An Efficient Classification method of Uncertain Data with Sampling

Jinchao Huang, Yulin Li, Kaiyue Qi and Fangqi Li

Abstract Current research on classification for uncertain data mainly focuses on the structural changes of classification algorithms. Existing methods have achieved encouraging results, however, they do not take an effective tradeoff between accuracy and running time, and they do not have good portability. This paper proposed a new framework to solve the classification problem of uncertain data from data processing point. The proposed algorithm represents the distribution of raw data by a sampling method, which means that the uncertain data are converted into determined data. The proposed framework is suitable for all classifiers, and then XGBoost is adopted as a specific classifier in this paper. Experimental results show that the proposed method is an effective way of handling classification problem for uncertain data.

Key words: classification; uncertain data; sampling; XGBoost

Jinchao Huang
Shanghai Jiaotong University, School of Cyber Security, Shanghai 200240, e-mail: hjc2015@sjtu.edu.cn

Yulin Li
School of Computer Engineering, University of Illinois at Urbana-Champaign, 508 E University Ave, Champaign, IL, 61820, U.S, e-mail: yulinli2@illinois.edu

Kaiyue Qi
School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China,e-mail: tommy-qi@sjtu.edu.cn

Fangqi Li
School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China,e-mail: solour_lfq@sjtu.edu.cn

1 Introduction

In recent years, with rapid developments of Internet and intelligent, the efficiency of social has been unprecedentedly improved. The effect, that machine learning and artificial intelligence can achieve, is far more than the effect of manual operation. In logical thinking, e-commerce, medical, financial and some other fields, business decisions become ever more reliant on machine learning and artificial intelligence. After the optimization, artificial intelligence systems can help to deliver goods and information to destinations at a faster speed and a lower price. As an important part of artificial intelligence, the performance of classification algorithm has receiving increasing attention. More accurate classification means a better sense of user experience, which also brings a lot of benefits to businesses.

Most of existing classification methods only apply to determined data, such as decision tree classification [10], K-Nearest Neighbor (KNN) classification [7], Nave Bayes classification [5], artificial neural network classification, support vector machine (SVM) classification [13] and so on. However, in practical applications, data often have characteristics of randomness, fuzziness and multiple uncertainty. Specifically, in many cases, we cannot get accurate data, but obtain the probability distribution of data, where traditional classification methods no longer apply.

There has been a growing attention on uncertain data mining. Zhang and Bi first proposed classification algorithms for uncertain data [1], and they extended SVM to process the data affected by noise. Qin et al. [8] and Smith [12] both improved C4.5, and presented uncertain decision tree algorithms DTU and UDT separately. There are also some scholars who modified nave Bayes classifiers to deal with the uncertain data [6] [9] [11].

The above methods mainly pay attention to improvements of model itself. Most of them need complicated computing, and therefore cannot satisfy the requirement of real-time in practical applications. In this paper, we propose a sampling-based classification algorithm for uncertain data. The new algorithm focuses on the transformation of data form, in particular to taking advantage of sampling to convert probability distribution form to determined data. When sample times are suitable, the data obtained can represent distribution of raw data perfectly, and then the original problem can be transformed into a traditional classification problem. Empirical comparisons conducted on a set of datasets demonstrate the effectiveness of our proposed approach.

The rest of this paper is organized as follows. In section 2, we give a brief introduction about the presence of data uncertainty. The new method is described in section 3. Descriptions of datasets and experiment results are presented in section 4. Section 5 concludes the paper and discusses the prospect of further research.

2 Modeling in the presence of uncertainty

Like traditional methods, a dataset of classification algorithm for uncertain data consists of d training tuples, represented as $\{t_1, t_2, ..., t_d\}$, and each tuple is associate with a feature vector $V_i = (f_{i,1}, f_{i,2}, ..., f_{i,k})$ and a class label $c_i \in C$. The difference is that each member $f_{i,j}$ of the feature vector is represented by probability distribution rather than deterministic value, as in Table 1. Table 1 shows a sequence of uncertain samples about credit card transaction, and each row represents a record.

