Predictive Modeling Framework for Diabetes Classification Using Big Data Tools and Machine Learning

Jangam J. S. Mani [a], **Sandhya Rani Kasireddy** [b]

[a,b] Department of Computing Science, Sri Padmavati Mahila University, Tirupati, India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: Diabetics is now a days a common and most threatening disease irrespective of age, gender and becoming a human threat. As the Internet of Things (IoT) environment is growing rapidly in health sector and continuously gathering the data from smart health care which directly reflects the growth of the big data. Predictive modeling helps doctors and physicians to the identify the growth of diabetics in patients from early ages and create an alarm such to make the patient more attentive towards the diabetics. Based on the previous approaches on diabetics prediction over big data related diabetic prediction yields in better understanding from the patient perspective. The approach for the proposed system is much wider in term of predicting the diabetic model with enough feature variables explaining the patient historical data and diet habits. The Frame work has been carried based on extensive machine learning methods in association for processing of the data over spark RDD. The Random Forest and Ada Boost algorithm showed us a prominent values in terms of predicting the results.

Keywords: Big data, Diabetics, Health care, Machine Learning, Predictive, Frame work, Spark, Random Forest, Ada Boost.

INTRODUCTION

Due to vast advancement in the technology and heavy usage of the data, Digital world is running over Big Data Framework. As the concept big data defines heavy data which need to manageable stored and processed effectively. Smart world is the major concern for the growth of the big data in terms of huge volumes of data which is represented in the form text, images, audio files and video files. As the data grows in volume data storage and processing are major problems which need to addressed. As the data is in huge volumes it is difficulty to process, untill it been well organized.

The good the data structured, the higher is always the that your data usage. Substantial data traits contain of quantity, speed, number, price and veracity. Volume linking into the total amount of this advice that essentially informs the optimal/optimally method to handle higher adaptability info collections along with three-dimensional information foundations along with its own particular management demands. Velocity simplifies the eternal Look of data streams Where invaluable information obtained. Veracity determines the nature of data from different places. Variety describes how to deliver various forms of data. The source data can be tabular data, text, sensor data, audio, video, graph and many more types called structured, quasi, semi-structured and unstructured. Value is essential to get meaningful information about different data which varies significantly.

Based on the challenges of the big data, we proposed our model in spark. A framework which is associated for big data tools helps us to resolve the processing capabilities of the data generated in the health sector of diabetics. This job tries to develop an analytical framework work that forecasting the clear current presence of diabetes at the individual. The Aim of the function would be to forecast Whether that the consumer is influenced by diabetes based on the association data of diet and blood pleasure of the users.

1.1 Diabetics

Nowadays, Diabetes is one of the quickest growing conditions on earth. Diabetes mellitus, popularly called diabetes, obesity as stated by WHO “Diabetes can be a serious, metabolic disorder characterised by elevated blood heights of blood glucose sugar (or blood glucose levels), that contributes as time passes for you and energy to acute harm to one's heart, arteries and uterus, nerves and kidneys". Diabetes may categorize three different kinds: Type 1, Type 2 and breast feeding diabetes. Type-1 diabetes can be also referred to as juvenile diabetes. Type-1 diabetes takes place whenever the body does not generate insulin. Type-2 diabetes takes place whenever the body doesn't create decent utilization of their insulin it delivers. Gestational diabetes is made of elevated blood sugar and this happens in females while pregnant and it is related to complications into the child and mother. GDM usually vanishes following pregnancy however, girls changed and also their kids tend to be at Higher Risk of developing type Two diabetes later in daily lifestyle.

The works flow of the research model.

1. The research proposes a prediction model for diabetics which is a common using big data analytic.
2. The feature selection has become easily the most essential measure, since it reduces the time and computational sophistication of their analysis.
3. The characteristic selection period selects the best features out of your data and also moves the info into your system learning algorithm.
4. The performance of the machine is analyzed Regarding sensitivity, specificity, precision, and period ingestion.

The article is categorized in following way - Section 2 presents review of literature. Section 3 presents technique used for the proposed model. Section 4 presents proposed methods and data processing Section 5 is concluded with the business logic with results and conclusion.

