FUZZY REGRESSION BASED CLUSTERING ON UNCERTAIN DATA

R.Ramya
Department of Computer Science & Engineering
IFET College of Engineering, Villupuram, India
ramyaramkavi@gmail.com

Dr.R.Kalpana ,Professor
Department of Computer Science & Engineering
IFET College of Engineering, Villupuram, India

ABSTRACT:

In recent centuries, a number of indirect data collection methodologies have led to the proliferation of uncertain data. Such data points are often represented in the form of a probabilistic function, since the corresponding deterministic value is not known. The modeling of imprecise and qualitative knowledge, as well as handling of uncertainty at various stages is possible through the use of fuzzy sets. Fuzzy logic is capable of supporting to a reasonable extent, human type reasoning in natural form by allowing partial membership for data items in fuzzy subsets. Integration of fuzzy logic and kl divergence in data mining has become a powerful tool in handling natural data. Introduce the concept of fuzzy clustering and also the benefits of incorporating fuzzy logic with kl divergence in data mining. Finally it provides a comparative analysis of fuzzy clustering algorithms namely association rule based fuzzy.

Keywords- *Regression testing, Test case selection, Fuzzy logic, Selection probability.* I.

INTRODUCTION:

Fuzzy regression is the common term which is required for the proper functioning of the system. Maintenance of the fuzzy cluster is mainly concerned with the related modifications to the system. These modifications may be due to changing user needs, error correction, improved performance, adaptation to changed environment, optimization etc. This adaptation of the fuzzy cluster system to data mining makes a completely modified fuzzy cluster system. Modified system breaks the previously verified functionalities of the system, which causes faults. This requires fuzzy cluster regression testing for detecting such faults. Studies show that fuzzy cluster maintenance activities on an average account for two third of the overall fuzzy cluster cost. Fuzzy cluster maintenance is frequently required to fix defects, enhance or adapt the existing functionalities of the fuzzy cluster. One necessary maintenance activity is regression testing, which is the process of validating modified fuzzy cluster in order to provide confidence t h a t the fuzzy cluster behaves correctly and the modification has not lead to degradation Of fuzzy cluster quality. The dominant strategy for performing regression testing is to rerun the test cases that are available from the earlier version of the fuzzy cluster. Regression testing is expensive, often accounts for almost one-half of the total cost of fuzzy cluster maintenance [1]. Running all the test cases in a test suite requires a large amount of effort and time. A report shows that it took 1000 machine hours to execute approximately 30,000 functional test cases. Hundreds of man-hours are spent by test engineers to monitor the process of regression testing [2]. For this reason minimization of regression testing effort for reducing fuzzy cluster maintenance costs has become an issue of considerable practical importance. After development and release, fuzzy cluster undergo regress maintenance phase [3].

A. Regression Testing (RT)

It is an integral part of the fuzzy cluster development method. RT is defined as "the process of retesting the modified parts of the fuzzy cluster and ensuring that no new regression errors have been introduced into previously unmodified part of the program". Regression test end up forming a safety net that makes refactoring easier and maintenance work less scary. It is associated with system testing only when there is the change in the code.

There are various RT techniques shown in fig.1:

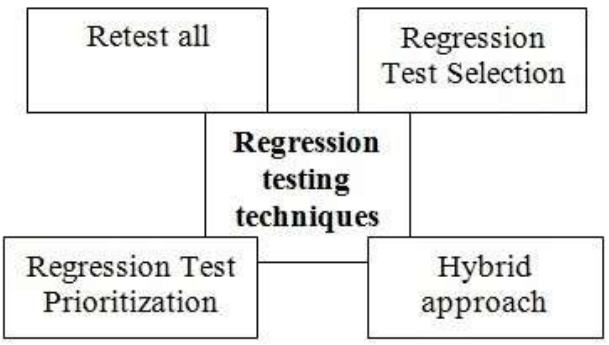

Fig 1: Techniques of Regression Testing

1) ***Retest all:*** It is one of the conventional methods of regression techniques. This method reruns all the test cases in the test suite. But, it consumes excessive time and resources as compared to other techniques.

