Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 4, pp. 1085 ~ 1090

Development of Software for Coarse Classifying[1)]

Ki Mun Jung[2)] · Myung Cheol Kim[3)] · Joon Keun Yum[4)]

Abstract

In general, the coarse classifying procedure splits the values of a continuous characteristic into bands and the values of a discrete characteristic into groups of values. Coarse classifying improves the robustness of the credit scoring system but it is complicate and troublesome procedure. Thus, in this paper, we develop a software for coarse classifying by using Visual Basic Language. By using the developed software, we can find the best split easily. Also, this software will help learners to study credit scoring.

Keywords : Coarse classifying, Credit scoring, Visual basic

1. 서 론

대부분의 금융기관에서는 개인의 신용위험(credit risk) 관리에 대한 필요성을 인지하고 이를 관리할 수 있는 시스템을 도입하여 운영하고 있다. 최근 들어 이러한 신용위험을 예측하는 문제는 통계학 및 확률모형화 분야에서도 중요한 문제로 연구되고 있으며(Thomas, 2000), 판별분석(discriminant analysis), 로지스틱 회귀분석(logistic regression analysis), 생존분석(survival analysis) 등이 활용되고 있다.

이러한 개인신용평가(credit scoring)를 위한 모형에서 각 특성들을 상대적으로 작은 몇 개의 그룹으로 묶어줌으로써 로버스트한 신용평가시스템이 구축되도록 하는데, 이를 항목그룹화(coarse classifying)라고 한다. 즉, 항목그룹화는 특성변수들의 항목을 유사한 성질을 가진 항목들로 그룹화를 하는 것을 의미한다. 그런데, 항목그룹화를 수행하는 과정이 복잡하기 때문에 이를 위한 소프트웨어의 개발이 필요하다. 그리고, 이러한 소프트웨어를 개발함으로써 금융통계와 관련된 학부 강좌에서 교육용으로도 이용할 수 있을 것으로 생각한다. 따라서, 본 논문에서는 개인신용평가를 위한 모형개발

---

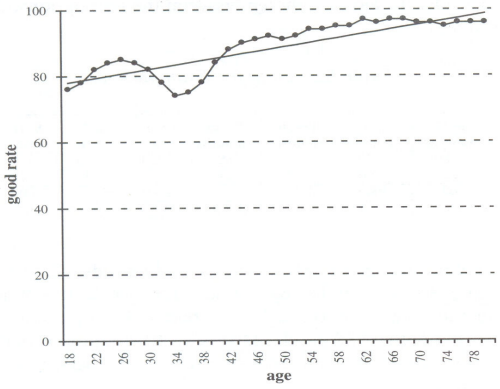
1) 이 연구는 2004학년도 동국대학교 연구년 지원에 의하여 이루어졌음.
2) (교신저자) 608-736 부산광역시 남구 대연동 110-1 경성대학교 정보통계학과 조교수
E-mail : kmjung@ks.ac.kr
3) 245-711 강원도 삼척시 중앙로 1 강원대학교 삼척캠퍼스 산업공학과 부교수
4) 100-715 서울시 중구 필동 3가 26 동국대학교 통계학과 교수

에서 항목그룹화를 위한 소프트웨어를 개발하고자 한다. 특히, 비주얼 프로그램으로써 소프트웨어를 개발하는데 많이 사용되고 있으며, 통계학과 학부과정에서 개설되고 있는 통계소프트웨어개발과 관련된 강좌에서 흔히 사용되고 있는 비주얼 베이직을 이용하여 개발하고자 한다. 개발된 소프트웨어의 모든 절차는 메뉴방식으로 실행이 가능하여 프로그램에 관한 전문적인 지식이 없는 일반 사람들도 쉽게 사용할 수 있다.

본 논문은 다음과 같이 구성된다. 2장에서는 항목그룹화의 이론적 배경에 대해서 자세히 살펴본다. 그리고 3장에서는 항목그룹화를 위해서 비주얼 베이직으로 개발된 소프트웨어를 소개하고 개발된 소프트웨어를 사용하는 방법에 대해서 예제를 통하여 설명하고자 한다.

2. 항목그룹화

항목그룹화는 특성변수들의 항목을 유사한 성질을 가진 항목들로 그룹화 하는 것을 의미하는데, 이는 각 특성변수들을 상대적으로 작은 몇 개의 그룹으로 묶어줌으로써 로버스트한 신용평점시스템을 구축하도록 해준다. 즉, 항목그룹화를 통해서 명목형 특성의 경우에는 항목의 수가 너무 많은 관계로 특정 속성에 대한 표본이 충분하지 않아 로버스트한 결과를 얻을 수 없는 문제점을 해결할 수 있으며, 연속형 특성의 경우에는 신용평점시스템을 이용하여 예측을 하기 위해서 비선형이 아닌 선형을 고려할 경우의 문제점을 해결할 수 있다. 예를 들어, 그림 1에서처럼 우량고객의 비율이 나이에 대해 단조증가하지 않는 경우에 일반적인 회귀선은 이를 반영하지 못하는 문제점이 있고, 이러한 문제점을 항목그룹화를 통해서 해결할 수 있다(Thomas 등, 2002).



