전체 클러스터를 검사하고, 이상치 클러스터를 제거하여 전송하는 단계;

상기 데이터 훈련, 분류 및 클러스터 정보를 병합한 결과를 검사하되, 최적의 임계값이 유도되도록 임계값을 조정하는 단계;

를 포함하는 그리드 기반 대용량 데이터 분류 방법.

청구항 8

제7항에 있어서,

상기 이상치 클러스터는 해당 클러스터가 데이터를 대표할 수 없거나 또는 생성된 전체 클러스터 중에서 각 클러스터에 소속된 데이터 수가 일정 기준 이하인 비중이 크지 않은 클러스터인 것을 특징으로 하는 그리드 기반 대용량 데이터 분류 방법.

청구항 9

제7항에 있어서,

상기 데이터 분류 방법은

대용량 데이터를 입력 및 관리하는 단계;

동적으로 할당된 그리드 자원으로 전송된 임계값을 기반으로 데이터가 분류되도록 클러스터를 생성하여 훈련하는 단계;

클러스터를 생성하는 대표 데이터를 관리 및 갱신하는 단계;

생성된 전체 클러스터를 검사하여 이상치 클러스터를 제거하는 단계;

를 포함하여 이루어지는 것을 특징으로 하는 그리드 기반 대용량 데이터 분류 방법.

청구항 10

제7항에 있어서,

데이터 훈련 방법은

대용량 데이터로부터 임계값을 기반으로 클러스터를 생성하는 단계;

생성된 클러스터의 패턴을 임계값과 비교하여 입력된 데이터의 패턴과 매칭시키는 단계;

생성된 클러스터의 패턴이 입력된 데이터의 패턴에 포함되면, 상기 데이터를 상기 생성된 클러스터에 소속시키는 단계;

생성된 클러스터의 패턴이 입력된 데이터의 패턴에 포함되지 않으면, 새로운 패턴에 따른 클러스터를 생성하는 단계;

클러스터를 생성시킨 데이터를 대표 데이터로 지정하여 갱신하는 단계;

를 포함하여 이루어지는 것을 특징으로 하는 그리드 기반 대용량 데이터 분류 방법.

청구항 11

제7항에 있어서,

이상치 클러스터 제거 방법은

생성된 전체 클러스터를 검사하되, 각 클러스터에 소속된 데이터의 수가 일정 기준 이하 또는 해당 클러스터가 데이터를 대표할 수 없는 이상치 클러스터를 제거하는 단계;

를 포함하여 이루어지는 것을 특징으로 하는 그리드 기반 대용량 데이터 분류 방법.

명세서

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

- <16> 본 발명은 그리드 기반 대용량 데이터 분류 장치 및 방법에 관한 것으로, 더욱 상세하게는 대용량 데이터가 분류되도록 클러스터를 생성시켜 혼련하고, 생성된 클러스터 중 데이터를 대표할 수 없거나 해당 클러스터에 소속된 데이터가 일정 수 이하인 비중이 작은 이상치 클러스터를 제거함으로써, 이에 따른 클러스터의 감소는 데이터 분류의 처리 시간을 단축시키는 효과가 있으며, 분류가 완료된 결과를 기반으로 최적의 임계값을 유도하도록 조정함으로써, 최적의 인식률을 근접할 수 있으며, 이에 따라 계산 시간을 감소시켜 작업 처리 시간을 단축시킬 수 있고, 동적으로 그리드 자원에 데이터 분류 작업을 할당함으로써, 대용량의 데이터 운용 및 처리가 가능해짐과 동시에 중앙 서버의 병목 현상을 방지하여 안정성을 증가시킬 수 있는 그리드 기반 대용량 데이터 분류 장치 및 방법에 관한 것이다.
- <17> 일반적으로, 그리드 컴퓨팅(Grid Computing)은 네트워크로 다수의 컴퓨터를 연결하여 데이터 처리 능력을 극대화시키는 방법으로, 지리적으로 분산된 컴퓨터, 저장 장치 등의 자원을 네트워크로 연결하여 상호 공유하고, 이를 이용할 수 있도록 이루어진다.
- <18> 그리고, 상기 그리드 컴퓨팅은 일종의 가상 컴퓨터로써, 일정 시간 동안 유휴 상태의 컴퓨터 자원을 검색 및 파악하여 각각의 컴퓨터가 보유한 자원을 공유하기 때문에, 데이터를 처리함에 있어 작업 속도를 증가시킬 수 있다.
- <19> 그러나, 문자열 및 실수로 구성된 대용량 데이터를 분류함에 있어서, 대표 패턴에 대한 분류가 이루어짐으로써, 대표 패턴이 한정적 및 제한적이므로 이에 대한 유동적인 그룹화가 실행되지 못했고, 이에 따라 분류의 정확성이 감소되었으며, 대용량의 데이터에 대한 최적의 인식률을 산출하기 위한 계산 시간이 증가하였고, 그리드 자원에 대한 분류가 유연하게 이루어지지 않아 중앙 서버의 병목 현상 등 안정성이 감소함과 동시에, 자원의 활용도 및 응답 시간이 증가하는 등의 문제점이 있었다.

발명이 이루고자 하는 기술적 과제

- <20> 본 발명은 상기한 문제점을 해결하기 위하여 안출한 것으로, 대용량 데이터가 분류되도록 클러스터를 생성시켜 혼련하고, 생성된 클러스터 중 데이터를 대표할 수 없거나 해당 클러스터에 소속된 데이터가 일정 수 이하인 비중이 작은 이상치 클러스터를 제거함으로써, 이에 따른 클러스터의 감소는 데이터 분류의 처리 시간을 단축시키는 효과가 있으며, 분류가 완료된 결과를 기반으로 최적의 임계값을 유도하도록 조정함으로써, 최적의 인식률에 근접할 수 있으며, 이에 따라 계산 시간을 감소시켜 작업 처리 시간을 단축시킬 수 있고, 동적으로 그리드 자원에 데이터 분류 작업을 할당함으로써, 대용량의 데이터 운용 및 처리가 가능해짐과 동시에 중앙 서버의 병목 현상을 방지하여 안정성을 증가시킬 수 있는 그리드 기반 대용량 데이터 분류 장치 및 방법을 제공하는 것을 목적으로 한다.

