데이터 마이닝 2020

서울 시립대 데이터 마이닝 속도 개선 과제 완료 보고

2020 데이터 마이닝 속도 개선 발표 내용

1. 데이터 마이닝 과제 진행 사항 발표
2. 데이터 마이닝 최종 결과 발표
3. 데이터 마이닝 검증 결과 발표
4. 데이터 마이닝 향후 과제 진행 방향

2020 데이터 마이닝 개선 산학(서울시립대)

작성자: 서울시립대학교
전종준

내용	일정, 방안
데이터마이닝 전처리모듈 개발 - 인바디, 스트레스, 양자데이터 전처리 모듈 개발, 조건 확립 - 수시 업데이트 가능한 모듈 개발	- 2020.4.30 - 시립대 과제 추진
데이터마이닝 개인별 알고리즘 속도 개선(약 10초 이내) - 데이터 DB로 변경 (기존은 파일) - 유저 그룹화 조건 확립, 결과 검증 - 비교 데이터 검증	- 2020.4.30 - 시립대 과제 추진
AI 컨설팅 데이터마이닝 부분 오류 사항 개선 - AI 컨설팅 페이지별 오류사항 리스트업 - 오류사항 정리 및 수정	- 2020.2 ~ 2020.4 오류사항 리스트업

개발범위

- 데이터마이닝 전처리모듈 개발
  - 인바디, 스트레스, 양자데이터 외 **데이터마이닝 알고리즘에 사용되는 데이터의 배치 (Batch) 병합 모듈의 작성 (파이썬 코드)**
- 쥬비스 데이터 마이닝 알고리즘 속도 개선
  - 고객 입력 정보를 이용한 통계량 작성방식 개선
  - 파이썬 코드를 이용한 알고리즘 작성
- 데이터 마이닝 알고리즘 오류 검증
  - 데이터 마이닝 알고리즘 오류 확인 및 수정

개발비용

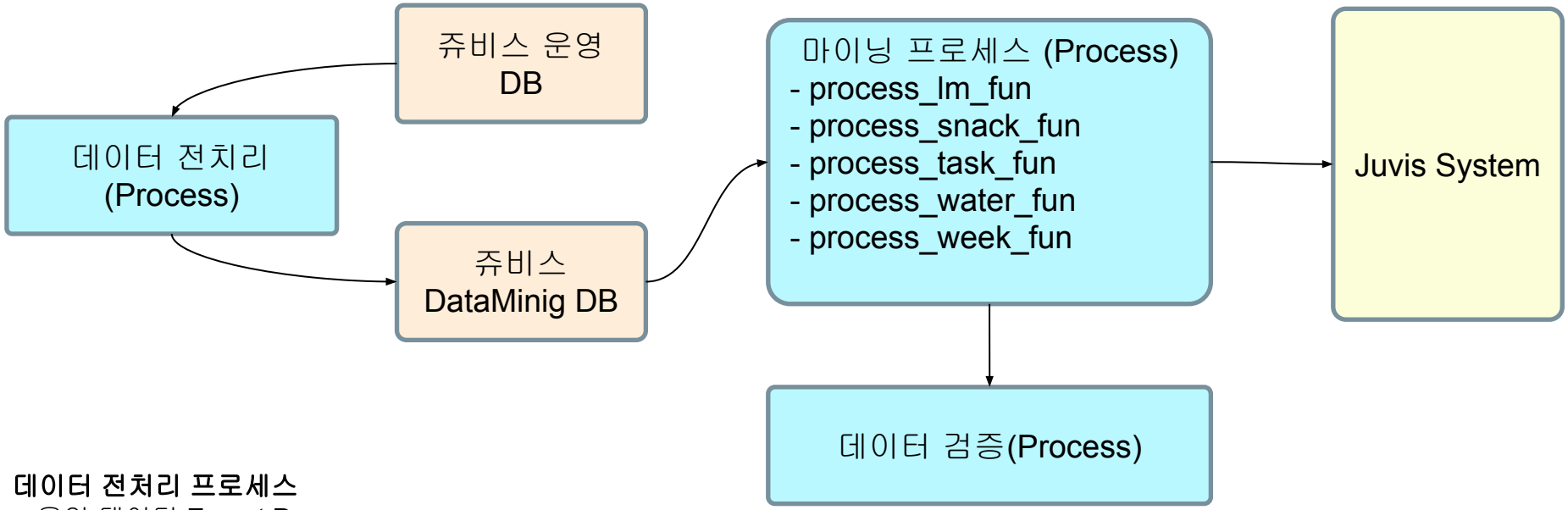
- 인건비 (총 2200만원)
  - 인건비: 1900만원

구분	2월	3월	4월
연구책임자	250만	350만	400만
연구원 (석사)	100만 (1명)	-	-
연구원(박사)	100만 (1명)	300만 = 150만*2명	400만 = 200만*2명