<그림 1> 나이(age)에 따른 우량고객의 비율

따라서, 이러한 항목그룹화에서의 주된 관심은 어떠한 항목들을 하나의 그룹으로 만들겠느냐 하는 것이다. 이때, 가장 잘 분류된 항목그룹화를 최적분할(best split)이라고 할 수 있으며, 이러한 최적분할을 판단하기 위해서 χ^2-통계량, 정보(Information) 통계량, Somer's의 일치 통계량(concordance statistic) 등이 사용된다. 그리고 일반적으로 다음과 같이 정의되는 χ^2-통계량이 가장 널리 사용된다.

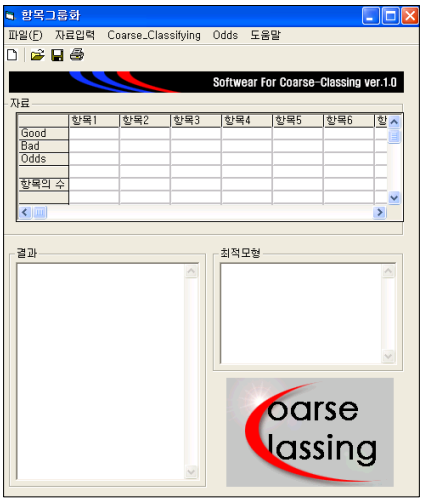
$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}.$$

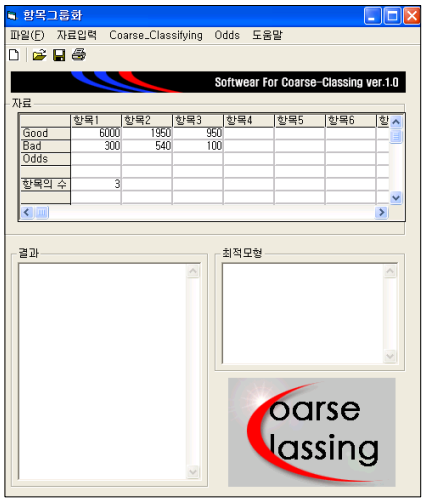
여기서, $O_{ij}=n_{ij}$로 i,j셀의 관측도수이고, $E_{ij}=\frac{n_{i\cdot}\times n_{\cdot j}}{n}$ 는 i,j셀의 기대도수이다.

그러나, 이러한 통계량을 사용하여 최적분할을 판단하기 위해서는 여러 번 반복해서 통계량을 구하고 이를 비교해야 하는 작업이 필요하다. 예를 들어, 어떤 특성변수의 항목이 A, B, C의 3개인 경우에 두 그룹으로 항목그룹화를 수행하려면 (A, B+C), (B, A+C), (C, A+B)에 대한 χ^2값을 구해서 이를 비교하여야 한다. 그리고 특성변수의 항목이 A, B, C, D의 4개이면 (A, B+C+D), (B, A+C+D), (C, A+B+D), (D, A+B+C), (A+B, C+D), (A+C, B+D), (A+D, B+C)에 대한 χ^2값을 구해서 이를 비교하여야 한다. 따라서 항목의 수가 많아지면 상당히 많은 항목그룹화의 경우에 대하여 반복적으로 χ^2값을 구하고 이를 비교해야 한다. 그러므로 단순하지만 반복적인 작업이 필요한 항목그룹화를 수행하는 과정을 위한 소프트웨어의 개발이 필요하다고 할 수 있다.

3. 항목그룹화를 위한 소프트웨어

이 절에서는 항목그룹화를 위해서 비주얼 베이직으로 개발된 소프트웨어를 소개하고자 한다. 본 연구에서 개발된 소프트웨어는 메뉴선택 방식으로 초보자들도 쉽게 사용할 수 있다. 먼저, 그림 2는 항목그룹화를 위한 소프트웨어의 첫 화면이다. 항목그룹화를 위해서 자료를 불러오거나 소프트웨어에서 직접 입력할 수 있다. 그리고 각 항목에 대한 승산비(odds ratio)를 구할 수 있으며, 카이제곱 통계량을 기준으로 최적의 항목그룹화를 찾기 위한 절차가 메뉴 방식으로 실행되며 빠른 사용을 위해 단축아이콘도 만들었다. 항목그룹화 소프트웨어를 이용하여 표 1의 특성변수에 대하여 항목그룹화를 수행하는 과정을 살펴보자.