Table 1 A Sequence of Uncertain Samples

ID	Age	Income(million yuan)	Maximum purchase(million yuan)	Purchase time	Label
s1	18−23	9.3−11.4	0.2−0.3	19:00−23:00	0
s2	27−34	13.9−17.0	0.6−0.7	16:00−20:00	1
s3	50−61	13.1−16.1	0.1−0.2	16:00−19:00	0
s4	54−67	10.0−12.2	0.2−0.3	13:00−16:00	0
s5	27−33	17.0−21.6	0.6−0.7	10:00−13:00	1

The common used method is to build a model that maps each feature vector composed of pdf to a probability distribution P on C. Specifically, when given a tuple $t_0 = (f_{0,1}, f_{0,2}..., f_{0,k}, c_0)$, the trained model will predict the class label c_0 based on $P_0 = M(f_{0,1}, ..., f_{0,k})$ with high accuracy. And it is defined that P_0 predicts c_0 if $c_0 = \arg\max_{c \in C}\{P_0(c)\}$. Especially in [12], before starting training model, a pdf would be stored as a set of s sample points by approximating the associated value $f_{i,j}$ with a discrete distribution, and this algorithm has achieved quite good results on accuracy. However, existing methods do not take an effective tradeoff between accuracy and running time. And there is still plenty of room for development.

3 The new method

As discussed above, most of existing methods change the classifier structure to make it applicable for uncertain data, where the feature form of input data is a probability distribution. These approaches only apply to some specific classifiers, which means poor portability. Hence, in this paper, we propose a new method that is suitable for all classifiers.

3.1 Framework of the new method

The framework of new method is shown in Fig.1. As we can see in Fig.1, the new method mainly contains two parts: the training phase and the testing phase. And

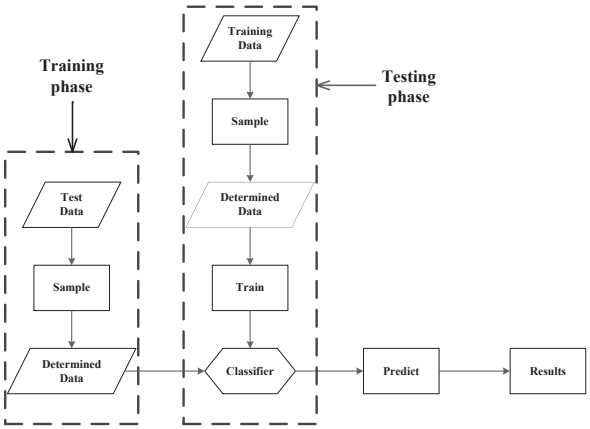

Fig. 1 Framework of the new method

the training part can also be divided into data processing section and model training section.

In data processing, we propose a sampling method to convert uncertain data to determined data. When facing the feature form of input data is a probability distribution, the sampling method samples each feature of every tuple according to its distribution, and then we can obtain a determined tuple composed of sampling values. By taking repeated sampling method, the raw data are expanded by a series of sampling tuples, which also means the distributions of raw data are represented by these sampled values. After that, the determined data are used as input to train the classifier. Then the same operation is performed on the test data. It is easy to see that the original problem is transformed into traditional classification problem, and this framework is suitable for all classifiers.

3.2 Classifier fitting

Gradient boosting algorithm is a powerful machine-learning technique and has produced some state-of-the-art results in data mining. It learns an ensemble of trees which makes a tradeoff between classification accuracy and model complexity. In this paper, we adopt XGBoost[2], a boosting algorithm proposed recently, as classifier. When facing a large-scale dataset, XGBoost is able to obtain the trained model

with good accuracy in a short time. The above advantage of XGBoost ensures that our method always achieves good effect.

4 Experiment and results

4.1 Datasets

To evaluate the performance of the new method, we conduct experiments on several standard benchmark datasets. The standard datasets are obtained from the UCI machine learning repository [3], including Ionosphere, Satellite, Wavefore-21 and Wavefore-40. The details are shown in Table 2. And because there is no benchmark uncertain dataset that can be found, we model the data uncertainty with a Gaussian distribution and a controllable parameter w, given in [12]. Supposing that the value of features is x_i, and $|X_i|$ is the threshold of original feature X_i. Then we define that $a_{i,t} = x_{i,t} - |X_i|$ and $b_{i,t} = x_{i,t} + |X_i|$, and they are regarded as mean and variance of the Gaussian distribution $f_{i,j}$, separately. From these, we can see that parameter w plays the role of determining the degree of injected uncertainty.

