ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Knowledge Discovery in Big Data Soft Set Environemnt Using Bijective Soft Set: Rectify Misclassified Data

Jyoti Arora
Dept. Computer Engineering & Technology
Guru Nanak Dev University
Amritsar, Punjab

Kamaljit Kaur
Dept. Computer Engineering & Technology
Guru Nanak Dev University
Amritsar, Punjab

Abstract: In this information age, most of the studies have focused on analyzing the complexities of whole datasets. Data mining process works for finding patterns in whole datasets using computerized techniques and machine learning. Most of the researchers have worked on it and followed data elimination approach. This leads to loss of information and raise misclassification. Mining of data directly does not have feature to provide useful information about individual attribute. In this paper, we have considered dataset of Breast Cancer from UCI repository. The proposed method utilized bijective soft set theory to extract knowledge based on parameters of datasets. Here, problems occurs during classification has been considered. As, we need to know about which elements are misclassified and how they can help to improve the accuracy and quality of dataset. So, we have proposed method for rectification of misclassification to increase the accuracy. The aim of thi papaer is to better understand the true positive and false negative terms by analyzing and rectifying the misclassified data.

Keywords: Big data, Data mining, Boundary Values, Misclassification, Bijective Soft Sets

I. INTRODUCTION

In this massive data age, data which is not treated directly by human beings or software applications is surrounding to all of us. Various technologies like healthcare, engineering, social sites, business services and many more are creating data in high speed. Now a days, data analyses and extracting knowledge is not an easy task due to huge growth of data. These premises leads to data mining[1] of big data, which is becoming most thrust topic in data growing world. Huge volume of data has surpasses the traditional machine learning approaches like supervised[2], unsupervised[3] and reinforcement[4] approaches, of data mining[5]. And data analyses is becoming a big challenge in this big data world. The formation of big data is increasing because of the usage of new technologies like internet, portable electronic devices and cloud computing. Big data has various characteristics like huge volume, high speed, different values and variety[6]. Due to these features big data requires advanced machine learning approaches like deep learning[7], kernel learning[8], active learning[9] and many more[10-12]. Excerpting the useful information does not dependent on the design of the approach but mainly rely on the suitability and efficiency of data. Various negative flaws like missing values, mislabeled values, hard to classify values, boundary values and noisy values effects the data which is used to extract information. It is obvious that this will leads to low accuracy and high misclassification.[13] Thus rectifying this misclassification is considered as a challenging task

Misclassification is the main concept to consider while extracting knowledge. Dataset considered to be classify can have some boundary values based on the range values. That values are difficult to categorize and consider as misclassified elements. Most of the researchers have defined misclassification in their own way. Healy et. al[15] explains misclassification occurs whenever single instance or items to be classify are not classified into right category. Researchers need to remonstrate this misclassification concept. They have focused on learning models without understanding the datasets attributes. To deal with fragmented learning and questionable data, Rough sets, Fuzzy sets, Soft sets turn out to be a viable theory. In this paper, we have used bijective soft set theory as the extension of soft set theory. We propose a novel approach for classification in view of bijective soft sets[14]. Rank based approach is used to identify misclassified values within attributes. Highest ranked misclassified attribute is rectified by the use of boundary value analysis.

The main contributions of this paper are:

- To maintain the quality of data.
- To reduce the misclassification based on parameters like true negative, False Positive and recall.
- and to increase the accuracy based on true positive and false negative using bijective soft set theory.
- Rectify the misclassified data in addition to maintaining the quality of data by using rank based approach and boundary value approach.

Figure 1 describes the outline of the paper and all the sections have been organized as the following: Second section discusses related work. Third section describes the some concepts named as Bijective soft set, Naive Bayes and support vector machine. Fourth section defines the problem and discusses objectives. Fifth section discusses the proposed work. Sixth section describes the dataset and setup. Seventh section shows the comparative results and Last section discusses conclusion and future challenges.