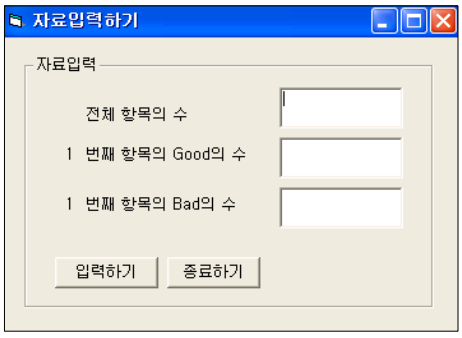
<그림 2> 항목그룹화 소프트웨어의 초기화면 <그림 3> 자료입력 후의 화면

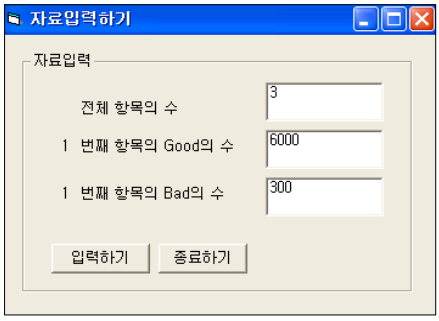
<표 1> 특성변수 주거형태별 신용상태

항목 / 신용상태	자가	월세	부모와 거주	계
양호(Good)	6,000	1,950	950	8,900
불량(Bad)	300	540	100	940
계	6,300	2,490	1,050	9,840

▶ 자료입력

「파일」메뉴의 「불러오기」메뉴를 이용하거나 「자료입력」메뉴를 이용하면 항목그룹화를 하고자하는 특성변수에 대한 자료를 쉽게 입력할 수 있다. 다음은「자료입력」메뉴를 이용하여 표 1의 자료를 입력하는 방법을 보여준다. 전체 항목 3개에 대하여 Good의 수와 Bad의 수를 입력한 후 종료하기 버튼을 클릭하면 그림 3과 같이 입력된 자료가 자료 창에 나타나게 된다.



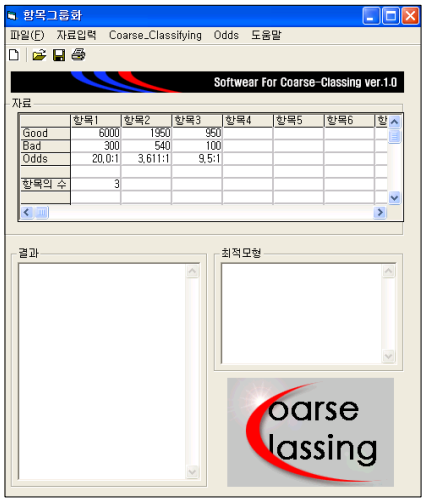


▶ 승산비의 계산

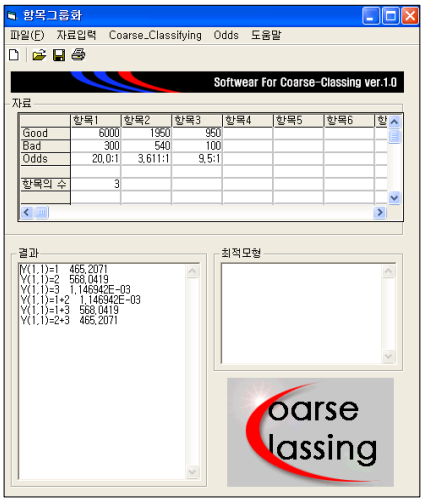
승산비를 계산하기 위해서는 「Odds」메뉴의 「Odds비 계산(Good:Bad)」메뉴를 선택하면 된다. 그러면 그림 4와 같이 소프트웨어의 자료 창에 승산비가 계산되어 나타나는 것을 볼 수 있다.

▶ χ^2통계량 값의 계산

최적의 분할을 찾기 위해서 χ^2통계량 값을 계산해야 하는데, 이를 위해서는 「Coarse_Classifying」메뉴의 「카이제곱 통계량 계산」메뉴를 선택하면 된다. 그림 5의 결과 창에는 「카이제곱 통계량 계산」메뉴를 이용해서 얻게 되는 각 그룹에 대한 카이제곱 통계량 값을 보여주고 있다. 결과 창의 Y(1,1)=1은 항목1(자가), 항목2(월세)+ 항목3(부모와 거주)를 의미하는 것으로 이때의 χ^2통계량 값이 465.2071라는 의미이다. 마찬가지로 Y(1,1)=2는 항목2(월세), 항목1(자가)+ 항목3(부모와 거주)를 의미하고 Y(1,1)=3은 항목3(부모와 거주), 항목1(자가)+ 항목2(월세)를 의미하는 것이다.