2) ***Regression Test Selection (RTS):*** Due to expensive nature of "retest all" technique an alternative approach called Regression Test Selection or we can say selective retest is performed [4].In this technique instead of rerunning the whole test suite, it selects a subset of valid test cases from an initial test suite that are necessary to test the modified program [5].It attempts to reduce the time required to retest a modified program and also reduces the testing costs in environment where the program undergoes frequent modifications. Formally, RTS problem is defined as follows:

Let P be an application program and P' be a modified version of P. Let T be the initial test suite for testing P. An RTS technique aims to select a subset of test cases T' **subset of** T to be executed on P', such that every error detected when P' is executed with T is also detected when P' is executed with T' [6].

RTS consists of two major activities:

i) Identification of the affected part.

ii) Test Case Selection.

RTS divides the existing test suite into Obsolete, Retestable, and Reusable test cases [7].

- Obsolete test cases are not valid for the modified program.
- Retestable test cases execute the modified and the affected parts of the program and need to be rerun during regression testing.
- Reusable test cases execute only the unaffected parts of the program.

RTS techniques are broadly classified into three categories:

i. Coverage-based selection technique:

It locates program components that have been modified or affected by modifications, and select test cases that exercise those components.

ii. Minimization-based selection technique:

Similar to coverage techniques except that they select the smallest subset of test cases that can satisfy some minimum coverage criteria for the modified parts of the code [8][9][10][11].

iii. Safe Selection Technique:

Minimization techniques omit some fault-revealing test cases. To eliminate the possibility of missing faults, safe selection technique was introduced. It selects every test in T that can expose one or more faults in P'. It guarantees that the discarded test cases do not reveal faults [12] [13].

3) ***Regression Test Prioritization (RTP):*** It orders the test cases in such a way that the overall rate of fault detection increases. Test cases having higher fault detection capability are given higher priority and are taken up for execution earlier. It is very much advantageous as the errors are detected and reported to the development team earlier.

4) ***Hybrid Approach:*** It is the combination of both RTS and RTP.

B. Fuzzy Logic

Fuzzy logic is a convenient way to map an input space to output space through fuzzy inference process. It is basically a multivalued logic which permits intermediate values to be defined between conventional evaluations. Fuzzy logic gives the ability to quantify and reason with words having ambiguous meanings. That is why it is the best choice for managing contradicting, doubtful and ambiguous opinions.

Fuzzy logic is formed with the combinations of four concepts as shown in fig.2:

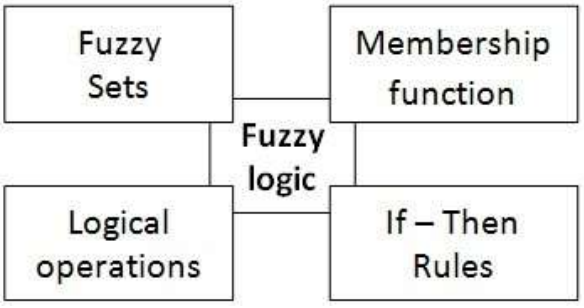

Fig.2: Fuzzy Concepts

Fuzzy sets are expressed as the set of ordered pairs [14] as shown in equation (1):

$$A = \{(x, \mu_A(x)) \mid x \varepsilon X, \mu_A(x): X \text{ Æ } [0, 1]\} \quad (1)$$

Where A is the fuzzy set, $(x, \mu_A(x))$ is the membership function and rest is the universe of discourse. Example: Word like **Good.** There is no single value which can define the term well; it differs from person to person. It has no clear boundary.

Simple way of forming membership functions is using straight lines. In this paper we have used triangular membership function, which is a simplest form using straight lines. It is the collection of three points forming a triangle. Logical operations are used to combine more than one inputs and conditions together for inference. There are three main logical operators namely AND, OR and NOT. For all possible combinations of the inputs If-Then rules are framed. The overall fuzzy process is shown in fig.3.

