Possibilistic Discrimination Analysis: Medical Diagnosis from Patient Records

A. Sikharulidze
Computer Software and Informational Technologies Chair, Iv. Javakhishvili Tbilisi State University

Abstract

In the present work a method of decision classification is described which constructs a numerical tabular knowledge base from historical cases and is the variation of discrimination analysis. The method processes the data described by the doctor and enables to effectively employ full information available.

Introduction

Among many expert medical diagnostic systems which use numerical-tabular base the most popular method is probably the method based on the Bayesian inference technique. But in many cases it turned out that Bayesian analysis demonstrates some difficulties. First of all this is the difficulty to calculate so called "prior" probabilities, and the second: Bayesian analyses can be useful only in such situations when the data is objective by its nature (for example we don't need an expert to determine the sex of patient: male or female), but in medicine very often there arises such situations that we need an expert (doctor) to determine some other characteristics of patient (symptoms) are present or not and how strongly they are exhibited. In such cases Bayesian method is helpless and its certainty is very low. This means that other ways must be searched for.

One of the known alternative methods is called the discrimination analysis [1] and uses the theory of fuzzy sets. The method is briefly described in the following section.

1. Discrimination Analysis

The knowledge base represents the list of historical patient records, where the symptoms, exhibited by these patients along with their proven diagnosis are recorded. From this information a new frequency distribution table is established, where i denotes the i-th symptom and j denotes the j-th decease, and - f_{ij} proportion of those recorded as suffering from disease j who exhibited symptom i. In the following table D_j denotes j-th decease and S_i - i-th symptom.

	D_1	...	D_{C_D}
S_1	f_{11}	...	f_{1C_D}
...	...	...	...
S_{C_s}	f_{C_s1}	...	$f_{C_sC_D}$

Based on this table two other tables are constructed: positive discrimination table $\{p_{ij}\}$ and negative $\{n_{ij}\}$ discrimination table, which are calculated as follows:

$$p_{ij} = \sum_{\substack{k \in D \\ K \neq j}} \{\chi_{Large-ratio}(f_{ij} / f_{ik})\} / (C_D - 1), \quad (1)$$

$$n_{ij} = \sum_{\substack{k \in D \\ K \neq j}} \{\chi_{Large-ratio}(f_{ik} / fj)\} / (C_D - 1), \quad (2)$$

where $p_{ij}, n_{ij} \in [0,1]$, and C_D denotes the cardinality of the disease set. Large-ratio is defined as a fussy set with membership characterizing function:

$$\chi_{Large-ratio} = R^+ \to [0,1].$$

An explanation of the positive and negative discrimination measures is that p_{ij} represents the accumulated belief that symptom i is more indicative of disease j than any of the remaining diseases, whilst n_{ij} represents the belief that symptom i is more indicative of not disease j than any of the other diseases.

When the records from new patients arrive, set of symptoms S exhibited by him is entered into the system. The simple technique for procuring a diagnosis is to select from the tables $\{p_{ij}\}$ and $\{n_{ij}\}$ only those rows corresponding to *S*, giving new tables $\{p'_{ij}\}$ and $\{n'_{ij}\}$. A diagnosis can be defined as a distribution over the diseases $\{\delta_j\}$ as follows:

$$\delta_j = \frac{1}{2}\{\chi_{Large}(\pi_j) + \chi_{Small}(\nu_j)\}, \quad j \in D, \quad (3)$$

where

$$\pi_j = \left\{\sum_i p'_{ij}\right\} \Big/ C_S \quad , \quad \nu_j = \left\{\sum_i n'_{ij}\right\} \Big/ C_S$$

and C_S denotes the cardinality of S.

π_j and ν_j represent the average of the positive and negative discrimination measures respectively, for disease j. The fuzzy sets Large and Small have characteristic membership functions:

$$\chi : [0,1] \to [0,1],$$

such that χ_{Large} is monotonic increasing and χ_{Small} is monotonic decreasing in its argument.

The disease j with maximum magnitude in $\{\delta_j\}$ can be interpreted as the most believable diagnosis.

This method was successfully used in Psychiatry [2], however there were several problems that arose. First of all that was the difficulty to calculate f_{ij}. In the patient records there wasn't directly stated weather the patient exhibited the symptom or not. Instead there was indicated the doctor's estimation of how strongly he believed that patient exhibited the symptom. This belief was ranked from 0 to 5, 0 meaning that the patient didn't exhibit the symptom at all, 5 meaning that the exhibition of symptom was very strong, 1 meaning that patient exhibited symptom very weakly, etc. First we assumed that if the exhibition estimation was not zero, it should be believed that the person exhibited the symptom, and the discrimination analysis was processed this way. But on one hand very week exhibition of some symptom can be viwed as not exhibiting that symptom at all rather then exhibiting it and on the other hand some diseases

are characterized by only strong exhibition of some symptoms or vice versa, and for such diseases the accuracy of our expert systems was not satisfactory. Because the calculation of frequencies became impossible in such situations, some other characterizing factors should have been searched for.

