UDC: 004.891

Making decision with fuzzy data

Anna Sikharulidze
Tbilisi State University, Department of Applied Mathematics and Computer Science,
Chair of Software and Information technologies, Univreisty St,2.

Abstract:
In the following work two fuzzy methods of creating numerical-tabular knowledge base and inferring from it are researched – discrimination and connectivity analyses[1]; their positive and negative aspects are discussed and according to this information the new method is elaborated - generalized discrimination analysis, which tries to process all information which was neglected in other two methods for some reason. The new method is generalization for classical discrimination analysis for fuzzy subsets of activities. The characteristic - frequency, which was used in classic discrimination analysis is replaced by fuzzy characteristic – Most Typical Value (MTV) [2], such as Fuzzy Expected Value (FEV)[3-5], Weighted Fuzzy Expected Value (FWEV)[6] and Generalized Weighted Fuzzy Expected Value (GWFEV)[7-8]. Respectively the method of decision-making is generalized, while the main idea remains the same – positive discrimination represents the belief that the activity is more indicative of the decision, than any of the remaining decisions, and negative – vice-versa. The proposition of generalization in the case when MTV=FEV and fuzzy measure is sampling distribution is proved. An example of use of generalized discrimination analysis is demonstrated.

Keywords*: Discrimination analysis, connectivity analysis, fuzzy sets, expert systems.*

1.INTRODUCTION

Recently many expert systems have been created in different scientific branches. Generally expert systems differ by the methods of inferring and creation of knowledge-base. There exist two different kinds of expert systems according to how the knowledge base is compiled:

- Expert systems based upon numerical-tabular knowledge base,
- Rule-based expert systems.

The expert knowledge in rule-based expert systems typically is in the form of a set of IF-THEN rules. Creating rule-based system means collecting the information by interviewing experts, and in many cases this causes reconciling partially inconsistent data. Incompleteness of knowledge rules is the main problem for such approach. Main advantage of rule-based systems over those, which use numerical-tabular knowledge base, is that they naturally reflect the semantics of problem and they can be relatively easily extended by adding dialogue and comments.

Although, it must be mentioned here, that despite of the existence of expert system shell, active human involvement is needed for creation of rule-based expert system. These humans are at least two experts. One of them is knowledge engineer, and the other (one or several) is the expert from the scientific branch, for which the expert system is created. These humans must have mutual understanding and desire of collaboration. Both of them must spend enough time for the problem, elaborate rules (which often is not that easy), test the rules, etc. Moreover, often experts (for example doctors) have no desire to formalize their experience. Because of lack of time long interview with experts is not possible. Explanation may be inconsistent and uncertain. The knowledge engineer must gain knowledge in area, which was studied by expert during years. In such cases usefulness of expert system depends on experience of knowledge engineer. This means that modeling of the knowledge and experience of one expert is performed using the knowledge and experience of another expert.

The work of knowledge engineer much eases when he creates expert system based on numerical-tabular knowledge base. In this case the expert gives him documents where cases already solved by expert are recorded. In medicine these may be the patient historical records where exhibited symptoms along with proven diagnoses are recorded. The expert gives additional explanation only when needed. Elaborating these documents is the prerogative of expert system shell, and thus, expert system shell itself creates numerical-tabular knowledge base.

Among many expert systems which use numerical-tabular knowledge 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. One of these is that Bayesian analysis is useful only in such situations when the data is objective by its nature, but in some scientific branches we need expert to determine if the activity was exhibited and how strongly. When data becomes subjective, i.e. fuzzy, certainty of Bayesian method is low and other ways must be searched for.

Alternative methods [1] were adopted by D.Norris, P.Pitsworth and J.baldwin called – discrimination analysis and connectivity analysis. In discrimination analysis activities are ranked according to how well they discriminate for each decision compared with other decisions. In the connectivity analysis sets of activities are established for each decision which represents an ideal pattern indicative of that disease.

In the present work discrimination analysis is discussed and its generalized version – Generalized Discrimination analysis is elaborated for fuzzy subsets of activities and compared to classic Discrimination analysis. The proposition of generalization is proved. Furthermore an example is shown demonstrating the use of Generalized Discrimination analysis.