This paper proceeds by describing the related work and previous works on regression testing and fuzzy logic in the next section. Section 3 discusses factors for test case selection probability estimation. Section 4 presents the proposed. Section 5 presents the implementation of the model, and

Section 6 discusses overall conclusion and future work.

Fig.3: Stages of Fuzzy logic processing

II. RELATED WORK

Fuzzy cluster testing tells whether a program is correct, by showing that it produces correct output over some finite subset of input. When we develop fuzzy cluster we use development testing, when we modify fuzzy cluster we retest it, which is called as RT. It serves many purposes with the primary one to increase confidence in the correctness and locate errors in the modified program. Development testing and RT is different from each other in many aspects:

- Development test requires creation of test suites, whereas regression test uses existing test suites.
- Development test requires testing of all fuzzy cluster components whereas regression test only test modified part and the part which is affected by the modification.
- Development test gets time for testing whereas regression test is performed in crisis situation, under time constraints.
- Development test is costly, performed only once whereas RT is performed many times.

Research on RT spans a wide variety of topics. The issue that has seen the greatest amount of research, however, is the selective retest problem. The process of finding minimal subset of test cases that can cover each element of the system started in the year 1977 by K. Fischer in his paper "A Test Case Selection Method for the Validation of Fuzzy cluster maintenance modifications" [15]. Later this technique was extended in the year 1981 by Fischer, Raji and Chruscicki. They had given a methodology for retesting modified fuzzy cluster [16]. Yau and Kishimoto had given "A method for revalidating modified programs in the maintenance phase" in the year 1987 [17]. In this a selection method was presented that depends on input partitions, and uses symbolic execution to determine tests that traverse modified blocks. Their method relies on knowledge of modifications, and is computationally expensive due to the use of symbolic execution. Ostrand and Weyuker had done analysis for regression techniques using dataflow-based regression testing methods in September 1988 [18]. Lewis, Beck and Hartmann had proposed a tool to support regression testing in September 1989 [19]. The greatest drawback of these methods is that they require prior knowledge of modifications. Leung and White provided insights into regression testing in October 1989

[7] and given a model to compare regression testing strategies in 1991. Binkley used semantic differencing to reduce the cost of regression testing in November 1992 [20]. Chen, Rothermel and Vo provided a system for selective regression testing in 1994 [21]. Rothermel and Harrold had analyzed the regression testing selection techniques in August 1996 [22] and given a safe, efficient regression test selection technique in 1997 [23]. In 2001 Rothermel, Untch, Chu and Harrold had provided a technique for prioritizing test cases for regression testing. Yoo and Harman had given method for multi-objective test case selection in 2007 and in 2012 they had discussed open problems and given potential directions of future research in their survey of regression testing minimization, selection and prioritization techniques [24]. Engstrom, Runeson and Skoglund systematically reviewed the regression test selection techniques in January 2010. Amir Ngah had proposed a model for RTS using the decomposition slicing technique in his PhD thesis on RTS by exclusion in

May 2012 [25]. In July, 2012 Siavash Mirarab et.al

H a d given a multicriterion s basedmization for size-constrained RTS [26]. In the same month Jianchun Xing et.al had given a safe RTS based on program dependence graphs of a program and its modified version [27]. Prevus research in the field of RTS has not focused on industrial contexts. Alex Augustsson in 2012 had introduced a framework for evaluating RTS techniques in industry [28].

Research on Fuzzy was first proposed by L. Zadeh in his paper "Fuzzy Sets" in the year 1965 [29]. More information about fuzzy logic was given by Klir and Folger in 1988. They had given uncertainty and information on fuzzy sets [30]. W. Pedycz had given fuzzy control and fuzzy systems in 1993 [31]. In the very next year Driankon, Hellendour and Reinfark had added to fuzzy control [32]. Finally in the 1999, Novak, Perilieva and Mockor had given the mathematical principles of fuzzy logic [33].