2 Related Work:

A lot of model are being developed and recommended to the users for their diet maintain.

Now there've been lots of recommendation strategies did actually offer end people beneficial health tips for executing a particular task that'll boost their wellness, dependent in their own specified health state along with also place of comprehension originated by the current history of their users along with also different users very similar to these. Predicated on the algorithms Utilised from the Advice programs, we now categorize them into three different classes, namely system learning-based, collaborative filtering-based, along with rule-based Processes

Sajida et al. in [1] Discusses the function of Adaboost and also Bagging outfit machine learning processes employing the J48 decision tree while the foundation for bettering the Diabetes Mellitus and sufferers as parasitic or non-diabetic, dependent on diabetes risk facets. Results attained following the experimentation demonstrates that Adaboost Device learning how outfit strategy works nicely relatively bagging as nicely because of a J48 decision tree.

Naveen Kishore et al. [2] Used a group of regulations methods of this apparatus controlling to forecast diabetes. Five system learning calculations particularly SVM, assortment Tree, Naïve Bayes, Logistic Regression and KNN are traditionally utilized to strike cardiovascular disease. This could possibly be with the capacity of forecasting the opportunity amounts of diabetes and supplies the outstanding dealing with learn pair of regulations better accuracy relatively different calculations. Better precision rate attained using arbitrary woods almost 75 percent.

Sneha et al. [3] Used important attributes, intended a forecast algorithm utilizing Device learning and also found the best classifier to provide the nearest effect when compared with clinical impacts. The procedure centered on picking out the features that ail in premature discovery of Diabetes Mellitus making use of Predictive investigation. Your selection tree algorithm and also the Random woods gets got the maximum specificity of 98.2percent and 98% respectively stays most readily useful for its investigation of parasitic statistics. Naïve Bayesian results says that the optimal/optimally precision of 82.3 percent. The study also generalizes that the Range of best attributes from Your Data Set to enhance the classification precision.

Nazim Razali et al. [4] presented an investigation of diabetic statistics utilizing purification methods founded on Naive Bayes, SMO, REPTree and Straightforward Logistic Regression. Logistic regression has realized the maximum speed of precision, accuracy and remember in contrast to additional three-dimensional calculations even though Naive Bayes is marginally lesser compared to just three calculations. It's a good idea to make employ of a bigger measurement of data sets and features and utilize much far better element selection solutions to enhance the design operation. Combo of classification Methods or hybrid vehicle classification such as piling, fostering and bagging can Enhance the Operation of classification.

Quan zu et al. [5] Used decision tree, random forest and neural network to forecast diabetes. It comprises 14 features. Fivefold cross endorsement used to test these units. As a consequence of advice unbalance, randomly extracted five times advice. And the final result is the fact the common of those five experiments. Used main component analysis (PCA) and nominal yield maximum relevance (mRMR) to diminish the dimensionality. Prediction with random forest could reach the highest precision (ACC = 0.8084) if all of the characteristics were utilized.

Manal Alghamdi1 et al. [6] The comparative performance of many machine learning processes such as Conclusion Tree, Naïve Bayes, Logistic Regression, Logistic Model Tree and Random Forests for predicting episode diabetes with clinical records of cardio respiratory fitness. The data set included 62 features categorized in to four categories: demographic traits, illness history, drug usage history, and also stress test key signs. Manufactured an Ensemble-based predictive version with 1 3 features that were chosen based on their own clinical value, Multiple Linear Regression, and Information Gain Ranking techniques. The negative consequence of the imbalance type of this constructed version was handled by Artificial Minority Oversampling Technique

(SMOTE). The typical performance of the predictive version classifier has been enhanced with the Ensemble machine learning procedure working with the Vote system using three Decision Trees (Naïve Bayes Tree, Random Forest, along with Logistic Model Tree) and reached high accuracy of prediction (AUC = 0.92)

3 Techniques:

Hadoop:

Hadoop is a frame work which is been written in java to handle huge amount of data known as big data. The frame work has a best distributed platform data storing and processing. Hadoop performs parallel processing on the data sets. A viable platform for interacting with other ecosystems using API.

