Learning from Uncertainty for Big Data

Future Analytical Challenges and Strategies

by Xizhao Wang
and Yulin He

Big data refers to data sets that are so large that conventional database management and data analysis tools are insufficient to work with them [1]. Presently, we are in an era of big data, which exists in various fields, such as social media, telecom, finance, medicine, bioinformatics, and power networks. Big data results mainly from the evolutionary development of data storage and data collection techniques in recent years. Big data has become a bigger-than-ever problem with the quick developments of data collection technologies. In fact, the word *big* is a fuzzy concept. So far, we do not have a mathematical definition of big data. But we can use several features or coordinates to describe it, for example, the well-known 5V characteristics.

The 5V characteristics of big data refer to huge *volume*, high *velocity*, much *variety*, low *veracity*, and high *value* [2].

1) Huge *volume* indicates that the size of data is extremely high. Currently it is very common to have the storage systems with TB ($\approx 10^3$ GB) and PB ($\approx 10^6$ GB) grade

Digital Object Identifier 10.1109/MSMC.2016.2557479
Date of publication: 23 August 2016

2333-942X/16©2016IEEE

© ISTOCKPHOTO.COM/TUNGSTENBLUE

levels for enterprises due to an exponential growth in the data storage.

2) High *velocity* means that not only the speed of data collection but also the speed of analyzing and utilizing data are fast. Big data is often available in real time and is batch oriented.
3) Much *variety* is also called *multimodality* of big data, which means that the types of data can be very complex. For example, consider big data in the clinical and health applications whose features include numerical data, symbolic data, text, image, video, and time series, among others.
4) Low *veracity* corresponds to the *changed uncertainty* and the large-scale missing values of big data. Sometimes, along with the growing size of datasets, the uncertainty of data itself often changes sharply, which definitely makes the traditional processing tools unavailable. Except for the changed uncertainty of data itself, the uncertainty in data modeling and data processing are also changing very notably.
5) High *value* is to say that the value hidden in big data can help users obtain "big" results. Mining the value of big data will reveal insights that can lead enterprises to obtain huge economic benefits.

This article will focus on the fourth V, the *veracity*, to demonstrate the essential impact of modeling uncertainty on learning performance improvement.

Challenge of Big Data Analytics

The big data application chain generally includes four phases: data generation, data management, data analytics, and data application. Big data analytics, which is considered the most important phase in the whole chain, refers to the process of discovering patterns from data. In this phase, there are six main challenges (shown in Figure 1), making big data analytics much more difficult and complicated than normal-sized data analytics.

- *Complex data representation.* How to uniformly represent various types of features is a great challenge for big data with multimodality [3]. We need to process the data in a uniform framework. The uniform/structured representation of data is the first step of data

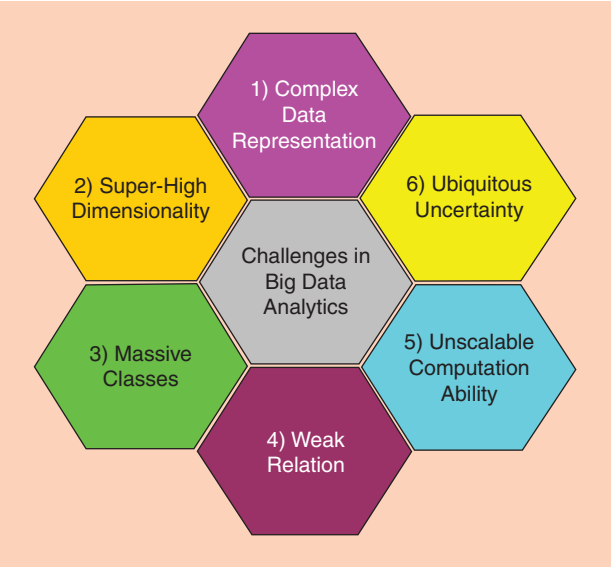

Figure 1. The six main challenges in big data analytics.

processing; it is indispensable. But due to big data's multimodality, it is very difficult to uniformly represent various types of data. It means that using existing methodologies to handle big data is almost impossible. It brings the first challenge of big data analytics.

