氏名	楊（ヨウ） 振君（シン クン）
博士の専攻分野の名称	博士（医学）
学位記番号	第 17665 号
学位授与年月日	平成15年3月25日
学位授与の要件	学位規則第4条第1項該当 医学系研究科生体統合医学専攻
学位論文名	Similar Cases Retrieval From the Database of Laboratory Test Results （検体検査結果データベースからの類似症例の検索）
論文審査委員	（主査） 教授 武田 裕 （副査） 教授 網野 信行　教授 森本 兼曩

論文内容の要旨

Objectives

In medical practice doctors are often confronted with difficult cases and want to refer to some helpful information. Initially, rule-based systems had been developed and proved to be effective to some extent. However, doctors felt it was difficult to express diagnostic logic as generally applicable rules required by a decision support system. This is because doctors make a diagnosis not only by the rules clearly demonstrated in textbooks, but also by evoking their memory of similar patients encountered before. These days, a huge volume of information about patients and medical processes has been stored in clinical databases, which can be compared to doctors' memory. The aim of this study is to find the suitable method to retrieve similar cases from the clinical database. The subjects of this study are the laboratory test data which are basically numerical and ordinal data. There are some problems to handle laboratory test data. First, there are some variables whose distributions can't be identified, thus appropriate data transformations can't be chosen. Second, some variables contain numerous data like "<5", ">100", which can't be calculated directly. Third, because some variables correlated each other, simple similarity measure may not be proper. In this study, a new method was proposed, by which these problems were resolved.

Methods and results

1. The method of retrieving the similar cases and its validation

The values for each variable were sorted in ascending order. A rank was assigned to each value accordingly and in progressive order. The rank r_{if} of the i th case in the fth variable was converted to Z_{if} ; $Z_{if}=(r_{if}-1)/(M_f-1)$, where M_f is the highest rank for variable f. Z_{if} lied between 0 and 1. By using this score, a distance between two cases was calculates.

Blood count data (WBC, RBC, hemoglobin, hematocrit and platelet) of 3000 cases were used to validate the method. From the data set, 100 sample cases were selected at random. The distance between each sample

case and other cases were calculates. Euclidean distance and Mahalanobis distance using the score were compared with Mahalanobis distance using raw data which is criterion of similarity measure.

Among the most similar 20 cases retrieved by using Mahalanobis distance calculated form the rank scores, 95% were coincident with those of the criterion. However, when using Euclidean distance, the coincident rates was 70%.

2. Evaluation of the usefulness of the method

The data relevant to thyroid diseases (TSH, FT4, FT3, Tg, TrAb, TgAb, and McAb) of 1655 cases was used to evaluate the usefulness of the method. From the data set, 96 cases with abnormal data were selected at random. For each sample cases, the distance between the sample case and the other cases were calculated.

To study the proper number of cases to be retrieved, 5, 10, 15 and 20 of the most similar cases were retrieved by using Mahalanobis distance and we checked whether the diagnosis of the retrieved cases were consistent with that of the sample case. According to the increase in the number of retrieved cases, the "hit rate" (percentage of the sample cases that there was at least one case in the retrieved ones whose diagnosis was consistent with the sample case) increased. On the other hand, the "consistent rate" (the rate of the cases whose diagnosis was consistent with the sample case in the retrieved ones) decreased. The "consistent rate" was 32.4% and the "hit rate" was 76.0% when the number of retrieved cases was 10, which was adopted as suitable number.

To study if using Mahalanobis distance as a similarity scale yields better result than Euclidean distance, the consistent rate was compared. The "consistent rate" when Euclidean distance was used was 27.7% which was less than the rate arrived at by using Mahalanobis distance (32.4%). Thus the Mahalanobis distance was superior to Euclidean distance as a similarity measure.

To study the significance of distance value, we compared the "consistent rates" when the cut off points of the distance value were set at 5, 3, and 2. The "consistent rates" were 31.0%, 28.9%, and 25.4%, when the cut off points of the distance value were set at 5, 3, and 2, respectively. There results suggest that the distance value itself does not have significant meaning.

Conclusions

The new method for retrieving the similar cases from the database of laboratory test results was proposed. The data of laboratory test results include numerical data and ordinal data. All the raw data were transformed into ranks and further into the scores ranged from 0 to 1, and Mahalanobis distance was calculated as the similarity measure. It was verified by using numerical data that this data transformation did not affect the result of similar case retrieval. This data transformation make it possible to calculate the distance in any laboratory test results data. Calculation of Mahalanobis distance is more complicated than Euclidean distance. But, the results using Mahalanobis distance were superior to that using Euclidean distance. As to the similar case retrieval, the distance value itself did not have a significant meaning thus the cut off point of distance need not be set. This method provides the basis for the retrieving the similar cases from the clinical database which consist of semi-quantitative data set.

論文審査の結果の要旨

本研究は、データベースから類似症例を検索することにより、診断支援として有用な情報を提供することを目指したシステムの開発の、基礎となる方法論を確立したものである。本研究では、検体検査結果のデータベースからの類似例の検索を課題としている。検体検査結果データには、分布が同定しにくいものが含まれていること、数字データだけでなく順序データも含まれていること、項目間に相関関係がある場合があること等の問題を含んでいる。これら

の問題を、全てのデータを順序データに置き換え、更にランクスコアに変換し、この値からマハラノビス距離を求める方法により解決した。この方法の検証を、全てが数字データである末梢血液像のデータを用いて行い、ランクスコアへの変換は、類似検索として利用する場合には影響が少ないことが確認された。本法を、甲状腺疾患に関連する項目のデータに対して適応し、本法による類似検索の有用性が評価された。本研究により提示された方法は、類似検索の方法として斬新であり、また、評価方法としても適当であり、学位に値するものと認める。