퍼지 *K*-Nearest Neighbor에 의한 정보검색시스템의 성능 향상

Performance Improvement of Information Retrieval System using Fuzzy K-Nearest Neighbor

현우석
한국성서대학교 정보과학부

Woo-Seok Hyun
Dept. of Information Science, Korean Bible University
E-mail : wshyun@bible.ac.kr

요 약

현대인들이 계속 쏟아지는 정보로부터 자신에게 필요한 정보만을 제한된 시간 안에 검색하는 일은 쉬운 일이 아니다. 컴퓨터를 이용하여 제한된 시간 내에 원하는 정보를 검색하고자 하는 정보검색 분야에서는 성능을 향상시키기 위한 연구가 활발히 진행되어 오고 있다.

본 논문에서는 정보검색 시스템의 성능을 향상시키고자 퍼지 *K*-Nearest Neighbor에 의한 정보검색시스템(IRS-FKNN:Information Retrieval System using Fuzzy *K*-Nearest Neighbor)을 제안한다. 제안하는 시스템은 기존의 시스템과 비교했을 때 검색결과의 신뢰성을 높이게 되어 시스템의 성능을 향상시키게 되었다.

1. 서론

넘쳐나는 정보의 홍수로 휩싸인 현대 사회에서 현대인들은 정보를 수집, 분류, 습득하는데 있어서 엄청난 부담을 지니고 있다. 따라서 현대인들에게 요구되는 가장 중요한 능력 중에 하나는 수많은 정보 가운데서 자신에게 필요한 정보를 정확하고 빠르게 검색하는 것이다. 그러나 계속 쏟아지는 정보로부터 자신에게 필요한 정보만을 한정된 시간 안에 검색하는 것은 쉬운 일이 아니다. 1960년대 초에 원하는 정보를 컴퓨터를 이용하여 제한된 시간 내에 검색하고자 하는 정보검색(information retrieval)[1]이라는 학문 분야가 탄생되었다.

정보검색 분야에서 성능을 향상시키기 위한 연구가 활발히 진행되어 오고 있다. 사전적 정보를 결합하거나 문서간의 링크 정보를 이용하는 방법 등이 있으며, 복합어를 검색에 이용하여 성능을 개선하고자 한 연구도 있다[2]. 또한 클러스터링을 사용하여 효율적인 검색을 하고자 하는 연구도 진행되어 오고 있다[3-4].

본 논문에서는 정보검색 시스템의 성능을 향상시키기 위하여 퍼지 K-Nearest Neighbor[3]를 이용한 정보검색시스템(IRS-FKNN:Information Retrieval System using Fuzzy *K*-Nearest Neighbor)을 제안한다. 제안하는 시스템은 기존의 시스템과 비교했을 때 검색결과의 신뢰성을 높이게 되어 시스템의 성능을 향상시키게 되었다.

2. 퍼지 *K*-Nearest Neighbor

기존의 crisp NN 클래스 분류 방법에서는 클래스 라벨을 임의의 입력 패턴에 대해 가장 가까

운 하나의 패턴의 클래스 라벨에 따라 할당하고 있다[3,5]. 그러나 이러한 방법의 단점으로는 선택된 K개의 패턴들이 주어진 패턴의 클래스를 분류할 때 모두 동일한 정도로 기여한다는 것을 들 수 있다. 이것은 패턴들이 복잡하게 겹쳐있는 지역에서의 클래스를 분류하는데 있어서 오분류의 원인을 제공하기도 한다. 이러한 단점을 보완하기 위해서 퍼지 K-NN[3]가 제안되었다. 이 방법에서는 K-NN 방법을 수행하기 전에 각 패턴들에 주변 K개의 가장 가까운 패턴들의 클래스에 따라 멤버쉽 값을 할당하게 된다. 이러한 방법에 의해 퍼지 집합으로 확장된 패턴에 대하여 K-NN방법을 수행하게 된다. 이러한 퍼지 집합으로의 확장은 서로 다른 클래스의 패턴들이 겹쳐서 분포되어 있는 부분에서의 오분류율의 감소를 가져오게 된다[6].