발명의 구성 및 작용

- <21> 상기한 바와 같은 목적을 달성하기 위하여 본 발명은 대용량의 데이터 분류에 사용될 이용가능한 그리드 자원을 분류하는 자원 관리 모듈과, 상기 데이터 분류를 위한 임계값을 초기화하여 데이터를 분류하되, 분류가 완료된 데이터의 결과에 따라 임계값을 조정하는 임계값 조정 모듈과, 상기 임계값 조정 모듈로 분류가 완료된 데이터의 결과를 검사하여 결과를 전달하는 결과 테스트 모듈을 포함하는 데이터 분류 관리부; 상기 그리드 자원 및 임계값을 전달하고, 분류가 완료된 데이터를 병합하여 상기 결과 테스트 모듈로 전송하는 다수의 자원 중개 모듈을 포함하는 데이터 중개부; 및 상기 그리드 자원 및 임계값을 수신하여 대용량 데이터를 클러스터로 분류하

여 훈련시키되, 이상치 클러스터는 제거하여 상기 데이터 중개부로 전송하는 다수의 데이터 분류부; 를 포함한다.

- <22> 그리고, 상기 자원 관리 모듈은 이용 가능한 그리드 자원을 수집하여 저장 공간 및 작업 처리량에 따라 분류하는 것을 특징으로 한다.
- <23> 또한, 상기 이상치 클러스터는 생성된 클러스터 중에서 데이터를 대표할 수 없거나 또는 포함된 데이터의 비중이 크지 않은 클러스터인 것을 특징으로 한다.
- <24> 여기서, 상기 자원 중개 모듈은 상기 그리드 자원 및 임계값을 데이터 분류부에 전달하는 정보 전달기; 및 다수개의 데이터 분류부로부터 출력된 데이터 훈련, 분류 결과 및 클러스터 정보를 병합하여 상기 결과 테스트 모듈로 전송하는 결과 통합기; 를 포함하는 것을 특징으로 한다.
- <25> 더불어, 상기 데이터 분류부는 상기 데이터를 입력 및 관리하는 훈련 데이터 관리 모듈; 상기 그리드 자원을 사용하고, 임계값을 기반으로 데이터가 분류되도록 클러스터를 생성시켜 훈련하는 데이터 훈련 모듈; 상기 클러스터를 형성하는 대표 데이터를 관리 및 갱신하는 클러스터 관리 모듈; 및 생성된 클러스터 중 이상치 클러스터를 제거하는 클러스터 제거 모듈; 을 포함하는 것을 특징으로 한다.
- <26> 그리고, 상기 클러스터 제거 모듈은 생성된 전체 클러스터를 검사하되, 각 클러스터에 소속된 데이터의 수가 일정 기준 이하 또는 해당 클러스터가 데이터를 대표할 수 없는 이상치 클러스터를 제거하는 것을 특징으로 한다.
- <27> 한편, 대용량 데이터 분류를 위해 수집된 그리드 자원을 저장 공간 및 작업 처리량에 따라 분류하고, 임계값을 초기화하는 단계; 상기 분류된 그리드 자원으로 동적으로 작업을 할당받아 임계값 기반으로 데이터가 분류되도록 클러스터를 생성하여 훈련하고, 전체 클러스터를 검사하고, 이상치 클러스터를 제거하여 전송하는 단계; 상기 데이터 훈련, 분류 및 클러스터 정보를 병합한 결과를 검사하되, 최적의 임계값이 유도되도록 임계값을 조정하는 단계; 를 포함한다.
- <28> 여기서, 상기 이상치 클러스터는 해당 클러스터가 데이터를 대표할 수 없거나 또는 생성된 전체 클러스터 중에서 각 클러스터에 소속된 데이터 수가 일정 기준 이하인 비중이 크지 않은 클러스터인 것을 특징으로 한다.
- <29> 또한, 상기 데이터 분류 방법은 대용량 데이터를 입력 및 관리하는 단계; 동적으로 할당된 그리드 자원으로 전송된 임계값을 기반으로 데이터가 분류되도록 클러스터를 생성하여 훈련하는 단계; 클러스터를 생성하는 대표 데이터를 관리 및 갱신하는 단계; 생성된 전체 클러스터를 검사하여 이상치 클러스터를 제거하는 단계; 를 포함하여 이루어지는 것을 특징으로 한다.
- <30> 그리고, 데이터 훈련 방법은 대용량 데이터로부터 임계값을 기반으로 클러스터를 생성하는 단계; 생성된 클러스터의 패턴을 임계값과 비교하여 입력된 데이터의 패턴과 매칭시키는 단계; 생성된 클러스터의 패턴이 입력된 데이터의 패턴에 포함되면, 상기 데이터를 상기 생성된 클러스터에 소속시키는 단계; 생성된 클러스터의 패턴이 입력된 데이터의 패턴에 포함되지 않으면, 새로운 패턴에 따른 클러스터를 생성하는 단계; 클러스터를 생성시킨 데이터를 대표 데이터로 지정하여 갱신하는 단계; 를 포함하여 이루어지는 것을 특징으로 한다.
- <31> 또한, 이상치 클러스터 제거 방법은 생성된 전체 클러스터를 검사하되, 각 클러스터에 소속된 데이터의 수가 일정 기준 이하 또는 해당 클러스터가 데이터를 대표할 수 없는 이상치 클러스터를 제거하는 단계; 를 포함하여 이루어지는 것을 특징으로 한다.
- <32> 이하, 본 발명에 따른 실시예를 첨부된 예시도면을 참고로 하여 상세하게 설명한다.
- <33> 도 1은 본 발명에 따른 그리드 기반 대용량 데이터 분류 장치를 개략적으로 도시한 블록도이다. 도면에서 도시하고 있는 바와 같이, 본 발명에 의한 그리드 기반 대용량 데이터 분류 장치(1)는 데이터 분류 관리부(10)와 데이터 중개부(30)와 다수개의 데이터 분류부(50)를 포함하여 이루어진다.
- <34> 여기서, 상기 데이터 분류 관리부(10)는 대용량의 데이터를 분류하기 위하여, 보유한 그리드 자원을 저장 공간 및 작업 처리량에 따라 필터링하고, 대용량의 데이터 분류 작업을 위한 임계값을 초기화하여 상기 그리드 자원 및 초기화된 임계값을 상기 전달하며, 대용량의 데이터 분류 작업이 완료된 데이터를 테스트하여 이에 따라 임계값을 조정하고, 조정된 임계값으로 최적의 인식률에 도달할 때까지 상기 임계값을 조정 및 갱신한다.
- <35> 이를 위하여, 상기 데이터 분류 관리부(10)는 자원 관리 모듈(11)과 임계값 조정 모듈(13)과 결과 테스트 모듈(15)로 이루어지는데, 상기 자원 관리 모듈(11)은 대용량의 데이터를 분류하기 위하여, 데이터를 저장할 수 있는 저장 공간과 상기 데이터를 처리할 수 있는 작업 처리 능력에 따라, 그리드 컴퓨팅 기반 그리드 자원들을 분

