University of Texas at El Paso
DigitalCommons@UTEP

Departmental Technical Reports (CS) | Department of Computer Science

3-1-2013

Towards Fuzzy Method for Estimating Prediction Accuracy for Discrete Inputs, with Application to Predicting At-Risk Students

Xiaojing Wang
University of Texas at El Paso, xwang@utep.edu

Martine Ceberio
University of Texas at El Paso, mceberio@utep.edu

Angel F. Garcia Contreras
University of Texas at El Paso, afgarciacontreras@miners.utep.edu

Follow this and additional works at: http://digitalcommons.utep.edu/cs_techrep

 Part of the Computer Sciences Commons

Comments:
Technical Report: UTEP-CS-13-22
To appear in *Proceedings of the Joint World Congress of the International Fuzzy Systems Association and Annual Conference of the North American Fuzzy Information Processing Society IFSA/NAFIPS'2013*, Edmonton, Canada, June 24-28, 2013.

Recommended Citation
Wang, Xiaojing; Ceberio, Martine; and Garcia Contreras, Angel F., "Towards Fuzzy Method for Estimating Prediction Accuracy for Discrete Inputs, with Application to Predicting At-Risk Students" (2013). *Departmental Technical Reports (CS)*. Paper 754.
http://digitalcommons.utep.edu/cs_techrep/754

This Article is brought to you for free and open access by the Department of Computer Science at DigitalCommons@UTEP. It has been accepted for inclusion in Departmental Technical Reports (CS) by an authorized administrator of DigitalCommons@UTEP. For more information, please contact lweber@utep.edu.

Towards Fuzzy Method for Estimating Prediction Accuracy for Discrete Inputs, with Application to Predicting At-Risk Students

Xiaojing Wang, Martine Ceberio,
and Angel F. Garcia Contreras
Department of Computer Science
University of Texas at El Paso
El Paso, Texas 79968, USA
xwang@utep.edu, mceberio@utep.edu,
afgarciacontreras@miners.utep.edu

***Abstract*—In many practical situations, we need, given the values of the observed quantities $x_1, \ldots, x_n$, to predict the value of a desired quantity y. To estimate the accuracy of a prediction algorithm $f(x_1, \ldots, x_n)$, we need to compare the results of this algorithm's prediction with the actually observed values.**

The value y usually depends not only on the values $x_1, \ldots, x_n$, but also on values of other quantities which we do not measure. As a result, even when we have the exact same values of the quantities $x_1, \ldots, x_n$, we may get somewhat different values of y. It is often reasonable to assume that for each combinations of x_i values, possible values of y are normally distributed, with some mean E and standard deviation σ. Ideally, we should predict *both* E and σ, but in many practical situations, we only predict a single value $\widetilde{y}$. How can we gauge the accuracy of this prediction based on the observations?

A seemingly reasonable idea is to use *crisp* evaluation of prediction accuracy: a method is accurate if $\widetilde{y}$ belongs to a k_0-sigma interval $[E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$, for some pre-selected value k_0 (e.g., 2, 3, or 6). However, in this method, the value $\widetilde{y} = E + k_0 \cdot \sigma$ is considered accurate, but a value $E + (k_0 + \varepsilon) \cdot \sigma$ (which, for small $\varepsilon > 0$, is practically indistinguishable from $\widetilde{y}$) is not accurate. To achieve a more adequate description of accuracy, we propose to define a *degree* to which the given estimate is accurate.

As a case study, we consider predicting at-risk students.

I. Formulation of the Problem

Predictions are needed. In many practical situations, we want to be able to predict the value y of some quantity based on the values $x_1, \ldots, x_n$ of several measurable quantities x_i.

Examples. For example, we may want to predict tomorrow's weather based on today's observations and based on the weather records of this and previous years.

Another example is that at a university, it is important to be able to predict first-year performance of students, so that special attention can be applied to students who face a risk of failing, to prevent this failure.

Estimating prediction accuracy: a general problem. How can we gauge the accuracy of different prediction methods $f(x_1, \ldots, x_n)$?

General idea of estimating prediction accuracy: compare predictions with actual results. A natural way to estimate the prediction accuracy is consider cases $k = 1, \ldots, K$ in which we already know the corresponding values $y^{(k)}$, and to compare theses actual values with the results $\widetilde{y}^{(k)} = f\left(x_1^{(k)}, \ldots, x_n^{(k)}\right)$ of applying the given prediction method $f(x_1, \ldots, x_n)$ to the corresponding inputs.