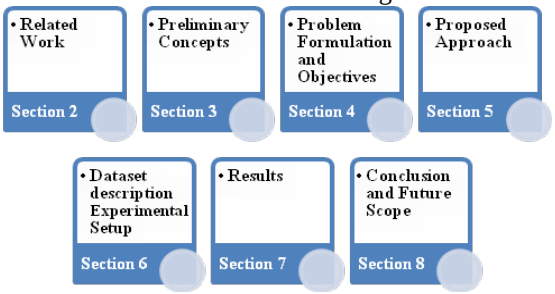


Figure 1. Outline of the Paper.

II. RELATED WORK

This section discusses the literature survey. Most of researchers have worked in the area of data mining, big data, n, bijective soft set theory and misclassification concepts. Their work has been discusses as follows:

Ke Gong et. al[16] introduces a tool bijective soft set(bss) theory to deal with uncertain problems. In this paper, Ke Gong et. al has introduced BSS concept and some of the operations performed on bijective soft set theory. Amit et.al [21]discusses the techniques to find out the missing information in datasets. As it leads to low accuracy. Author has used four datasets and demonstrated high classification accuracy. et. al has proposed maximum distance instance selection (MDIS) approach for completing missing information. Ke Gong et. al[22] has used bijective soft set theory for evaluating supply chain information sharing and management. Varun et. al[25] has used BSS theory in product design. This theory has performed mapping between customer requirement and design concepts. In this paper, Simon et.al [18]has proposed a feature selection method for analyzing big datasets. Proposed method has perfromed clustering and separation of parameters using measure of dispersion. This method is one of the enhancement of traditional feature selection methods. Sowmya et. al[17]has discusses some future challenges and opportunities regarding big data. Author has presented a framework for big data mining. Liao Lung et. al[24] has proposed a decision tree method for extracting knowledge from faulty data of sensor devices in cloud computing environment. This method has given high accuracy and efficiency of dataset. Charu et. al[19] has presented candidate generate and test algorithm to deal with data uncertainty. This paper has given improved pattern mining as compared to traditional pattern mining. Michael et. al[20] has deal with understanding the problem of instances that are misclassified in whole data. In this paper, he has identified instance hardness using various datasets. In this paper Lianmeng et. al[23] has proposed a approach HBRBCS known as hybrid belief rule based classification system. This proposed method is based on uncertain dta and is combination of data driven belief rule base and a knowledge driven belief rule base. In this paper Elena et. al[26] has proposed RSF(reliability structure function) based on uncertain data. Here, fuzzy decision tree termed as FDT has been used for creating RS function. In this paper, Alan et. al [32] has developed mathematical model for misclassification. The presented model is similar to Markov model type. Udhaya et. al[28] has given a classification method for electrocardiogram arrhythmia. This paper has used a classification method bijective soft set theory. Here, signals are divided into five different classes and given a comparative study based on precision, recall and f-measure. Leszek et. al[31] has given a advance method known as Gini Index for decision tree. This method has been used for big data mining based on misclassification. This paper has resulted high accuracy. Rendall et. al[30] discusses the opportunities and challenges related to big data tools. This paper has given different approaches for analyzing health data. Saket et. al[29] has proposed a technique to find the relationship between the different categories of document. This paper has resulted improvement in precision and recall parameter. Udhaya et. al [27] has used a feature selection technique for classification of Gene Expression data. In this paper three datasets has been taken and comparative study has been presented based on accuracy.

III. PRELIMINARRY CONCEPT

In this paper bijective soft set theory has been used for classification and Proposed approach has been compared with naive bayes and support vector machine. This section discusses the basic concepts of these approaches:

A. Bijective Soft Set

Bijective soft set is an extension of soft set. It is defined as:

Definition1: [34]Consider U is universal set and Y is set of parameters. (X,Y) is known as soft set over U iff X is a mapping of Y into power set of U where X is a mapping given by X:Y->P(U) where ∀ φ. then (X,Y) is BSS, if (X,Y) is such that

- All the instances are mapped to each attribute. i.e. Union(X(e)) = U for each e □ y.
- Every instance has unique attribute. i.e. any two parameters e_1 and e_2 □ y, $e_1 \neq e_2$, $X(e_1) \cap X(e_2) = \varphi$

In our proposed work we have used bijective soft set for classification.