류하는데, 기 정해진 기준 저장 공간 및 기준 작업 처리 능력 미만인 그리드 자원들은 필터링으로 제외되도록 여과시키고, 이에 따른 수식은 하기와 같다.

수학식 1

<36> $LOC(X[j + 1]) = LOC(X[j]) + c$

<37> $LOC(X[j]) = L_0 + c_j$

<38> 여기서, LOC 는 그리드 자원에 대하여 노드에 따른 순서를 나타내며, c 는 그리드 자원의 데이터 저장 능력인 저장 공간 및 작업 처리 능력을 나타내고, j 는 그리드 자원의 노드 위치를 나타내며, L_0 은 노드 위치의 기본값을 나타낸다.

<39> 다시 말하면, 상기 수학식 1에 따라 선별된 그리드 자원에 대하여 사용가능한 그리드 자원을 정렬한 후, 순차적인 방법에 의하여 작업 처리 능력이 우수한 그리드 자원이 동적으로 할당된다.

<40> 그리고, 상기 임계값 조정 모듈(13)은 기준 저장 공간 및 기준 작업 처리량 이상인 그리드 자원에 할당된 대용량 데이터의 분류가 완료되면, 이를 테스트하여 테스트 결과를 기반으로 임계값을 재조정하여 변경할 수 있으며, 변경된 임계값은 상기 데이터 중개부(30)로 전달되어 대용량의 데이터를 분류하는데 이용된다.

<41> 또한, 상기 결과 테스트 모듈(15)은 기준 저장 공간 및 기준 작업 처리량 이상인 그리드 자원에 할당된 대용량 데이터의 분류가 완료되면, 상기 다수개의 데이터 분류부(50)의 분류 결과는 상기 데이터 중개부(30)에서 각 결과가 병합되는데, 상기 병합된 결과를 이용하여 데이터 분류 작업이 완료된 데이터에 대하여 분류 정확도를 테스트하기 위해 구비된다.

<42> 상기 데이터 중개부(30)는 상기 데이터 분류 관리부(10)로부터 전달받은 그리드 자원 및 임계값 정보를 상기 다수개의 데이터 분류부(50)로 전송하고, 상기 그리드 자원 및 임계값 정보를 이용하여 데이터 분류가 완료된 각각의 결과를 병합하여 상기 데이터 분류 관리부(10)로 전달한다.

<43> 이를 위하여, 상기 데이터 중개부(30)는 다수개의 자원 중개 모듈(31)로 이루어지는데, 상기 자원 중개 모듈(31)은 정보 전달기(31a)와 결과 통합기(31b)를 포함하여 이루어진다.

<44> 여기서, 상기 다수개의 자원 중개 모듈(31)은 상기 데이터 분류 관리부(10)로부터 전달받은 그리드 자원 정보 및 임계값 정보를 데이터 분류부(50)로 전송하고, 상기 다수개의 데이터 분류부(50)의 대용량 데이터 분류 결과를 데이터 분류부(50)의 상위 노드인 데이터 중개부(30)의 자원 중개 모듈(31)에서 병합하며, 이를 데이터 중개부(30)의 상위 노드인 데이터 분류 관리부(10)로 전달한다.

<45> 그리고, 상기 정보 전달기(31a)는 데이터 분류 관리부(10)로부터 전달되는 그리드 자원 정보 및 임계값 정보를 데이터 분류부(50)로 전송하기 위하여 구비되며, 다수개의 자원 중개 모듈(31) 내에 각각 구비되어 하위 노드인 데이터 분류부(50)로 전달되는 정보를 전송한다.

<46> 또한, 상기 결과 통합기(31b)는 하위 노드인 데이터 분류부(50)에서 분류가 완료된 대용량의 데이터에 대한 결과를 상위 노드인 데이터 중개부(30)의 다수개의 자원 중개 모듈(31) 내 결과 통합기(31b)를 통하여 병합하여 상기 데이터 분류 관리부(10)로 전달하는 기능을 가지며 구비된다.

