김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제3호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 62 No. 3 JUCHE105(2016).

KNN알고리듬에 기초한 루실자료의 한가지 처리방법

장 심 철

사회경제생활과 과학연구부문에서는 많은 자료들을 수집하여 처리를 진행하게 되며 이때 관측설비의 오동작이나 조작공의 조작오유 등 여러가지 원인으로 하여 자료의 루실이 존재하게 된다.

이러한 자료루실은 정보의 손실을 일으킬뿐아니라 부정확한 자료모형을 만들어내고 결심채택에 부정적영향을 미칠수 있다. 이로부터 루실자료의 처리를 비롯한 여러가지 자료예비처리방법들이 연구되였다.[1－3]

우리는 자료발굴의 주요한 연구령역인 자료분류에서 광범히 리용되고있는 KNN알고리듬에 기초한 한가지 루실자료의 처리방법에 대한 연구를 하였다.

1. KNN알고리듬에 기초한 루실자료의 처리방법

최근방(KNN: K Nearest Neighbors)알고리듬은 실례에 기초한 분류알고리듬으로서 자료발굴령역에서 분류문제를 해결하는데 광범히 리용되고있다.

이 알고리듬에서는 어떤 실례표본 X에 대하여 분류를 진행할 때 X로부터의 거리가 최소인 K개의 실례표본들을 선택하고 이 K개의 실례표본들의 보편적인 류형을 X의 류형으로 한다.

이것을 수학적으로 표현하면 다음과 같다.

$$c(X)=\arg\max_{c\in C}\sum_{i=1}^{K}\delta(c,\ c(y_i)) \tag{1}$$

여기서 $c(X)$는 X의 류형, $y_i(i=1,\ 2,\ \cdots,\ K)$는 X의 K개 근방, $c(y_i)$는 y_i의 류형, c는 분류류형, C는 c의 값모임이며 δ함수는

$$\delta(c,\ c(y_i))=\begin{cases}1 & (c=c(y_i))\\ 0 & (c\neq c(y_i))\end{cases} \tag{2}$$

이다.

먼저 원천자료속의 모든 기록을 두가지 부류 즉 모든 속성값이 완전한 완전자료묶음 D_c와 적어도 어느 한 속성값이 빈값인 루실자료묶음 D_i로 나눈다.

다음 루실자료값을 그것과 가장 가까운(거리가 최소인) K개의 완전자료에 의하여 결정하는데 이러한 완전자료를 루실자료의 가장 가까운 K개 근방(KNN)이라고 부른다.

완전자료를 A_i로 표시하면 루실자료값 Q는 다음과 같이 계산할수 있다.

$$Q=\begin{cases}\sum_{i=1}^{K}A_i\Big/K & \text{속성 } A\text{가 련속형일 때}\\ A_i & \text{속성 } A\text{가 리산형일 때}\end{cases} \tag{3}$$

그러나 K개의 근방들과 표본실례 X사이의 거리는 서로 다르며 거리가 작을수록 루실자료에 대한 작용은 더 크다. 그러므로 K개의 근방들에 서로 다른 무게를 주어 거리가 작을수록 무게값도 크게 하면 KNN알고리듬의 효과를 높일수 있다.

이러한 무게를 고려한 KNN알고리듬을 W-KNN알고리듬이라고 부른다.

완전자료 A_i에 대응하는 무게를 W_i로 표시하면 루실자료값 Q는 다음과 같이 계산할 수 있다.

$$Q=\begin{cases}\dfrac{\sum_{i=1}^{K}W_iA_i}{\sum_{i=1}^{K}W_i} & \text{속성 } A\text{가 련속형일 때}\\ A_i & \text{속성 } A\text{가 리산형일 때}\end{cases} \tag{4}$$

KNN알고리듬을 적용하는데서 K의 값과 거리함수의 선택이 중요한데 여기서 K값은 알고리듬의 정확도와 공간복잡도에 영향을 준다. 만일 K값을 크게 취하면 알고리듬의 우연성이 커지게 되고 K값을 작게 취하면 표본수가 부족하여 통계적인 의미를 잃게 될수 있다. 경험적으로는 자료가 작을 때 일반적으로 3～6이 비교적 적합하다는것을 보여주었다.

한편 거리는 기록자료들사이의 접근정도를 직접적으로 결정하며 알고리듬의 정밀도에 중요한 영향을 미친다.

일반적으로 절대값거리, 유클리드거리, 체브쉐브거리 등을 선택할수 있으며 유클리드거리가 가장 많이 리용된다. 그것을 수학적으로 표현하면 다음과 같다.