2. Possibilistic Discrimination Analysis

In the situations when the information is obscure and obtained or described by some expert (the doctor in our case), such notion as Fuzzy Expected Value (*FEV*)[3] is believed by many authors to be one of the best characterizing value for the population set. According to [3] *FEV* is defined as follows:

DEFINITION 1[3]: FEV of a compatibility function $\chi_{\tilde{A}}$ of the fuzzy subset $\tilde{A}$ with respect to the fuzzy measure g is Sugeno's integral over X:

$$FEV(\chi_A) = \int_X \chi_{\tilde{A}} \circ g(\cdot) \equiv \sup_{T\in[0,1]} \{T \wedge g(H_T)\}, \qquad (4)$$

where $\wedge$ indicates a minimum of two arguments.

Consider the situation where $X = \{x_1, x_2, ..., x_n\}$ is a finite set arranged in the following way: $\chi_{\tilde{A}}(x_1) \le \chi_{\tilde{A}}(x_2) \le \cdots \le \chi_{\tilde{A}}(x_n)$. Denote: $X_i = \{x_i, ..., x_n\}, i = 1,2,...n.$ As known, the *FEV* can be calculated so [4]:

$$FEV = \max_i \{\chi_{\tilde{A}}(x_i) \wedge g(X_i)\} = \min_i \{\chi_{\tilde{A}}(x_i) \vee g(X_i)\}, \qquad (5)$$

where $\vee$ - is a maximum of two arguments.

According to this procedure for the finite set for every disease and symptom *FEV* can be easily calculated, where X will be the set of patients that suffered from given disease, the uniform distribution can be used in the case of fuzzy measure g, and $\chi_{\tilde{A}}(x_i)$ shall be the compatibility values estimated by the expert.

Following this technique the new table will be obtained, which we call the *FEV* Distribution Table.

	D_1	...	D_{C_D}
S_1	FEV_{11}	...	FEV_{1C_D}
...	...	...	...
S_{C_s}	FEV_{C_s1}	...	$FEV_{C_sC_D}$

Different from the discrimination analyses the new patient that arrives along with his symptom pattern will also have the compatibility values of the symptoms that means that this values must somehow participate in final diagnosis as well as in calculation of positive and negative probabilistic discrimination values. For a given a patient with a particular compatibility values for each symptom ($\mu_1, ..., \mu_n$), the positive and negative probabilistic discrimination values will be calculated as follows:

$$pp_{ij}(\mu_i) = \sum_{\substack{k\in D \\ K\neq j}} \{\chi_{Large-ratio}\left(\frac{FEV_{ik} - \mu_i}{FEV_{ij} - \mu_i}\right)\}/(C_D - 1), \qquad (6)$$

$$pn_{ij}(\mu_i) = \sum_{\substack{k \in D \\ k \neq j}} \{ \chi_{Large-ratio}\left(\frac{FEV_{ij} - \mu_i}{FEV_{ik} - \mu_i} \right) \} / (C_D - 1), \qquad (7)$$

The final diagnosis can be calculated as it is done in discrimination analysis:

$$\delta_j = \frac{1}{2} \{ \chi_{Large}(\pi_j) + \chi_{Small}(\nu_j) \}, \quad j \in D, \qquad (8)$$

where

$$\pi_j = \left\{ \sum_i pp_{ij}(\mu_i) \right\} \Big/ C_S, \quad \nu_j = \left\{ \sum_i np_{ij}(\mu_i) \right\} \Big/ C_S .$$

3. EXAMPLE

Suppose we have only two diseases D_1 and D_2, both are characterized by only two symptoms S_1 and S_2. Also the following information is available: five patients who suffered from D_1 exhibited S_1 with compatibility value 0.8, three with 0.6 and two with 0.9. Six of these patients exhibited S_2 with compatibility value 0.1, two with 0.3 and other two 0.4. Six patients who suffered from D_2 exhibited S_1 with compatibility value 0.1, three with 0.2 and one with 0.4. Five of these patients exhibited S_2 with compatibility value 0.2, four with 0.1 and one with 0.3.

Suppose the new patient arrives exhibition of S_1 for him is evaluated as 0.9, and S_2 as 0.1. It is obvious that first disease is characterized by higher exhibition of S_1 then D_2 that means that this patient must have suffered from first disease. But if we try to use here discrimination analysis, we can't get any result. Both symptoms were actually exhibited during both diseases and frequencies for each equal to 1. That means that with discrimination analysis we will obtain results $\delta_1 = 0.5$ and $\delta_2 = 0.5$ meaning none of the diseases can be preferable.

Now lets apply the possibilistic discrimination analysis and calculate *FEV*s.