2.DISCRIMINATION ANALYSIS

For general purposes let us consider that our objective is to reason according general set of activities. The information is read from general database, where historical data with exhibited activities and correct decisions are recorded.

>From the information in database the frequency distribution table is established, where i denotes the i-th activity and j denotes the j-th decision, and - f_{ij} proportion of cases when j-th decision was correctly stated and i-th activity was exhibited. In the following table D_j denotes j-th decision and A_i - i-th activity, C_D denotes cardinality of the set of decisions and C_A denotes cardinality of the set of activities.

	D_1	...	D_{C_D}
A_1	f_{11}	...	f_{1C_D}
...	...	...	...
A_{C_A}	f_{C_A1}	...	$f_{C_AC_D}$

For each activity and decision positive discrimination and negative discrimination is calculated according to the formulas:

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

$$n_{ij} = \sum_{\substack{k\in D \\ k\neq j}} \left\{ \chi_{Large-ratio}\left(\frac{f_{ik}}{f_{ij}}\right)\right\} /(C_D - 1),$$

where $p_{ij}, n_{ij} \in [0,1]$; Large-ratio is the fuzzy subset with compatibility function: $\chi_{Large-ratio} : R^+ \to [0,1]$, mapping the positive real numbers, representing ratios, into the interval [0,1].

The explanation of the positive and negative discrimination measures is that p_{ij} represents belief that activity i is more indicative of decision j than any of the remaining decisions, whilst

n_{ij} represents the belief that activity i is more indicative of not decisions j than any of the other decisions.

Given a case with a particular activity subset { A_j } we select from the tables { p_{ij} } and { n_{ij} } only those rows corresponding to { A_j }, producing new tables { p'_{ij} } and { n'_{ij} }. A decision can be defined as a distribution over decisions { δ_j } as follows:

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

where

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

and C_A denotes cardinality of { A_j }.

π_j and ν_j represent the average of the positive and negative discrimination measures respectively, for decision j. The Fuzzy sets Large and Small have characteristic compatibility functions: $\chi : [0,1] \to [0,1]$, where $\chi_{L\arg e}$ is monotonic increasing, and χ_{Small} – monotonic decreasing in its argument.

The decision j with maximum magnitude in { δ_j } can be interpreted as the most believable decision.

3. GENERALIZED DISCRIMINATION ANALYSIS

As described above, in discrimination analysis it is necessary to build the frequency distribution table according to the information from general purpose database. Each element of table f_{ij} is the frequency of cases where *i*-the decision was correctly stated and *j*-th activity was exhibited. To calculate this frequency, we must know definitely was the activity performed or not. But in many real situations this cannot be stated definitely. For example in medical diagnostics if the activity is the symptom “severe pain”, it is difficult to determine was it exhibited or not. In such cases it necessary that patient or doctor estimates the symptom. For example in Psychiatry [9] the doctor has scale from 0 to 5 and he estimates the level of exhibition of symptoms by the patient. Obviously such information is fuzzy and symptoms exhibited by patient represent the fuzzy subset of the set of symptoms.

In such cases calculation of frequencies becomes difficult. This problem can be solved one way, if we consider every activity exhibited if it wasn’t estimated as 0 (but, this way we lose part of information). On the other hand, it is difficult to consider exhibited activity which was estimated by 1, and activity estimated by 5, moreover, some diseases may be characterized by strong exhibition of some symptom.

Obviously, for such cases the use of discrimination analysis is possible only if we generalize it for fuzzy subsets of activities.

At first sight, this problem is solved by connectivity analysis. The initial information is considered as fuzzy; the values in incidence matrix[1] are expert estimations of exhibited activity for particular case (doctor’s estimation of symptom exhibited by patient). But it must be also mentioned that in connectivity analysis, as well as in discrimination analysis during inference process the authors require that subset of symptoms, exhibited by patient, for which decision should be made, be crisp, i.e. Above mentioned { A_j } is crisp subset. This is difficult to be achieved in many situations. It must be also said, that during connectivity analysis so called chains of connection are established, which consist of some amount of activities. That is, symptoms that are less connected are neglected. In discrimination analysis, if we had some information about each activity in positive and negative discrimination tables, in connectivity analysis there will be several symptoms, that are not included in connectivity chains, and, correspondingly, we don't have any information about them. Thus, again, we loose part of information.

