A survey paper on Frame work for classification of uncertain data

Shruti Sharma, Jigyasu dube

Abstract— **The classification of uncertain data have a need to pay more attention in recent years due to the appearance of more and more database with uncertainties, such as sensor database, location database and biometric information systems. The objects in the uncertain database are vague and imprecise.Often we assume that data values are exact or precise in the database but data is sometimes inherently indecisive. There are some reasons owing to that errors are creep inside – 1. Data obtained from physical device are often imprecise due to measurement errors 2. Quantization error introduced by the digitization process. 3. Applications like sensor network data values are continuously changing and recorded information is always stale 4. Indecision also comes from repeated measurements.**

Due to above reasons traditional data mining techniques cannot be applied directly on uncertain database.

So there is a need to apply a novel technique that will be able to handle the uncertain database. In this paper we design a framework for classification of uncertain data using support vector machine and remove the problem in SVM using novel fuzzy C-means clustering algorithm (NFCC)

Index Terms— **Uncertain data, SVM, Fuzzy C-means clustering, active and supervised learning, multy spectral data , SMO**

I.INTRODUCTION

Data that contain specific uncertainty if data is uncertain so automatically such kind of uncertainty comes in Database and it becomes uncertain database. Data uncertainty is universal in real word function. In some situations, the data may have errors or may only be partially complete. For example, Database like sensor networks create big volume of uncertain data sets. In other cases, the data points that are related to objects can also have some sort of probabilistic value. Like this, surveys and imputation techniques create data which is uncertain in nature. This has created a need for classification of uncertain data. Classification of this kind of data become a big challenge. Data is not a simple point in space but represent by uncertainty reason. Formally we consider a set of n objects in a D- dimensional space. There exists a lot of work related to the classification of uncertain data. Among them SVM is widely used classification approach based on class conditional density estimation and class prior probability. The key problem in SVM is class conditional density estimation. For uncertain classification problems, we have to learn the class conditional density from probabilistic data objects represent by probability distributions using fuzzy c-means clustering algorithms. In order to extend the naïve Framework using Novel fuzzy c-means clustering algorithms, In this research, we develop a new classification method for large dataset. The uncertainties of the individuals attribute are modled by a probability density function or other statistical parameters such as the variance.takes the compension of the fuzzy C-means clustering and SVM.The algorithm proposes in this research has a similar idea as the sequential minimum optimization in order to work with large datatsets .

Manuscript received May, 2014.

Shruti Sharma,Artificial Intellegance (IT) ,Rajive Gandhi Prodougiki,Bhopal (M.P.)/Shri Vaishnav Institute of Technology,Indore/, Indore, India, / 9302822147

***Jigyasu dubee**, Artificial Intellegance (IT) ,Rajive Gandhi Prodougiki,Bhopal (M.P.)/Shri Vaishnav Institute of Technology,Indore/India ,*

II.RELATED WORKS

A. Meena , K. Raja,[1] In this research work PET-SFCM(Special Fuzzy C-Means) is used on PET scan image datasets. In the field of medical, PET (Positron Image Tomography) helps to identify the disease and localize it. PET also helps to plan the right treatement .

In this research work PET-SFCM(Special Fuzzy C-Means) is used on PET scan image datasets. Experimental results are also compared with K- means clustering The performance of the SFCM provides satisfactory results . In future, to calculate objective based quality assessment that could analyze images and report their quality without human involvement. *Ming-yen lin, Cheng-tai fu, Sue-chen hsuesh,*[2]Data bases like Sensor network and Location based services collect or produce data with an existential probability that is known as Uncertainty. So mining data from Uncertain Datasets become a new challenge

The frequency or support of a dataset or itemset can be obtained by counting the number of occurrence in the dataset but in uncertain dataset it is difficult to obtain because the occurance of a item set is inexact.

item set is *frequent* if its expected support is exceeds the user specified minium support threshold.

The above problem is known as Increamental Update problem . In this paper Ming-yen lin provides an algorithm known as P-FUP algorithm (Probabilistic Fast Update) . The experimental results shows that P-FUP algorithm is efficient and 2.8 times faster then other algorithms.

Le Li Y, Zhiwen u, Zijian Feng,[3]*Xiaohang Zhangl,*As we know that dataMining becomes a huge field hence classification of data is also becomes a big chellange in at al [2].There are various classification technics like hard classifier, Soft classifier etc. In this research work *Le Li Y, Zhiwen u, Zijian Feng,Xiaohang Zhangl provides* automatic soft classifier with the help of Fuzzy C Means Clustering with Fuzzy Distance Function algorithm. This algorithm

provides Fuzzy objects with into its clusters then clusters splits automatically based on the objective function until this function reaches the threshold given by the user. According to experimental results the automatic soft classifier works well in different types of database with uncertain data.