Researches using fuzzy logic for the purpose of regression test case selection and prioritization are very scant. Xu, Gao and Khoshgoftaar had firstly shown the application of fuzzy expert system in regression test selection in 2005 [34]. Later in 2011 Praveen, Sirish and Raghurama of BITS pilani had given fuzzy criteria for assessing the fuzzy cluster testing effort [35]. In 2012 Ali M. Alakeel had proposed a fuzzy test cases prioritization technique for regression testing with assertions [36] and also fuzzy logic was used for prioritizing test cases for GUI based fuzzy cluster [37]. Recently, in 2013 H.B. Gupta et.al used fuzzy logic for regression technique [38].In our paper we use fuzzy rule base for the test case selection probability estimation [39].

III.FACTORS FOR TEST CASE SELECTION PROBABILITY ESTIMATION

We have taken three main factors to calculate the selection probability of a test case:

A. Code covered

B. Execution

C. Class covered

There may be many more factors which may be taken up for this calculation but these are the three main factors which have the most effect.

A. Code covered

This indicates the portion of the code covered by a particular test case. This may be number of lines covered by the test case, number of statements covered by the test case, number of functions covered and number of program branches covered. Test cases with highest level of code coverage are run first. We selected code covered as a factor for estimating the test case selection probability because it is believed that the test cases which cover more code have higher rate of fault detection

B. Execution time

This indicates the time required for a particular test case to complete its execution [41]. It may or may not include the loading time. Test cases having minimum execution time are given weightage and are executed first. We selected execution time as a factor for estimating the test case selection probability because it helps in selecting and reordering the execution of test cases ensuring that defects are revealed earlier in the test execution phase. Hence, ET $\propto$ 1/ test case selection probability

C. Faults covered

Similar to code covered it indicates the number of faults covered by a particular test case. Basically, it tells about the number of uncovered faults. Test cases which are capable of detecting or we may say covering more number of faults are taken up for execution first. We selected faults covered as a factor for estimating the test case selection probability because it is the most important parameter to be considered during testing [42]. More the number of faults is detected by a particular test case more effective will be that test case. Hence, FC test case selection probability

IV. PROPOSED MODEL

This paper gives a Regression Test Case Selection Technique based on fuzzy model, which uses the factors listed under section 3 for estimating the test case selection probability. Individual value of any factor may not provide the appropriate value for selection probability. So we use

fuzzy rule based approach which considers all the factors and their relative values simultaneously for estimating the selection probability. The basic fuzzy inference is expressed as shown in fig.4.

The model used in this paper contains three inputs namely code covered, execution time, faults covered and one output i.e. selection probability. For all the inputs and output, membership functions are chosen and the values for each membership functions corresponding to each inputs and output is defined.

In this concept clustering of an object can be established with the fuzzy based on the different inputs. The membership function which allows partial cluster and this increases the efficiency also. When compared to distance based clustering rule based clustering gives powerful result and thus probability estimation gives in this model

A. Membership functions and values for Inputs

Here, Code covered = CC; Execution time = ET;
Faults covered = FC

$$\mu_A(CC) = \begin{cases} \text{LOW;} & 0 \le CC \le 0.38 \\ \text{MEDIUM;} & 0.32 \le CC \le 0.65 \\ \text{HIGH;} & 0.62 \le CC \le 1 \end{cases}$$

$$\mu_A(ET) = \begin{cases} \text{LOW;} & 0 \le 1 \\ \text{MEDIUM;} & 0.24 \le ET \le 0.59 \\ \text{HIGH;} & 0.52 \le ET \le 1 \end{cases}$$

$$\mu_A(FC) = \begin{cases} \text{LOW;} & 0 \le FC \le 0.34 \\ \text{MEDIUM;} & 0.29 \le FC \le 0.56 \\ \text{HIGH;} & 0.51 \le FC \le 0.71 \\ \text{VERY HIGH;} & 0.65 \le FC \le 1 \end{cases}$$

$$\mu_A(SP) = \begin{cases} \text{VERY LOW;} & 0 \le SP \le 0.14 \\ \text{LOW;} & 0.12 \le SP \le 0.33 \\ \text{MEDIUM;} & 0.29 \le SP \le 0.56 \\ \text{HIGH;} & 0.51 \le SP \le 0.71 \\ \text{VERY HIGH;} & 0.65 \le SP \le 1 \end{cases}$$