According to the above discussion, our problem was to elaborate method, which would consider initial information, as well as information, for which decision must be made, as fuzzy. Alongside, it was preferred that information is retained for each activity. These means, that we should maximally use all information available.

As already mentioned, one of positive sides of discrimination analysis is that information is retained about every activity, from one side how much they are indicative of some decision compared with other decisions, on the other hand – how contradictory. But to represent how indicative the activity is for decision, in discrimination analysis frequency is used, but for fuzzy subsets calculation of frequencies is a problem. Respectively frequency – as the characteristic of representativeness of activity must be changed by other characteristic. For such can be chosen Most Typical value, so called MTV[2], which indicates how much the activity is typical for the decision.

Instead of frequency distribution table, other table is constructed which we call MTV distribution table:

$$
\begin{array}{cccc}
 & D_1 & \dots & D_{C_D} \\
A_1 & MTV_{11} & \dots & MTV_{1C_D} \\
\dots & \dots & \dots & \dots \\
A_{C_A} & MTV_{C_A 1} & \dots & MTV_{C_A C_d}
\end{array}
$$

Generalized positive and negative discrimination values are calculated like classical discrimination analysis:

$$gp_{ij} = \sum_{\substack{k \in D \\ K \neq j}} \left\{ \chi_{L\arg e-ratio}\left(\frac{MTV_{ij}}{MTV_{ik}} \right) \right\} / (C_D - 1),$$

$$gn_{ij} = \sum_{\substack{k \in D \\ K \neq j}} \left\{ \chi_{L\arg e-ratio}\left(\frac{MTV_{ik}}{MTV_{ij}} \right) \right\} / (C_D - 1),$$

where MTV_{ij} it the most typical value of i-th activity for j-th decision. The process of decision-making is consists of following steps:

1. FUZZIFICATION**:** Suppose, given case is represented with the fuzzy subset with the following compatibility values:

$$\{\mu_1, \mu_2 \dots, \mu_{C_A}\}$$

2. INFERENCE**:** Positive and negative discriminations specific to this case must be calculated for each decision:

$$cgp_{ij}(\mu_i) = gp_{ij} \circ \mu_i$$

$$cgn_{ij}(\mu_i) = gn_{ij} \circ \mu_i \,,$$

where $\circ$ is operation of minimum or product.

3. COMPOSITION**.** The final decision can be defined as a distribution over the set of decisions as follows:

$$\delta_j = \frac{1}{2}\left(\chi_{L\arg e}(\pi_j) * \chi_{Small}(\nu_j)\right), \quad j \in D,$$

where

$$\pi_j = \left\{ \sum_i cgp_{ij}(\mu_i) \right\} \Big/ C_S \,, \qquad \nu_j = \left\{ \sum_i cgn_{ij}(\mu_i) \right\} \Big/ C_S \,,$$

and $*$ is the operation of maximum or sum.

4. DEFUZZIFICATION**.** Like the authors of classic discrimination analysis we will use the maximum method of defuzzification and interpret the decision with maximum magnitude in $\{\delta_j\}$ as most believable.

The most Popular MTV is Fuzzy Expected Value (FEV) [3-5].

When $MTV = FEV$ and g fuzzy measure is sampling distribution, following proposition of generalization can be proved:

Proposition: The generalized discrimination analysis is the extension of classic discrimination analysis, when for crisp (“non-fuzzy”) subset of activities generalized and classic discrimination values coincide.