Operations Performed on BSS: The attribute values derived from Y are compared against the range derived from corresponding attribute values. In case of successful classification Attribute is classified within set (X,Y), otherwise considered as misclassification.

Example 1: Let R1(Min) and R2(Max) define a range in which attribute values must lie. Then set is constructed as

If R1 < A and R2>=A then

$$\text{Build } (X_i, A_i) \wedge (X_{i+1}, A_{i+1}) \text{.....Equation(1)}$$

Equation(1)is for buildup of bijective soft set. This set contains all the attribute values which are classified. In case attribute value is lying at boundary or not in derived range then degree of misclassification will increase.

Sets are build of every distinct attribute and AND operation is applied to combine set together. Let (W,Y) is another set formed through classification on some another attribute. Result of both the operation are merged by the use of AND operation.

(X,Y) AND (W,Y). By repeating the operation on every distinct attribute bijective soft set is obtained and misclassification has been computed.

B. Naive Bayes Theorem

This classifier is based on probability theory named bayes theory. It is one of the easy classifier in terms of understanding. It is applicable for large dataset because it doesn't require any iterative attribute. This classifier is not only followed probabilistic approach but it follows conditional probabilistic approach. This classifier works slow in case of exponential data growth.

Conditional Probabilistic equation is given as

$$P(X|Y) = P(X)\,P(Y|X) / P(Y) \text{.....Equation(2)}$$

where P(X|Y) = Fraction of data in which Y is true for the values for that X is true. Y= {y1,y2,.....yn}. Equation (2) is for computing conditional probability.

This process of classification can be used for classifying text data, Spam email filtration and many more for online applications. [37]

C. Support Vector Machine

Cortes and Vapnik has given a classifier named Support Vector Machine. This method defines the hyper plane to classify the elements into various classes. The equation for designing the hyper plane is:

$$g(y) = \alpha_0 + \alpha^t y \text{ .....Equation(3)}$$

The optimal value of hyper plane is taken as

$$|\alpha 0 + \alpha t\ y| = 1\text{.....Equation(4)}$$

The working principle of support vector machine is to group the data into classes and further splits into classes. The main of this classifier is to maximize the hyper plane margin between true and false samples. The prediction strategy of SVM is performed by using sides of hyper plane. It can handle all type of problems like linear as well as non linear.[36]

IV. PROBLEM FORMULATION

Molodstov[33] has first introduced the soft set theory. In 2016, Ke Gong[34] has enhanced this theory and used it for getting shoreline resources. Bijective soft set theory have parametric dependency feature. In this paper, To achieve optimal performance BSS has been used with big data set to determine misclassification degree. Further boundary value analysis is performed to rectify misclassification.

A. Objectives

The main objectives for this proposed approach is to analyze the big data in efficient manner. Main goal of this paper is to introduce a fault handling classifier for huge data analysis. The classifier perform various functions like partitioning of data based on ranges for big data analysis, analyses of attribute those create misclassification, rectification of misclassification, improve accuracy.

V. PROPOSED APPROACH

Our proposed approach is based on classification of bijective soft set. Bijective soft set is classified based on rules formed on the basis of boundary value analysis. Algorithm 1 describes the steps of proposed approach is listed as under:

Algorithm1 BSSCLASSIFICATION
Input: *Dataset*
Output: *Misclassification, Accuracy, TP, TN, FP, FN, Precision, Recall.*

A. Obtain dataset and construct bijective soft set for all the attributes in the form (F,B) using Equation (1).
B. Compute range from dataset using Max and Min values obtained from dataset attributes.
C. Compare attribute values against range(<Max,Min>)
   If(Value(Attribute) not in Range<Max,Min>) or (Value(Attribute)= =<Max,Min> then
   Misclassification computed.
D. Repeat step B and C for every attribute
E. Perform arrangement of attributes on the basis of degree of misclassification and assign highest Rank to Maximum misclassification giving attribute.
F. Choose highest Ranked attribute and perform Boundary value analysis.
G. Rectify the misclassified value corresponding to misclassification giving attribute.
H. Perform step E to G for every value of corresponding attribute those have misclassification.
I. Calculate misclassification, accuracy