C. Rule

It is basically a storage space associated with the model, which stores the knowledge related to the subject in the form of "If-Then" rules. Rules are formed with the composition of inputs and output, and each rule individually represents a condition-action statement in human understandable format. In this paper we consider all possible combinations of inputs getting a total of $3^3 = 27$ sets. Based on these 27 sets of combinations a total of 27 rules are formed to constitute a complete rule base for the model. Some of the rules are as shown in fig.5.

The membership values are defined based on the data collected from the classroom projects. The rules are formed on the basis of the collected data and expert advice.

In this model we use 'Mamdani' style for inference, one of the two available fuzzy inference systems. For combining together all the obtained results we use MAX method.

D. Working of the model

This model works as:

Step 1) Inputs corresponding to each factor is taken in crisp format, and is converted into fuzzy form.

Step2) Based on the membership functions value corresponding to each input factors, appropriate rule is fired.

Step 3) All inputs are taken together simultaneously, for this we use AND operator in order to combine the inputs together.

Step 4) MIN method is used for evaluating AND operator.

Step 5) All the results obtained is aggregated using MAX method

Step 6) Finally, the aggregated result is defuzzified using centroid method.

Step 7) Step 1 to step 6 is repeated for different inputs. Model one by one. The appropriate rule is fired based on the input values and the output for the selection probability is produced for each pair of input values.

Considering an example, let the inputs be:

Code covered = 0.63; Execution time = 0.25; Faults covered = 0.54

These are the crisp inputs. So, firstly we convert these crisp values into fuzzy values.

A.fuzification

It is seen from fig.6 that the value 0.63 belongs to both MEDIUM and HIGH set. So we need to get the fuzzy values corresponding to both MEDIUM and HIGH set.

Let 'x' represents the crisp values and 'y' represents the fuzzy values. Here, x = 0.63 and y =???

For **MEDIUM**, end-points of the corresponding line is: [(0.45, 1) and (0.65, 0)]

So, equation (2) gives the equation of the line as:

$(y - y1) = \{(y_2 - y_1) / (x_2 - x_1)\} * (x - x_1)$ (2)

Here, (x_1, x_2) and (y_1, y_2) are the end-points of the line. so, y = [{(0-1) / (0.65-0.45)} * (0.63-0.45)] + 1
or, y = [(-5) * (0.18)] +
1 or, **y = 0.1**

For **HIGH**, end-points of the corresponding line is: [(0.62,0) and (0.75,1)]

So, equation (3) gives the equation of the line as:

$(y - y_1) = \{(y_2 - y_1) / (x_2 - x_1)\} * (x - x_1)$ (3) Here, (x_1, x_2) and (y_1, y_2) are the end-points of the line. So, y = [{(1-0) / (0.75 – 0.62)} * (0.63-0.62)] +0
Or, y = [(7.6923) * (0.01)]+
0 or, **y = 0.0769**

Similarly, the fuzzy value for the other two inputs is found.

***2) Execution Time**:* It is 0.0588 in MEDIUM set and 0.3333 in LOW set.

***3) Faults Covered**:* It is 0.4545 in MEDIUM set and 0.1995 in HIGH set.

B. Rule Selection

Based on these values the rules fired are:

*Here, **VL** = Very Low, **L** = Low, **M** = Medium, **H** = High and **VH** = Very High*

Code covered == M) & (execution time == L) & (faults covered == M) => (selection probability = H).

Code covered == M) & (execution time == L) & (faults covered == H) => (selection probability = VH).

Code covered == M) & (execution time == M) & (faults covered == M) => (selection probability = M).

Code covered == M) & (execution time == M) & (faults covered == H) => (selection probability = H).

Code covered == H) & (execution time == L) & (faults covered == M) => (selection probability = H).