  - 간접비: 300만원

구성도 목표

마이닝 프로세스
- 기존 과제의 R Script로 작성된 모든 Mining Process
- 각 프로세스의 결과 값이 10초 이내



데이터 전처리 프로세스
- 운영 데이터 Export Process
- 전처리 데이터 생성 Import Process
- 쥬비스 DataMinig DB Table Schema

쥬비스 운영 DB / 쥬비스 Dataminig DB는 쥬비스에서 제공

각 function 별 데이터 검증 프로세스
- 각 function 별 해당 나이별/성별/kg대/.. 등 고객 카테고리 별 데이터 검증 방법

데이터 기준

1.유사 고객 추출 기준

성별 키('wc_ht'), 체중('bca_wt'), 나이('wc_age')를 이용하여 입력 고객과 다른 동성별의 고객간의 거리를 계산함.

: 동일한 성별의 고객을 먼저 추출한 뒤에, 정규화 과정을 통해 변형된 데이터를 가지고 거리를 계산합니다.

정규화 과정은 정확히 (데이터 - 평균) / (표준편차) 로 진행됩니다.

그러므로 키 1센치와 나이 1의 값의 차이는 동일하지 않습니다. 그리고 거리를 계산하는 기준은 유클리디안 거리로 (입력 데이터 - 고객 데이터)^2 로 정의됩니다.

입력 고객과의 거리가 작은 25%의 동성별의 사람을 유사 고객으로 정의함

2.성공한 사람과 실패한 사람 구분 기준

A.성공의 기준을 나누는 열에 따라 기준이 달라짐.

i.str_SDNN 열에는 8주차와 1주차의 차이를 기준으로 함 : SDNN 8주차와 1주차의 차이라 함은 얼마나 작아졌는지 기준

ii.mnt_count 열에는 8주차까지의 평균을 기준으로 함

iii.양자 데이터인 경우 range_data.csv의 N, A, B을 기준으로 각각 적용함 : range_data.csv의 N, A, B을 기준으로 각각 적용함 이라 함은 normal_range_start / normal_range_end

iv.다른 열의 경우 1주차 대비 8주차의 비를 기준으로 함

-> 성공 실패의 기준이 체중의 감량 달성은 전혀 상관없나요?

성공 실패의 기준이 체중인 경우에는 체중의 감량 달성과 관련있으며, iv의 경우에 해당됩니다.

또, i의 str_SDNN에서의 차이는 정확히 (8주차 수치) - (1주차 수치)로 계산됩니다.

양자 데이터의 경우는 말씀하신 기준이 맞습니다.

다른 열의 경우 (8주차 수치) / (1주차 수치) 를 기준으로 하며 이 때 다른 열은 체중, 체지방량 등을 포함한 경우로, i, ii, iii 경우를 제외한 경우입니다.

예를 들어, group_col에 인바디, 스트레스의 열 중 하나가 입력되었을 때 str_SDNN, mnt_count, 양자 데이터열 을 제외하면 (8주차 수치)/(1주차 수치)를 기준으로 한다는 의미이며

데이터 기준

3.성공 상위 중위 하위 구분 기준

A.성공의 기준을 나누는 열에 따라 기준이 달라짐

i.str_SDNN 열

1.차이가 60이하이면 성공 하위

2.차이가 40이하이면 성공 중위

3.차이가 20이하이면 성공 상위

ii.양자 데이터

1.range_data.csv의 N, A, B에따라 구분

iii.그 외 열

1.기준의 상위 25%면 성공 하위

2.기준의 상위 18.75%면 성공 중위

3.기준의 상위 6.25%면 성공 상위

그외 열 기준의 상위 / 하위의 기준은 뭔가요?

2.번 성공/실패 기준의 기준으로 나머지 수치가 얼마나 작아졌나가 기준인가요?

그러면 기준 데이터에 따라 성공 상위/중위/하위 사람이 달라 진다는건가요?