A. Example of Proposed Approach:

Let $X=\{X_1,X_2,X_3,\text{---------}X_N\}$ indicates attributes present within the dataset. $T=\{t_1,t_2,\text{----------}t_m\}$ are tuples or attribute values. $BX=\{B_1,B_2,\text{----------},B_n\}$ are the Boundary Values. Rank allocation process corresponds to R={A,B,C,D}. Ranks determine importance corresponding to attributes. Ranks are allocated following the mapping process.
The mapping process is given as under

$$(X,T)\text{->}BX$$

In case above mapping process hold, misclassification is detected. BSS is constructed in case tuple is found for a particular class.

$$\{BS,X_{i+1}\}\text{->}BSX_i \cup BSX_{i+1}$$

If BS□B then Misclassification corresponding to attribute set T is added to M={M1,M2,-----Mm} where M1<M2<M3-----<Mm.
If M1=1 then M2 is obtained by adding 1 to M1. This process continues with highest misclassification degree Mm. Degree of importance depends upon degree of misclassification.
In case Attribute X1□Mm then rank 'A' is allotted to that attribute. This process in descending order of Misclassification. In general, if X□M then mapping process $\{X_i, M\}$-> R is validated. Mechanism of rectification to dataset begins after identifying misclassified attributes. Attribute with maximum misclassification is identified using

$$\{X_i, M\}\text{->}Max(R)$$

Tuple corresponding to $\{X_i,M\}$->Max(R) is removed from boundary and placed at some optimal position selected randomly.

$$(X_i, T_i\}\text{->}random(T_i)$$

This process is repeated until all the attribute values are removed from the boundary. The process of classification is performed again to determine misclassification.

VI. DATASET DESCRIPTION AND EXPERIMENTAL SETUP

A. *Dataset Description*

For our experiment dataset has been collected of breast cancer from UCI learning repository. This data contains11 attributes and 500+ elements. To extract useful information, Algorithm 1 has been followed This dataset is created by University of Wisconsin Hospital. Data has been collected from Nov. 1991 to January 1989. Data considered is in numeric data type. Medical data is critical to consider as it contains critical attributes clump thickness, Cell size, Cell shape and many more.[35]

B. *Experimental Setup*

In this experiment, simulated environment has been created in Netbeans 8.1. This environment works as classifier, in which dataset name and attribute to classify has been considered as input and further degree of misclassification, accuracy after rectification of misclassified elements are considered as output.

VII. RESULTS

We have proposed a method in which classification is based on bijective soft set and further misclassification rectification is performed. In our point of view, there is no any existing classifier which provides high accuracy in addition to rectifying misclassification in providing dataset. In this paper, Various parameters have been considered like Precision, Recall, F-measure, True positive, True negative, False Positive and False Negative and results are obtained by comparing with various existing algorithm like Naive Bayes Approach and Support Vector Machine.

Table 1 has given comparative results based on precision, recall and F-measure. Figure 2 represents the comparative study of proposed approach with support vector machine and naive bayes classifier graphically. Figure 3 represents the True Positive, True Negative, False Positive and False Negative values of these three approaches. These all the results have been computed on the same dataset in simulated environment.

A. *Accuracy Metrices*

In this section, metrics has been discussed. Based on these metrics comparative study has been taken.

Precision: This parameter indicates the correct classification.

Precision =true positive / (true positive + false positive).

Recall: It indicates the degree with which classification is made by the model being used.

Recall = true positive / (true positive + false negative)

F-measure: It is the combination of precision and recall. It computes harmonic mean of both the terms.

F-measure = (2*precision* recall) / (precision + recall).

True Positive(TP): This metric describes the instances that follows the condition correctly. ***True Negative(TN):*** It describes the instances that doesn't follow the condition in the absence of condition. ***False Positive(FP):*** It describes the instances that follows the condition in the absence of condition. ***False Negative(FN):*** It describes the instances that doesn't follows the condition in the presence of condition.