<47> 여기서, 상기 상위 노드 및 하위 노드에서 상위 및 하위는 데이터 분류 관리부(10), 데이터 중개부(30), 데이터 분류부(50) 간 비교 시의 상대적인 개념을 나타낸다.

<48> 상기 다수개의 데이터 분류부(50)는 상기 데이터 분류 관리부(10)에서 출력되는 그리드 자원 정보 및 임계값 정보를 데이터 중개부(30) 내 다수개의 자원 중개 모듈(31)의 결과 통합기(31b)를 통하여 전달받고, 대용량의 데이터를 동적으로 할당된 그리드 자원으로 배분하여 훈련 및 분류하며, 상기 데이터가 입력되어 분류되도록 클러스터를 형성하는데, 기 생성된 클러스터와 패턴이 유사한 데이터가 입력되면 해당 클러스터에 상기 입력된 데이터를 소속시키고, 기 생성된 클러스터와 패턴이 다른 새로운 패턴의 데이터가 입력되면, 새로운 패턴의 클러스터를 생성하며, 상기 데이터를 대표 데이터로 지정하여 갱신시키고, 모든 데이터가 분류 완료되면, 생성된 전체의 클러스터를 검사하되, 각 클러스터에 소속된 데이터의 수가 일정 기준 이하의 비중이 작거나 또는 해당 클러스터가 데이터를 대표할 수 없는 이상치 클러스터를 제거한다.

<49> 이를 위하여, 상기 다수개의 데이터 분류부(50) 각각에 훈련 데이터 관리 모듈(51)과 클러스터 관리 모듈(53)과

데이터 훈련 모듈(55)과 클러스터 제거 모듈(57)을 포함한다.

- <50> 여기서, 상기 훈련 데이터 관리 모듈(51)은 대용량의 데이터를 입력받아 관리한다.
- <51> 그리고, 상기 데이터 훈련 모듈(55)은 상기 훈련 데이터 관리 모듈(51)에서 관리중인 데이터에 대하여, 입력과 동시에 클러스터를 생성시키는데, 다음 입력되는 데이터에 대하여 데이터 중개부(30)에서 전송된 임계값을 기반으로 생성된 클러스터와 패턴이 유사한지의 여부를 확인하고, 패턴이 유사한 데이터의 경우에는 생성된 클러스터에 상기 데이터를 소속시키되, 생성된 클러스터와 패턴이 유사하지 않은 새로운 패턴의 데이터일 경우에는 새로운 클러스터를 생성시켜 데이터가 분류되도록 훈련을 완료한다.
- <52> 또한, 상기 클러스터 관리 모듈(53)은 상기 데이터 훈련 모듈(55)에서 생성된 새로운 클러스터를 형성하게 한 새로운 패턴의 데이터를 대표 데이터로 지정하여 관리하고, 계속해서 다음 데이터가 입력되어 상기 데이터 훈련 모듈(55)에서 분류되도록 클러스터를 형성할 시, 새로운 패턴의 클러스터를 형성하는 데이터를 대표 데이터로 지정 및 갱신하도록 구비된다.
- <53> 여기서, 상기 클러스터 제거 모듈(57)은 상기 데이터 훈련 모듈(55)에서 생성된 전체 클러스터를 검사하되, 각 클러스터에 소속된 데이터를 카운트하여 각 클러스터에 소속된 데이터의 수가 일정 기준 이하인 비중이 작은 클러스터 또는 해당 클러스터가 데이터를 대표할 수 없는 클러스터를 이상치 클러스터로 정의하여, 상기 이상치 클러스터일 경우에는 이상치 클러스터를 제거하고, 다음 분류시에 입력된 데이터가 생성된 클러스터와 패턴을 비교할 시, 검색될 클러스터의 수를 감소시켜 검색 시간을 감소시켜, 처리 속도를 높일 수 있다.
- <54> 마지막으로, 상기 다수개의 데이터 분류부(50)로부터 산출된 결과는 상위 노드인 데이터 중개부(30)에서 병합하여 데이터 분류 관리부(10)의 결과 테스트 모듈(15)로 전달되며, 상기 결과 테스트 모듈(15)의 분석 및 테스트 결과에 따라 임계값 조정 모듈(13)의 임계값이 조정되고, 상기에서 서술된 바와 같이 조정된 임계값으로 인식률이 최적화되도록 임계값을 조정한다.
- <55> 도 2는 본 발명에 따른 그리드 기반 대용량 데이터 분류 방법을 개략적으로 도시한 흐름도이다. 도면에서 도시하고 있는 바와 같이, 본 발명에 의한 그리드 기반 대용량 데이터 분류 방법은 그리드 컴퓨팅을 기반으로 데이터를 분류함에 있어서, 대용량의 데이터 분류를 위하여 자원 관리 모듈에서 이용가능한 그리드 자원을 수집하고, 각기 다른 저장 공간과 작업 처리 능력을 가진 그리드 자원을 저장 공간 및 작업 처리 능력에 따라 순차적으로 나열시키며, 기준 저장 공간 및 기준 작업 처리 능력 미만인 그리드 자원은 제거되도록 분류하면서 시작된다(S10).
- <56> 그리고, 대용량의 데이터 분류 작업을 위한 임계값을 임계값 조정 모듈에서 초기화시키고(S20), 이에 따른 임계값과 그리드 자원을 데이터 중개부 내 다수개 존재하는 자원 중개 모듈의 정보 전달기를 통하여 다수개의 데이터 분류부로 각각 전송한다(S30).
- <57> 또한, 데이터 분류부로 할당된 그리드 자원 및 임계값으로 대용량 데이터가 분류되도록 클러스터를 생성하여 훈련시키고, 생성된 전체 클러스터에 대하여 이상치 클러스터를 제거하여 이를 상기 데이터 분류부의 상위 노드인 데이터 중개부 내 결과 통합기로 전송하는데, 상기 결과 통합기는 다수개의 데이터 분류부에서 전달받은 결과 정보를 병합한다(S40).
- <58> 여기서, 다수의 데이터 분류부에서 전송된 데이터의 분류, 훈련 결과와 클러스터 정보를 결과 통합기에서 병합하여 데이터 분류 관리부 내 결과 테스트 모듈로 전송하며(S50), 데이터의 분류, 훈련 결과와 클러스터 정보를 수신한 결과 테스트 모듈은 분석 및 테스트한다(S60).
- <59> 여기서, 최적의 인식률을 가지도록 임계값이 유도되었는지의 여부를 확인하여(S70), 최적의 임계값이 유도될 때까지 테스트 결과를 기반으로 임계값 조정 모듈에서 분류를 위한 임계값을 재조정하여 갱신하여(S80) 계속해서 상기 단계(S10 - S60)를 반복구동하고, 최적의 임계값이 유도되면 본 발명에 따른 그리드 기반 대용량 데이터 분류 방법을 종료시킨다.
- <60> 도 3은 본 발명에 따른 그리드 기반 대용량 데이터 분류 방법 중 데이터 분류 방법을 상세하게 도시한 흐름도이다. 도면에서 도시하고 있는 바와 같이, 본 발명에 의한 그리드 기반 대용량 데이터 분류 방법 중 데이터 분류 방법은 도 2의 단계(S50)를 상세히 설명하는데, 데이터 분류부로 동적으로 할당된 그리드 자원에 대용량 데이터를 배분하고, 훈련 데이터 관리 모듈로 대용량 데이터를 입력 및 관리한다(S50-1).
- <61> 그리고, 상기 훈련 데이터 관리 모듈에서 관리중인 대용량 데이터 분류를 위하여 할당된 그리드 자원에게 상기 데이터 분류부로 전송된 임계값을 전달하고, 이를 기반으로 대용량 데이터가 분류되도록 클러스터를 형성시켜