- *Super-high dimensionality*. Big data in specific domains, especially in bio-informatics or life science computing areas, is often extra-high dimensional. The problem is that existing algorithms are not well-scalable to high-dimensional data. Usually, with the increase of the data dimension, the required amounts of time or memory go up exponentially. This is the so-called *curse of dimensionality*. Zhai et al. [4] gave a detailed description of the rapid change of the data set's dimensions in the field of scientific research over the past 25 years. Many machine learning and data mining algorithms are designed based on a distance measure in a metric space, for instance, the popular k-nearest neighbor. Studies [5] and [6] show that, in a high-dimension space, the distance measure has a very strange phenomenon; that is, some fixed points are the nearest neighbors of every case in the space. It is called *hubness*, which indicates that the distance formula has been ineffective and invalid.
- *Massive classes*. In the big data era, we have to deal with classification tasks with thousands of classes, such as the large-scale recognition problem. The existing classifiers seem to be qualified for the classification tasks, but their performance is seriously downgraded. Study [7] clearly describes the scale of the problem.
- *Weak relation*. A relation is more general than a mapping [8], [9], and finding a relation is more difficult than finding a mapping when conducting big data analytics. For example, the labels may be missing or cases may be labeled erroneously in classification tasks. The high expense for labeling cases leads to the weakly supervised problem. Traditionally, we need to find a mapping from a set of cases to another set. In most situations in a big data setting, we only need to find a relation between two subsets of cases. This is because sometimes in a big data setting, we may not need an exact mapping, and often, it is impossible to find such a precise mapping.
- *Unscalable computation ability*. The current computational ability is not scalable to the big data problem. Existing learning algorithms cannot adapt themselves well to the new big data settings. It means both the problem complexity and computational ability increase remarkably in the big data era, but the increase of computational ability does not match well against the increase of problem complexity. When a data set is changing from a regular size to a large size with many type attributes, some frequently used data mining and machine learning algorithms, such as a *support vector machine*, a *neural network*, a *decision tree*, *C-means*, and *C-modes*, will not work well. In many domains, a learning/mining algorithm is recognized as being effective for big data only if its complexity is linear or quasi-linear.
- *Ubiquitous uncertainty*. Uncertainty exists in every phase of big data learning [10]. For example, big data often has much noise, and most attribute values of a case in big data are missing (e.g., there are 80%~90% missing links in social networks and over 90% missing attribute values for a doctor diagnosis in clinic and health fields). Some traditional learning algorithms have obviously not been valid for processing the data with 90% missing values, and, therefore, how to design the new learning algorithm to tackle the large-scale missing data is difficult. Moreover, there are many models that can be selected for big data processing. Due to the growing uncertainty existing in the selection process, choosing an appropriate model based on the formulated uncertainty is another big challenge. The third difficulty is how to well represent the data uncertainty and how to take it into the mining process in the data analytics phase. From normal-sized data to big data, does the uncertainty increase or decrease? It depends. For example, for the mean of a random variable, uncertainty will decrease due to the large numbers theorem, but for the model selection problem, it will increase.

Current Strategies of Big Data Analytics

Fundamental strategies (shown in Figure 2) for big data analytics may include divide-and-conquer, parallelization, incremental learning, sampling, granular computing, feature selection, and hierarchical classes.

- *Divide-and-conquer*. Just as M. Jordan highlighted in [11], divide-and-conquer is one of the fundamental strategies of processing big data. It has three basic procedures: going from big to small, processing in every

small block, and fusing separate results together. In fact, in the fields of high-performance computing and very large database, this strategy has been used for many years.