Mehdi adda, Petko valtchav ,[4] *Rokia Missoui*,Data mining is also applicable on Web related data that is known as Web mining. In this datamining technic tackle processing semantically annotated data *Mehdi adda, Petko valtchav , Rokia Missoui* tackle fundamental mining problem and provides its solution with the help of mining method known as xPminer. This method performs a complete and nonredundant traversal of the pattern space in the sense that it discovers all the frequent patterns while generating any candidate at once.

Hailong Xu ,[5] *Yong Liao , Xiaodan wang,* SVM is a Learning algorithm which solves two class pattern recognition problem. It is assume that SVM provides straightforward labeled data but SVM gives some sort of probabilistic data to overcome this problem in this research *Hailong Xu , Yong Liao , Xiaodan wang propose an* Uncertainty Based active learning approach for SVM that reduses annotation effort during the learning process of SVM. In this research two threshold are used to restrict uncertainty Figures

As said, to insert images in Word, position the cursor at the insertion point and either use Insert | Picture | From File or Aimin Wang , wenying Ge,[6]At al in [2] & [5] training set of the SVM contains probabilistic information. Tto solve the above mentioned prolem, In this paper an algorithm of nonlinear Support Vector Classification Machine based on fuzzy theory is defined by Aimin Wang & wenying Ge . Due to the restriction of the confidence $\lambda\ (0 < \lambda \leq 1)$, Aimin Wang & wenying Ge are using the classification method in fuzzy theory to solve the problem of constraining programming of uncertain chance. By establishing a chain like this: constraining programming of uncertain chance → clearlyequivalent programming → programming of antithesis, the universal algorithm of nonlinearly dividable Support Vector Classification Machine based on Fuzzy theory can be deduced.

Bjorn wake,[7] John Atli , Benediktonson,As we see in al at [2] SVM and its classification approach and problem related to it. In this research work two SVM's are fused to each other in this manner that each datasource treated separately and classified by a SVM .Instead of fusing the final classification output the original output of each SVM discriminant function are used in the subsequent fusion process.The aim of this fusion is to improve the result of the single SVM..In this approach two thresholds are used to restrict the uncertainty .The advantage of this fusion is to accelerate performance of SVM and extend accuracy of SVM by using extended kernel function.

Janez Demsar,[8]In this article Janez Demsar recommend a set of simple, yet safe and robust non-parametric tests for statistical comparisons of classifiers: the Wilcoxon signed ranks test which is used in comparison of two classifiers and the Friedman test with the corresponding post-hoc tests which is used in comparison of more classifiers over multiple data sets.

Gustavo Camps,[9]In this paper, we have addressed the framework of kernelbased methods in the context of classification of hyperspectral data. In particular, we have analyzed and compared four kernel-based techniques both theoretically and experimentally. the use of SVMs is more beneficial, yielding better results than the other kernel-based methods, ensuring sparsity, and at a much lower computational cost.

Kang H. Lee & Byeong H. Kang2,[10]In supervised learning approaches of text classification have a need of large number of labeled examples to get a high level of effectiveness.This causes a large number of burden on human experts. There are two common approaches present ,that reduce the amount of labeled examples : (1) uncertain examples for human-labeling and (2) unlabeled data with a small number of labeled examples. While previous work in text classification focused only on one approach, *Kang H. Lee & Byeong H. Kang2 provide* a framework that combine both approaches in similarity-based text classification. By implementing this new thresholding strategy (RinSCut) for uncertainty sampling, *Kang H. Lee & Byeong H. Kang2* gives a new framework which automatically selects informative uncertain data that should be presented to human expert for labeling and positive-certain data that are directly used for learning without human-labeling.With our similarity-based learning algorithm (KAN), experiments have been conducted on Reuters-21578 data set. This proposed scheme has been compared with random sampling and previous conventional uncertainty sampling, based on micro and macro-averaged F1.

III. Comparative study

S No	Name of auther	Tech-nic Used	Descrip-tion	Advant-age	Limita-tion
1	A Meena M Raju	Spati-al Fuzy C Mean Clust-erin g	Spatial Fuzy C Mean Clustering used on PET scan image data set and experimen tal result are also compared with K means clustering.	SFCM provides the satisfacto-ry result and this algorithm is very effective to calculate objective based quality assessme nt that could analyse images and provide their report without human involvme-nt	The above Work is totally based on Alzieme r's disease is an eurodege nerative disorder associate-d with memory loss in this type of work volumnat-ric analysis of hyppoca mpus is necessay. The analysis and