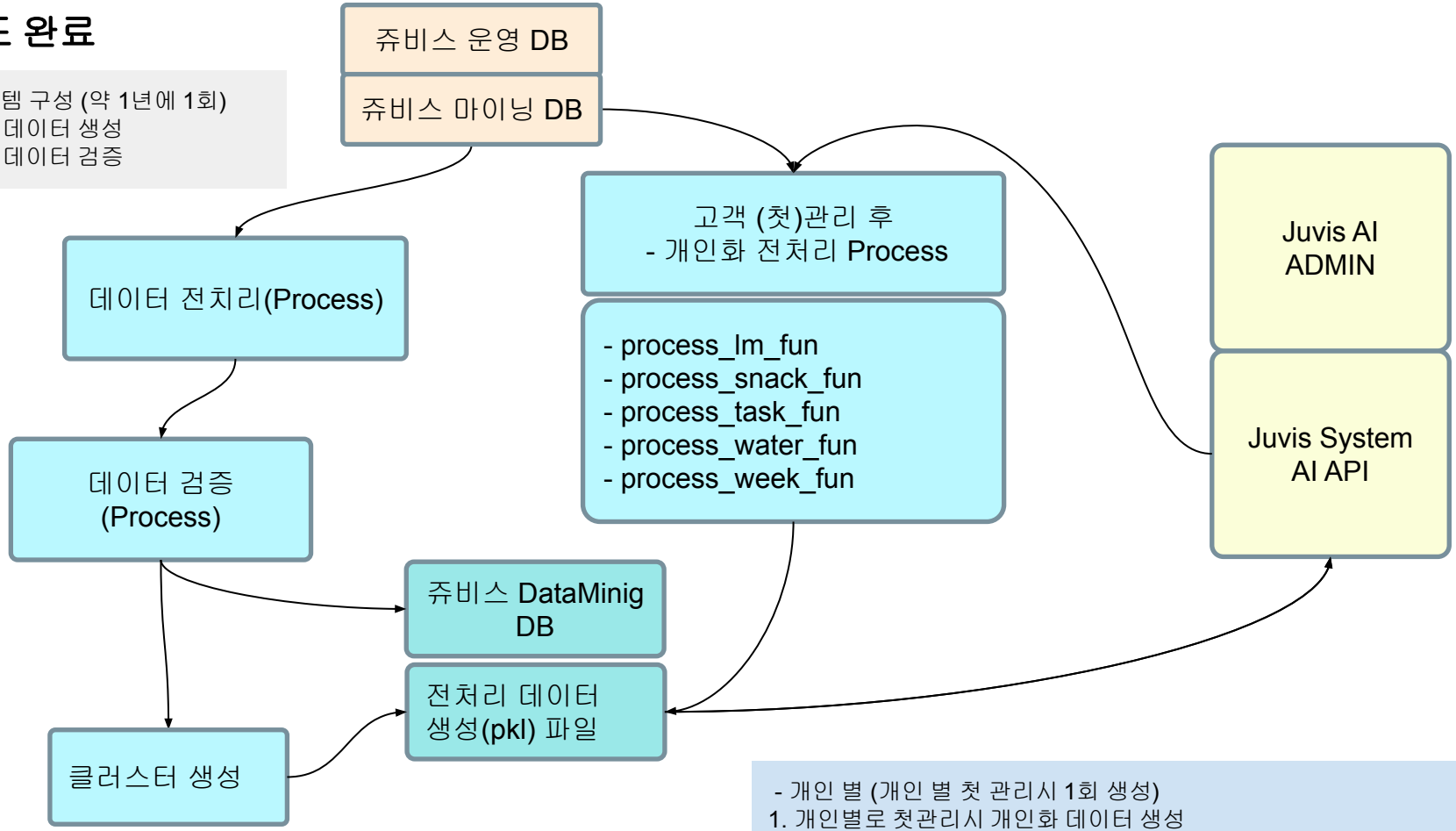
: 그 외의 열은 2번에서 작성한 기준과 같습니다. (mnt_count의 경우 8주차까지의 평균, 그 외 열의 경우 ((8주차 수치) / (1주차 수치)))

2. 성공한 사람과 실패한 사람 구분 기준이 가진 값에 따라서 성공과 실패가 나뉘게 됩니다.

따라서 말씀하신대로 선택한 열의 기준 값에 따라 성공 상위/ 중위/ 하위에 사람이 달라지게 됩니다.