훈련한다(S50-2).

- <62> 또한, 상기 과정에서 데이터 분류를 진행하며 생성된 클러스터를 관리 및 갱신하는데, 기 생성된 클러스터와 유사하지 않은 패턴의 데이터는 새로운 패턴의 클러스터를 생성하게 되는데, 이 과정에서 유사하지 않아 새로운 패턴의 클러스터를 형성하는 데이터를 대표 데이터로 지정하여 이를 관리 및 갱신한다(S50-3)
- <63> 더불어, 데이터 분류가 완료된 후, 생성된 전체 클러스터를 검사하되, 각 클러스터에 소속된 데이터를 카운트하여 각 클러스터에 소속된 데이터의 수가 일정 기준 이하인 비중이 작은 클러스터 또는 해당 클러스터가 데이터를 대표할 수 없는 클러스터를 이상치 클러스터로 정의하고, 상기 이상치 클러스터가 발견될 경우에는 이를 제거하여 다음 분류시에 입력된 데이터가 기 생성된 클러스터와 패턴을 비교할 시, 검색될 클러스터의 수를 감소시켜 검색 시간을 줄일 수 있도록 실행된다(S50-4).
- <64> 도 4는 도 3의 데이터 분류 방법 중 데이터 훈련 방법을 상세하게 도시한 흐름도이다. 도면에서 도시하고 있는 바와 같이, 데이터 중개부로부터 전송된 임계값을 기반으로 대용량 데이터를 할당된 그리드 자원으로 입력하여 클러스터를 생성한다(S50-2a).
- <65> 그리고, 다음에 입력된 데이터를 기 생성된 클러스터와 패턴을 비교하는데, 상기 임계값을 기반으로 매칭시켜(S50-2b), 기 생성된 클러스터의 패턴에 입력된 데이터의 패턴이 포함될 경우(S50-2c), 상기 기 생성된 클러스터에 입력된 데이터를 소속시킨다(S50-2d).
- <66> 한편, 상기 단계(S50-2c)에서 기 생성된 클러스터의 패턴에 입력된 데이터의 패턴이 포함되지 않을 경우, 새로운 패턴에 따른 클러스터를 생성하여 이를 갱신한다(S50-2e).
- <67> 더불어, 상기 입력된 데이터가 모두 기 생성된 클러스터에 포함되거나 또는 새로운 클러스터를 생성하여 훈련이 완료되었는지의 여부를 확인하고(S50-2f), 모든 데이터에 대한 훈련이 완료되지 않은 경우에는 데이터를 기 생성된 클러스터에 임계값을 기준으로 비교하여 매칭하는 단계(S50-2b)로 이동하며, 모든 데이터에 대한 훈련이 완료되었을 경우에는 본 발명에 따른 데이터 훈련 방법을 종료시킨다.
- <68> 도 5는 도 3의 데이터 분류 방법 중 이상치 클러스터 제거 방법을 상세하게 도시한 흐름도이다. 도면에서 도시하고 있는 바와 같이, 훈련 데이터 모듈에서 생성이 완료된 전체 클러스터를 검사하면서 시작된다(S50-4a).
- <69> 여기서, 전체 클러스터를 검사함에 있어서, 완료된 클러스터에 대하여 각 클러스터에 소속된 데이터의 수를 카운트하고(S50-4b), 상기 카운트된 데이터의 수가 일정 기준 이하인지 또는 해당 클러스터가 데이터를 대표할 수 있는지의 여부를 확인한다(S50-4c).
- <70> 그리고, 해당 클러스터에 소속된 데이터의 수가 일정 기준을 초과한 즉, 비중이 작지 않은 클러스터로 분류되고, 해당 클러스터가 데이터를 대표하여 클러스터의 속성을 유지할 때 검사를 완료하는데, 각각의 클러스터를 검사하면서 전체 클러스터에 대한 검사가 완료되었는지의 여부를 확인하고(S50-4d), 전체 클러스터에 대한 검사가 완료되었으면 종료시키는데, 아직 이상치 클러스터의 여부를 검사받지 못한 클러스터가 있을 경우에는 생성된 클러스터를 검사하는 단계(S50-4a)로 이동한다.
- <71> 또한, 해당 클러스터에 소속된 데이터의 수가 일정 기준을 이하, 즉 비중이 작은 클러스터로 분류되거나 또는 해당 클러스터가 데이터를 대표하지 못하여 클러스터의 속성을 유지하지 못하는 클러스터는 이상치 클러스터로 정의하여 이를 삭제하는데(S50-4e), 이는 다음 분류 시 기 생성된 클러스터와 입력되는 데이터의 패턴을 비교할 시, 이상치 클러스터의 제거로 인한 클러스터의 수 감소로 검색 및 비교의 시간이 줄어들고, 이에 따라 처리 속도가 증가한다.
- <72> 여기서, 전체 클러스터에 대한 검사가 완료되었는지의 여부를 묻고(S50-4d), 완료되었으면 종료시키는데, 아직 이상치 클러스터의 여부를 검사받지 못한 클러스터가 있을 경우에는 생성된 클러스터를 검사하는 단계(S50-4a)로 이동한다.
- <73> 도 6는 도 3의 데이터 분류 방법 중 클러스터를 그룹화한 인식 방법을 개략적으로 도시한 개념도이다. 도면에서 도시하고 있는 바와 같이, 본 발명에 의한 클러스터를 그룹화한 인식 방법은 분류 작업이 진행되면, 임계값에 따라 클러스터의 수가 정해진다.
- <74> 그리고, 생성된 클러스터를 그룹핑에 의한 방법으로 분류 작업을 수행하는데, 상기 생성된 클러스터를 A, B 두 분류 그룹으로 그룹화하여, 상기 그룹 내의 클러스터 유닛들이 정상적인 분류 위치에 있지 않은 Case1 과 Case2 와 같은 경우에는 잘못 분류된 경우로써, 일부 유닛이 다른 클러스터에 위치된다.