					segment--ation of hippoca-mpus is very complicat-ed and time consumin g
2	Mieng Yen Lien, cheng Tai Fu	PFU P algori thm	Ming yenLin Provides an P-FUP Algorithm to overcome Uncertaint y of sensor networks P-FUP is 2.8 times faster then Other algorithms .	P-FUP is 2.8times Faster then Other algorithm s.	**1**. No of candidate generated by PFUP is less than generated by MBP. **2**. Mieng YenLien, Cheng Tai Fu have not verify accuracy.
3	Leli , Zhiwe n, Yu Zijian Feng	Fuzz y C Mean Clust -ring With Fuzy dista-nce Funct -ion	In this research work a Soft Classifier is defined with the Help of Fuzzy-C Means Clustering and Fuzzy distance Function which can classify All types of data from any type Of database.	This Soft classifier can work on any type of database with uncertain data.	It is time consum--ing because the algorithm first assigns the fuzzy objects into their correspon -ding clusters This algorithm provides Fuzzy objects with into its clusters then clusters splits automatic -ally based on the objective function until this function reaches the threshold given by the user
4	Mehdi adda ,Petko Valtch ev	Sequ-antial l Pat-te rn Mini n-g	In this research work MehdiAdd a provides a new Classificat	The method XPminer can mined all of the pattern space in	It is not work on Depth First Method For pattern
			ion technic known As XPminer.	the sense that it discover all the frequent patterns.	mining.
5	Hailon g, Xu Yong Lio	Actie Learn -ing Appr oach	Hailong Xu Yong Lio Provides Uncertain ty Based active learning approach for SVM that reduses annotation during the learning process of SVM.	This method reduses the learning cost while acheiving desired performa-nce.	In this research the Sample Distribu -tion should be taken into account and selected instances should not be represent --able and similar to the instances have been selected i..e. more informa-t ive for learning.
6	Aimig wang, Wengi-ng,,Zhi ming yang	Fuzy Logic on SVM	this article discusses an algorithm of non linear SVM based on fuzzy theory	It is efficient method.	It is comparat ively complex and time consumin g method.
7	Bjornw askeJo n Atli	Fusio n of two SVM .	In this research work two SVM's are fused together inThis manner that each data source treated separately to accelerate SVM's Performan -ce.	Due to fusion of two SVM's it is more accurate than using single SVM.	The fusion of two SVM is compara--tively costly than A single SVM.

8	Janez Demsar	Statistical Comparisons of Classifiers Over Multiple Data Sets	In this research work the Wilcoxen And Friedmen tests are used.	This method works on real word data	There is however no golden standard for making such comparisons and the tests performd often have dubious and lead to unwarranted and unverified conclusions.
9	Gustavo Camps-valls ,Lorenzo Bruzz	Kernal Based methods	This research work Addressed the frame work of kernel based methods in the context of classification of hyperspectral data. In particular, we have analyzed and compared four kernel-based techniques	It gives batter results than other classifica--tion technics. It has lower computational cost.	the solution offered by SVMs is sparse in the sense that there is a high number of null multipliers and hence the correspoding training samples are considered irrelevant for the classification.
10	Kang h lee, Byeong Hkang	A Fraework to Combine 1.Selecting informative uncertain exam-ples for human labe-ling .2.Using many inexpensive unlabeled data With a small no of Labeled example	Kang h lee, Byeong H kang Propose a new framework which automatically selects informative data. .	Using more positive-certain examples (i.e., 500 for the Reuters-21578) works slightly better than the smaller number of positive-certain ones(250 examples).	Using more positive-certain examples (i.e., 1,000 examples for the 20-News groups and 500 for the Reuters-21578) works slightly better than the smaller.

IV. OUR RESEARCHE METHODOLOGY

This Novel fuzzy C-means clustering algorithm NFCC works for discriminative patterns directly and efficiently. Our proposed approach works for this uncertainty, i.e. indecisive data. NFCC directly mine the different patterns based on probability function because uncertain data fields' attributes have no longer confident values. Our propose approach mines the most discriminative patterns directly and effectively on uncertain data NFCC is less time consuming as it directly mines the patterns the time consumed in pattern mining and feature selection is reduced. NFCC address the problem of extending traditional SVM model to the classification of uncertain data. The key problem in at al [5]current approach is class conditional probability estimation, and kernel density estimation is a common way for that. We will extend the kernel density estimation method to handle uncertain data. This reduces the problem to the evaluation of double-integrals. For particular kernel functions and probability distributions, the double integral can be analytically evaluated to give a closed-form formula, allowing an efficient formula-based algorithm. In my research work propose a new framework for classifying sequential data based on discriminative iterative patterns. Appling feature Fuzzy C-means Clustering Algorithms proposed to identify highly discriminative patterns which distinguish the failing traces from normal ones. in my research to extend the naïve SVM method to handle uncertain data I develop a new classification method for large data sets. We will use for uncertain data classification fuzzy C-means clustering and SVM. The algorithm proposed in order to work with large data sets, partition the original data set into several clusters and we will illustrate through the experiment results that the number of support vectors obtain using the SVM classification based on fuzzy clustering . We will carry out a comprehensive experiment using many public datasets under varying uncertain parameters.