구성도 완료

전체 시스템 구성 (약 1년에 1회)
1. 전처리 데이터 생성
2. 전처리 데이터 검증



- 개인 별 (개인 별 첫 관리시 1회 생성)
1. 개인별로 첫관리시 개인화 데이터 생성
2. 개인화 생성 데이터를 활용
 - 미리 생성 해놓음으로 데이터 사용시 소요 시간이 없음 (10초 미만)

주 사용 패턴

task	mem_id	group_col	interest_col
task_fun	mem_id	bca_bfm	비타민B6
task_fun	mem_id	bca_bfm	비타민C
task_fun	mem_id	bca_bfm	트립토판
task_fun	mem_id	bca_wt	bca_bfm
task_fun	mem_id	bca_wt	bca_wt
task_fun	mem_id	bca_wt	ED_ABD
task_fun	mem_id	bca_wt	ED_THIGHR
task_fun	mem_id	bca_wt	str_healthpoint
task_fun	mem_id	bca_wt	str_LF
task_fun	mem_id	bca_wt	str_SDNN
task_fun	mem_id	bca_wt	str_vascularage
task_fun	mem_id	bca_wt	wc_vfa
task_fun	mem_id	bca_wt	고밀도_지질_단백질
task_fun	mem_id	bca_wt	마그네슘
task_fun	mem_id	bca_wt	비타민B3
task_fun	mem_id	bca_wt	비타민B6
task_fun	mem_id	bca_wt	아연
task_fun	mem_id	bca_wt	칼슘
task_fun	mem_id	bca_wt	트립토판
task_fun	mem_id	str_LF	bca_wt
task_fun	mem_id	str_rpower	ed_fed
task_fun	mem_id	str_rpower	str_VEI
task_fun	mem_id	str_SDNN	bca_wt
task_fun	mem_id	str_SDNN	mfa_whr
task_fun	mem_id	str_SDNN	str_rpower
task_fun	mem_id	str_vascularage	wc_vfa
task_fun	mem_id	str_VEI	아라키돈_산
water_fun_cache	mem_id		cht_successRate
water_fun_cache	mem_id		str_LF
water_fun_cache	mem_id		wc_vfa
pie_fun	mem_id	bca_wt	has_snack
pie_fun	mem_id	bca_wt	mnt_count

Admin / Api

JUVIS AI ADMIN

DASHBOARD

Dashboard

MENU

데이터 마이닝

고객별 데이터 마이닝

고객별 데이터 클러스터

감량 예측

상담 예측

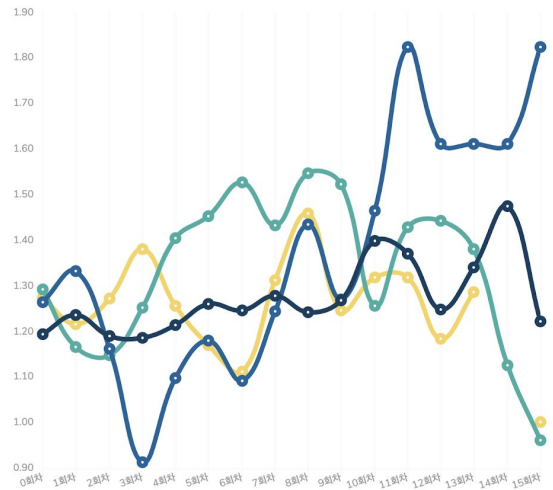
Hi, Juvis

고객 데이터 마이닝

Dashboard / 데이터마이닝 / 고객별

고객 정보

그래프 0723okmi



지수 나 그룹1 그룹2 그룹3 bca_bfm비타민B6 조회

회차	나	성공 1	성공 2	성공 3
1	1.194	1.2649714286	1.294	1.275
2	1.23775	1.333	1.167	1.2165
3	1.191	1.162	1.1495	1.2733333333
4	1.18675	0.9136666667	1.2525	1.381
5	1.215	1.0995	1.4046666667	1.257
6	1.262	1.18	1.4536666667	1.17
7	1.248	1.091875	1.5286666667	1.113
8	1.279	1.2445	1.4328	1.313
9	1.243	1.435	1.549	1.4593333333
10	1.2685	1.2715	1.5244	1.2465
11	1.4	1.466	1.2575	1.320125
12	1.3705	1.824	1.4305	1.3195
13	1.249	1.6125	1.443	1.1845
14	1.3415	1.6125	1.381	1.288
15	1.476	1.6125	1.126	null
16	1.224	1.825	0.963	1.0015

고객 리스트

-지점- mem_id 검색

	고객ID	지점	계약번호	계약일	시작->종료	시작->계약kg	관리횟수	관리기간	프로그램이름	마이닝
211679	0723okmi	인천점	26607726 (10046145)	1977-05-30 [여] 1980-	2020-04-13 - 2020-06-12 2020-05-07	76.49 - 69.69	10	8	감량포텐 듀얼 이벤트 8주 (JPI12746)	Mining