Advantage of discrete inputs. In general, the accuracy of a prediction method depends on the inputs. For example, methods of weather prediction are usually more accurate when predicting typical weather and become less accurate when the weather switches to rare unusual patterns. Therefore, ideally, we should estimate prediction accuracy for give values of the inputs $x = (x_1, \ldots, x_n)$.

This is not easy to do for continuous inputs, since when the inputs are continuous, the values of each input x_i are different in different situations; so, strictly speaking, for each observed combination of inputs $x = (x_1, \ldots, x_n)$, there are no other observations with the exact same combination.

From this viewpoint, there is a definite advantage in having discrete inputs, in which each variable x_i has finitely many possible values. In this case, there are only finitely many possible combinations $x = (x_1, \ldots, x_n)$, and thus, when we have sufficiently many observations, we will have several observations corresponding to the each input combination.

Estimating prediction accuracy: case of discrete inputs. In this paper, we concentrate on discrete inputs. For discrete inputs, if we have sufficiently many observations, we can estimate the prediction accuracy for each combination of inputs $x = (x_1, \ldots, x_n)$.

In the following text, we will assumed that a combination $x = (x_1, \ldots, x_n)$ is fixed. For this combination, we have a value $\widetilde{y}$ predicted by the prediction methods, and we have values $y^{(1)}, \ldots, y^{(K)}$ observed for different situations with the given input values $x_1, \ldots, x_n$.

II. Case Study: predicting At-Risk Students

Case study: description. In this paper, we analyze the performance of first-time undergraduate students who started in

fall semester at our university. The sample contains 14,558 first-time, degree-seeking undergraduate students who entered the institution in 2003 to 2008 fall semesters.

Students data were collected by the authors' institution's Center for Institutional Evaluation, Research & Planning (CIERP). Some of the data comes the centralized database used by the institution and some from the data from a survey that had originally been used to gather information about the inc0oming students.

What we want to predict. A usual measure of a student performance is his or her average grade. In the US academic system, such an average is known as Grade Point Average (GPA). For each class i, a student gets a numerical grade g_i which is usually equal to 0, 1, 2, 3 or 4, with 4 being the best. Each class is characterized by the number c_i of "credit hours", usually the number of contact hours per week in a regular-length semester. The GPA g is then defined as a weighted average

$$g = \frac{\sum\limits_i c_i \cdot g_i}{\sum\limits_i c_i}.$$

Comment. Since we want to predict a student's GPA, we have to exclude students who dropped all the classes during their first semester without earning any grades. After excluding students without first-year GPA, we are left with a total of 12,062 students in the sample data set.

Which parameters x_i are used to predict y. The main objective of our study was to improve the prediction methods which are currently used by CIERP. Because of this objective, for our prediction, we used the same four quantities that CIERP currently uses in their prediction model (based on [7]) Each student' record in the data set contains the following information:

- the variable x_1 related to the student's score on the math placement exam; this score can have five different values values 0, 1, 2, 3, and 4; it is known that this score is correlated with the student success; see, e.g., [4];
- the variable x_2 represents a student's high school percentile; it can take any of the 101 values 0, 1, …, 100; high school performance is also known to be a strong predictor of first-year college performance; see, e.g., [2];
- the variable x_3 represents the number of hours that a student plans to work outside the school; this number was taken from a survey, in which students had to mark one of the following five options:
  - not planning to work;
  - working for less than 20 hours per week;
  - working 20–29 hours;
  - working 30–39 hours; and
  - working 40 hours or more per week;
- the "yes"-"no" variable x_4 describes whether a student delayed his/her graduation from high school; this also affects the student success; see, e.g., [6].

Overall, there can be $5 \cdot 101 \cdot 5 \cdot 2 = 5{,}050$ possible combinations of these inputs. In practice, some combinations are rare, so we observed only 1,404 combinations. So, if we divide students into groups corresponding to different combinations $x = (x_1, \ldots, x_4)$ of input values, then, out of 12,062 students, we have, on average, 9 students in each group. Actually, some groups have only 1 student, while other have up to 45 students.

Comment. It is worth mentioning that the prediction can become slightly more accurate if we also take into account the student's gender and whether a student belongs to the under-represented minority group.

What we did. In [8], we developed non-linear models that use Choquet integrals to predict the first-year GPA.

III. Analysis of the Problem

Why do we have different values of y for the same input. In order to properly solve the problem of estimating prediction accuracy, it is important to first understand why for the exact same values of all the inputs $x_1, \ldots, x_n$, we observe different values of the quantity y.