Table 2 Statistics of Datasets

Datasets	Feature numbers	Number of training set	Number of test set	Class numbers
Ionosphere	34	200	151	2
Satellite	36	4435	2000	8
Wavefore-21	21	4000	1000	3
Wavefore-40	40	4000	1000	3

4.2 Evaluation Metrics and Comparison Baselines

To measure the performance of various methods, we use accuracy and running time to analyze the results.

We compare the new method with two existing algorithms: VFDT [4] and UDT [12]. VFDT is a decision tree algorithm with fast learning speed when values of experimental data are determined. And UDT is a decision tree algorithm for data with static uncertainty, which has achieved a high accuracy. Note that, in the experiments, except for the VFDT algorithm, all the other algorithms use the data injected uncertainty as input.

4.3 Results and analysis

We first examine the accuracy of the algorithms, and results are shown in Fig.2−Fig.5, which correspond to four datasets respectively. For each figure, the horizontal axis is the value of parameter w, and the vertical axis is accuracy. As mentioned above, VFDT is a special case, which runs on datasets without uncertainty, so the performance curve of VFDT is always a straight line. From these figures, it is easy to see that the new method performs much better than the others. In Fig.2, all accuracies of new method achieve more than 94% under 4 different w conditions, and accuracies of the other two algorithms only attain 90% or so. As for Waveform-40 dataset, the promotion is more obvious, where the precision is improved more than 10 percent when $w = 0.01$. Results of all datasets have fully verified that our method has a great effect on accuracy rate, which also indicate that our sampling strategy can describe the distribution of raw data well.

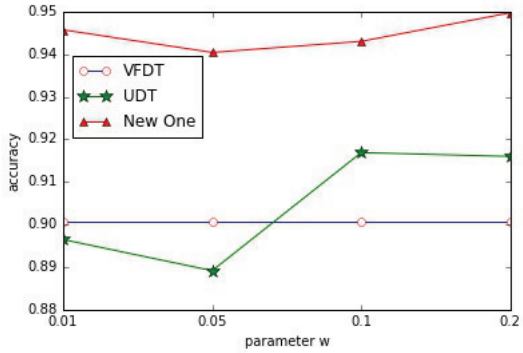

Fig. 2 Comparisons of accuracy (Ionosphere)

Then we study the execution time of the algorithms, and the results are represented in Table3. Because the running time of UDT is too long, we only give results of VFDT and our proposed method. From Table3, we can see that the running times of two algorithms are about the same. Experiments on ionosphere, waveform-21 and waveform-40 demonstrate that our method is faster, but result on satellite is reversed. In addition, the running time of our new method increases slowly with the increasing of uncertainty, meaning that it is more difficult to construct a proper model when there is more uncertainty. On the whole, our method has good performance on the execution time.

Such good results prove that the new method is an effective way of handling classification problem for uncertain data.

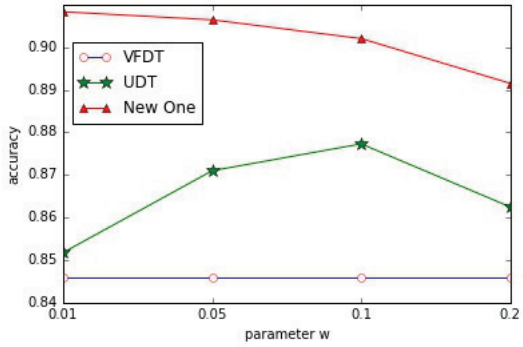

Fig. 3 Comparisons of accuracy (Satellite)