<75> 여기서, 소수 유닛이 소속된 클러스터인 Group A Outlier, Group B Outlier는 클러스터의 자격이 없으므로 제거한다.

<76> 또한, 대용량 데이터를 분류함에 있어서, 상기 대용량 데이터는 클러스터를 통하여 분류된 상태이므로, 생성된 클러스터에 대하여 그룹을 지정하는 방법에 따라서 다양한 분류 결과를 얻을 수 있다.

<77> 이때, 또 다른 분류를 위하여, 그 분류에 성격에 맞는 훈련과 식별에 대한 과정을 재시작하는 것이 아니라, 단지 그룹화를 다르게 하면 원하는 결과값을 산출할 수 있다.

<78> 이상에서는 본 발명의 바람직한 실시예를 예시적으로 설명하였으나, 본 발명의 범위는 이같은 특정 실시예에만 한정되지 않으며 해당 분야에서 통상의 지식을 가진자라면 본 발명의 특허 청구 범위내에 기재된 범주 내에서 적절하게 변경이 가능 할 것이다.

발명의 효과

<79> 이상에서 설명한 바와 같이 상기와 같은 구성을 갖는 본 발명은 대용량 데이터가 분류되도록 클러스터를 생성시켜 훈련하고, 생성된 클러스터 중 데이터를 대표할 수 없거나 해당 클러스터에 소속된 데이터가 일정 수 이하인 비중이 작은 이상치 클러스터를 제거함으로써, 이에 따른 클러스터의 감소는 데이터 분류의 처리 시간을 단축시키는 효과가 있으며, 분류가 완료된 결과를 기반으로 최적의 임계값을 유도하도록 조정함으로써, 최적의 인식률에 근접할 수 있으며, 이에 따라 계산 시간을 감소시켜 작업 처리 시간을 단축시킬 수 있고, 동적으로 그리드 자원에 데이터 분류 작업을 할당함으로써, 대용량의 데이터 운용 및 처리가 가능해짐과 동시에 중앙 서버의 병목 현상을 방지하여 안정성을 증가시킬 수 있는등의 효과를 거둘 수 있다.

도면의 간단한 설명

<1> 도 1은 본 발명에 따른 그리드 기반 대용량 데이터 분류 장치를 개략적으로 도시한 블록도.

도 2는 본 발명에 따른 그리드 기반 대용량 데이터 분류 방법을 개략적으로 도시한 흐름도.

〈3〉 도 3은 본 발명에 따른 그리드 기반 대용량 데이터 분류 방법 중 데이터 분류 방법을 상세하게 도시한 흐름도.

<4> 도 4는 도 3의 데이터 분류 방법 중 데이터 훈련 방법을 상세하게 도시한 흐름도.

<5> 도 5는 도 3의 데이터 분류 방법 중 이상치 클러스터 제거 방법을 상세하게 도시한 흐름도.

<6> 도 6는 도 3의 데이터 분류 방법 중 클러스터를 그룹화한 인식 방법을 개략적으로 도시한 개념도.