어떤 기록자료 X를 하나의 벡토르 $\{a_1(X),\ a_2(X),\ \cdots,\ a_n(X)\}$로 표시하자. 여기서 $a_1(X),\ a_2(X),\ \cdots,\ a_n(X)$는 X의 n개 속성값이다. 이때 2개의 기록자료 X_i와 X_j사이의 유클리드거리는 다음과 같다.

$$d(X_i,\ X_j)=\sqrt{\sum_{k=1}^{n}[a_k(X_i)-a_k(X_j)]^2} \tag{5}$$

시계렬자료에서 루실자료를 처리하는 경우 K개의 표본자료를 선택할 때 거리함수를 리용하여 반복계산하지 않고 직접 서로 이웃한 표본자료를 리용한다.

이때 루실자료와 가까운 표본자료일수록 더 큰 무게를 가지도록 하기 위하여 이웃한 표본자료들의 무게 W_i를 다음과 같이 설정한다.

$$W_i=\begin{cases}\dfrac{K+1-i}{b_K} & i=1,\ 2,\ \cdots,\ K\\ 0 & i=K+1,\ K+2,\ \cdots,\ n\end{cases} \tag{6}$$

여기서 $b_K=K(K+1)/2$이다.

2. 모의실험 및 결과분석

제품의 생산량은 시간변화에 따르는 시계렬이며 론문에서는 어느 한 공장의 생산실적자료기지로부터 세가지 제품의 1년간 생산량을 실례로 취급하였다.(표 1)

표 1. 세가지 제품의 1년간 생산량(×1 000대)

제품종류	월											
	1	2	3	4	5	6	7	8	9	10	11	12
X_1	45.9	48.9	50.3	53.1	54.9	56.2	56.2	56.7	54.5	57.7	58.3	57.9
X_2	48.7	53.1	56.1	59.3	60.2	62.8	63.9	64.9	57.7	63.7	63.7	62
X_3	28.9	31.8	33.6	36	38.3	39.4	39.1	39.2	37.9	39.7	39.7	39

알고리듬의 검증을 위하여 론문에서는 매 제품의 월생산량기록에서 1개 자료를 빼고 대신 《?》를 넣어 11개월의 생산량자료를 만든다.(표 2)

표 2. 루실자료를 포함한 생산량(×1 000대)

제품종류	월											
	1	2	3	4	5	6	7	8	9	10	11	12
X_1	45.9	48.9	?	53.1	54.9	56.2	56.2	56.7	54.5	57.7	58.3	57.9
X_2	48.7	53.1	56.1	59.3	?	62.8	63.9	64.9	57.7	63.7	63.7	62
X_3	28.9	31.8	33.6	36	38.3	39.4	?	39.2	37.9	39.7	39.7	39

이때 K=4로 취한다. 즉 루실자료점앞뒤로 각각 2개의 표본점을 선택한다.

X와 거리가 가장 가까운 4개 자료점의 무게는 식 (6)으로부터 각각 4/10, 3/10, 2/10, 1/10이다. KNN알고리듬과 회귀모형을 통하여 구한 루실자료의 값은 표 3과 같다.

표 3. 처리된 후 루실자료값의 비교

처리방법	x_{13}		x_{25}		x_{37}	
	값	절대오차	값	절대오차	값	절대오차
관측	50.3		60.2		39.1	
W-KNN	50.16	0.14	60.17	0.03	38.97	0.13
KNN	50.7	0.4	60.5	0.3	38.7	0.4
회귀모형	50.86	0.56	58.16	2.04	37.14	1.96

표 3으로부터 무게를 고려한 KNN알고리듬을 리용하여 추정한 루실자료의 값이 일반 KNN알고리듬이나 회귀모형에 비하여 정합정도가 더 높다는것을 알수 있다.

맺 는 말

무게를 고려한 KNN알고리듬을 리용하여 루실자료를 보상하는 한가지 방법을 제안하고 모의실험을 통하여 방법의 효과성을 검증하였다.

참 고 문 헌

[1] M. Hutter et al.; Computational Statistics Data Analysis, 48, 633, 2005.

[2] Wang Shuang Cheng et al.; Journal of Software, 15, 1, 1024, 2004.

[3] 刘明吉 等; 计算机科学, 27, 4, 54, 2000.

주체104(2015)년 11월 5일 원고접수

A Processing Method of Missing Data based on KNN Algorithm

Jang Sim Chol

In socio-economic life and scientific research field, people collect and process a lot of data, there are some missing data because of various factors such as failure of observation plane, operation error and so on. These missing data can be cause of information loss and generate inaccurate model, so it can have a negative influence on a decision-making. A method of processing missing data based on KNN algorithm with weight is proposed and verified effectiveness of the method through simulation experiment in this paper.

Key words: data mining, KNN algorithm, missing data, classification