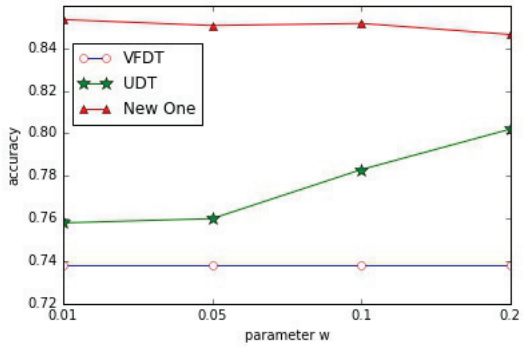

Fig. 4 Comparisons of accuracy (Wavefore-21)

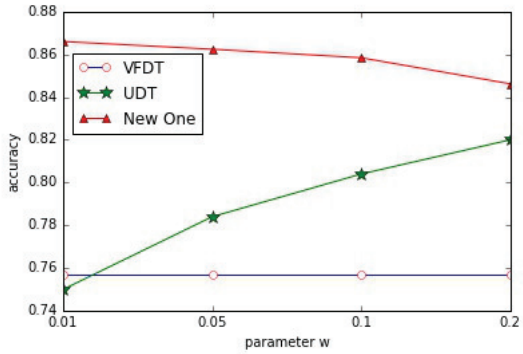

Fig. 5 Comparisons of accuracy (Wavefore-40)

Table 3 Comparisons of Running Time (unit: second)

Datasets	VFDT	New One($w = 0.01$)	New One($w = 0.05$)	New One($w = 0.1$)	New One($w = 0.2$)
Ionosphere	0.46	0.107	0.105	0.114	0.108
Satellite	4.12	14.11	14.65	16.62	18.04
Wavefore-21	5.62	4.96	5.19	5.5	6.12
Wavefore-40	10.62	7.24	7.71	8.07	8.72

5 Conclusion

In this paper, we proposed a new framework to solve the classification problem of uncertain data, from data processing point. The new framework is suitable for all classifiers, and we adopt XGBoost as classifier, specifically. Then we conduct a series of experiments on several standard benchmark datasets. Results show that the new method obtains a pretty good performance on accuracy and execution time, which fully proves the proposed method is an effective way of handling classification problem for uncertain data.

Acknowledgements This research work is funded by the National Key Research and Development Project of China (2016YFB0801003)

References

1. Bi, J., Zhang, T.: Support vector classification with input data uncertainty. Proc.of Neural Inf.proc.systems **17**, 161–168 (2004)
2. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system pp. 785–794 (2016)
3. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository (2017). URL http://archive.ics.uci.edu/ml
4. Domingos P, H.G.: Mining high-speed data streams. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 71–80 (2000)
5. Duda, R.O., Hart, P.E.: Pattern classification and scene analysis. Wiley (1973)
6. He, J., Zhang, Y., Li, X., Wang, Y.: Bayesian Classifiers for Positive Unlabeled Learning. Springer Berlin Heidelberg (2011)
7. Peterson, L.: K-nearest neighbor. Scholarpedia **4**(2) (2009)
8. Qin, B., Xia, Y., Li, F.: Dtu: A decision tree for uncertain data. In: Advances in Knowledge Discovery and Data Mining, Pacific-Asia Conference, PAKDD 2009, Bangkok, Thailand, April 27-30, 2009, Proceedings, pp. 4–15 (2009)
9. Qin, B., Xia, Y., Wang, S., Du, X.: A novel bayesian classification for uncertain data. Knowledge-Based Systems **24**(8), 1151–1158 (2011)
10. Quinlan, J.R.: Induction on decision tree. Machine Learning **1**(1), 81–106 (1986)
11. Ren, J., Lee, S.D., Chen, X., Kao, B., Cheng, R., Cheung, D.: Naive bayes classification of uncertain data. In: IEEE International Conference on Data Mining, pp. 944–949 (2009)
12. Tsang, S., Kao, B., Yip, K.Y., Ho, W.S., Lee, S.D.: Decision trees for uncertain data. IEEE Transactions on Knowledge Data Engineering **23**(1), 64–78 (2011)
13. Vapnik, Vladimir, N.: The nature of statistical learning theory. Technometrics **8**(6), 1564 (1997)