Code covered == H) & (execution time == L) & (faults covered == H) => (selection probability = VH).

Code covered == H) & (execution time == M) & (faults covered == M) => (selection probability = H).

Code covered == H) & (execution time == M) & (faults covered == H) => (selection probability = VH).

C. Rule Evaluation

As shown in fig.7, the selection probability medium, high and very high category. Therefore,

1) $\mu_{\text{selection probability=M}} = \max [\min \{\mu_{\text{code covered=M}} (0.63), \mu_{\text{execution time=M}} (0.25), \mu_{\text{faults covered=M}} (0.54)\}] = \max [\min \{0.1, 0.0588, 0.4545\}] = \mathbf{0.0588}$

2) $\mu_{\text{selection probability=H}} = \text{Max} [\min \{\mu_{\text{code covered=M}} (0.63), \mu_{\text{execution time=L}} (0.25), \mu_{\text{faults covered=M}} (0.54)\}, \min \{\mu_{\text{code covered=M}} (0.63), \mu_{\text{execution time=M}} (0.25), = \max [0.1, 0.0769, 0.0588] = \mathbf{0.1}$

D. Defuzzification

The above obtained fuzzy output is finally put to defuzzification in order to get the crisp value against the output variable selection probability. Out of several methods available for defuzzification we choose the Centroid method [44].

In this the Centre of Gravity (COG) is calculated for the area under the curve using equation:

$\mu_{\text{faults covered=H}} (0.54)\}, \min \{\mu_{\text{code covered=H}} (0.63), \mu_{\text{execution time=L}} (0.25), \mu_{\text{faults covered=M}} (0.54)\}, \min \{\mu_{\text{code covered=H}} (0.63), \mu_{\text{execution time=M}} (0.25), \mu_{\text{faults covered=M}} (0.54)\}]$

= max [min {0.1, 0.3333, 0.4545},

min {0.1, 0.0588, 0.1905},

min {0.0769, 0.3333, 0.4545},

min {0.0769, 0.0588, 0.4545}]
= max [0.1, 0.0588, 0.0769, 0.0588]

= **0.1**

3) $\mu_{\text{selection probability=VH}}$=max [min {$\mu_{\text{code covered=M}}$ (0.63), $\mu_{\text{execution time=L}}$ (0.25),

V. EVALUATION

The proposed rules were also processed with the designed fuzzy model in MATLAB; the selection probability against each input values were found (shown as modeled selection probability in table 1). For the purpose of evaluation we used the Root Mean Square Error (RMSE) method, which is used to measure the difference between the actual obtained value from the calculation that is being modeled and the value predicted by the model. RMSE is defined as the square root of the mean squared error as shown in equation

VI. CONCLUSION

This paper proposed a fuzzy model for estimation of the selection probability for regression test case. The model estimates the probability based on three important factors namely code covered, execution time and faults covered. The fuzzy approach is used to combine these inputs and reach at the estimation of the probability for selecting a test case. Fuzzy logic is a powerful tool which gives the ability to quantify with the contradicting, doubtful and ambiguous opinions. The selection of factors has been made based on some expert advice. However the results obtained are very close to the actual results. The only limitation is that, in this paper we have considered only three important test case selection factors. However there may be few more factors, which may be added. There may be the number of

extensions of the model by using techniques like artificial neural network and neuro fuzzy approach. This is left as future work.

VII.REFERENCES

[1] S. Abiteboul, P. Kanellakis, and G. Grahne, "On the Representation and Querying of Sets of Possible Worlds," *Proc. ACM SIGMOD*, 1987.

[2] *Managing and Mining Uncertain Data*, C. Aggarwal, ed. Springer, 2009.