Apache Spark:

Apache Spark has jumped together look up engine to get big machine and data understanding. It's an open minded and spread computing process. It works by using in-memory caching and optimized question implementation for quickly inquiries against information of virtually any measurements.

Spark core is your principal portion of this Apache Spark frame. It's the entire implementation engine to get Spark stage which each of the functionality characterized is assembled up on. It Supplies in Built memory computing along with assigning data collections saved inside storage methods. It uses RDD data structure which is a special data structure. Spark core performs all the basic input-output functions, scheduling, monitoring etc., and other important functions are fault recovery and effective memory management.

Logistic regression

Logistic regression is a machine algorithm for classification. Inside this algorithm, the possibilities describing the probable results of one trial have been modelled utilizing a specified functionality. Logistic regression was fashioned for classification reasons, also so is useful for comprehending that the effect of numerous different factors on a single outcome factor. Works just once the called factor is binary, supposes all predictors are separate of one another and supposes info is liberated from lost worth.

Important features to the logistic regression model are DR1TWS_19.0 : Tap water source = Other, MCQ053_9.0 : Been on treatment for anemia in last 3 months = Don't know, BPAARM_8.0 : Arm selected for Blood Pressure Measurement = Could not obtain, DUQ240_9.0 : Ever used cocaine / methamphetamine / heroin = Don't know, INQ030_7.0 : Income from Social Security or Railroad Retirement = Refused to answer, PAQ635_9.0 : Walk or use a bicycle 10 minutes continuously to go places = Don't know, DR1TWS_10.0 : Tap water source = Other, DR1TTFAT : Dietary intake (one day) total fat (g), DMDMARTL_77.0 : Marital Status = Refused to answer, IND235_23.0 :Monthly family income = Other

Decision tree

No need to use scaled data for the decision tree. Pull out continuous data to concatenate with dummies. Supplied A-Data of features with its own categories, an alternative shrub delivers a succession of regulations which may be utilised to categorize the information. Decision-tree isn't hard to comprehend and visualise, necessitates modest data prep, also also certainly will manage both numeric and numerical info. Choice trees may produce intricate trees Which Do not generalise nicely, and trees might be shaky as little variants Important features to the Decision tree model are DIQ010_2.0: Doctor ever said you have diabetes = NO, RIDAGEYR: Age (yr), PHAFSTHR: Fasting time before blood draw for labs (hr), BMXWAIST: Waist circumference (cm), LGXSGTSI: Gamma glutamyl transferase (U/L), LBXSOSSI = Osmolality (mmol/Kg), LBXMCHSI: Mean cell hemoglobin (pg), URXUMA: Albumin, urine (ug/mL), LBXSTR: Triglycerides (mg/dL), BPXSYA: Systolic Blood pressure (avg of 3 rdgs) mm Hg

Random forest

Random forest is an adaptable, user friendly machine learning algorithm which produces, despite hyper-parameter pruning, a fantastic result the majority of the moment. It's also perhaps probably one among the most famous calculations, as a result of its diversity and simplicity. Random forest has not exactly the exact hyper parameters as being a decision tree or even a bagging classifier. Luckily, there isn't any requirement to unite an alternative tree using a canning classifier as you can readily utilize the classifier-class of random woods. With arbitrary forest, it's possible to even manage regression activities using the algorithm's regressor. Even the hyper parameters in random forest are used to raise the predictive capability of this model or even to produce the version faster. Let us go through the hyper parameters of all sklearns builtin random forest function.

Important features to the tuned random forest model are DIQ010_2.0: Doctor ever said you have diabetes = NO, RIDAGEYR: Age in (years), PHAFSTHR: Total fasting time before blood draw for labs (hr), BMXWAIST: Waist circumference (cm), RXDCOUNT: Number of prescription medications currently taking, DIQ050_2.0: Taking insulin now = NO, LBXSGTSI: Gamma glutamyl transferase (U/L), URXUMA: Albumin, urine (ug/mL), LBXSOSSI: Osmolality (mmol/Kg), LBXSTR: Triglycerides (mg/dL).