퍼지 K-NN 알고리즘은 주어진 패턴 데이터들에 대해서 각각 클래스별로 멤버쉽을 할당하게 된다. 이렇게 할당된 멤버쉽들은 클래스 분할에 사용될 클래스 멤버쉽을 결정할 때, 주어진 패턴이 얼마나 기여할 것인지를 나타내게 된다. 즉, K개의 가장 가까운 패턴들로부터 해당 입력 패턴의 거리 함수와 선택된 K개의 패턴들이 각 클래스에 대해서 갖는 초기 멤버쉽을 기초로 해서 클래스 분할에 사용될 멤버쉽을 할당하게 된다[3]. 입력 패턴 x에 할당된 멤버쉽은 다음과 같다.

$$u_i(x) = \frac{\sum_{j=1}^{K} u_{ij}\left(1/\|x-x_j\|^{2/(m-1)}\right)}{\sum_{j=1}^{K}\left(1/\|x-x_j\|^{2/(m-1)}\right)} \quad (1)$$

여기서 u_{ij}는 클래스 라벨이 붙은 패턴 데이터들 중에서 j번째 패턴의 i번째 클래스에 대한 멤버쉽을 나타낸다. 식 (1)에서 보여진 것처럼 패턴 x에 할당된 멤버쉽은 주어진 패턴에 가장 가까운 K개의 패턴들과의 거리의 역수와 이들의 클래스 멤버쉽에 의해 영향을 받는다. 즉, 입력 패턴으로부터 거리가 가까울수록 거리의 역수는 입력 패턴의 멤버쉽에 좀더 많은 가중치를 제공하게 된다. 식 (1)에서 사용된 클래스 라벨이 붙은 초기 패턴들에 대한 멤버쉽을 결정하기 위해서는 초기화 과정이 필요한데, 식 (2)는 여기서 제공되는 초기 클래스 멤버쉽을 제공하기 위한 합리적인 방법을 나타낸다[3].

$$u_j = \begin{cases} 0.51+(n_j/K)*0.49, & if \quad j=i \\ (n_j/K)*0.49, & if \quad j\neq i \end{cases} \quad (2)$$

여기서, n_j는 K개의 가장 가까운 패턴들 중 j번째 클래스로 라벨이 붙은 패턴들의 개수이다.

3. 퍼지 K-Nearest Neighbor를 이용한 정보검색시스템

기존의 crisp NN 클래스 분류 방법에서는 클래스 라벨을 임의의 입력 패턴에 대해 가장 가까운 하나의 패턴의 클래스 라벨에 따라 할당하여 선택된 K개의 패턴들이 주어진 패턴의 클래스를 분류할 때 모두 동일한 정도로 기여한다는 것을 문제점으로 들 수 있다. 이것은 패턴들이 복잡하게 겹쳐있는 지역에서의 클래스를 분류하는데 있어서 오분류의 원인이 되기도 하는데 이러한 단점을 보완하기 위해서 퍼지 K-Nearest Neighbor[3]를 이용한 정보검색시스템(IRS-FKNN:Information Retrieval System using Fuzzy K-Nearest Neighbor)을 제안한다.

3.1 시스템의 구조

퍼지 K-Nearest Neighbor[3]를 이용한 정보검색시스템(IRS-FKNN:Information Retrieval System using Fuzzy K-Nearest Neighbor)의 구조는 그림 1 과 같다.

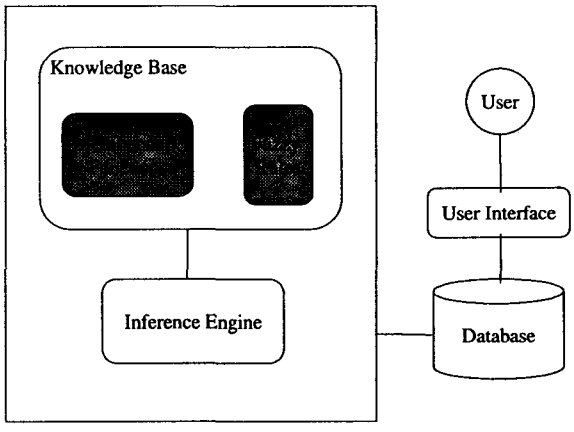

그림 1 IRS-FKNN의 구조

3.2 지식베이스

지식베이스는 사용자가 필요로 하는 정보를 검색하기 위한 퍼지 규칙과 각 패턴들에 주변 K개의 가장 가까운 패턴들의 클래스에 따라 멤버쉽 값을 할당한 멤버쉽 함수로 구성되어 있다. 본

시스템에서는 추론 엔진이 제공하고 있는 규칙에 대한 명세에 따라 퍼지 규칙을 작성하였다.