Accuracy: It is defined as correctly classified instances. It has been calculated as:

Accuracy = (TP + TN) /(TP+TN+FP+FN)

Misclassification: It is defined as incorrectly classified instances. It has been calculated as:

Misclassification = (FP +FN) / (TP+TN+FP+FN)

Table I. Result in terms of distinct parameters is givem as under

Accuracy Metrices	*Naive Bayes Approach*	*Support Vector Machine*	*Proposed Appraoch*
Precision	0.98	0.966	1.000
Recall	0.967	0.969	1.000
F-measure	0.993	0.963	1.000

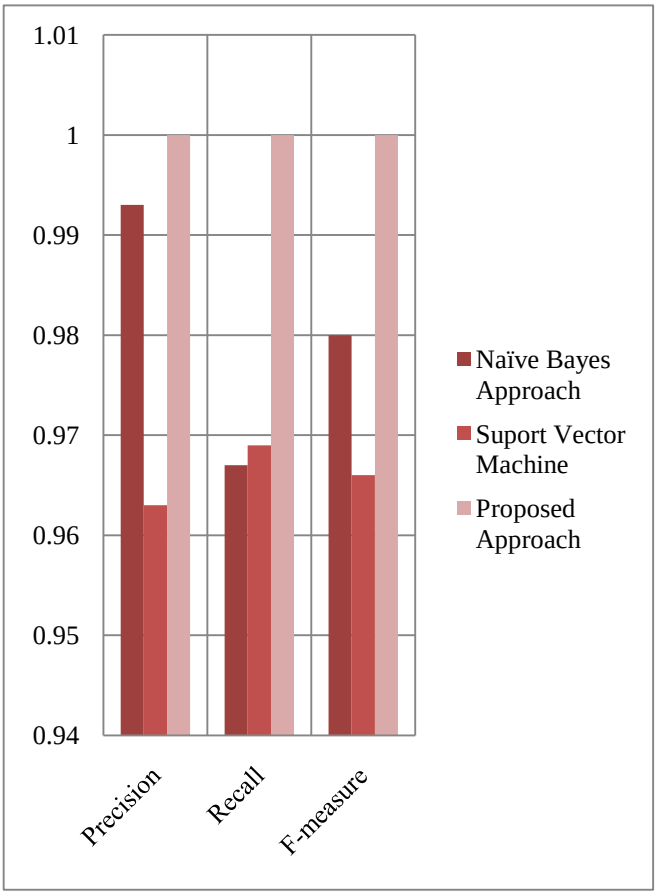
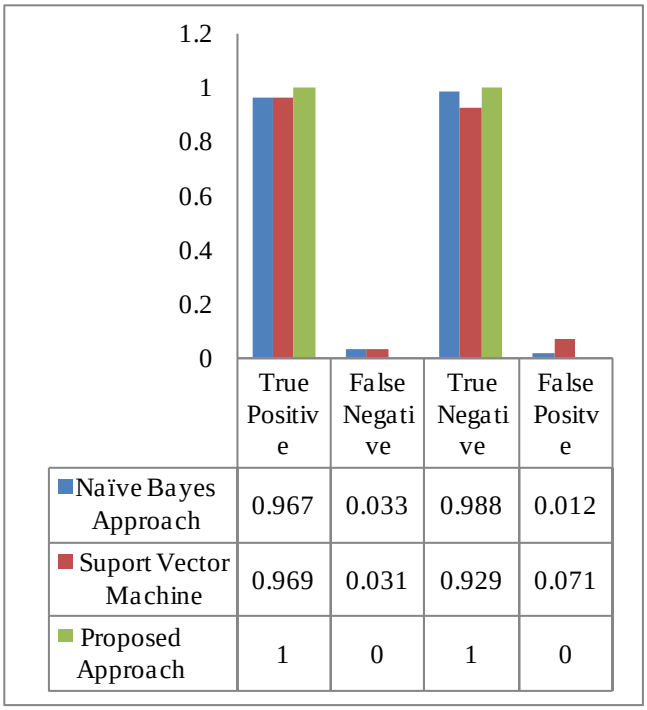