Both above examples clearly show why the observed values $y^{(k)}$ $(1 \le k \le K)$ are, in general, different: because in reality, the value y depends not only on the values of $x_1, \ldots, x_n$, but also on many other values. For example, first-year university success also depends:

- on family support (which is often lower for first-generation students),
- on whether a student him/herself has children to take care of,
- on how far away from the university the student lives,
- etc.

How to describe the difference in y. Many different factors influence the prediction. In other words, the difference between different actual values $y^{(k)}$ corresponding to the same combination of inputs $x = (x_1, \ldots, x_n)$ is caused by the joint effect of many independent factors – factors each of which has a relatively small effect on this difference. In statistics, such a situation is captured by the Central Limit Theorem, according to which such joint effects lead to normal (Gaussian) distribution; see, e.g., [5].

It is therefore reasonable to conclude that the values $y^{(k)}$ $(1 \le k \le K)$ corresponding to the same inputs $x = (x_1, \ldots, x_n)$ are normally distributed. It is known that to describe a 1-D normal distribution, it is sufficient to know the mean E and the standard deviation σ; in this case, the corresponding probability density function has the form

$$\rho_{E,\sigma}(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\frac{(y-E)^2}{2\sigma^2}\right). \qquad (1)$$

Ideal prediction vs. real prediction. In view of the above, ideally, we should predict *both*:

- the mean value E of the desired quantity y, and

- the standard deviation σ that describes how the observed values $y^{(k)}$ differ from this mean E.

In practice, however, most prediction methods predict only one value: the "typical' value $\widetilde{y}$. In this case, a natural question is: how good is this prediction?

IV. Estimating Prediction Accuracy: Crisp Approach

Main idea. From the purely theoretical viewpoint, the probability density (1) corresponding to a normal distribution is always positive, which means that it is theoretically possible to observe values which are far away from the mean E.

In practice, however, it is known that with a very high probability, the random value y lies within a k_0-sigma interval $[E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$ for an appropriate $k_0 = 2, 3, 6$, etc. For example:

- for $k_0 = 2$, we have $y \in [E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$ with probability $\approx 95\%$;
- for $k_0 = 3$, we have $y \in [E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$ with probability $\approx 99.7\%$;
- for $k_0 = 6$, we have $y \in [E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$ with probability $\approx 1 - 10^{-8}$.

It is therefore reasonable to select k_0 and check whether the estimate $\widetilde{y}$ is within the corresponding k_0-sigma interval:

- If the value $\widetilde{y}$ is within the k_0-sigma interval, we consider the prediction to be accurate.
- If the value $\widetilde{y}$ is outside the k_0-sigma interval, we consider the prediction to be inaccurate.

Comment. The above crisp criterion describes whether a given prediction algorithm $f(x_1, \ldots, x_n)$ is accurate for a given input $x = (x_1, \ldots, x_n)$. To gauge how accurate the method is in general, we can use, e.g., the percentage of inputs for which the predictions as accurate in the above sense.

Mathematical comment. It is worth mentioning that for a group of small size K, we have more strict limitations on the number of samples within a k_0-sigma interval. Indeed, the standard deviation is usually estimated as

$$\sigma^2 = \frac{1}{K-1} \cdot \sum_{k=1}^{K} \left(y^{(k)} - E\right)^2. \tag{2}$$

Since the sum is greater than or equal than the largest value

$$\Delta \stackrel{\text{def}}{=} \max_k \left|y^{(k)} - E\right|, \tag{3}$$

we thus conclude that

$$\sum_{k=1}^{K} \left(y^{(k)} - E\right)^2 \geq \Delta^2,$$

and therefore,

$$\sigma^2 \geq \frac{1}{K-1} \cdot \Delta^2. \tag{4}$$

Multiplying both sides of this inequality by $K - 1$, we get

$$\Delta^2 \leq (K-1) \cdot \sigma^2, \tag{5}$$

and therefore,

$$\Delta \leq \sqrt{K-1} \cdot \sigma. \tag{6}$$

By definition (3) of the maximum Δ, all the values $y^{(k)}$ lie within the interval $[E - \Delta, E + \Delta]$. So, all these values lie within the interval

$$[E - \sqrt{K-1} \cdot \sigma, E + \sqrt{K-1} \cdot \sigma] \tag{7}$$

corresponding to $k_0 = \sqrt{K-1}$. Thus:

- when $K \leq 5$, we have $\sqrt{K-1} \leq 2$ and thus, all observed values lie within the two-sigma interval;
- when $K \leq 10$, we have $\sqrt{K-1} \leq 3$ and thus, all observed values lie within the three-sigma interval;
- when $K \leq 37$, we have $\sqrt{K-1} \leq 6$ and thus, all observed values lie within the six-sigma interval.