Proof: As known for discrete cases FEV can be calculated, using following formula [10]:

Consider finite set $\mathrm{X} = \{x_1, x_2, ..., x_n\}$ and its some fuzzy subset $\underset{\sim}{A} \subset X$, such that compatibility values are ordered following way: $\chi_{\tilde{A}}(x_1) \le \chi_{\tilde{A}}(x_2) \le \cdots \le \chi_{\tilde{A}}(x_n)$, then Fuzzy Expected Value (FEV) of compatibility function $\chi_{\tilde{A}}$ (FEV) with regard to fuzzy measure g equals:

$$FEV \overset{def}{=} \max_i \{\chi_{\tilde{A}}(x_i) \wedge g(\mathrm{X}_i)\} = \min_i \{\chi_{\tilde{A}}(x_i) \vee g(\mathrm{X}_i)\}$$

where $\mathrm{X}_i = \{x_i, ..., x_n\}, i = 1,2,...n.$, and $\vee$ denotes maximum of two elements.

Its known [11], that crisp subset is concrete case of fuzzy subset, when the compatibility function has only to values – 0 or 1. Correspondingly, for each decision population will divide into two groups. For j-th decision and i-th activity the process of calculation of FEV can be represented by the following table:

group #	χ_i	n_i	$n^{(i)}$	$g_i = \frac{n^{(i)}}{n}$	$\chi_i \wedge n_i$
1	0	$n_{\chi=0}$	$n = n_{\chi=0} + n_{\chi=1}$	$\frac{n}{n} = 1$	0
2	1	$n_{\chi=1}$	$n_{\chi=1}$	$\frac{n_{x=1}}{n} = f_{ij}$	f_{ij}

In this table χ_i represents compatibility values, n_i- quantity of cases in group, $n^{(i)}$ - aggregated quantities; $n_{\chi=0}$ represents the quantity of cases, where j-th decision was correctly stated, but i-th activity was not exhibited; $n_{\chi=1}$ represents the quantity of cases where j- th decision was correctly stated, and i - th activity was exhibited.

As seen from the table, the value of FEV will equal to $\frac{n_{\chi=1}}{n} = f_{ij}$, which is the ratio of cases where i-th activity was exhibited and total amount of cases, which equals to the frequency of exhibition of this activity. Respectively, for crisp subsets MTV distribution table will transform to frequency distribution table.

Generalized positive and negative discrimination values will be calculated as follows:

$$gp_{ij} = \sum_{\substack{k\in D \\ k\neq j}} \left\{ \chi_{L\arg e-ratio}\left(\frac{MTV_{ij}}{MTV_{ik}}\right)\right\} /(C_D - 1) = \sum_{\substack{k\in D \\ k\neq j}} \left\{ \chi_{L\arg e-ratio}\left(\frac{f_{ij}}{f_{ik}}\right)\right\} /(C_D - 1) = p_{ij} ,$$

$$gn_{ij} = \sum_{\substack{k\in D \\ k\neq j}} \left\{ \chi_{L\arg e-ratio}\left(\frac{MTV_{ik}}{MTV_{ij}}\right)\right\} /(C_D - 1) = \sum_{\substack{k\in D \\ k\neq j}} \left\{ \chi_{L\arg e-ratio}\left(\frac{f_{ik}}{f_{ij}}\right)\right\} /(C_D - 1) = n_{ij} ,$$

and

$$cgp_{ij}(\mu_i) = gp_{ij} \circ \mu_i = p_{ij} \circ \mu_i$$

$$cgn_{ij}(\mu_i) = gn_{ij} \circ \mu = n_{ij} \circ \mu_i ,$$

will give us p_{ij} and n_{ij}, when $\mu_i = 1$, and 0, when $\mu_i = 0$, this practically means choosing discriminations of those activities which were exhibited during this concrete case. The following procedure totally coincides with classical discrimination analysis, and thus results will also coincide with results of classic discrimination analysis.

It must be mentioned here, that FEV doesn’t always represent the Most Typical Value of Population and in its place Weighted Fuzzy Expected Value (WFEV) [6] and Generalized Weighted Fuzzy Expected Value (GWFEV) [7-8] can be used. Correspondingly, in MTV distribution table, FEV will be replaced by these values.

4. 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 let us apply the generalized discrimination analysis and calculate *FEV*s.

As described in [12] 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}=0.7.