- *Parallelization.* Parallelization indicates that large problems are divided into smaller ones, which can then be solved individually at the same time. There are several different forms of parallel computing, such as bit level, instruction level, and task parallelism. It is noteworthy that parallelization cannot decrease workload but can reduce working hours. It is not such a case that each problem/algorithm can be parallelized well. It depends strongly on the nature and structure of the problem.
- *Incremental learning.* Incremental learning gradually improves the parameters in learning algorithms by using only new cases rather than using all available cases (existing ones plus new ones). Incremental learning is a step-by-step learning process. Training is conducted only on the new incoming data blocks. One data block is used for training only once. It is focusing on batch data or streaming data. The major defect of incremental learning is that the algorithm is required to have good memory. For the data blocks trained already, its knowledge is considered as being remembered well and saved within the model. It is an obviously a limitation of this strategy [12].
- *Sampling.* Sampling is an old technique in probability and statistics. There are many typical results of sampling, theoretically and technically. Commonly used sampling methods include simple random sampling, systematic sampling, stratified sampling, cluster sampling, quota sampling, minimum–maximum sampling, etc. [13]. Essentially, sampling technology is to study the relation between a sample and the population. A traditional sampling course does not focus on the large-scale data set. With the coming of the big data era, many new difficulties emerge.
- *Granular computing.* A recent study [14] reveals that granular computing (GrC) [15] is a general computation theory for effectively using granules such as classes, clusters, subsets, groups, and intervals to build an efficient computational model for complex applications with huge amounts of data. Intuitively, GrC is to reduce the data size into different levels of granularity. Under certain circumstance, some big data problems can be readily solved in such a way.
- *Feature selection.* Feature selection [16] is a kind of dimensionality reduction method that aims to obtain a representative subset that has fewer features in comparison with the original feature space. High-dimensional data belongs to the big data area. When the scale of features is too large (for example, over 100 trillion features), some unexpected difficulties may emerge during the process of feature selection. The latest study [17] introduced how to scale to ultrahigh dimensional feature selection task on big data.

Figure 2. Seven fundamental strategies for big data analytics.

Uncertainty-Based Big Data Learning

During recent years, one can view a rapid growth in the hybrid study that integrates uncertainty and learning from data (e.g., [18]–[24]). The representation, measure, modeling, and handling of uncertainty embedded in the entire process of data analytics have a significant impact on the performance of learning from big data. Without properly dealing with these uncertainties, the performance of learning strategies may be greatly degraded.

Uncertainty Definition

Presently, there is no general definition for *uncertainty* that fits any situation. We usually consider the uncertainty under a specific background. Five types of uncertainty are mentioned: *Shannon entropy* (*SE*) [21], *classification entropy* (*CE*) [23], *fuzziness* [18] [19], *nonspecificity* [22], and *rough degree* [24].

- *Shannon entropy.* Given a random variable $X = \{x_1, x_2, \ldots, x_n\}$ and its probability distribution $P = \{p_1, p_2, \ldots, p_n\}$, the random uncertainty is measured by Shannon entropy:

$$SE(P) = -\sum_{i=1}^{n} p_i \log_2(p_i).$$

When $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$, $SE(P)$ attains its maximum of 1.

- *Classification entropy.* For a two-class problem, there is a data set S of which each sample can be defined as positive class or negative class. Classification entropy means the impurity of the class distribution in S and is defined as

$$CE_2(P) = -\left[\frac{|S_+|}{|S|}\log_2\frac{|S_+|}{|S|} + \frac{|S_-|}{|S|}\log_2\frac{|S_-|}{|S|}\right],$$

where $|S|$ is the number of all samples in S and $|S_+|$ and $|S_-|$, respectively, denote the numbers of positive-class and negative-class samples in S. When $|S|=|S_+|$ or $|S|=|S_-|$, $CE_2(P)$ reaches the minimum of 0; when $|S_+|=|S_-|$, $CE_2(P)$ reaches the maximum of 1. Similarly, classification entropy for a C-class problem is defined as

$$CE_C(P)=-\sum_{k=1}^{C}\frac{|S_k|}{|S|}\log_2\frac{|S_k|}{|S|},$$

where $|S_k|$ is the number of the kth class samples in S.

