Interval and Fuzzy Techniques in Business-Related Computer Security: Intrusion Detection, Privacy Protection

Mohsen Beheshti[1], Jianchao Han[1], Luc Longpré[2],
Scott A. Starks[2], J. Ivan Vargas[2], Gang Xiang[2]

[1] Computer Science Dept., California State University, Dominguez Hills
Carson, CA 90747, USA, mbeheshti@csudh.edu
[2] NASA Pan-American Center for Earth and Environmental Studies
University of Texas, El Paso, TX 79968, USA, gxiang@utep.edu

Abstract. E-commerce plays an increasingly large role in business. As a result, business-related computer security becomes more and more important. In this talk, we describe how interval and fuzzy techniques can help in solving related computer security problems.

1 Interval Techniques in Computer Security: Motivations

Importance of computer security. E-commerce plays an increasingly large role in business. As a result, business-related computer security becomes more and more important; see, e.g., [2, 16].

Why interval techniques. In computer security, interval uncertainty comes from the lack of knowledge. One of the reasons for this lack of knowledge is that the users are reluctant to provide the businesses with the exact information because they do not want this information to be misused. For example, a user may be reluctant to provide his or her exact date of birth but willing to provide an age interval (e.g., 30–40).

To be successful, an electronic business needs to process the user data. It is therefore important to develop efficient algorithms for statistical processing of such interval-valued data.

2 How to Extend Statistical Techniques to Situations with Interval Uncertainty

Traditional approach to data processing: statistical analysis. One of the main objectives of computer security is to predict the user's behavior, so that we will be able to stop malicious intrusions without interfering with the legitimate use of the computer systems.

To be able to make these predictions, we must find the relation between the desired difficult-to-observe characteristics of the user behavior – such as maliciousness – and the observable characteristics. Situations when we must be able to predict difficult-do-directly-observe characteristics based on easier-to-measure ones are typical in engineering and science. For example, in engineering, we must predict the building's stability based on the observed characteristics; in medicine, it is desirable to check whether a person has a certain disease (such as cancer) based, ideally, only on non-invasive tests such as ultrasonic and X-ray imaging.

The traditional way to find this dependence is (see, e.g., [20]):

- to find the make several observations of different characteristics,
- to compute statistical characteristics of the corresponding measurement results, such as the population mean $E_x = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$, the population variance $V_x = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x)^2$, the population covariance
$$C_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x) \cdot (y_i - E_y),$$
and then
- to use these statistical characteristics in the design of an (approximate) model (such as linear regression $z = a_0 + a_x \cdot x + a_y \cdot y + \ldots$) that predicts the value of the desired characteristic z as a function of easier-to-measure quantities x, y, etc.

Statistical analysis under interval uncertainty: a problem. We have mentioned that in many real-life situations, we only know the intervals $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ of possible values of x_i. For different possible values $x_i \in \mathbf{x}_i$, we get different values of E_x, V_x, C_{xy}, etc. In such situations, it is desirable to find the ranges of possible values of these characteristics:
$$\mathbf{E}_x = \left\{ \frac{1}{n} \cdot \sum_{i=1}^{n} x_i : x_i \in \mathbf{x}_i, i = 1, \ldots, n \right\},$$
$$\mathbf{V}_x = \left\{ \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x)^2 : x_i \in \mathbf{x}_i, i = 1, \ldots, n \right\},$$
$$\mathbf{C}_{xy} = \left\{ \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E_x) \cdot (y_i - E_y) : x_i \in \mathbf{x}_i, y_i \in \mathbf{y}_i, i = 1, \ldots, n \right\}.$$
The practical importance of the problem of computing population variance under interval uncertainty was emphasized, e.g., in [17, 18] on the example of processing geophysical data and in [6] on the example of processing environmental data.

The problem of computing the range of a given function under interval uncertainty is known as the problem of *interval computations*; see, e.g., [11, 15]. It is known that in general, this problem is computationally difficult (NP-hard). Specifically, even computing the range of the variance is, in general, NP-hard [7, 8].