Limitations of the crisp approach. As usual, the problem with above crisp approach is that we get a "yes"-"no" characterization of the prediction accuracy, and this does not adequately express the intuitive idea of accuracy. For example, if we select $k_0 = 2$, then:

- we classify the estimate $\widetilde{y} = E + 2\sigma$ as accurate, while
- a nearby value $E + (2 + \varepsilon) \cdot \sigma$ is not accurate, no matter how small the value $\varepsilon > 0$ we take.

From this practical viewpoint, when the value ε is sufficiently small, there is no practical difference between the estimates $E + 2\sigma$ and $E + (2 + \varepsilon) \cdot \sigma$, so different conclusions about prediction accuracy make no sense.

Natural idea. A natural idea is to take into account that whether a prediction method is accurate or not is a matter of degree. In other words, a natural idea is to use *fuzzy* techniques, techniques which were specifically designed to capture such degrees; see, e.g., [1], [3], [9].

V. Estimating Prediction Accuracy: Fuzzy Approach

Idea. We would like to estimate the degree $\mu_{E,\sigma}(\widetilde{y})$ to which, for given E and σ, an estimate $\widetilde{y}$ is a "typical" representative of the corresponding Gaussian random variable.

It is reasonable to require that when $\widetilde{y} = E$, then this degree is the largest, i.e. (since fuzzy sets are usually calibrated in such a way that the largest degree is 1), we should have

$$\mu_{E,\sigma}(E) = 1. \tag{8}$$

It is also reasonable to require that the smaller the probability that a value $\widetilde{y}$ can actually appear as a random outcomes $y^{(k)}$, the smaller the degree to which this value is typical. The simplest way to satisfy this requirement is to make the degree $\mu_{E,\sigma}(y)$ proportional to the corresponding probability density $\rho_{E,\sigma}(y)$, i.e., to take

$$\mu_{E,\sigma}(y) = C \cdot \rho_{E,\sigma}(y), \tag{9}$$

for some constant C. This constant must be determined from the previous requirement (8), which for the expression (9) takes the form

$$C \cdot \rho_{E,\sigma}(E) = 1. \tag{10}$$

Thus,

$$C = \frac{1}{\rho_{E,\sigma}(E)}, \tag{11}$$

and so, the formula (9) takes the form

$$\mu_{E,\sigma}(y) = \frac{\rho_{E,\sigma}(y)}{\rho_{E,\sigma}(E)}. \tag{12}$$

Substituting the expression (1) into this formula (12), we arrive at the following conclusion.

Resulting formula.

$$\mu_{E,\sigma}(y) = \exp\left(-\frac{(y-E)^2}{2\sigma^2}\right). \tag{13}$$

Discussion. The corresponding Gaussian membership function is actively used in fuzzy applications; see, e.g., [1], [3]. Our experience shows that it indeed leads to an intuitively reasonable estimates of prediction accuracy [8].

REFERENCES

[1] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[2] G. D. Kuh, J. Kinzie, J. Buckley, J. A. Bridges, and J. C. Hayek, "What matters to student success: A review of the literature", *Proceedings of the National Symposium on Postsecondary Student Sucess: Spearheading a Dialog on Student Success*, Washington, DC, November 1–6, 2006.

[3] H. T. Nguyen and E. A. Walker, *First Course In Fuzzy Logic*, CRC Press, Boca Raton, Florida, 2006.

[4] M. Parker, "Placement, retention, and success: A longitudinal study of mathematics and retention", *The Journal of General Education*, 2005, Vol. 54, No. 1, pp. 22–40.

[5] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2007.

[6] L. S. Stratton, D. M. O'Toole, and J. N. Wetzel, "Are the factors affecting dropout behavior related to initial enrollment intensity for college undergraduates?" *Research in Higher Education*, 2007, Vol. 48, No. 4, pp. 453–485.

[7] L. S. Stratton, D. M. O'Toole, and J. N. Wetzel, "A multinomial logit model of college stopout and dropout behavior", *Economics of Education Review*, 2008, Vol. 27, No. 3, pp. 319–331.

[8] X. Wang, *Extracting Fuzzy Measures From Sample Data: Optimization Algorithms and Applications*, PhD Dissertation, Department of Computer Science, University of Texas at El Paso, 2012.

[9] L. A. Zadeh, "Fuzzy sets", *Information and control*, 1965, Vol. 8, pp. 338–353.