- *Fuzziness.* Uncertainty always exists in our human language, e.g., young and old. Then, what is the boundary between young and old? Fuzzy subsets are used to measure this kind of uncertainty in human language. For a universe $U=\{x_1,x_2,\ldots,x_n\}$, a fuzzy subset A of U is defined as

$$A=\{\mu_A(x_1),\mu_A(x_2),\ldots,\mu_A(x_n)\},$$

where μ_A, called the membership function of A, is a mapping function from U to [0, 1]. Assume there are three fuzzy subsets: $A_1=\{0.7,0.4,0.1\}$, $A_2=\{0.8,0.2,0\}$, and $A_3=\{1,0,0\}$. Fuzziness is a measure that can help us determine which one is more fuzzy or less fuzzy. The definition of fuzziness of a fuzzy subset A is

$$Fuzz(A)=-\frac{1}{n}\sum_{i=1}^{n}[\mu_A(x_i)\log_2\mu_A(x_i)+[1-\mu_A(x_i)]\times\log_2[1-\mu_A(x_i)]].$$

Thus, we can know fuzzy subset $A1$ is the most fuzzy because of $Fuzz(A_1)>Fuzz(A_2)>Fuzz(A_3)$.

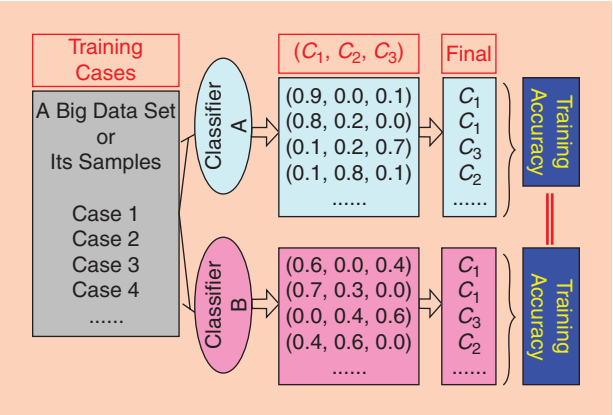

Figure 3. The general framework of uncertainty-based learning for big data. Classifier A has the same training accuracy as Classifier B, but Classifier A has a smaller uncertainty (e.g., fuzziness or ambiguity) than Classifier B. We say, for some types of big data (*not* for all), Classifier A has the better generalization than Classifier B, which provides quite a different viewpoint to design the learning algorithm in comparison to the traditional pattern-recognition viewpoint.

- *Nonspecificity.* Nonspecificity is also known as *ambiguity*, which is another measure to evaluate the uncertainty of fuzzy subset $A=\{\mu_1,\mu_2,\ldots,\mu_n\}$. The nonspecificity or ambiguity of fuzzy subset A is defined as

$$\text{Ambig}(A)=\sum_{i=1}^{n}[(\mu_i^*-\mu_{i+1}^*)\text{In}i],$$

where $A^*=\{\mu_1^*,\mu_2^*,\ldots,\mu_n^*\}$ is a permutation of membership degree distribution of A such that for any i, $\mu_i^*\le\mu_{i+1}^*$ and $\mu_{n+1}^*=0$.

- *Rough degree.* For the rough set $(\underline{R}X,\overline{R}X)$ of X, its rough degree is defined as

$$RD(X)=1-\frac{|\underline{R}X|}{|\overline{R}X|},$$

where $\underline{R}X=\{x\in U|[x]_R\subseteq X\}$ is the lower approximation of X, $\overline{R}X=\{x\in U|[x]_R\cap X\neq\varnothing\}$ is the upper approximation of X, U is the universe of discourse, R is an equivalence relation, X is a subset of U, and $[x]_R=\{y|\,y\in U,yRx\}$ is an equivalence class.

Some Studies on Learning from Uncertainty for Big Data

Here, we briefly introduce two studies regarding uncertainty-based learning for big data. One is fuzziness-based semisupervised learning and the other is ambiguity-based model tree (AMT) handling mixed attributes. The first study is basically within the following general framework [19] of uncertainty-based learning for big data (as shown in Figure 3).

Fuzziness-Based Semisupervised Learning

Assume that A is a big data set in which most cases have no labels. B is a small part of A, and each case in B has a label. We can train a classifier from B, but we cannot expect a good prediction performance on A-B. Based on the prediction of each case in A-B, we would like to select some cases from A-B and then add them (together with their predicted labels) into B. It is expected to have the improved prediction accuracy on A-B after retraining on B. Here, the key problems are what requirements the trained classifier should meet and how to select cases from A-B. Theoretically, the trained classifier is required to have an accuracy of more than 0.5. We focus on the sample selection strategy from uncertainty view as shown in Algorithm 1.