<7> <도면의 주요 부분에 대한 도면 부호의 간단한 설명>

<8> 1: 그리드 기반 대용량 데이터 분류 장치

<9> 10: 데이터 분류 관리부 11: 자원 관리 모듈

<10> 13: 분류 임계값 조정 모듈 15: 결과 테스트 모듈

<11> 30: 데이터 중개부 31: 자원 중개 모듈

<12> 31a: 정보 전달기 31b: 결과 통합기

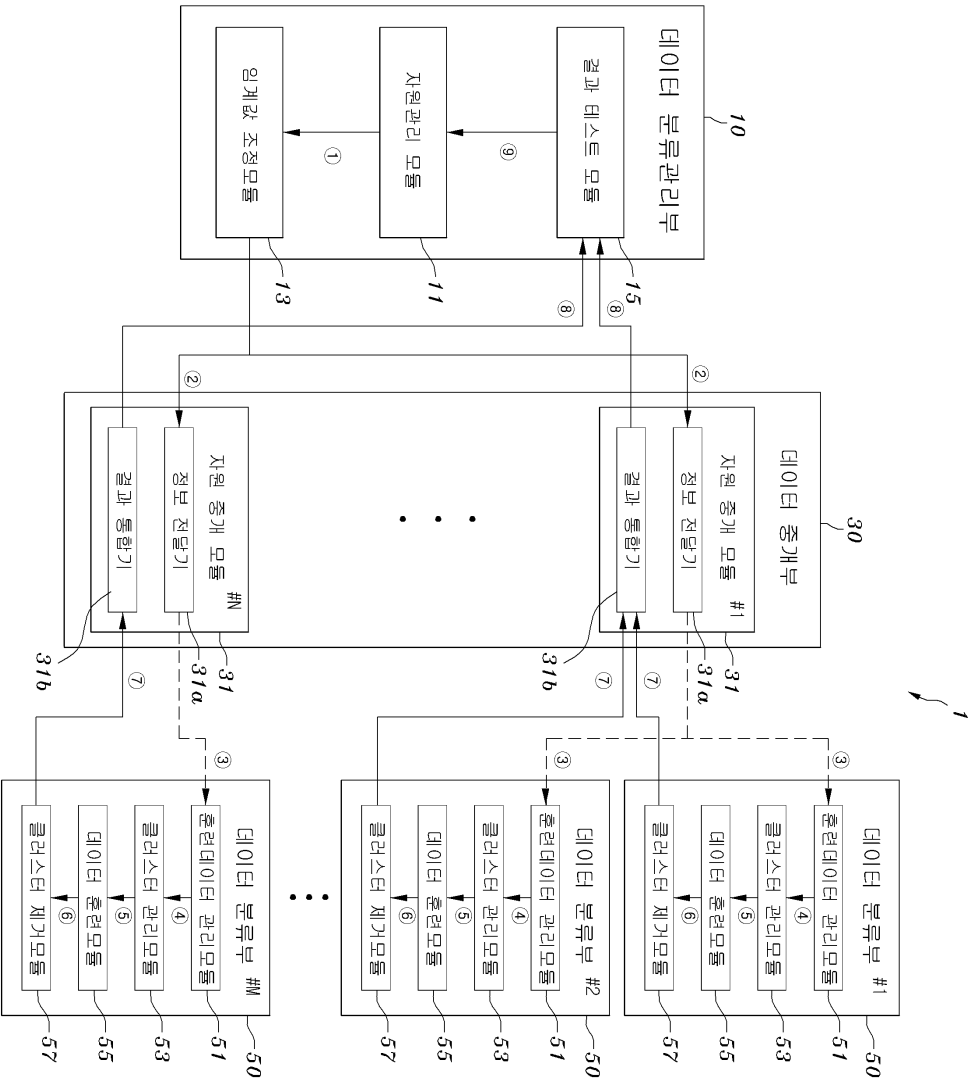
<13> 50: 데이터 분류부 51: 훈련 데이터 관리 모듈