$\mu_{\text{faults covered=H}}$ (0.54)}, min {$\mu_{\text{code covered=H}}$ (0.63),

$\mu_{\text{execution time=L}}$ (0.25),

$\mu_{\text{faults covered=H}}$ (0.54)}, min {$\mu_{\text{code covered=H}}$ (0.63),

$\mu_{\text{execution time=M}}$ (0.25),

$\mu_{\text{faults covered=H}}$ (0.54)}]= max [min {0.1, 0.3333, 0.1905}, min {0.0769, 0.3333, 0.1905}, min {0.0769, 0.0588, 0.1905}]

[3] P. Andritsos, A. Fuxman, and R.J. Miller, "Clean Answers over Dirty Databases: A Probabilistic Approach," *Proc. 22nd IEEE Int'l Conf. Data Eng. (ICDE)*, 2006.

[4] L. Antova, C. Koch, and D. Olteanu, "From Complete to Incomplete Information and Back," *Proc. ACM SIGMOD*, 2007.

[5] L. Antova, C. Koch, and D. Olteanu, "$10^ {(10^{6})}$ Worlds and Beyond: Efficient Representation and Processing of Incomplete Information," *Proc. 23rd IEEE Int'l Conf. Data Eng. (ICDE)*, 2007.

[6] L. Antova, T. Jansen, C. Koch, and D. Olteanu, "Fast and Simple Relational Processing of Uncertain Data," *Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE)*, 2008.

[7] C.C. Aggarwal and P.S. Yu, "Outlier Detection with Uncertain Data," *Proc. SIAM Int'l Conf. Data Mining (SDM)*, 2008.

[8] C.C. Aggarwal, "On Unifying Privacy and Uncertain Data Models," *Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE)*, 2008.

[9] C.C. Aggarwal, "On Density Based Transformations for Uncertain Data Mining," *Proc. 23rd IEEE Int'l Conf. Data Eng. (ICDE)*, 2007.

[10] C.C. Aggarwal and P.S. Yu, "A Framework for Clustering Uncertain Data Streams," *Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE)*, 2008.

[11] C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu, "A Framework for Clustering Evolving Data Streams," *Proc. 29th Int'l Conf. Very Large Data Bases (VLDB)*, 2003.

[12] M. Arenas, L. Bertossi, and J. Chomicki, "Consistent Query Answers in Inconsistent Databases," *Proc. 18th ACM Symp. Principles of Database Systems (PODS)*, 1999.

[13] M. Ankerst, M.M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering Points to Identify the Clustering Structure," *Proc. ACM SIGMOD*, 1999.

[14] D. Barbara, H. Garcia-Molina, and D. Porter, "The Management of Probabilistic Data," *IEEE Trans. Knowledge and Data Eng.*, vol. 4, no. 5, pp. 487-502, Oct. 1992.

[15] D. Bell, J. Guan, and S. Lee, "Generalized Union and Project Operations for Pooling Uncertain and Imprecise Information," *Data and Knowledge Eng.*, vol. 18, no. 2, 1996.