It is highlighted in our learning scheme that, traditionally, only group G_3 is mentioned for learning performance improvement, while both G_3 and G_1 are used.

For demonstration, we collect a big data set for the Chinese chess game scene classification. The file size is 1.86 GB, including more than 10^7 records of playing a chess game and more than 10^9 scenes of a chess game. This is a typical semisupervised learning with unstructured data: there are numerous scenes that need to be labeled. Complicated scene labeling usually requires senior

experts (i.e., chess masters). It is a very costly process. Traditionally, the scene evaluation function can be used to compute a value, and based on this value one can give a class. But the accuracy in this way is really poor. The experimental results based on Chinese chess game scene classification data show that our fuzziness-based semisupervised learning algorithm can achieve very high prediction accuracy. It further confirms our statement that the appropriate processing of uncertainty can significantly improve the classification system performance.

Ambiguity-Based Model Tree Handling Mixed Attributes

A model tree is an effective way to process the mixed-attributes (a special case of big data's multimodality) classification problem in which the mixed attribute mainly means the mixture of symbolic data and numerical data. Globally, a model tree is a tree structure, but in each leaf node, the particular model is built. In our AMT, the decision tree is established based on the reduction of ambiguity generated during the dividing process from a father node to its child nodes, and the leaf node of model tree is a three-layer feed-forward neural network that is trained with an extreme learning machine (ELM) algorithm [25]–[27]. In AMT, a decision tree and ELM are used to deal with the discrete and continuous attributes, respectively. Algorithm 2 gives a brief description to the generalization process of AMT. It is worth noting that AMT can be extended to the image and text attribute by incorporating it into the deep learning [28], which is a very hot topic in recent years. Deep learning essentially is an automatic feature-selection strategy that was originally developed for image feature extraction and image classification. For a big data classification problem with image-valued attributes, the model tree combined with deep learning will be very effective. Some recent studies [29], [30] reveal that the ELM autoencoder can outperform various state-of-art deep-learning methods.

The experimental results on several big data sets (more than 2 million samples) show a good performance of parallelization of our methodology. The training time of parallel AMT demonstrates a decreasing trend with the increase of computers, which indicates feasibility of parallelization in reducing the computational time. The experimental results also demonstrate good performance of our AMT's generalization ability. A comparison of 15 data sets shows that AMTs achieve higher testing accuracies than functional trees [31], naïve Bayesian trees [32], and logistic model trees [33], [34] on most datasets.

Algorithm 1. Fuzziness-based sample selection.

Step 1:	Randomly divide data set A as training set B and testing set A-B
Step 2:	Train a base classifier based on set B
Step 3:	For each sample, in both the training set and the testing set, obtain the fuzzy vector output based on the base classifier
Step 4:	Compute the fuzziness for each output
Step 5:	Sort the samples based on the quantity of fuzziness in the training set and in the testing set, respectively
Step 6:	Based on the sorting, categorize the training set (and the testing set) into three groups: high-fuzziness group G_1, mid-fuzziness group G_2, and low-fuzziness group G_3
Step 7:	Groups G_1 and G_3, together with their predicted labels, will be added in B for the next round of learning

Algorithm 2. AMT.

Input:	A big data set S with a mixed-attribute set $A = (D_1, D_2, \ldots, D_m; C_1, C_2, \ldots, C_n)$, where $D_i (i = 1, 2, \ldots, m)$ are discrete attributes and $C_j (j = 1, 2, \ldots, n)$ are continuous attributes
Output:	An AMT
Step 1:	Select the attribute D_i with minimal ambiguity as the root node of model tree
Step 2:	Split a parent node into K child nodes, $S_1, S_2, \cdots, S_K$, according to the values of discrete attribute D_i
Step 3:	For every child node S_k, select the discrete attribute whose ambiguity is smaller than D_i as the split attribute
Step 4:	Repeat Steps 2 and 3 until the maximal ambiguity of the child node is smaller than a given threshold
Step 5:	Treat this child node as a leaf node, and train an ELM with continuous attributes on this leaf node.