AdaBoost

It makes N-number of decision designs throughout the practice amount of data. Whilst the very initial decision tree/model was created, the listing that's wrongly categorized throughout the very first version is given greater resolution. These records are shipped as input to your next version. The method will proceed until we define several base students you would like to make. Bear in mind, the replica of documents is enabled together with boosting methods. The very first version consists of and the errors out of the very first version are noticed by the algorithm, that the first album that's wrongly categorized is provided the input to the second version. This approach is repeated until the stated condition is met. It's known as Adaptive reinforces whilst the weights have been re assigned to every case, together with high weights to erroneously classified cases.

Important features to the tuned adaboost classification model are LBXSOSSI: Osmolality (mmol/Kg), LBXSNASI: Sodium (mmol/L), RIDAGEYR: Age (yr), LBXSCLSI: Chloride (mmol/L), BMXWAIST: Waist circumference (cm), LBXMCHSI: Mean cell hemoglobin (pg),

LGXSGTSI: Gamma glutamyl transferase (U/L), PHDSESN_1.0: Examination session = Morning, LBDHDD: Direct HDL-Cholesterol (mg/dL), BYXSYA: Systolic blood pressure (mm Hg).

Gradient boosting

Gradient boosting, the name its self suggests the boosting mechanism for the machine algorithms. The is most popular because of its accuracy and speed in solving the logics based on the highest number of features to be processed along with more complex data. This works on the next best possible model, so works it combines the previous model to current model and make a note of the accuracy and the speed with prediction error. The best solution will be generated with the outcome variables and minimum error.

Important features to the tuned Gradient boosting classification model are DIQ010_2.0 : Doctor ever said you have diabetes = NO, RIDAGEYR :Age (yr) , PHAFSTHR : Total fasting time before blood draw for labs (hr), BMXWAIST : Waist circumference (cm), LBXSOSSI : Osmolality (mmol/Kg) , PHDSESN_1.0 : Examination session = Morning, LBXSNASI : Sodium (mmol/L), LGXSGTSI : Gamma glutamyl transferase (U/L), LBXMCHSI : Mean cell hemoglobin (pg), DIQ010_3.0 : Doctor ever said you have diabetes = BORDERLINE.

XGBoost

XGBoost is short for Extreme Gradient Boosting. This uses a boosting framework from gradient algorithm. This works well with unstructured data.

Important features to the XGboost classification model are DIQ010_2.0: Doctor ever said you have diabetes = NO, RIDAGEYR: Age (yr), PHAFSTHR: Total fasting time before blood draw for labs (hr). BMXWAIST: Waist circumference (cm), DIQ010_3.0: Doctor ever said you have diabetes = BORDERLINE, LBXSGTSI: Gamma glutamyl transferase (U/L), DIQ050_2.0: Taking insulin now = NO, BPACSZ_3.0: Blood pressure cuff size = Adult, LBXRDW: Red cell distribution width (%), PHDSESN_1.0: Examination session = morning.

Tune the XGBoost model:

Now, tune XGBoost model with RandomizedSearchCV. With tuned XGboost no-diabetes precision value is 0.78, recall value is 0.83 and f1-score value is 0.81. Pre-diabetes precision value is 0.60, recall value is 0.55 and f1-score value is 0.57. Diabetes precision value is 0.76, recall value is 0.62 and 1-score value is 0.68. Specificity of No Diabetes is 0.73, Prediabetes is 0.81 and Diabetes is 0.98. Tuned XGBoost is clearly the best tree-based model. It would be interesting to tune the Logistic Regression model.

Important features to the tuned XGboost classification model are DIQ010_2.0 : Doctor ever said you have diabetes = NO, SMAQUEX_2.0 : Smoking recent use questionnaire flag = >= 18 yrs old, RIDAGEYR : Age (yr), DIQ050_2.0 : Taking insulin now = NO, SMAQUEX2_2.0 : Smoking cigarette use questionnaire flag = 12-17 yrs old, DIQ010_3.0 : Doctor ever said you have diabetes = BORDERLINE, PHAFSTHR : Total fasting time before blood draw for labs (hr), PHDSESN_2.0 : Examination session = afternoon, PHDSESN_1.0 : EDF ZZRIAGENDR_2.0 = Gender = Female

4 Proposed Predictive Model

Proposing a framework which has the following major components: data collection, preprocessing, performance testing, ML model fitting, aiding prediction of diet and quality of diabetes class. The framework flow has following steps:

Data collection and storage: NHANES data are made available in small files, each containing data relating to one topic for each two-year survey cycle. The files are SAS transport format. For this analysis, data elements were limited to those available during the entire ten-year span. A total of 314 files were downloaded for this analysis.