<14> 53: 클러스터 관리 모듈 55: 데이터 훈련 모듈

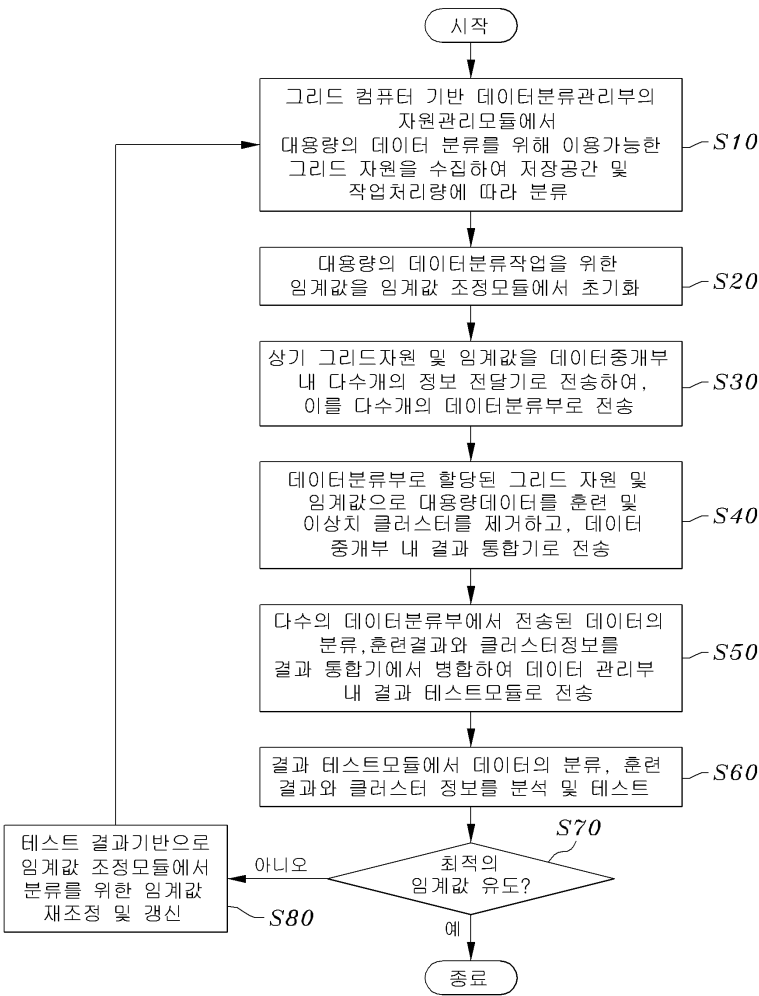
<15> 57: 클러스터 제거 모듈

도면

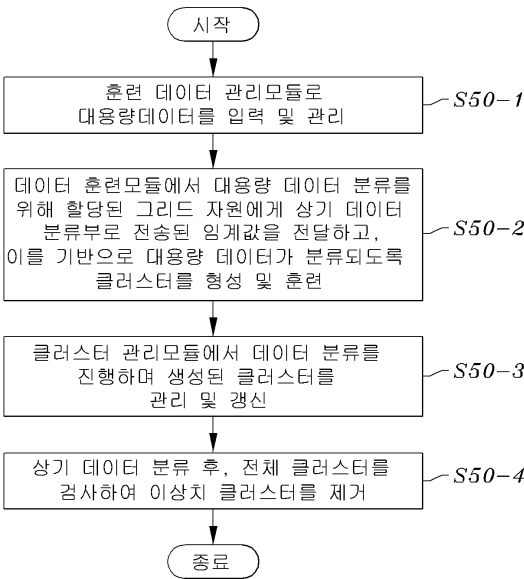
도면1



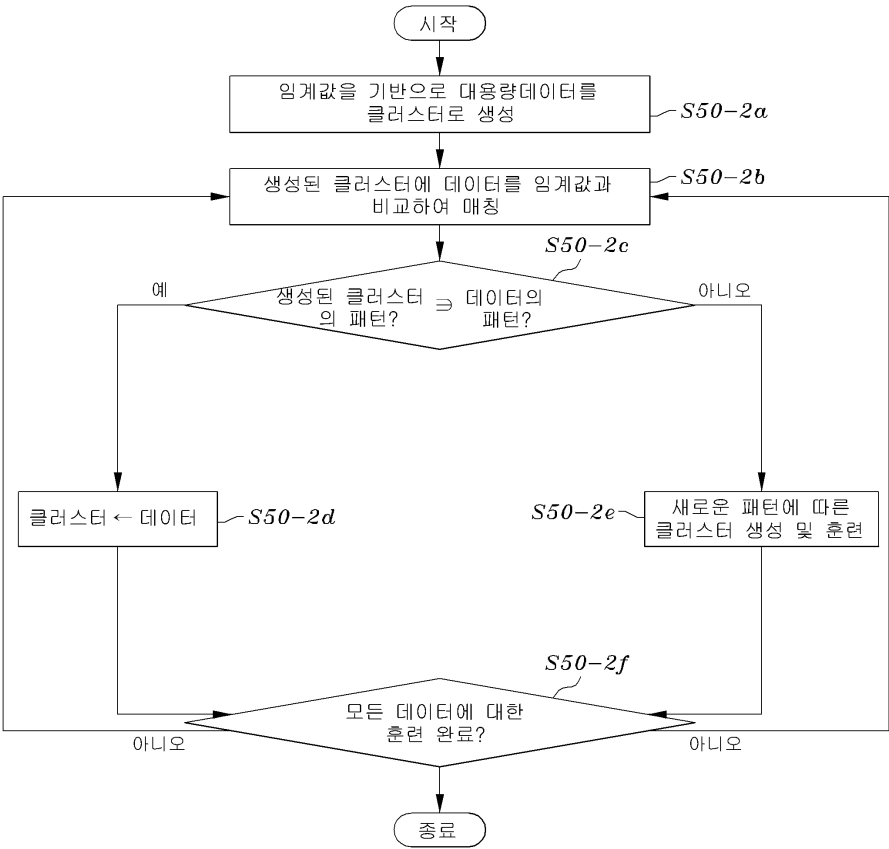
도면2



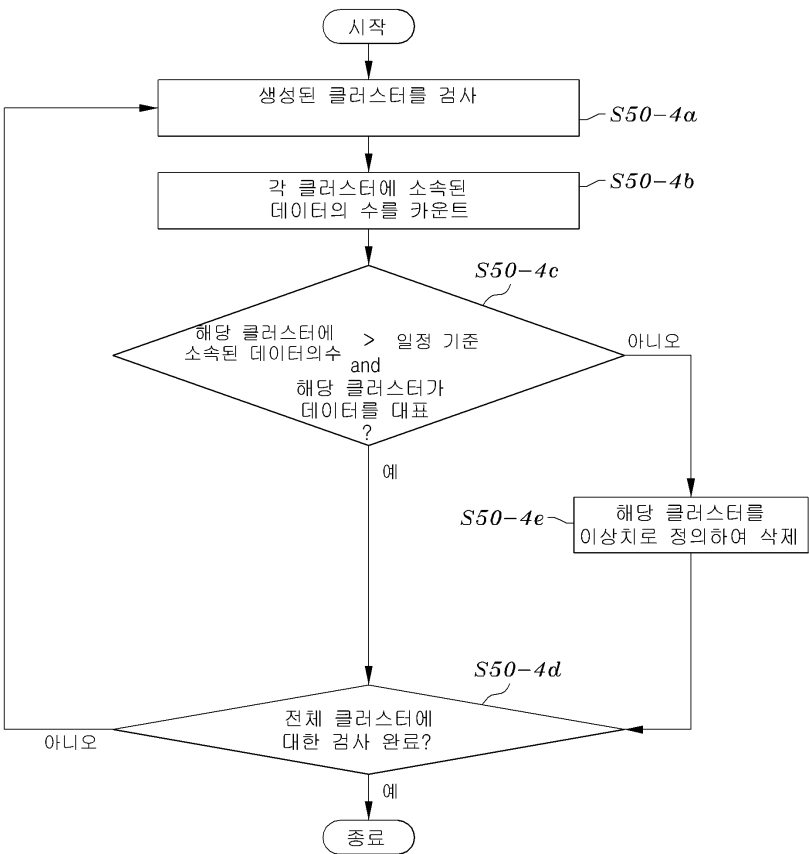
도면3



도면4



도면5



도면6