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	0.7	0.2
S_2	0.3	0.2

We can easily calculate the generalized positive and negative discrimination values (Suppose $\chi_{L\arg e-ratio}(x) = x/3.5$):

$$gp_{11} = gn_{12} = \chi_{L\arg e-ratio}\left(\frac{0.7}{0.2}\right) = \chi_{L\arg e-ratio}(3.5) = 1,$$

$$gn_{11} = gp_{12} = \chi_{L\arg e-ratio}\left(\frac{0.2}{0.7}\right) = \chi_{L\arg e-ratio}(0.286) = 0.082,$$

$$gp_{21} = gn_{22} = \chi_{L\arg e-ratio}\left(\frac{0.3}{0.2}\right) = \chi_{L\arg e-ratio}(1.5) = 0.429,$$

$$gn_{21} = gp_{22} = \chi_{L\arg e-ratio}\left(\frac{0.2}{0.3}\right) = \chi_{L\arg e-ratio}(0.667) = 0.190.$$

For our patient with $\mu_1 = 0.9$ and $\mu_2 = 0.1$, we can calculate positive and negative discriminations specific to this case. Corresponding tables will have the following view:

Positive discrimination

	D_1	D_2
S_1	0.9	0.0738
S_2	0.0429	0.019

Negative discrimination

	D_1	D_2
S_1	0.0738	0.9
S_2	0.019	0.0429

Afterwards,

$$\pi_1 = \nu_2 = \frac{0.9 + 0.0429}{2} = 0.471,$$

$$\pi_2 = \nu_1 = \frac{0.0738 + 0.019}{2} = 0.046.$$

Now we can perform the following calculations (let $\chi_{L\,arg\,e}(x) = x$ and $\chi_{Small}(x) = 1 - x$).

$$\delta_1 = \frac{1}{2}(0.471 + 0.0.954) \approx 0.71,$$

$$\delta_2 = \frac{1}{2}(0.046 + 0.529) \approx 0.29\,.$$

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

5. CONCLUSION

In real situations there often arise cases when available information is fuzzy by its nature. Elaboration of methods of decision making with fuzzy data is actual and demanding. Presented method - generalized discrimination analysis makes it possible to reason in cases of fuzzy subsets of activities and effectively uses all information available.

REFERENCES

1. Norris D., Pilsworth B., Baldwin J., Medical Diagnosis form Patient records – A Method Using Fuzzy Discrimination and Connectivity Analysis., Fuzzy Sets and Systems 21, 1989, pp. 37-45.
2. Friedman M., Henne M. and Kandel A., Most typical values for fuzzy sets, Fuzzy sets and Systems 87, 1997, pp. 27-37.
3. Kandel A. and Byatt W., Fuzzy sets, fuzzy algebra and fuzzy statistics, Proc. IEEE 66, 1978, pp.1619-1639.
4. Kandel A., Fuzzy Statistics and forecast evaluation, IEEE Trans. Systems Man Cybernet 8, 1978, pp. 396-401.
5. Kandel A., Fuzzy techniques in pattern recognition, John Wiley, New York, 1982.
6. Friedman M., Schneider M., and Kandel A., The use of weighted fuzzy expected value (WFEV) in fuzzy expert systems, Fuzzy Sets and Systems 31, 1989, pp.37-45.
7. Sirbiladze G. and Sikharulidze A., Weighted fuzzy averages in fuzzy environment, part I. Insufficient expert data and fuzzy averages, International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 11, No.2, 2003, pp. 139-158.
8. Sirbiladeze G. and Sikharulidze A., Weighted fuzzy averages in fuzzy environment, part II. Generalized weighted fuzzy expected values in fuzzy environment, International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 11, No.2, 2003, pp. 159-172.
9. Sikharulidze A., Application of discrimination and connectivity analysis in psychiatry, Bulletin of the Georgian Academy of Sciences, Vol. 164 , No.1, 2001, pp.39-40.
10. Kandel A., On the control and evaluation of uncertain processes, IEEE Trans. on Automatic Control AC-25 No.6 , 1980, pp.1182-1187.
11. Кофман А., Введение в теорию нечетких множеств, изд. Радио и Связь, Москва, 1982.
12. Sirbiladze G. and Sikharulidze A., Insufficient Data and Fuzzy Averages, Journal of Applied Mathematics and Informatics, Vol. 6, No. 2, 2001, pp.76-95.

Received: 2004-05-17