The data set which has been dealing with the nutrition data which is of CSV format has been extracted from HDFS, there after its been pre-processed as per the requirements. Once the data has been processed perfectly the parameters are been passed to the algorithm

Extract features: Take labs, demographics, Examinations, diet, and questionnaire data to create a rule based algorithm to create diabetes class labels as No Diabetes-0, Pre Diabetes-1 and Diabtes-2.

Data Metrics and Statistical assessment:

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

The Precision value should be 1 (high), to achieve best classifier. The value becomes 1 if only that numerator along with denominator must be equal i.e

TP = TP +FP,

Which makes the FP to be zero?

Precision=TP/(TP+FP)

The second important Recall which is also called as sensitivity else true positive rate. Recall will become 1 if only the numerator along with denominator are equal i.e

TP = TP +FN,

this also means FN is zero.

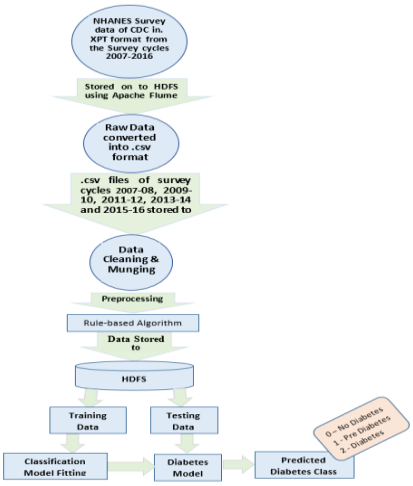
Recall=TP/(TP+FN)

Therefore, ideally for good classifier, we need both precision and recall.

F1-score is a metric which takes into account both precision and recall.

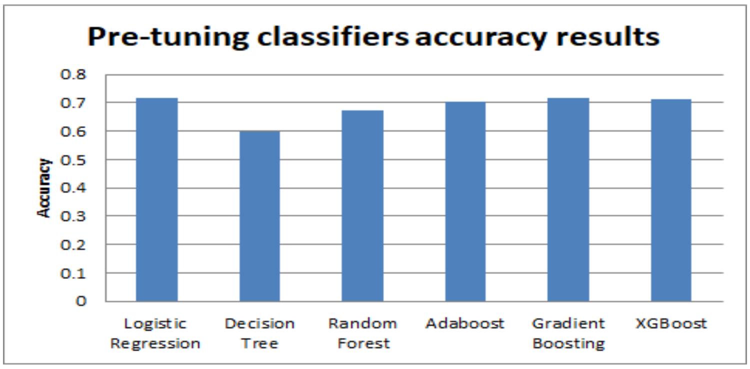
F1-Score=2*[(Precision*Recall) / (Precision + Recall)]

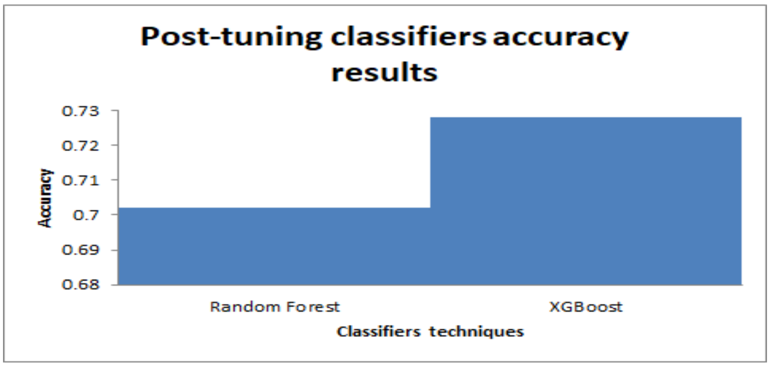
F1 Score becomes 1 only when precision and recall are both 1. F1 score becomes high only when both precision and recall are high. F1 score is the harmonic mean of precision and recall and is a better measure than accuracy.