Concluding Remarks

Big data so far does not have a mathematical definition but can be described by several features such as its 5V features. This article focuses on the fourth feature—veracity—trying to indicate that 1) some problems of uncertainty processing, such as over 80% values of each case in a data set missing, appear and are possibly tackled only in the big data setting; and 2) processing of uncertainty embedded in the entire process of data analytics has a significant impact on the performance of learning from big data. We summarize most strategies (as shown in Figure 4) of big data computing and highlight the theme: going from big to small.

Acknowledgments

This work is financially supported by the China Postdoctoral Science Foundation (2015M572361), the Basic Research Project of Knowledge Innovation Program in Shenzhen (JCYJ 20150324140036825), and the National Natural Science Foundations of China (71371063, 61503252, 61473194, and 61170040).

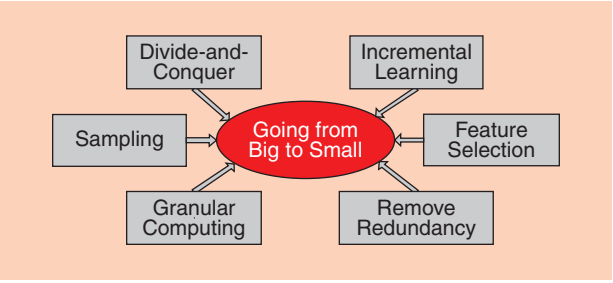

Figure 4. The theme for big data analytics is to change big into small. The uncertainty model and processing play a key part for these methodologies of going from big to small.

About the Authors

Xizhao Wang (xzwang@szu.edu.cn) earned his Ph.D. degree in computer science from the Harbin Institute of Technology, China, in 1998. He is currently a professor in the College of Computer Science and Software Engineering at Shenzhen University, China. His research interests include learning from examples with fuzzy representation, fuzzy measures and integrals, neuro-fuzzy systems and genetic algorithms, feature extraction, multiclassifier fusion, and the recent learning from big data. He is a Distinguished Lecturer of the IEEE Systems, Man, and Cybernetics Society and an IEEE Fellow.

Yulin He (yulinhe@szu.edu.cn) earned his M.S. degree in computer science and his Ph.D. degree in optical engineering from Hebei University in 2009 and 2014, respectively. He is currently a postdoctoral fellow in the College of Computer Science and Software Engineering at Shenzhen University, China. His research interests include the Bayesian network, artificial neural networks, evolutionary optimization, probability density estimation, and approximate reasoning. He is an IEEE Member.

References

[1] National Research Council, *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press, 2013.

[2] Gartner.com, "Big data," *IT Glossary*. [Online]. Available: http://www.gartner.com/it-glossary/big-data/

[3] M. Yang, P. F. Zhu, F. Liu, and L. L. Shen, "Joint representation and pattern learning for robust face recognition," *Neurocomputing*, vol. 168, pp. 70–80, Nov. 2015.

[4] Y. Zhai, Y. S. Ong, and I. W. Tsang, "The emerging 'Big Dimensionality'," *IEEE Comput. Intell. Mag.*, vol. 9, no. 3, pp. 14–26, 2014.

[5] D. Francois, V. Wert, and M. Verleysen, "The concentration of fractional distances," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 7, pp. 873–886, 2007.

[6] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *J. Mach. Learning Res.*, vol. 11, pp. 2487–2531, Sept. 2010.

[7] M. R. Gupta, S. Bengio, and J. Weston, "Training highly multiclass classifiers," *J. Mach. Learning Res.*, vol. 15, pp. 1461–1492, Apr. 2014.

[8] W. Zhu and S. P. Wang, "Matroidal approaches to generalized rough sets based on relations," *Int. J. Mach. Learning Cybern.*, vol. 2, no. 4, pp. 273–279, 2011.

[9] W. H. Xu, S. H. Liu, and W. X. Zhang, "Lattice-valued information systems based on dominance relation.," *Int. J. Mach. Learning Cybern.*, vol. 4, no. 3, pp. 245–257, 2013.

[10] X. Z. Wang and J. Huang, "Editorial: Uncertainty in learning from big data," *Fuzzy Sets Syst.*, vol. 258, pp. 1–4, Jan. 2015.

[11] M. I. Jordan, "Divide-and-conquer and statistical inference for big data," presented at the 14th Computing in the 21st Century Conf., Tianjin, China, Oct. 2012.