Statistical analysis under interval uncertainty: privacy case. NP-hardness means that there is no general algorithm for computing $\mathbf{V}$ in all possible cases. As shown in [12], there are practically useful cases when a feasible algorithm for computing $\mathbf{V}$ is possible. One such case is the case of privacy, when we have a fixed partition of the real line, and all intervals come from this partition. In precise terms, we fix values $x_{(1)} < x_{(2)} < \ldots < x_{(m)}$, and we are only allowing intervals of the type $[x_{(k)}, x_{(k+1)}]$ – e.g., the age of 10–20, 20–30, etc.

Statistical analysis in the privacy case: a sample algorithm. The corresponding algorithm for computing the range $[\underline{V}, \overline{V}]$ for the variance V is based on the following idea. Since no two intervals are proper subsets of one another, we can sort them in lexicographic order – i.e., in the order where $\mathbf{x}_i \le \mathbf{x}_j$ if and only if either $\underline{x}_i < \underline{x}_j$ or ($\underline{x}_i = \underline{x}_j$ and $\overline{x}_i \le \overline{x}_j$). After the sorting, it can then be shown that the maximum of V is attained at one of sequences of the type $(\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n)$. Thus, $\overline{V}$ is equal to the largest of the values V corresponding to $n+1$ such sequences.

For each value k, for the corresponding sequence, computing $V = M_2 - E$, where $M_2 = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i^2$, requires linear time. Actually, we only need to spend this time to compute E, M_2, and V for $k = 0$. After we computed M_2 and E for some k, going from the case k to the case $k+1$ means replacing only one term in E and in M_2 ($\overline{x}_{k+1}$ by $\underline{x}_{k+1}$), and thus, requires constant number of steps:

$$E \to E + \frac{1}{n} \cdot (\underline{x}_{k+1} - \overline{x}_{k+1}); \; M_2 \to M_2 + \frac{1}{n} \cdot ((\underline{x}_{k+1})^2 - (\overline{x}_{k+1})^2).$$

So, after sorting (which requires time $O(n \cdot \log(n))$; see, e.g., [4]), we only need linear time to compute V for all $n+1$ candidates for an optimal sequence – and after that, linear time to find the largest of these n values as $\overline{V}$. Thus, the overall computation time for computing $\overline{V}$ is equal to

$$O(n \cdot \log(n)) + O(n) = O(n \cdot \log(n)).$$

Similarly, the minimum of V is attained at one of sequences of the type $(\overline{x}_1, \ldots, \overline{x}_k, \underline{x}_{k+1}, \ldots, \underline{x}_n)$. Thus, $\underline{V}$ is equal to the smallest of the values V corresponding to $n+1$ such sequences, and we can also compute $\underline{V}$ in time $O(n \cdot \log(n))$; for details and proofs, see $[1, 5, 9, 12, 23]$.

General case. The problem of computing the range for the variance is particular case of a general class of problems in which we need to combine probabilistic and interval uncertainty. The correspongind problems and algorithms are described, e.g., in $[6, 12, 13, 21, 22]$.

3 Rough Set Techniques – An Alternative to Statistical Techniques

Need for alternative techniques. In the traditional statistical approach, each variable x can, in principle, take an arbitrary value from its continuous range:

e.g., from the entire real line, or from the interval [0,120] for age. In the case of privacy-type interval uncertainty, in effect, for each variable x, we have a *finite* number of possible interval values: 0–10, 10–20, etc. Therefore, instead of trying to adjust continuous techniques to this discrete case, it may make sense to design new discrete techniques for handling the corresponding data.

Rough set approach: main idea. Some such techniques have been designed as part of Pawlak's *rough set theory* – a theory that, since its appearance in the early 1980s, has been successfully used in data analysis, pattern recognition, data mining, and knowledge discovery; see, e.g., [10, 14, 19, 24].