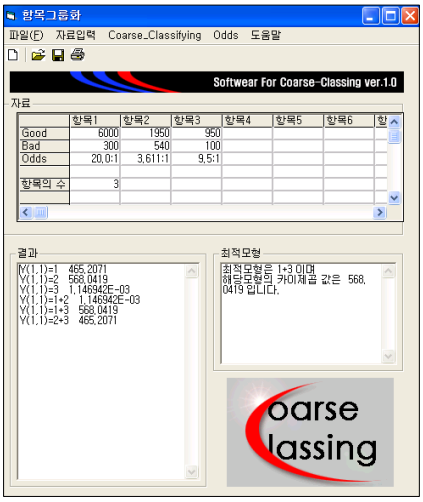
<그림 4> 입력된 자료의 승산비 계산



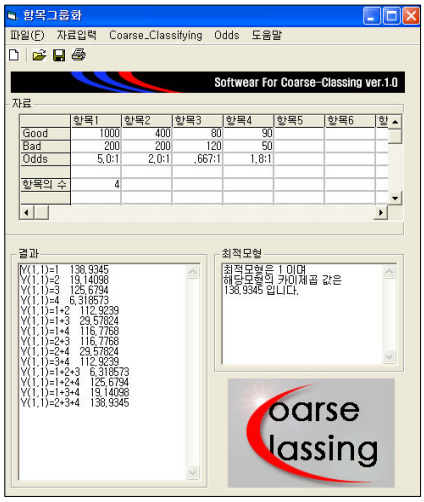
<그림 5> χ^2통계량 값의 계산

▶ 최적분할을 통한 항목그룹화

최적분할을 통한 항목그룹화를 수행하기 위해서는 「카이제곱 통계량 계산」메뉴를 통해서 얻게 되는 각 항목그룹별 카이제곱 통계량 값을 비교해야 한다. 이를 위해서는 「Coarse_Classifying」메뉴의 「카이제곱 통계량에 근거한 항목 그룹화」메뉴를 선택하면 된다. 그러면, 그림 6과 같이 최적모형 창에 선택된 최적분할이 나타나게 된다. 즉, 항목1(자가)과 항목3(부와 거주)을 하나의 그룹으로 만드는 것이 최적분할이 된다는 것이다.



<그림 6> 최적분할을 통한 항목그룹화



<그림 7> 최적분할을 통한 항목그룹화
(항목의 수가 4인 경우)

▶ 항목의 수가 4개인 경우의 최적분할을 통한 항목그룹화

그림 7은 항목의 수가 4개인 경우의 최적분할을 통한 항목그룹화의 결과이다. 앞에서 설명한 방법에 의해서 동일하게 수행하면 그림 7과 같은 결과를 얻을 수 있다. 즉, 항목1과 나머지 항목들로 구분하는 것이 최적분할이 됨을 알 수 있다.

4. 결론

본 연구에서는 비주얼 베이직을 활용하여 개인 신용평가에서의 항목그룹화를 위한 소프트웨어를 개발하였다. 모든 절차는 메뉴방식으로 실행 가능하며, 초보자들도 쉽게 사용할 수 있으며, 빠른 사용을 위해 단축아이콘도 제공하였다. 이러한 항목그룹화를 위한 소프트웨어는 매우 많은 경우를 고려해야만 하는 항목그룹화의 비효율적인 절차들을 간단히 해결할 수 있도록 해준다. 뿐만 아니라 금융통계와 관련된 강좌에서 교육용으로도 이용할 수 있으리라 생각된다. 다만, 이번에 개발된 소프트웨어에서는 두 그룹만 분할하는 경우로 제한하였기 때문에 이를 확장시켜 두 그룹 이상의 여러 그룹으로 나누는 기능을 추가하고 입력 자료의 파일형식을 다양화 시키는 작업이 필요할 것으로 생각된다.

참고문헌

1. 강만기 (2002). *Visual Basic에 의한 전산통계학*, 교우사.
2. Thomas, L. C. (2000). A Survey of Credit and behavioral scoring : forecasting financial risk of lending to consumers, *International Journal of Forecasting*, Vol. 16, 149-172.
3. Thomas, L. C., Edelman D. B., Crook, J. N. (2002). *Credit scoring and its application*, SIAM.

[2006년 8월 접수, 2006년 10월 채택]