In previously done work there were a lot of work on finding discrimination among patterns, but they all are time consuming, as they have to first mine the complete set of frequent patterns using some association classification

technique. We will extend the kernel density estimation method to handle uncertain data. This reduces the problem to the evaluation of double-integrals. For particular kernel functions and probability distributions, the double integral can be analytically evaluated to give a closed-form formula, allowing an efficient formula-based algorithm. This work propose a new framework for classifying sequential data based on discriminative iterative patterns. Appling feature Fuzzy C-means Clustering Algorithms proposed to identify highly discriminative patterns which distinguish the failing traces from normal ones demonstrate that theses two uncertain method have a improved performance on reducing the consequence of uncertain information and significantly improve the classification accuracy.

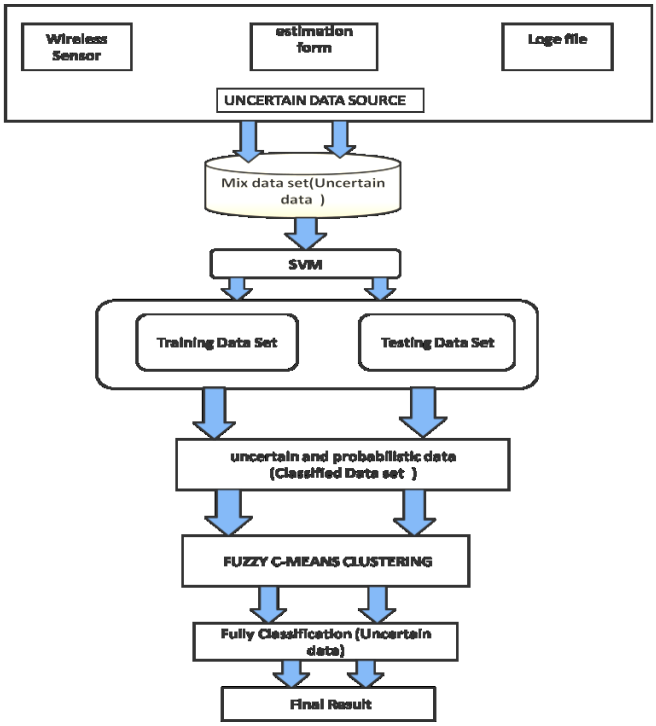

Figure 1: framework for classification uncertain data

V. CONCLUSION

In this paper survey We design a efficient framework for classification uncertain data using Support vector machine and survey of uncertain data classifications and remove the problem in SVM Using proposed Novel Fuzzy C-means Clustering Algorithms (NFCC) to get uncertain support vector machines, which are fuzzy support vector machine and Fuzzy C-means Clustering Algorithms. and the principle of these uncertain methods reducing the effect of outliers is explained. We want to perform experiment Real life data set data set demonstrate that theses two uncertain method have a improved performance on reducing the consequence of uncertain information and significantly improve the classification accuracy.

REFERENCES

[1] *A. MeenaK. Raja "Spatial Fuzzy C-Means PET Image Segmentation of Neurodegenerative Disorder ISSN : 0976-5166 Vol. 4 No.1 Feb-Mar 2013*

[2] Ming-Yen Lin,Cheng-Tai Fu, Sue-Chen Hsueh," Incremental Update on Probabilistic Frequent Itemsets in Uncertain Databases" ICUIMC'12, February 20–22, 2012, Kuala Lumpur, Malaysia.

[3] *Le Li Zhiwen Yul2* Zijian Fengl Xiaohang Zhangl" Automatic Classification of Uncertain Data by Soft Classifier" Proceedings of the 2011 International Conference on j\fachine Learning and Cybernetics, GuiJin, 10-13 July, 2011*

[4]Hailong Xu, Xiaodan Wang, Yong Liao, Chunying Zheng," An Uncertainty sampling-based Active Learning Approach For Support Vector Machines" 978-0-7695-3816-7/09 2009 IEEE.

[5]*Björn Waske, Student Member, IEEE, and Jón Atli Benediktsson, Fellow, IEEE," Fusion of Support Vector Machines for Classification of Multisensor Data" IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 45, NO. 12, DECEMBER-200.*

[8] *Janez Demsar "Statistical Comparisons of Classifiers over Multiple Data Sets" Journal of Machine Learning Research 7 (2006) 1–30*

[9] Gustavo *Camps, Lorenzo Bruzzone "Kernel-Based Methods for Hyperspectral Image Classification" IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 43, NO. 6, JUNE 2005*

[10] L. Antova, C. Koch, and D. Olteanu, "10ð106Þ Worlds and Beyond: Efficient Representation and Processing of Incomplete Information," Proc. 23rd IEEE Int'l Conf. Data Eng. (ICDE), 2007.