As described in [5] for calculation of FEV_{11} we can build the following table:

# of group	n_i	χ_i	$n^{(i)}$	$g_i = n^{(i)}/n$	$\chi_i \wedge g_i$
1	3	0.6	10	1	0.6
2	5	0.8	7	0.7	0.7
3	2	0.9	2	0.2	0.2

where n_i is the number of people in i-th group: $n^{(i)} = \sum_{j=1}^{n} n_j$, $i = 1,2,...,n$, $n = 5$. Thus the most typical is the second group and FEV_{11}=0.8.

For calculation of FEV_{21} we have the following table:

# of group	n_i	χ_i	$n^{(i)}$	$g_i = n^{(i)}/n$	$\chi_i \wedge g_i$
1	6	0.1	10	1	0.1
2	2	0.3	7	0.4	0.3
3	2	0.4	2	0.2	0.2

Thus the most typical is the second group and FEV_{21}=0.3.

For calculation of FEV_{12} we have the following table:

# of group	n_i	χ_i	$n^{(i)}$	$g_i = n^{(i)}/n$	$\chi_i \wedge g_i$
1	6	0.1	10	1	0.1
2	3	0.2	4	0.5	0.2
3	1	0.4	1	0.1	0.1

Thus the most typical is the second group and FEV_{12}=0.2.

For calculation of FEV_{22} we have the following table:

# of group	n_i	χ_i	$n^{(i)}$	$g_i = n^{(i)}/n$	$\chi_i \wedge g_i$
1	4	0.1	10	1	0.1
2	5	0.2	6	0.6	0.2
3	1	0.3	1	0.1	0.1

Thus the most typical again is the second group and FEV_{22}=0.2.

So *FEV* Distribution Table will look this way:

	D_1	D_2
S_1	0.8	0.2
S_2	0.3	0.2

For our patient with $\mu_1 = 0.9$ and $\mu_2 = 0.1$, we can easily calculate the positive possibilistic and negative possibilistic values (Suppose $\chi_{Large-ratio}(x) = x/10$):

$$pp_{11}(\mu_1) = np_{12}(\mu_1) = \chi_{Large-ration}\left(\frac{|0.9-0.2|}{|0.9-0.8|}\right) = \chi_{Large-ration}(7) = 0.7,$$

$$np_{11}(\mu_1) = pp_{12}(\mu_1) = \chi_{Large-ration}\left(\frac{|0.9-0.8|}{|0.9-0.2|}\right) = \chi_{Large-ration}(0.142) = 0.0142,$$

$$pp_{21}(\mu_2) = np_{22}(\mu_2) = \chi_{Large-ration}\left(\frac{|0.2-0.1|}{|0.3-0.1|}\right) = \chi_{Large-ration}(0.5) = 0.05,$$

$$np_{21}(\mu_2) = pp_{22}(\mu_2) = \chi_{Large-ration}\left(\frac{|0.3-0.1|}{|0.2-0.1|}\right) = \chi_{Large-ration}(2) = 0.2.$$

Afterwards,

$$\pi_1 = \nu_2 = \frac{0.7-0.05}{2} = 0.375,$$

and

$$\pi_2 = \nu_1 = \frac{0.0142-0.2}{2} = 0.1071.$$

Now, using (8) we can do the following calculations (let $\chi_{Large}(x) = x$ and $\chi_{Small}(x) = 1-x$).

$$\delta_1 = \frac{1}{2}(0.375 + 0.9029) = 0.63895,$$

$$\delta_2 = \frac{1}{2}(0.1071 + 0.625) = 0.36605.$$

These results give us the possibility to judge that it's more believable that given patient suffered from first disease.

4. Conclusion

It should be underlined that this method uses full information that is available. But for further analysis it should be also mentioned that here *FEV* distinguishes the only group with the chosen compatibility value. This way *FEV* avoids the other groups that also may not be very

convenient. For this case such notion as Weighted Fuzzy Expected Value (*WFEV*) [4] or Generalized Weighted Fuzzy Expected Value (*GWFEV*) [5] can be used.

REFERENCES

1. D.Norris, B.W.Pilsworth, J.F.Baldwin, *Medical Diagnosis form Patient records – A Method Using Fuzzy Discrimination and Connectivity Analysis*., Fuzzy Sets and Systems 21 (1989), 37-45.
2. A.Sikharulidze, *Application of Discrimination and Connectivity Analysis in Psychiatry* – Bulletin of Georgian Academy of Sciences., 2001.
3. Kandel A. *On the Control and Evaluation of Uncertain Processes*. IEEE Trans. on Automatic Control , vol AC-25, No.6 (1980), 1182-1187.
4. Friedman M. Schneider M., Kandel A. *The use of Weighted Fuzzy Expected Value (WFEV) in Fuzzy Expert Systems*. Fuzzy Sets and Systems 31 (1989) 37-45.
5. Sirbiladze G., Sikharulidze A. *Insufficient Expert data and Fuzzy Averages*, Applied Mathematics and Informatics, 2001.