Figure 2. Comparative Study of Proposed Approach



	True Positive	False Negative	True Negative	False Positve
Naïve Bayes Approach	0.967	0.033	0.988	0.012
Suport Vector Machine	0.969	0.031	0.929	0.071
Proposed Approach	1	0	1	0

Figure 3. Comparative results in terms of TP, FP, FN, TN

VIII. CONCLUSION AND FUTURE SCOPE

This paper discusses the knowledge extraction process in terms of big data. As in this data growing age, data is becoming critical and difficult to classify. For that issue, a new approach has been proposed to handle classification and misclassification based on maintaining the quality of data. In Proposed approach, classification of bijective soft set with Rank based approach efficiently handles misclassification of data. In order to handle misclassification, boundary value analysis is conducted. Rank based approach identifies most important attribute in terms of misclassification. Parameters evaluated to prove worth of study include F-measure, precision and recall. All these accuracy related parameters shows high performance as compared to other existing naive biased and support vector machine. Proposed approach can be applied to unstructured data for future enhancement.

IX. REFERENCES

[1] CC. Aggarwal, " Data Mining: The Textbook". Berlin, Germany: Springer; 2015.

[2] O. Okun, G. Valentini, "Supervised and Unsupervised Ensemble Methods and their Applications Studies in Computational Intelligence", Springer, vol. 126,2008.

[3] Nelles, Oliver "Unsupervised Learning Techniques." In Nonlinear System Identification, Springer Berlin Heidelberg, pp: 137-155, 2001.

[4] Burch, Carl,"A survey of machine learning." Tech. report, Pennsylvania Governor's School for the Sciences,2001

[5] D. Fisher, R.DeLine, M. Czerwinski, and S. Drucker. "Interactions with big data analytics." interactions 19, pp: 50-59, 2012

[6] N. Khan, I. Yaqoob, I. A. T. Hashem, Z. Inayat, W. K. M. Ali, M. Alam, M. Shiraz, A. Gani. "Big data: survey, technologies, opportunities, and challenges.", The Scientific World Journal, 2014

[7] I. Arel, DC Rose, TP Karnowski, "Deep machine learning-a new frontier in artificial intelligence research" IEEE Comput Intell Mag 5(4), pp: 13–18 , 2010

[8] G Ding, Q Wu, YD Yao, J Wang, Y Chen, "Kernel-based learning for statistical signal processing in cognitive radio networks" , IEEE Signal Proc Mag , pp: 126–136 , 2013

[9] Y Fu, B Li, X Zhu, C Zhang, "Active learning without knowing individual instance labels: a pairwise label homogeneity query approach", IEEE Trans Knowl Data Eng , pp: 808–822, 2014

[10] Y. Bengio, A. Courville, P. Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35, vol no. 8 , pp: 1798-1828, 2013.

[11] B. C. Ooi, K. L Tan, S. Wang, W. Wang, Q. Cai, G. Chen, J. Gao, "SINGA: A distributed deep learning platform", In Proceedings of the 23rd ACM international conference on Multimedia, pp. 685-688. ACM, 2015.

[12] SJ Pan, Q Yang, "A survey on transfer learning", IEEE Trans Knowl Data Eng , vol no 22(10), pp: 1345–1359, 2010.

[13] D. Pyle, "Data Preparation for Data Mining", San Francisco: Morgan Kaufmann Publishers Inc., 1999.

[14] D. Molodtsov, "Soft set theory-first results," Comput. Math. with Appl., vol. 37, no. 4, pp. 19–31, 1999

[15] J. D. Healy, "The effects of misclassification error on the estimation of several population proportions", Bell System Technical Journal 60, vol no. 5, pp: 697-705, 1981.

[16] K.Gong, Z. Xiao, X. Zhang, "The bijective soft set with its operations." Computers & Mathematics with Applications 60, vol no. 8, pp: 2270-2278, 2010.