To find the dependence between the discrete characteristics X, ..., Y, and the desired discrete characteristic Z, the rough set theory proposes the following idea. We start with the set U of objects (situations) for which we know the values $(x, \ldots, y, z)$ of all the characteristics X, ..., Y, and the value of $z \in Z$ of the characteristic Z. It may happen that the known characteristics are not sufficient to determine Z, i.e., that we have a combination of values $x \in X$, ..., $y \in Y$, of these characteristics for which there are two different patterns $(x, \ldots, y, z)$ and $(x, \ldots, y, z')$ with $z \neq z'$. It is reasonable to define the *degree of dependency* $\gamma_P(Z)$ of Z on the set $P = \{X, \ldots, Y\}$ as the ratio $\gamma_P(Z) \stackrel{\text{def}}{=} \dfrac{|CL_P(Z)|}{|U|}$, where $CL_P(Z)$ is the set of all patterns for which the values of all variables from P uniquely determine the value of Z, and $|A|$ denotes the total number of patterns in the set A.

Usually, we start by measuring a large number of different characteristics X, ..., Y, most of which may be irrelevant to the problem of predicting Z. One of our objectives is to try to come up with a smaller set of characteristics $C \subset P$ that would enable us to achieve the same predictive power. Ideally, we should find such subsets that enables to classify exactly the same number of patterns as before, i.e., for which $\gamma_C(Z) = \gamma_P(Z)$. Such subsets are called *reducts*.

Rough set approach: computational aspects. It is easy to check whether a given subset is a reduct – it is sufficient to go over all originally classifiable patterns and make sure that we do not have two patterns in which the values of all characteristics from C are the same, but the values of Z are different. However, finding an *optimal* reduct – e.g., a reduct consisting of the smallest possible number of characteristics – is known to be NP-hard. Crudely speaking, it means that the following:

- we can always find the optimal reduct by trying all $2^{|P|}$ possible subsets $C \subseteq P$ – which requires exponential time;
- NP-hard means, crudely speaking, that, in general, it is not possible to find an optimal reduct faster than in the exponential time.

In practice, when we start with a large number of characteristics, we cannot test all subsets, so we have to use heuristic algorithms that provide us with a possible sub-optimal ("good enough") reduct instead of the optimal one. Several such algorithms have been developed and successfully used in applications of rough set theory. In particular, such heuristics have been successfully used for intrusion detection [3].

Limitations of the interval approach to rough set classification. In the above approach, we implicitly assumed that we know exactly how to classify each pattern, and we are absolutely sure that all the information within each pattern is correct.

In practice, however, some patterns may be more dubious, and some cases of seemingly legitimate computer behavior may actually be intrusions. How can we take this uncertainty into account?

4 Towards a Combination of Fuzzy and Interval Techniques in Computer Security

Why fuzzy techniques in computer security. Usually, a human being can easily recognize a spam. However, in some practical situations, a user may not be sure whether a given message is a spam or a legitimate email. The best a user can do in such a situation is to say that a given email is most probably, legitimate, or that it is somewhat suspicious – but the user usually cannot provide us with a probability or other numerical characteristic of his or her degree of suspicion.

Similarly, system administrators can often detect intrusion, but frequently, they can only indicate that something suspicious is going on, without being 100% sure that what they observing is indeed a computer intrusion. A system administrator usually cannot describe this feeling in numerical terms; at best, he or she can tell that a given situation is somewhat suspicious, or very suspicious, or most probably legitimate.

Fuzzy logic was designed to describe these and similar qualitative terms from natural language. In the fuzzy logic, each term like "somewhat" is translated into a numerical degree – a number from the interval $[0, 1]$. Let us show how this approach can be used in computer security.

In this paper, we will consider the most important practical case when the predicted variable Z has two possible values "positive" and "negative" – such as "intrusion" or "no intrusion".

How to incorporate fuzzy techniques into our problem. In accordance with the above comment, for each object $u \in U$, we have a degree $d(u)$ to which this object is relevant and its data is reliable. Also, instead of the exact value $z \in Z$, we have the degree $g(u)$ to which this object u is a positive example.

In other words, we have the list of tuples each of which has the form $(x, \ldots, y, g(u), d(u))$, where:

- $x \in X$, ..., $y \in Y$, are the (discrete) values of the corresponding characteristics,
- $g(u)$ is the degree to which this object is a positive example, and
- $d(u)$ is the degree of importance of u.