[16] O. Benjelloun, A. Das Sarma, A. Halevy, and J. Widom,

"ULDBs: Databases with Uncertainty and Lineage," *Proc. 32nd Int'l Conf. Very Large Data Bases (VLDB)*, 2006.
[17] J. Bi and T. Zhang, "Support Vector Machines with Input Data Uncertainty," *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2004.
[18] C. Bohm, A. Pryakhin, and M. Schubert, "The Gauss-Tree: Efficient Object Identification of Probabilistic Feature Vectors," *Proc. 22nd IEEE Int'l Conf. Data Eng. (ICDE)*, 2006.
[19] C. Bohm, P. Kunath, A. Pryakhin, and M. Schubert, "Querying Objects Modeled by Arbitrary Probability Distributions," *Proc. 10th Int'l Symp. Spatial and Temporal Databases (SSTD)*, 2007.
[20]D.Burdick,P. Deshpande, T.S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, "OLAP over Uncertain and Imprecise Data," *Proc. 31st Int'l Conf. Very Large Data Bases (VLDB '05)*, pp.970-981, 2005.
[21] D. Burdick, A. Doan, R. Ramakrishnan, and S. Vaithyanathan, "OLAP over Imprecise Data with Domain Constraints," *Proc. 33rdInt'l Conf. Very Large Data Bases (VLDB)*, 2007.
[22] R. Cavello and M. Pittarelli, "The Theory of Probabilistic Databases," *Proc. 13th Int'l Conf. Very Large Data Bases (VLDB)*, 1987.
[23] A.L.P. Chen, J.-S. Chiu and F.S.-C. Tseng, "Evaluating Aggregate Operations over Imprecise Data," *IEEE Trans. Knowledge and Data Eng.*, vol. 8, no. 2, pp. 273-284, Apr. 1996.
[24] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. Vitter, "Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data," *Proc. 30th Int'l Conf. Very Large Data Bases (VLDB)*, 2004.
[25] R. Cheng, D. Kalashnikov, and S. Prabhakar, "Evaluating Probabilistic Queries over Imprecise Data," *Proc. ACM SIGMOD*, 2003.
[26] R. Cheng, S. Singh, S. Prabhakar, R. Shah, J. Vitter, and Y. Xia, "Efficient Join Processing over Uncertain-Valued Attributes," *Proc. 15th ACM Int'l Conf. Information and Knowledge Management (CIKM)*, 2006.
[27] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, J. Vitter, and Y. Xia, "Efficient Join Processing over Uncertain Data," Technical Report CSD TR# 05-004, Dept. of Computer Science, Purdue Univ., 2005.
[28] R. Cheng, D. Kalashnikov, and S. Prabhakar, "Querying Imprecise Data in Moving Object Environments," *IEEE Trans. Knowledge and Data Eng.*, vol. 16, no. 9, pp. 1112-1127, Sept. 2004.
[29] C.-K. Chui, B. Kao, and E. Hung, "Mining Frequent Item sets from Uncertain Data," *Proc. 11th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD)*, 2007.
[30] J.M. Ponte and W.B. Croft, "A Language Modeling Approach to Information Retrieval," Proc. 21st Ann. Int'l ACM SIGIR Co.
[31] A.D. Sarma, O. Benjelloun, A.Y. Halevy, and J. Widom, "Working Models for Uncertain Data," Proc. Int'l Conf. Data Eng. (ICDE), 2006.
[32] D.W. Scott, Multivariate Density Estimation: Theory, Practical, and Visualization. Wiley, 1992.
[33] B.W. Silverman, Density Estimation for Statistics and Data Analysis.Chapman and Hall, 1986.
[34] F. Song and W.B. Croft, "A General Language Model for Information Retrieval," Proc. Int'l Conf. Information and Knowledge Management (CIKM), 1999.
[35] Y. Tao, R. Cheng, X. Xiao, W.K. Ngai, B. Kao, and S. Prabhakar,"Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions," Proc. Int'l Conf. Very Large Databases (VLDB), 2005.
[36] P.B. Volk, F. Rosenthal, M. Hahmann, D. Habich, and W. Lehner,"Clustering Uncertain Data with PossibleWorlds," Proc. IEEE Int'lConf. Data Eng. (ICDE), 2009.
[37] J. Xu and W.B. Croft, "Cluster-Based Language Models for Distributed Retrieval," Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 1999.
[38] C. Yang, R. Duraiswami, N.A. Gumerov, and L.S. Davis, "Improved Fast Gauss Transform and Efficient Kernel Density Estimation," Proc. IEEE Int'l Conf. Computer Vision (ICCV), 2003.
[39] D. Florescu, D. Koller, and A. Levy, "Using Probabilistic Information in Data Integration," *Proc. 23rd Int'l Conf. Very Large Data Bases (VLDB)*, 1997.
[40] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer, "Learning Probabilistic Relational Models," *Proc. 16th Int'l Joint Conf. Artificial Intelligence (IJCAI)*, 1999.
[41] N. Fuhr and T. Rolleke, "A Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems," *ACM Trans. Information Systems*, 1997.
[42] A. Fuxman, E. Fazli, and R.J. Miller, "Conquer: Efficient Management of Inconsistent Databases," *Proc. ACM SIGMOD*, 2005.