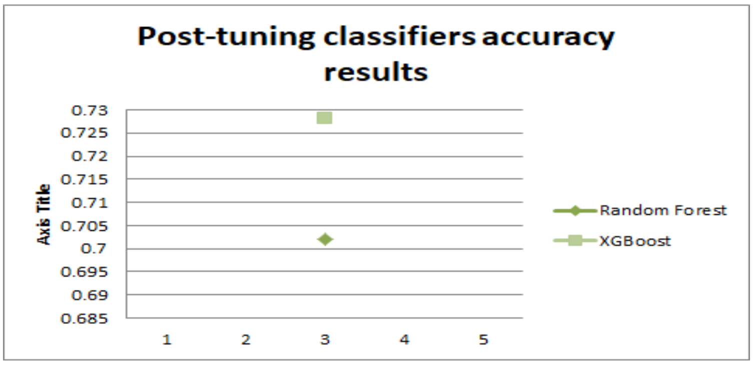

NHANES Survey data of CDC in. XPT format from the Survey cycles 2007-2016
Stored on to HDFS using Apache Flume
Raw Data converted into .csv format
.csv files of survey cycles 2007-08, 2009-10, 2011-12, 2013-14 and 2015-16 stored to
Data Cleaning & Munging
Preprocessing
Rule-based Algorithm
Data Stored to
HDFS
Training Data
Testing Data
Classification Model Fitting
Diabetes Model
Predicted Diabetes Class
0 – No Diabetes
1 - Pre Diabetes
2 - Diabetes

5 Experiment and Results:

The experimental results of the classification techniques are divided into pre-tuning and post -tuning results. The comparison of the accuracy of the pre-tuning and post-tuning classifiers are shown in fig.







From the above results it's clear that the post tuning classifiers has shown prominent and effective accuracy while applied with Random Forest and XGBoost.

6 Conclusion

The paper describes a big data frame work model associated with spark in order to predictive modeling and classifying the user diabetic or non-diabetic by employing the machine learning library from the spark MLLib. The results generated as prominent for the data. The model has been evaluated as pre tuning and post tuning with all the feature variable available in the data. The classifier used to predict owing to be faster and quicker in terms of learning capabilities. The performance seems to be very effective and work has executed as per the planned approach and gained enough and good insights.

REFERENCES:

[1] Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82, 115–121. doi: 10.1016/j.procs.2016.04.016.

[2] Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy, "Prediction Of Diabetes Using Machine Learning Classification Algorithms", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 01, JANUARY 2020, ISSN 2277-8616.

[3] N. Sneha, Tarun Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection", Journal of Big Data (2019) 6:13 https://doi.org/10.1186/s40537-019-0175-6.

[4] Nazim Razali, Syed Zulkarnain Syed Idrus and Mohd Helmy Abd Wahab, "Analyzing Diabetic Data using Classification" Journal of Physics: Conference Series, 1529 (2020) 022105.

[5] Quan zou, Kaiyang Qu, Yamei Luo and Hua Tang, "Predicting Diabetes Mellitus with Machine Learning Techniques", Frontiers in Genetics, Published: 06 November 2018, doi: 10.3389/fgene.2018.00515.

[6] Manal Alghamdi1, Mouaz Al-Mallah1, and Steven Keteyian, "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project" PLOS ONE | https://doi.org/10.1371/journal.pone.0179805 July 24, 2017.

[7] M. Wang, C. Lee and Hani Hagras, "A Type-2 Fuzzy Ontology and Its Application to Personal Diabetic-Diet Recommendation," IEEE Transactions on Fuzzy Systems, Vol. 18, No. 2, 2010.

[8] C. Lee, M. Wang, H. Hagras, Z. Chen, S. Lan, C. Hsu, S.Kuo, H. Cheng and H. Lee, "A novel genetic fuzzy markup language and its application to healthy diet assessment," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 20, pp. 247-278, 2012.

[9] C. Lee, M. Wang and Z. Chen, "Genetic fuzzy markup language for diet application", Proceedings of the 2011 IEEE International Conference on Fuzzy Systems.