From the traditional fuzzy logic to interval-valued fuzzy logic. It is worth mentioning that we may want to distinguish between situations in which we have an equal number of arguments for and against positivity of a given object u, and situations when we know nothing about u. In the traditional fuzzy logic, both situations are characterized by the value $g(u) = 0.5$. To distinguish between such situations, it is reasonable to use *intuitionistic fuzzy logic* in which we have tow different degrees: the degree $g^+(u)$ to which u is positive and the degree $g^-(u)$ to which u is negative.

- In the case of equal number of arguments, we have $g^+(u) = g^-(u) = 0.5$;
- in the case of complete ignorance, we have $g^+(u) = g^-(u) = 0$;
- in general, we have $g^+(u) + g^-(u) \leq 1$.

In this approach, our degree of belief that an object u is positive is no longer described by a single value $g(u)$. Actually, the only information that we have about this value is that it is somewhere between $g^+(u)$ and $1 - g^-(u)$. So:

- in the case of equal number of arguments, we have $g^+(u) = g^-(u) = 0.5$ and thus, the interval $[g^+(u), 1 - g^-(u)]$ consists of a single value 0.5;
- in the case of complete ignorance, we have $g^+(u) = g^-(u) = 0$, so the interval $[g^+(u), 1 - g^-(u)]$ coincides with the entire interval $[0, 1]$;
- in general, we have an interval $[g^+(u), 1 - g^-(u)] \subseteq [0, 1]$.

How to incorporate fuzzy techniques into the rough set approach. In the traditional rough set approach, any rule with one counter-example is immediately dismissed as invalid. Since we now allow the possibility that tuples may be described wrong or misclassified, we are no longer sure that this counter-example actually disproves the rule.

In such situation, for each rule, we can only talk about the degree to which it is consistent with the given data. Let us assume that we restrict ourselves to a subset $C = \{X, \ldots, Y\} \subseteq P$ of the original set of characteristics. Each combination of values $c = (x, \ldots, y)$ of characteristics from C can be then classified as either positive or negative. Let S_c denote all the tuples with this combination of values. Then, it is reasonable to define the certainty to classify S_c as a positive region as $C_c^+ = \dfrac{\sum\limits_{u \in S_c} d(u) \cdot g^+(u)}{\sum\limits_{u \in S_c} d(u)}$, and the certainty to classify S_c as a negative region as $C_c^- = \dfrac{\sum\limits_{u \in S_c} d(u) \cdot g^-(u)}{\sum\limits_{u \in S_c} d(u)}$. We then:

- classify c as positive if $C_c^+ \geq t^+$ for some pre-defined threshold t^+,
- classify c as negative if $C_c^- \geq t^-$ for some pre-defined threshold t^-.

The set of all classified c will be denoted by CL. For all combinations $c \notin CL$, the information presented in c is not sufficient for the classification.

We can then define the degree of dependency $\gamma_C(Z)$ as the relative weight of all classified tuples, i.e., as the ratio

$$\gamma_C(Z) \stackrel{\text{def}}{=} \frac{\sum\limits_{c(u)\in CL} d(u)}{\sum\limits_{u\in U} d(u)}.$$

A reduct is then defined as a subset $C \subseteq P$ for which $\gamma_C(Z) = \gamma_P(Z)$.

Preliminary results. The existing heuristics for finding sub-optimal reducts can be naturally extended to this fuzzy situation, and our preliminary results of using these heuristics are encouraging.

5 Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, NSF grants EAR-0225670 and DMS-0532645, Star Award from the University of Texas System, and Texas Department of Transportation grant No. 0-5453.