[17] R. Sowmya, K. R. Suneetha, "Data Mining with Big Data." In Intelligent Systems and Control (ISCO), 2017 11th International Conference, IEEE, pp. 246-250, 2017.

[18] S. Fong, P. Robert, B. Aghai, Y.W. Si, "Lightweight Feature Selection Methods Based on Standardized Measure of Dispersion for Mining Big Data." In Computer and Information Technology (CIT), 2016 IEEE International Conference,IEEE, pp. 553-559, 2016.

[19] C.C. Aggarwal, Y. Li, J. Wang, J. Wang, "Frequent pattern mining with uncertain ", In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 29-38, 2009.

[20] M. R. Smith, T. Martinez, C. G. Carrier. "An instance level analysis of data complexity." Machine learning 95,vol no. 2, pp: 225-256, 2014

[21] A. Dhurandhar, K. Sankaranarayanan. "Improving classification performance through selective instance completion." Machine Learning 100, vol no. 2-3, pp: 425-447, 2015.

[22] K.Gong, X. Zhang, "Applying bijective soft set in assessment of supply chain information sharing", In Information Management, Innovation Management and Industrial Engineering (ICIII), 2010 International Conference, IEEE, vol. 1, pp. 168-171, 2010.

[23] L. Jiao, T. Denoeux, Q. Pan, "A Hybrid Belief Rule-Based Classification System Based on Uncertain Training Data and Expert Knowledge." IEEE Transactions on Systems, Man, and Cybernetics: Systems46, vol no. 12, pp: 1711-1723, 2016

[24] L. Lang, H. Yonghong, L. Xingming. "Study on the mining method for specific fault data of multimedia sensor networks in cloud computing environment." Multimedia Tools and Applications, pp: 1-16, 2016.

[25] V. Tiwari, P. K. Jain, P. Tandon. "A bijective soft set theoretic approach for concept selection in design process." Journal of Engineering Design, pp:1-18, 2017.

[26] E. Zaitseva, V. Levashenko, "Construction of a Reliability Structure Function Based on Uncertain Data", IEEE Transactions on Reliability 65, vol no. 4, pp: 1710-1723.

[27] S. U. Kumar, H. H. Inbarani, S. S. Kumar. "Improved bijective-soft-set-based classification for gene expression data." In *Computational Intelligence, Cyber Security and Computational Models*, pp. 127-132. Springer India, 2014.

[28] S. U. Kumar, H. H. Inbarani, S. S. Kumar "Classification of ECG cardiac arrhythmias using bijective soft set." In *Big Data in Complex Systems*, pp. 323-350. Springer International Publishing, 2015.

[29] S. Mengle, N. Goharian, A. Platt. "Discovering relationships among categories using misclassification information." In *Proceedings of the 2008 ACM symposium on Applied computing*, pp. 932-937. ACM, 2008.

[30] M. Herland, T. M. Khoshgoftaar, R. Wald. "A review of data mining using big data in health informatics." *Journal of Big Data* 1, no. 1 pp:2, 2012

[31] L. Rutkowski, M. Jaworski, L Pietruczuk, P. Duda. "A new method for data stream mining based on the misclassification error." *IEEE transactions on neural networks and learning systems* 26, no. 5 pp: 1048-1059, 2015

[32] Gross, J. Alan . "An Approach to the Minimization of Misclassification in the Repair of Equipment." *IEEE Transactions on Reliability* 19, no. 1 pp: 10-13, 1970

[33] D. Molodtsov, "Soft set theory—first results." *Computers & Mathematics with Applications* 37, no. 4-5, pp: 19-31, 1999

[34] K. Gong, P. Wang, Y. Peng. "Fault-tolerant enhanced bijective soft set with applications." *Applied Soft Computing* ,2016.

[35] https://archive.ics.uci.edu/ml/datasets.html

[36] Xuegong, Zhang. "Introduction to statistical learning theory and support vector machines." *Acta Automatica Sinica* 26, no. 1 , pp: 32-42, 2000

[37] I. Rish. "An empirical study of the naive Bayes classifier." In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, pp. 41-46, 2001.