[12] H. B. He, S. Chen, K. Li, and X. Xu, "Incremental learning from stream data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 22, no. 1, pp. 1901–1919, 2011.

[13] S. L. Lohr, *Sampling: Design and Analysis*, 2nd ed. Boston, MA: Cengage Learning, 2009.

[14] C. L. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Inform. Sci.*, vol. 275, pp. 314–347, Aug. 2014.

[15] W. Pedrycz, *Granular Computing: Analysis and Design of Intelligent Systems*. Boca Raton, FL: CRC Press/Francis Taylor, 2013.

[16] Z. X. Zhu, S. Jia, and Z. Ji, "Towards a memetic feature selection paradigm," *IEEE Comput. Intell. Mag.*, vol. 5, no. 2, pp. 41–53, 2010.

[17] M. K. Tan, I. W. Tsang, and L. Wang, "Towards ultrahigh dimensional feature selection for big data.," *J. Mach. Learning Res.*, vol. 15, pp. 1371–1429, Apr. 2014.

[18] X. Z. Wang, R. A. R. Ashfaq, and A. M. Fu, "Fuzziness based sample categorization for classifier performance improvement," *J. Intelligent Fuzzy Syst.*, vol. 29, no. 3, pp. 1185–1196, 2015.

[19] X. Z. Wang, H. J. Xing, Y. Li, Q. Hua, C. R. Dong, and W. Pedrycz, "A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 5, pp. 1638–1654, 2015.

[20] X. Z. Wang, R. Wang, H. M. Feng, and H. C. Wang, "A new approach to classifier fusion based on upper integral," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 620–635, 2014.

[21] X. Z. Wang, Y. L. He, and D. D. Wang, "Non-naive Bayesian classifiers for classification problems with continuous attributes," *IEEE Trans. Cybern.*, vol. 44, no. 1, pp. 21–39, 2014.

[22] X. Z. Wang, L. C. Dong, and J. H. Yan, "Maximum ambiguity based sample selection in fuzzy decision tree induction," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp. 1491–1505, 2012.

[23] X. Z. Wang and C. R. Dong, "Improving generalization of fuzzy if-then rules by maximizing fuzzy entropy," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 3, pp. 556–567, 2009.

[24] Z. B. Xu, J. Y. Liang, C. Y. Dang, and K. S. Chin, "Inclusion degree: A perspective on measures for rough set data analysis," *Inform. Sci.*, vol. 141, no. 3-4, pp. 227–236, 2002.

[25] G. B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: A survey," *Int. J. Mach. Learning Cybern.*, vol. 2, no. 2, pp. 107–122, 2011.

[26] G. B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B*, vol. 42, no. 2, pp. 513–529, 2012.

[27] A. M. Fu, C. R. Dong, and L. S. Wang, "An experimental study on stability and generalization of extreme learning machines," *Int. J. Mach. Learning Cybern.*, vol. 6, no. 1, pp. 129–135, 2015.

[28] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[29] L. L. C. Kasun, H. Zhou, G. B. Huang, and C. M. Vong, "Representational learning with extreme learning machine for big data," *IEEE Intell. Syst.*, vol. 28, no. 6, pp. 31–34, 2013.

[30] J. Zhang, S. F. Ding, N. Zhang, and Z. Z. Shi, "Incremental extreme learning machine based on deep feature embedded," *Int. J. Mach. Learning Cybern.*, vol. 7, no. 1, pp. 111–120, 2015.

[31] J. Gama, "Functional trees," *Mach. Learning*, vol. 55, no. 3, pp. 219–250, 2004.

[32] R. Kohavi, "Scaling up the accuracy of naïve Bayes classifiers: A decision-tree hybrid," in *Proceedings of KDD'96*, pp. 202–207, 1996.

[33] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Mach. Learning*, vol. 59, no. 1, pp. 161–205, 2005.

[34] M. Sumner, E. Frank, and M. Hall, "Speeding up logistic model tree induction," in *Proceedings of PKDD'05, Lecture Notes in Computer Science*, vol. 3721, pp. 675–683, Oct. 2005.

SMC