References

1. Aló, R., Beheshti, M., Xiang, G.: Computing Variance under Interval Uncertainty: A New Algorithm and Its Potential Application to Privacy in Statistical Databases. In: *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, July 2–7, 2006 (to appear).
2. Bace, R. G.: *Intrusion Detection*, Sams Publishing, 1999.
3. Cai, Z., Guan, X., Shao, P., Peng, A., Sun, G.: A rough set theory based method for anomaly intrusion detection in computer network systems. Expert Systems **20** (2003) 251–260.
4. Cormen, Th. H., Leiserson, C. E., Rivest, R. L., Stein, S.: *Introduction to Algorithms*, MIT Press, Cambridge, MA, 2001.
5. Dantsin, E. Kreinovich, V. Wolpert, A. Xiang, G.: Population Variance under Interval Uncertainty: A New Algorithm. Reliable Computing **12** (2006) 273–280.
6. Ferson S.: *RAMAS Risk Calc 4.0: Risk Assessment with Uncertain Numbers*, CRC Press, Boca Raton, Florida, 2002.
7. Ferson, S., Ginzburg, L., Kreinovich, V., Longpré, L., Aviles, M.: Computing Variance for Interval Data is NP-Hard. ACM SIGACT News **33** (2002) 108–118.
8. Ferson S., Ginzburg, L., Kreinovich, V., Longpré, L., Aviles, M.: Exact Bounds on Finite Populations of Interval Data", Reliable Computing **11** (2005) 207–233.
9. Granvilliers, L., Kreinovich, V., Müller, N.: Novel Approaches to Numerical Software with Result Verfication. In: Alt, R., Frommer, A., Kearfott, R. B., Luther, W., Eds., *Numerical Software with Result Verification*, Proceedings of the International Dagstuhl Seminar, Dagstuhl Castle, Germany, January 19–24, 2003, Springer Lectures Notes in Computer Science, 2004, Vol. 2991, pp. 274–305.

10. Han, J., Hu, X., Cercone, N.: Supervised Learning: A Generalized Rough Set Approach. In: *Proc. of International Conference on Rough Sets and Current Trend in Computing*, Banff, Canada, 2000, pp. 284–291.
11. Kearfott, R. B., Kreinovich, V. Eds.: *Applications of Interval Computations*, Kluwer, Dordrecht, 1996.
12. Kreinovich, V., Xiang, G., Starks, S. A., Longpré, L., Ceberio, M., Araiza, R., Beck, J., Kandathi, R., Nayak, A., Torres, R., Hajagos, J.: Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity. *Reliable Computing* (to appear).
13. Kuznetsov, V. P.: *Interval statistical models*, Moscow, Radio i Svyaz Publ., 1991 (in Russian).
14. Lin, T. Y., Yao, Y. Y., Zadeh, L. A.: *Data Mining, Rough Sets and Granular Computing*, Physica-Verlag, 2002.
15. Moore, R. E.: *Methods and Applications of Interval Analysis*, SIAM, Philadelphia, 1979.
16. Ning, P.: *Intrusion Detection in Distributed Systems: An Abstraction-Based Approach*, Springer, 2003.
17. Nivlet, P., Fournier, F., Royer, J.: A new methodology to account for uncertainties in 4-D seismic interpretation. In: *Proceedings of the 71st Annual International Meeting of the Society of Exploratory Geophysics SEG'2001*, San Antonio, Texas, September 9–14, 2001, pp. 1644–1647.
18. Nivlet, P., Fournier, F., Royer, J.: Propagating interval uncertainties in supervised pattern recognition for reservoir characterization", *Proceedings of the 2001 Society of Petroleum Engineers Annual Conference SPE'2001*, New Orleans, Louisiana, September 30 – October 3, 2001, paper SPE-71327.
19. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, 1991.
20. Rabinovich, S.: *Measurement Errors: Theory and Practice*, American Institute of Physics, New York, 2005.
21. Walley, P.: *Statistical reasoning with imprecise probabilities*, Chapman and Hall, N.Y., 1991.
22. Walster, G. W.: Philosophy and practicalities of interval arithmetic, In: Moore, R. E., Ed.: *Reliability in Computing*, 1988, pp. 307–323.
23. Xiang, G.: Fast algorithm for computing the upper endpoint of sample variance for interval data: case of sufficiently accurate measurements. Reliable Computing **12** (2006) 59–64.
24. Ziarko, W.: Variable Precision Rough Set Model. Journal of Computer and System Sciences **46**, No. 1 (1993).