3.3 추론엔진

추론엔진에서는 사용자로부터 질의가 들어왔을 때 추론 과정을 통해서 사용자에게 추론 결과를 제공하여 주는 부분이다. 다시 말해서 knowledge base에 사실과 규칙을 저장하고 사용자가 질의를 하면 주어진 knowledge base를 이용하여 질의에 대한 답을 추론하여 제공해준다.

3.4 사용자 인터페이스

사용자가 정보를 원활하게 검색할 수 있도록 지원해 주는 기능을 제공한다. 사용자가 필요로 하는 정보를 찾기 위하여 질의를 입력하게 하고 적절한 결과를 사용자에게 제시해 준다.

4. 평가

본 시스템의 성능을 평가하기 위하여 crisp K-NN 클래스 분류방법에 의한 정보검색시스템(IRS-KNN)과 퍼지 K-Nearest Neighbor에 의한 정보검색시스템(IRS-FKNN)에서 결과를 비교한다. 이를 위하여 "twoclass" 데이터를 사용하여 결과를 보인다. 초기 K={1, 2, 3, 4, 5, 6, 7, 8, 9}를 사용하여 각각 패턴 데이터에 대한 초기 멤버쉽을 할당하여 실험을 수행하였다. 그림 2는 IRS-KNN과 IRS-FKNN에서 평균 오분류의 개수를 나타낸다.

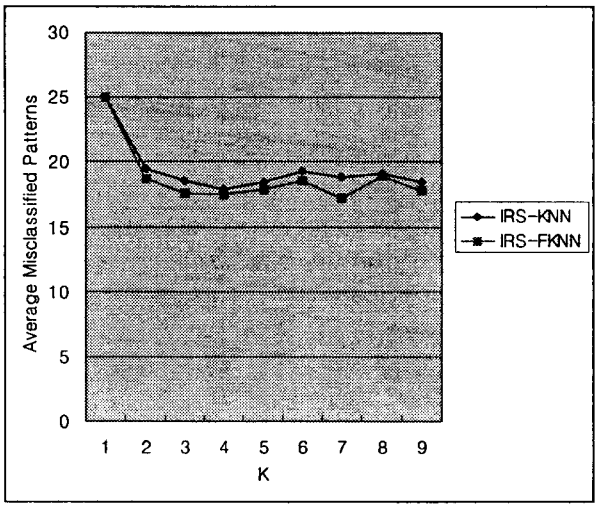

그림 2 시스템간의 "twoclass" 데이터의 평균 오분류 개수

그림 2에서 보듯이 IRS-KNN 보다는 IRS-FKNN에서 평균 오분류의 수를 감소시킴에 따라 검색결과의 신뢰성을 높이게 되어 시스템의 성능을 향상시키게 되었다.

5. 결론 및 향후과제

제안하는 시스템(IRS-FKNN)은 기존의 시스템(IRS-KNN)과 비교해 볼 때 평균 오분류의 수를 감소시킴에 따라 검색결과의 신뢰성을 높이게 되어 시스템의 성능을 향상시키게 되었다. 이것은 crisp K-NN에서 선택된 K개의 패턴들이 주어진 패턴의 클래스를 분류할 때 모두 동일한 정도로 기여하게 되어 패턴들이 복잡하게 겹쳐있는 지역에서의 클래스 오분류의 원인이 되기도 하는데, 이것을 제거하기 위하여 퍼지 K-NN 방법을 사용하였기 때문이다.

향후 연구과제로는 초기의 K의 선택이 신뢰성 저하의 원인이 될 수도 있는데 이 점을 개선시키고자 하는 차후 연구가 요구된다.

6. 참고문헌

[1] Rijsbergen, C. J. van, Information Retrieval, 2nd edition, butterworths, 1979.

[2] John Bear and David Martin, Using Information Extraction to Improve Document Retrieval, Text Retrieval Conference, 1996.

[3] J. Keller, M. Gray, and J. Givens, JR, "A fuzzy K-nearest neighbor algorithm," IEEE Trans. Syst., Man, Cybern., vol. 15, no. 4, pp. 258-263, August 1985.

[4] H. Choe, J. B. Jordan, On the Optimal Choice of Parameters in a Fuzzy C-Means Algorithm, IEEE International Conference on Fuzzy Systems, San Diego, pp. 349-354, 1992.

[5] J. Tou and R. Gonzalez, pattern Recognition Principles, Addison-Wesley, 1974.

[6] 황철, 이정훈, "Interval 제2종 퍼지 K-Nearest Neighbor, " 한국퍼지및지능시스템학회 2002 추계학술발표논문집, pp. 271-274, 2002.